NONLINEAR SYSTEMS ANALYSIS

SECOND EDITION

M. VIDYASAGAR

Centre for AI and Robotics, India

PRENTICE HALL, Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging-in-Publication Data

Vidyasagar, M. (Mathukumalli)
Nonlinear systems analysis / M. Vidyasagar. -- 2/e.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-623463-1
1. System analysis. 2. Differential equations, Nonlinear.
I. Title.
QA402.V53 1993
003'.75--dc20 92-5390
CIP

Acquisitions editor: ***Pete Janzow***
Production editor: ***Jennifer Wenzel***
Copy editor: ***Cami Goffi***
Cover designer: ***Joe DiDomenico***

Prepress buyer: ***Linda Behrens***
Manufacturing buyer: ***David Dickey***
Supplements editor: ***Alice Dworkin***
Editorial assistant: ***Phyllis Morgan***

A Simon & Schuster Company
Englewood Cliffs, New Jersey 07632

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-623463-1

ISBN 0-13-623463-1

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

To Charlie Desoer

आकाशात्पतितम् तोयम्
यथा गच्छति सागरम्
सर्वदेवनमस्कार:
केशवम् प्रति गच्छति

As all water falling from the sky
Eventually reaches the sea
So do salutations to various deities
Reach the same almighty

From *Sandhyavandanam* (A salute to the twilight)

లోకంబులు లోకేశులు
లోకస్థులు దెగిన తుది నలోకంబగు పెం
జీకటి కవ్వల నెవ్వం
డేకాకృతి వెలుగు నతనినే సేవింతున్

Beyond the worlds
Their rulers and their denizens
Beyond the unwordly void
The one Who shines alone
Him I worship

From *Andhra Maha Bhagavatam* by Bammera Potana
(c. 1400 A.D.)

CONTENTS

PREFACE

It is now more than a decade since I wrote the book *Nonlinear Systems Analysis*. Since that time, several developments have taken place in this area which have made it desirable to update the contents of the book. Accordingly, virtually the entire book has been rewritten. The most notable changes are the following:

1) During the past decade, there have been some significant advances in the area of nonlinear control system design based on the use of differential geometric methods. Thus it is imperative that anyone interested in nonlinear system theory should have at least a passing acquaintance with these methods. In this second edition, I have included a new chapter which discusses the differential geometric approach (Chapter 7). For ease of exposition, all systems are considered to evolve over an open subset of $\mathbf{R}^n$; thus the analysis is only local. Topics covered include reachability, observability, and feedback linearization (in both the input-state and input-output settings), zero dynamics, and the stabilization of linearizable systems. In addition to presenting the theory, I have also included some applications of the theory to problems in robotics. Motivated by this chapter, an interested and diligent student could pursue a more rigorous course of study with an advanced text.

2) Several significant results have been obtained in the "traditional" areas of Lyapunov stability and input-output stability since the writing of the first edition. Some of these results are included in the present edition, such as: observer-controller stabilization of nonlinear systems, and the stability of hierarchical systems (Section 5.8); relationships between Lyapunov stability and input-output stability (Section 6.3); and a useful class of transfer functions of distributed systems (Section 6.5). In addition to the above, Section 4.2, containing a rigorous analysis of the describing function method, is also new.

3) Various standard texts in stability theory have gone out of print, making their contents all but inaccessible to the student. Two examples of such books are: *Stability of Motion* by W. Hahn and *Feedback Systems: Input-Output Properties* by C. A. Desoer and myself. At the same time some of the techniques presented in these books are finding new and previously unsuspected applications. With this in mind, in the present edition I have included some relevant material from these and other classic books, such as the converse Lyapunov theory (Section 5.7), and the feedback stability of time-varying and/or nonlinear systems (Section 6.6).

4) In view of the increasing importance of digital computers, I have included a discussion of discrete-time systems in the chapters dealing with Lyapunov stability and input-output stability.

5) Three new appendices have been added. Appendix A describes a sixty year-old theorem due to Witold Orlicz, on the prevalence of differential equations with unique solutions. This paper is quite inaccessible, but its contents deserve wide dissemination. Appendix B gives a proof of the Kalman-Yacubovitch lemma, while Appendix C contains a proof of the Frobenius theorem. The contents of the last two appendices are of course readily available elsewhere, but their inclusion in the present text makes it more self-contained.

6) The original edition of this book contained examples which were mostly drill problems or exercises. During the recent years I have come to feel that nonlinear system theory is most useful in studying the behavior of an entire *class* of systems rather than a given *specific* system. Accordingly, several applications of nonlinear system theory have been included throughout the book. Most of them have to do with robotics in some form or other.

With these changes, the book is somewhat bigger than the first edition. It would be difficult to cover the entire book during a single semester. However, I hope its value as a reference has been enhanced by the changes. Chapter 2 contains basic material which should be covered in order to appreciate the remainder of the text. But a sincere attempt has been made to ensure that Chapters 3 through 7 are independent, so that an instructor can pick and choose material to suit his/her needs. Even within a chapter, it is possible to cover certain sections and omit others. A perusal of the Contents reveals the amount of flexibility available in putting together a suitable course from the contents of the text.

In spite of the enlargement in the size of the book, some topics which deserve the attention of system theorists are not included. Examples of such topics are chaotic motions, averaging analysis, Volterra series, bifurcation theory, and catastrophe theory. I have made a conscious decision to omit these topics, mainly to keep the length of the book within reasonable limits. But no study of nonlinear systems is complete without at least an introduction to these topics. Moreover, there are several excellent texts available addressing each of the above topics.

In the preface to the first edition, I wrote fancifully that the book could be used by "engineers, mathematicians, biologists *et cetera.*" Judging by the Science Citation Index, no biologists appear to have read the book (though two *social scientists* have, amazingly enough). More realistically, I would expect the present edition to be of interest primarily to engineers interested in a rigorous treatment of nonlinear systems, and to mathematicians interested in system theory. Though some aspects of control are covered in the book (especially in Chapter 7), the focus is still on analysis rather than synthesis. Hence I have retained the original title. I do expect that the book can be used not just in Electrical Engineering departments, but also in Mechanical Engineering departments, and perhaps in some departments of Applied Mathematics. Above all, I hope it will continue to serve as a reference source for standard results in nonlinear system analysis.

I would like to thank Toshiharu Sugie for his careful reading of early versions of Chapters 5 and 6. I would also like to thank those who reviewed the text, particularly Brian Anderson, Aristotle Arapostathasis, Ragu Balakrishnan, Joseph Bentsman, Alan Desrochers, Brad Dickinson, Ashok Iyer, Bob Newcomb, Charles L. Phillips, and Irwin Sandberg.

It is my pleasure and honor to dedicate to this book to Professor Charles A. Desoer of the University of California at Berkeley. Though I was not privileged to be one of his Ph.D. students, I was fortunate enough to have come under his influence while still at a formative stage in my career. Any instances of originality, creativity and clarity in my research and exposition are but pale imitations of his shining example.

NOTE TO THE READER

All items within each section are numbered consecutively, be they equations, theorems, definitions, or something else. A reference such as "(17)" refers to the 17-th item *within the same section.* When it is necessary to refer to an item from another section, the full citation is given, e.g., "Theorem (5.1.16)." All theorems, lemmas, and definitions are stated in *italics.* In a definition, the concept being defined is displayed in **bold face.** The same convention is used in the running text as well. The use of italics in the running text is reserved for *emphasis.* The box symbol ■ is used to denote the end of a proof. In cases where there might be some ambiguity, the same symbol is also used to denote the end of an example. Lower-case bold letters such as $\mathbf{x}$ denote vectors, upper-case bold letters such as $\mathbf{A}$ denote matrices, and italic letters denote scalars; however, there are a few exceptions to this convention. For example, the identity matrix is denoted by I.

Finally, the reader is urged to attempt all the problems, since they are an integral part of the text. Happy reading!

1. INTRODUCTION

The topic of this book is the analysis of nonlinear systems. The adjective "nonlinear" can be interpreted in one of two ways, namely: "*not* linear" or "not *necessarily* linear." The latter meaning is intended here.

Why should one study nonlinear systems? The fact is that virtually *all* physical systems are nonlinear in nature. Sometimes it is possible to describe the operation of a physical system by a linear model, such as a set of ordinary linear differential equations. This is the case, for example, if the mode of operation of the physical system does not deviate too much from the "nominal" set of operating conditions. Thus the analysis of linear systems occupies an important place in system theory. But in analyzing the behaviour of any physical system, one often encounters situations where the linearized model is inadequate or inaccurate; that is the time when the contents of this book may prove useful.

There are several important differences between linear systems and nonlinear systems: 1) In the case of linear systems described by a set of linear ordinary differential equations, it is often possible to derive *closed-form expressions* for the solutions of the system equations. In general, this is not possible in the case of nonlinear systems described by a set of nonlinear ordinary differential equations. As a consequence, it is desirable to be able to make some predictions about the behaviour of a nonlinear system even in the *absence* of closed-form expressions for the solutions of the system equations. This type of analysis, called **qualitative** analysis or **approximate** analysis, is much less relevant to linear systems. 2) The analysis of nonlinear systems makes use of a *wider variety* of approaches and mathematical tools than does the analysis of linear systems. The main reason for this variety is that no tool or methodology in nonlinear systems analysis is *universally* applicable (in a fruitful manner). Hence the nonlinear systems analyst needs a wide variety of tools in his or her arsenal. 3) In general, the level of mathematics needed to master the basic ideas of nonlinear systems analysis is higher than that for the linear case. Whereas matrix algebra usually occupies center stage in a first course in linear systems analysis, here we use ideas from more advanced topics such as functional analysis and differential geometry.

A commonly used model for a nonlinear system is

$$\mathbf{1} \qquad \dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t), \mathbf{u}(t)], \ \forall t \geq 0,$$

where t denotes time; $\mathbf{x}(t)$ denotes the value of the function $\mathbf{x}(\cdot)$ at time t and is an n-dimensional vector; $\mathbf{u}(t)$ is similarly defined and is an m-dimensional vector; and the function $\mathbf{f}$ associates, with each value of t, $\mathbf{x}(t)$, and $\mathbf{u}(t)$, a corresponding n-dimensional vector. Following common convention, this is denoted as: $t \in \mathbf{R}_+$, $\mathbf{x}(t) \in \mathbf{R}^n$, $\mathbf{u}(t) \in \mathbf{R}^m$, and $\mathbf{f}: \mathbf{R}_+ \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$. Note that (1) is a first-order vector differential equation. The quantity $\mathbf{x}(t)$ is generally referred to as the **state** of the system at time t, while $\mathbf{u}(t)$ is called the **input**

or the **control** function. It is clear that (1) represents a continuous-time system. Its discrete-time counterpart is

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{u}_k),\ k = 0, 1, 2, 3, \cdots, \tag{2}$$

which is a first-order vector difference equation. There is no loss of generality in assuming that the system at hand is described by a first-order (differential or difference) equation. To see this, suppose the system is described by the n-th order scalar differential equation

$$\frac{d^n y(t)}{dt^n} = h\left[t, y(t), \dot{y}(t), \cdots, \frac{d^{n-1} y(t)}{dt^{n-1}}, u(t)\right],\ \forall t \geq 0. \tag{3}$$

This equation can be recast in the form (1) by defining the n-dimensional state vector $\mathbf{x}(t)$ in the familiar way, namely

$$x_1(t) = y(t), x_2(t) = \dot{y}(t), \cdots, x_n(t) = \frac{d^{n-1} y(t)}{dt^{n-1}}. \tag{4}$$

Then (3) is equivalent to

$$\begin{aligned}
&\dot{x}_1(t) = x_2(t),\\
&\dot{x}_2(t) = x_3(t),\\
&\cdot\\
&\cdot\\
&\cdot\\
&\dot{x}_{n-1}(t) = x_n(t)\\
&\dot{x}_n(t) = h\left[t, x_1(t), x_2(t), \cdots, x_n(t), u(t)\right]
\end{aligned} \tag{5}$$

Now (5) is of the form (1) with

$$\mathbf{x}(t) = [x_1(t) \cdots x_n(t)]', \tag{6}$$

$$\mathbf{f}(t, \mathbf{x}, u) = [x_1\ x_2\ \cdots x_n\ h(t, x_1, \cdots, x_n, u)]'. \tag{7}$$

More generally, even coupled nonlinear differential equations can be put into the form (1). Analogous remarks apply also to difference equations. In fact, much of the power of "modern" control theory derives from the generality and versatility of the state-space descriptions (1) and (2).

In studying the system (1), one can make a distinction between two aspects,[1] generally referred to as analysis and synthesis, respectively. Suppose the input function $\mathbf{u}(\cdot)$ in (1) is

[1] Henceforth attention is focused on the continuous-time system (1), with the understanding that all remarks apply, *mutatis mutandis*, to the discrete-time system (2).

specified (i.e., fixed), and one would like to study the behaviour of the corresponding function $\mathbf{x}(\cdot)$; this is usually referred to as **analysis**. Now suppose the problem is turned around: the system description (1) is given, as well as the desired behaviour of the function $\mathbf{x}(\cdot)$, and the problem is to find a suitable input function $\mathbf{u}(\cdot)$ that would cause $\mathbf{x}(\cdot)$ to behave in this desired fashion; this is usually referred to as **synthesis**. Most of this book is devoted to the analysis of nonlinear systems.

The rest of this chapter is devoted to introducing several commonly used terms. The system (1) is said to be **forced**, or to have an input; in contrast, a system described by an equation of the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], \forall t \geq 0, \tag{8}$$

is said to be **unforced.** Note that the distinction is not too precise. In the system (1), if $\mathbf{u}(\cdot)$ is specified, then it is possible to define a function $\mathbf{f}_{\mathbf{u}}: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}^n$ by

$$\mathbf{f}_{\mathbf{u}}(t, \mathbf{x}) = \mathbf{f}[t, \mathbf{x}, \mathbf{u}(t)]. \tag{9}$$

In this case (1) becomes

$$\dot{\mathbf{x}}(t) = \mathbf{f}_{\mathbf{u}}[t, \mathbf{x}(t)], \forall t \geq 0. \tag{10}$$

Moreover, if $\mathbf{u}(\cdot)$ is clear from the context, the subscript $\mathbf{u}$ on $\mathbf{f}_{\mathbf{u}}$ is often omitted. In this case there is no distinction between (10) and (8). Thus it is safer to think of (8) as describing one of two possible cases: (i) there is no external input to the system, or (ii) there is an external input, which is kept fixed throughout the study.

11 Definition *The system (1) or (8) is said to be* **autonomous** *if the function* $\mathbf{f}$ *does not explicitly depend on its first argument t; it is said to be* **nonautonomous** *otherwise.*

Note that some authors use "time-invariant" instead of "autonomous" and "time-varying" instead of "nonautonomous."

Consider the system (1), and suppose it is autonomous, i.e., $\mathbf{f}$ is independent of t. Now suppose a *non-constant* input function $\mathbf{u}(\cdot)$ is applied. Then the corresponding function $\mathbf{f}_{\mathbf{u}}$ defined in (10) may in fact depend on t [since $\mathbf{u}(t)$ depends on t]. The point to note is that a system may be either autonomous or nonautonomous depending on the context.

The next concept is central to nonlinear system theory.

12 Definition *A vector* $\mathbf{x}_0 \in \mathbf{R}^n$ *is said to be an* **equilibrium** *of the unforced system (8) if*

$$\mathbf{f}(t, \mathbf{x}_0) = \mathbf{0}, \forall t \geq 0. \tag{13}$$

If $\mathbf{x}_0$ is an equilibrium of the system (8), then the differential equation

14 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], \forall t \geq t_0; \mathbf{x}(t_0) = \mathbf{x}_0,$

has the unique solution

15 $\mathbf{x}(t) = \mathbf{x}_0, \forall t \geq t_0.$

In other words, if a system starts in an equilibrium, it remains in that state thereafter.

Many features that are taken for granted in the case of linear systems do not hold for nonlinear systems. This is one of the major challenges of nonlinear systems analysis. To illustrate a few of these features, consider the system description (8). In order to represent a physical system, the model (8) should satisfy one of the following statements:

1. Equation (8) has at least one solution (existence of a solution).
2. Equation (8) has exactly one solution for all sufficiently small values of t (local existence and uniqueness of solution).
3. Equation (8) has exactly one solution for all t in the interval $[0, \infty)$ (global existence and uniqueness of solution).
4. Equation (8) has exactly one solution for all t in the interval $[0, \infty)$, and this solution depends continuously on the initial condition $\mathbf{x}(0)$ (well-posedness).

Statements 1 to 4 are progressively stronger. Ideally one would wish that the system description (8) exhibits the behaviour described in Statement 4. Unfortunately, without some restrictions on the nature of the function $\mathbf{f}$, *none* of these statements may be true, as illustrated by the following examples.

16 Example Consider the scalar differential equation

17 $\dot{x}(t) = -\operatorname{sign} x(t), \forall t \geq 0; x(0) = 0,$

where the "sign" function is defined by

$$\operatorname{sign} x = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases}$$

It is easy to verify that no *continuously differentiable* function $x(\cdot)$ exists such that (17) is satisfied. Thus even Statement 1 does not hold for this example.

18 Example Consider the scalar differential equation

$$\dot{x}(t) = \frac{1}{2x(t)}, \forall t \geq 0; x(0) = 0.$$

This equation admits *two* solutions, namely

$$x(t) = \pm t^{1/2}.$$

Thus Statement 1 is true, but Statement 2 is false.

19 Example Consider the scalar differential equation

$$\dot{x}(t) = 1 + x^2(t), \forall t \geq 0; x(0) = 0. \tag{20}$$

Then, over the interval [0, 1), this equation has the unique solution

$$x(t) = \tan t.$$

But there is no continuously differentiable function $x(\cdot)$ *defined over the entire interval* $[0, \infty)$ such that (20) holds. This is because, as $t \to \pi/2$, the solution $x(t) \to \infty$, a phenomenon known as "finite escape time." Thus Statements 1 and 2 are true for this system, but Statement 3 is false. ■

It is therefore clear that the questions of existence and uniqueness of solutions of (8), and their continuous dependence on the initial conditions, are very important. These questions are studied in Chapter 2.

The subject of Chapter 3 is second-order systems. Before attempting a study of n-th order systems in all of their generality, it is fruitful to begin with the special case of second-order systems, since many of the arguments are simplified in this special case.

In Examples (18) and (19), it was possible to derive closed-form expressions for the solutions of the differential equations under study, because the equations were of a very simple nature. However, this is not possible in general, and one must be content with approximate analysis methods. These are the subject of Chapter 4.

An important issue in nonlinear systems analysis is that of the *well-behavedness*, in a suitably defined sense, of the solutions to the unforced system (8) or the forced system (1). This is usually called the question of "stability." Ideally one would like to draw conclusions about the well-behavedness or otherwise of these solutions *without actually solving the system equations.* Chapter 5 is concerned with the stability of unforced systems of the form (8), while Chapter 6 is concerned with the stability of forced systems—so-called "input-output" stability. An added bonus in Chapter 6 is that the systems studied are more general than (1); in fact, the theory developed there applies equally well to delay systems, and systems described by partial (not ordinary) differential equations.

Chapter 7 focuses on a recent development in the study of nonlinear control systems, namely the use of differential-geometric methods. The general theme of this chapter is that many results from the theory of linear control systems can be extended to a broad class of autonomous nonlinear control systems.

2. NONLINEAR DIFFERENTIAL EQUATIONS

In this chapter, we undertake a systematic study of nonlinear ordinary differential equations (o.d.e.'s). As one can see from the examples given in Chapter 1, a nonlinear equation can in general exhibit very wild and unusual behavior. However, it is shown in this chapter that, for a practically significant class of nonlinear o.d.e.'s, it is possible to ascertain the existence and uniqueness of the solutions corresponding to each initial condition, as well as continuous dependence of the solution on the initial condition.

Except for very special cases which are usually "cooked" in advance, it is not possible to obtain a closed-form expression for the solution of a nonlinear o.d.e. Hence it is necessary to devise methods for analyzing the behavior of the solution of a given nonlinear o.d.e. *without* relying on being able to find a closed-form solution for it. The numerical solution of o.d.e.'s is a well-developed subject in its own right, and it is not covered in the present book; the interested reader is referred to any of the several excellent books on the topic, e.g., Gear (1971). In this chapter, we content ourselves with a method for obtaining bounds on the solution of a given equation without actually solving the equation. Using this method, it is possible to determine, at each instant of time, a region in $\mathbf{R}^n$ in which the solution of the given equation must lie. Such a method is useful for two reasons: (i) By obtaining bounds on the solution, one can draw conclusions about the qualitative behavior of the solution, and the labor involved is considerably less than that needed to find an exact solution. (ii) The bounds obtained by this method can serve as a check on approximate solutions obtained by other means, e.g., numerical solution using a computer.

The study of nonlinear o.d.e.'s in general terms requires rather advanced mathematical tools. The first two sections of this chapter are devoted to developing these tools.

2.1 MATHEMATICAL PRELIMINARIES

This section contains an introduction to several concepts that are used subsequently, such as linear vector spaces, normed linear spaces, Banach and Hilbert spaces, convergence, and continuity.

2.1.1 Linear Vector Spaces

This subsection is devoted to an axiomatic development of linear vector spaces, both real and complex. In most practical situations, it is enough to deal with real vector spaces. However, it is sometimes necessary to deal with complex vector spaces in order to make the theory complete. For example, a polynomial of degree n has n zeros only if one counts complex zeros.

Note that it is also possible to define a linear vector space over an arbitrary field (e.g., the binary field, the field of rational functions, etc.). However, such generality is not needed in this book.

1 Definition *A* **real linear vector space** (*respectively, a* **complex linear vector space**) *is a set V together with two operations: the addition operation* $+: V\times V \to V$ *and the multiplication operation* $\cdot: \mathbf{R}\times V \to V$ *(respectively* $\cdot: C\times V \to V$*), such that the following axioms hold:*

(V1) $x+y=y+x$, $\forall x, y\in V$ *(commutativity of addition).*

(V2) $x+(y+z)=(x+y)+z$, $\forall x, y, z\in V$ *(associativity of addition).*

(V3) *There is an element* 0_V *in V such that* $x+0_V=0_V+x=x$, $\forall x\in V$ *(existence of additive identity).*

(V4) *For each* $x\in V$*, there exists an element denoted by* $-x\in V$ *such that* $x+(-x)=0_V$ *(existence of additive inverse).*

(V5) *For each* $r_1, r_2\in\mathbf{R}$ *(respectively,* $c_1, c_2\in C$*), and each* $x\in V$*, we have that* $r_1\cdot(r_2\cdot x)=(r_1r_2)\cdot x$ *[respectively* $c_1\cdot(c_2x)=(c_1c_2)\cdot x$*].*

(V6) *For each* $r\in\mathbf{R}$ *(respectively* $c\in C$*) and each* $x, y\in V$*, we have* $r\cdot(x+y)=r\cdot x+r\cdot y$ *[respectively* $c\cdot(x+y)=c\cdot x+c\cdot y$*].*

(V7) *For each* $r_1, r_2\in\mathbf{R}$ *(respectively, for each* $c_1, c_2\in C$*) and each* $x\in V$*, we have* $(r_1+r_2)\cdot x=r_1\cdot x+r_2\cdot x$.

(V8) *For each* $x\in V$*, we have* $1\cdot x=x$.

This axiomatic definition of a linear vector space is illustrated by several examples.

2 Example The set $\mathbf{R}^n$, consisting of all ordered n-tuples of real numbers, becomes a *real* linear vector space if addition and scalar multiplication are defined as follows: If $\mathbf{x}=(x_1, \cdots, x_n)$, $\mathbf{y}=(y_1, \cdots, y_n)\in\mathbf{R}^n$ and r is a real number, then

$$\mathbf{x}+\mathbf{y}=(x_1+y_1, \cdots, x_n+y_n),$$

$$r\cdot\mathbf{x}=(rx_1, \cdots, rx_n).$$

In other words, the sum of two n-tuples is obtained by component-wise addition, while the product of a real number and an n-tuple is obtained by multiplying each component of the n-tuple by the real number.

As a limiting case, it is interesting to note that $\mathbf{R}^1=\mathbf{R}$, the set of real numbers, is itself a real linear vector space.

Now let C^n denote the set of all ordered n-tuples of complex numbers. By defining addition and scalar multiplication as above, one can make C^n into either a real linear vector space or a complex linear vector space, depending on the set of values to which the "scalar" r

is restricted to belong. This shows that whether a linear vector space is real or complex is determined, not by the nature of the elements of the space, but by whether the associated set of scalars is the field of real numbers or the field of complex numbers.

3 Example Let $F[a, b]$ denote the set of all real-valued functions defined over an interval $[a, b]$ in $\mathbf{R}$. Thus a typical element of $F[a, b]$ is a function $f(\cdot)$ mapping $[a, b]$ into $\mathbf{R}$. The set $F[a, b]$ becomes a real linear vector space if addition and scalar multiplication are defined as follows: Let $x(\cdot)$ and $y(\cdot)$ be two functions in $F[a, b]$ and let $r \in \mathbf{R}$. Then $x + y$ is the function defined by

$$(x+y)(t) = x(t) + y(t), \ \forall t \in [a, b],$$

$$(r \cdot x)(t) = rx(t), \ \forall t \in [a, b].$$

Thus the sum of two functions is obtained by point-wise addition and the multiple of a scalar and a function is obtained by point-wise multiplication.

If one thinks of an n-tuple as a function mapping the finite set $\{1, \cdots, n\}$ into $\mathbf{R}$, then one can see that the definition of addition and multiplication in $F[a, b]$ are entirely analogous to those in $\mathbf{R}^n$.

4 Example The set $F^n[a, b]$ consisting of all functions mapping the interval $[a, b]$ into the set $\mathbf{R}^n$ defined in Example (2) is a linear vector space if addition and scalar multiplication are defined as follows: Suppose $\mathbf{x}(\cdot)$ and $\mathbf{y}(\cdot)$ are functions in $F^n[a, b]$ and that $r \in \mathbf{R}$. Then

$$(\mathbf{x}+\mathbf{y})(t) = \mathbf{x}(t) + \mathbf{y}(t), \ \forall t \in [a, b],$$

$$(r \cdot \mathbf{x})(t) = r \cdot \mathbf{x}(t), \ \forall t \in [a, b].$$

Note that the addition and the scalar multiplication on the right side are in accordance with Example (2).

5 Example Let S denote the set of all complex-valued sequences $\{x_i\}_{i=0}^{\infty}$. Then S can be made into either a real or a complex linear vector space, by appropriate choice of the associated set of scalars, if addition and scalar multiplication are defined as follows: Let $x = \{x_i\}$ and $y = \{y_i\}$ be elements of the set S and suppose $r \in \mathbf{R}$. Then

$$(x+y)_i = x_i + y_i, \ \forall i,$$

$$(r \cdot x)_i = rx_i, \ \forall i.$$

If one thinks of a sequence as a function from the set of nonnegative integers into the set C, then one can see that the linear vector space in the present example is entirely analogous to both C^n and to $F[a, b]$.

6 Definition *A subset M of a linear vector space V is called a* **subspace** *of V if M satisfies two conditions:*

1. *If $x, y \in M$, then $x + y \in M$.*
2. *If $x \in M$, $r \in \mathbf{R}$ or C, then $r \cdot x \in M$.*

Roughly speaking, M is a subspace of V if it is a linear vector space in its own right.

7 Example Let $F[a, b]$ be as in Example (3). Let $t_0 \in [a, b]$, and let $F_{t_0}[a, b]$ denote the subset of $F[a, b]$ consisting of all functions $x(\cdot)$ in $F[a, b]$ such that $x(t_0) = 0$. In other words, $F_{t_0}[a, b]$ consists of all functions in $F[a, b]$ that vanish at t_0. Then $F_{t_0}[a, b]$ is a subspace of $F[a, b]$.

2.1.2 Normed Linear Spaces

The concept of a linear vector space is a very useful one, because in that setting it is possible to define many of the standard concepts that are useful in engineering such as linear operators, and linear dependence. It is also possible to study the existence and uniqueness of solutions to linear (algebraic) equations. However, the limitation is that there is no notion of distance or proximity in a linear vector space. Hence it is not possible to discuss concepts such as convergence or continuity. This limitation is the motivation for introducing the notion of a normed linear space, which is basically a linear vector space with a measure of the "length" of a vector.

8 Definition *A* **normed linear space** *is an ordered pair $(X, \|\cdot\|)$ where X is a linear vector space and $\|\cdot\| : X \to \mathbf{R}$ is a real-valued function defined on X such that the following axioms hold:*

(N1) $\|x\| \geq 0$, $\forall x \in X$; $\|x\| = 0$ *if and only if* $x = 0_X$.

(N2) $\|\alpha x\| = |\alpha| \cdot \|x\|$, $\forall x \in X$, $\forall \alpha \in \mathbf{R}$ *or* C.

(N3) $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in X$.

The norm on a normed linear space is a natural generalization of the length of a vector on $\mathbf{R}^2$ or $\mathbf{R}^3$. Thus, given a vector x in a normed linear space $(X, \|\cdot\|)$, the nonnegative number $\|x\|$ can be thought of as the length of the vector x. Axiom (N1) states that only the zero vector has zero length, and that every other vector has positive length. Axiom (N2) states that if a vector is "scaled" by multiplying it by a scalar, then the length of the vector gets "scaled" by multiplying it by the magnitude of the scalar. The condition in (N3) is known as the *triangle inequality,* and states that the length of the sum of two vectors is no larger than the sum of their lengths.

9 Example Consider the linear vector space $\mathbf{R}^n$, together with the function $\|\cdot\|_\infty : \mathbf{R}^n \to \mathbf{R}_+$ defined by

10 $\|\mathbf{x}\|_\infty = \max_{1 \le i \le n} |x_i|.$

(The reason for the subscript ∞ will become clear later.) The function $\|\cdot\|_\infty$ satisfies axioms (N1) through (N3), as can be easily verified. In fact, (N1) and (N2) can be verified by inspection. To verify (N3), suppose $\mathbf{x} = (x_1, \cdots, x_n)$ and $\mathbf{y} = (y_1, \cdots, y_n) \in \mathbf{R}^n$. We know, by the triangle inequality for real numbers, that

$$|x_i + y_i| \le |x_i| + |y_i|, \forall i.$$

Therefore

$$\|\mathbf{x} + \mathbf{y}\|_\infty = \max_i |x_i + y_i|$$

$$\le \max_i |x_i| + \max_i |y_i| = \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty,$$

so that (N3) is satisfied. Thus the pair $(\mathbf{R}^n, \|\cdot\|_\infty)$ is a normed linear space. The norm $\|\cdot\|_\infty$ is called the l_∞-norm on $\mathbf{R}^n$.

11 Example Consider once again the linear vector space $\mathbf{R}^n$, but this time with the function $\|\cdot\|_1 : \mathbf{R}^n \to \mathbf{R}_+$ defined by

12 $$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|.$$

Clearly $\|\cdot\|_1$ also satisfies (N1) and (N2). To verify (N3), suppose $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$. Then

$$\|\mathbf{x} + \mathbf{y}\|_1 = \sum_{i=1}^n |x_i + y_i| \le \sum_{i=1}^n (|x_i| + |y_i|)$$

$$= \sum_{i=1}^n |x_i| + \sum_{i=1}^n |y_i| = \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1.$$

Hence the pair $(\mathbf{R}^n, \|\cdot\|_1)$ is also a normed linear space. The norm $\|\cdot\|_1$ is called the l_1-norm on $\mathbf{R}^n$.

It is important to note that, even though the underlying linear vector space is the same in Examples (9) and (11), the normed linear space $(\mathbf{R}^n, \|\cdot\|_1)$ is a different entity from the normed linear space $(\mathbf{R}^n, \|\cdot\|_\infty)$.

13 Example Consider once again the linear vector space $\mathbf{R}^n$, together with the function $\|\cdot\|_p : \mathbf{R}^n \to \mathbf{R}_+$ defined by

14 $$\|\mathbf{x}\|_p = [\sum_{i=1}^n |x_i|^p]^{1/p},$$

where p is any number in the interval $[1, \infty]$. If $p = 1$, then $\|\cdot\|_p$ becomes the norm function

of Example (11), whereas if $p \to \infty$, then $\|\cdot\|_p$ approaches the norm function of Example (9). [This is the reason for the subscripts in Examples (9) and (11).] The function $\|\cdot\|_p$ clearly satisfies the conditions (N1) and (N2), and can be shown to satisfy (N3) whenever $1 \leq p \leq \infty$. Thus the pair $(\mathbf{R}^n, \|\cdot\|_p)$ is a normed linear space for each value of p in the interval $[1, \infty]$; of course, for distinct values of p we have distinct normed linear spaces. The norm $\|\cdot\|_p$ is called the l_p-norm on $\mathbf{R}^n$.

In particular, if $p = 2$, then

$$\mathbf{15} \qquad \|\mathbf{x}\|_2 = [\sum_{i=1}^{n} |x_i|^2]^{1/2},$$

which is generally called the *Euclidean norm* on $\mathbf{R}^n$. It is also called the l_2-norm on $\mathbf{R}^n$. The Euclidean norm is a particular example of a so-called inner product norm, which is defined in Section 2.1.3.

The norm $\|\cdot\|_p$ can also be defined on the set C^n in an entirely analogous fashion, simply by interpreting the quantity $|x_i|$ in (14) as the magnitude of the complex number $|x_i|$. Thus the pair $(C^n, \|\cdot\|_p)$ is also a normed linear space for each $p \in [1, \infty]$.

Both $\mathbf{R}^n$ and C^n are examples of *finite-dimensional* linear vector spaces. As a consequence, it can be shown that, given *any* two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbf{R}^n$, there exist constants k_1 and k_2 such that

$$k_1 \|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq k_2 \|\mathbf{x}\|_a, \ \forall \mathbf{x} \in \mathbf{R}^n \text{ (or } C^n).$$

For instance,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n \|\mathbf{x}\|_\infty, \ \forall \mathbf{x} \in \mathbf{R}^n,$$

and

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq n^{1/2} \|\mathbf{x}\|_\infty, \ \forall \mathbf{x} \in \mathbf{R}^n.$$

A similar relationship exists between any two norms on $\mathbf{R}^n$ and C^n. ■

Suppose $(X, \|\cdot\|)$ is a normed linear space, and that $x, y \in X$. Then one can think of the quantity $\|x - y\|$ as the distance between x and y. With the aid of this notion of distance (or proximity), it is possible to define the notion of convergence in a normed linear space setting.

16 Definition *A sequence $\{x_i\}_{i=0}^{\infty}$ in a normed linear space $(X, \|\cdot\|)$ is said to* **converge** *to $x_0 \in X$ if, for every $\varepsilon > 0$, there exists an integer $N = N(\varepsilon)$ such that*

$$\mathbf{17} \qquad \|x_i - x_0\| < \varepsilon, \ \forall i \geq N.$$

The basic definition of convergence can be interpreted in many ways. The sequence of "vectors" $\{x_i\}$ converges to x_0 if and only if the sequence of real numbers $\{\|x_i - x_0\|\}$ converges to 0. Alternatively, let $B(x_0, \varepsilon)$ denote the ball in X defined by

18 $$B(x_0, \varepsilon) = \{x \in X: \|x - x_0\| < \varepsilon\}.$$

Then the sequence $\{x_i\}$ converges to x_0 if and only if, for each positive ε, the ball $B(x_0, \varepsilon)$ contains all but a finite number of elements of the sequence $\{x_i\}$.

Definition (16) gives a means for testing whether or not a given sequence $\{x_i\}$ converges to a given element $x_0 \in X$. In other words, to test for convergence using Definition (16), it is necessary to have at hand a candidate for the limit of the sequence. However, in many cases we generate a sequence $\{x_i\}$ without knowing to what, if anything, it might converge. Thus it is desirable to have a criterion for convergence that does not involve a candidate for the limit in an explicit fashion. This is provided by the concept of a Cauchy sequence.

19 Definition *A sequence $\{x_i\}$ in a normed linear space $(X, \|\cdot\|)$ is said to be a* **Cauchy sequence** *if, for every $\varepsilon > 0$, there exists an integer $N = N(\varepsilon)$ such that*

20 $$\|x_i - x_j\| < \varepsilon, \text{ whenever } i, j \geq N.$$

Thus a sequence is *convergent* if its terms approach arbitrarily closely a *fixed* element, whereas a sequence is *Cauchy* if its terms approach *each other* arbitrarily closely. The relationship between convergent sequences and Cauchy sequences is brought out next.

21 Lemma *Every convergent sequence in a normed linear space is a Cauchy sequence.*

Proof Suppose $\{x_i\}$ is a convergent sequence in a normed linear space $(X, \|\cdot\|)$, and denote its limit by x_0. To prove that the sequence is also a Cauchy sequence, suppose $\varepsilon > 0$ is given; then pick an integer N such that

22 $$\|x_i - x_0\| < \varepsilon/2, \forall i \geq N.$$

Such an integer N exists, by Definition (16). Then, whenever $i, j \geq N$, it follows from the triangle inequality that

23 $$\|x_i - x_j\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus $\{x_i\}$ is a Cauchy sequence. ■

Lemma (21) shows that if the elements of a sequence are getting closer and closer to a fixed element, then in the process they must also be getting closer and closer to each other. One can ask whether the converse is true: If the elements of a sequence are getting closer and closer to each other, are they in fact getting closer and closer to a fixed element? In general, the answer is no. But some normed linear spaces have the special property that every Cauchy sequence in them is also convergent. This property is so important that such spaces

are given a special name.

24 Definition *A normed linear space* $(X, \|\cdot\|)$ *is said to be a* **complete** *normed linear space, or a* **Banach space** *if every Cauchy sequence in* $(X, \|\cdot\|)$ *converges to an element of* X.

Banach spaces are important for two reasons: (i) If $(X, \|\cdot\|)$ is a Banach space, then every Cauchy sequence is convergent. This property provides a means of testing whether a sequence is convergent without having at hand a candidate for the limit of the sequence. (ii) Even if a particular normed linear space $(X, \|\cdot\|)$ is not complete, it can be made into a Banach space by adding some elements; for obvious reasons, this process is known as "completing" the space. Thus, in most situations, it can be assumed without loss of generality that the normed space at hand is complete.

25 Example Let $[a, b]$ be a *bounded* interval in $\mathbf{R}$, and let $C[a, b]$ denote the set of all continuous functions mapping the interval $[a, b]$ into $\mathbf{R}$. Define a function $\|\cdot\|_C: C[a, b] \to \mathbf{R}_+$ as follows: If $x(\cdot) \in C[a, b]$, then

$$\mathbf{26} \qquad \|x(\cdot)\|_C = \max_{t \in [a, b]} |x(t)|.$$

Since the interval $[a, b]$ is assumed to be bounded, the maximum on the right side is well-defined and is finite for each $x(\cdot) \in C[a, b]$. Now it is easy to verify that the function $\|\cdot\|_C$ verifies axioms (N1) and (N2). To verify axiom (N3), suppose $x(\cdot), y(\cdot) \in C[a, b]$. Then

$$\|x(\cdot) + y(\cdot)\|_C = \max_t |x(t) + y(t)| \leq \max_t (|x(t)| + |y(t)|)$$

$$\leq \max_t |x(t)| + \max_t |y(t)| = \|x(\cdot)\|_C + \|y(\cdot)\|_C,$$

where all maxima are taken over $[a, b]$ Thus the pair $(C[a, b], \|\cdot\|_C)$ is a normed linear space. The norm $\|\cdot\|_C$ is called the "sup" norm (for "supremum").

Note that a sequence of functions $\{x_i(\cdot)\}$ in $C[a, b]$ converges to a function $x(\cdot) \in C[a, b]$ if and only if the sequence of real numbers $\{x_i(t)\}$ converges to $x(t)$ uniformly for all $t \in [a, b]$. Now we know from advanced calculus that if each of the original functions $x_i(\cdot)$ is continuous and the convergence is uniform, then the limit function is also continuous. Thus the space $C[a, b], \|\cdot\|_C)$ is a Banach space. ■

The notion of distance in a normed linear space enables us to define continuity of functions.

27 Definition *Let* $(X, \|\cdot\|_X)$ *and* $(Y, \|\cdot\|_Y)$ *be two normed linear spaces, and suppose* $f: X \to Y$. *Then the function* f *is said to be* **continuous at** $x_0 \in X$ *if, for every* $\varepsilon > 0$, *there exists a* $\delta = \delta(\varepsilon, x_0)$ *such that*

28 $\|f(x)-f(x_0)\|_Y < \varepsilon$, whenever $\|x-x_0\|_X < \delta$.

f is said to be **continuous** *if it is continuous at all $x \in X$. Finally, f is said to be* **uniformly continuous** *if, for every $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that*

29 $\|f(x)-f(y)\|_Y < \varepsilon$, whenever $\|x-y\|_X < \delta$.

The concept of a continuous function from one normed linear space to another is a natural extension of the concept of a continuous real-valued function of a real variable. In a general normed linear space setting, the norm plays the same role as the absolute value does in the set of real numbers. The important difference between continuity and *uniform* continuity is that in the latter case δ depends only on ε and not on x.

It is fairly easy to show that if $f: X \to Y$ is continuous at $x_0 \in X$, and if $\{x_i\}$ is a sequence in X converging to x_0, then the sequence $\{f(x_i)\}$ in Y converges to $f(x_0)$; see Problem 2.9.

The next example combines several of the concepts presented thus far.

30 Example Suppose $\|\cdot\|$ is a given norm on $\mathbf{R}^n$, and let $C^n[a, b]$ denote the set of all continuous functions mapping the interval $[a, b]$ into $\mathbf{R}^n$, where $[a, b]$ is a bounded interval in $\mathbf{R}$. Define the function $\|\cdot\|_C: C^n[a, b] \to \mathbf{R}_+$ as follows: If $\mathbf{x}(\cdot) \in C^n[a, b]$, then

31 $$\|\mathbf{x}(\cdot)\|_C = \max_{t \in [a, b]} \|\mathbf{x}(t)\|.$$

To show that $\|\cdot\|_C$ is a norm on $C^n[a, b]$, one proceeds exactly as in Example (25). Axioms (N1) and (N2) are readily verified. To verify (N3), suppose $\mathbf{x}(\cdot)$ and $\mathbf{y}(\cdot)$ belong to $C^n[a, b]$. Then

$$\begin{aligned}
\|\mathbf{x}(\cdot)+\mathbf{y}(\cdot)\|_C &= \max_t \|\mathbf{x}(t)+\mathbf{y}(t)\| \\
&\leq \max_t \{\|\mathbf{x}(t)\| + \|\mathbf{y}(t)\|\} \text{ from the triangle inequality on } \mathbf{R}^n \\
&\leq \max_t \|\mathbf{x}(t)\| + \max_t \|\mathbf{y}(t)\| \\
&= \|\mathbf{x}(\cdot)\|_C + \|\mathbf{y}(\cdot)\|_C,
\end{aligned}$$

where all maxima are taken as t varies over the interval $[a, b]$. Thus (N3) is satisfied and $\|\cdot\|_C$ is a norm on $C^n[a, b]$. By the same reasoning as in Example (25), one can see that the pair $(C^n[a, b], \|\cdot\|_C)$ is a Banach space.

In this example, it is essential to note the difference between $\|\cdot\|$ and $\|\cdot\|_C$; $\|\cdot\|$ is a norm on $\mathbf{R}^n$, while $\|\cdot\|_C$ is a norm on the space $C^n[a, b]$. The former has an n-vector as its argument, while the latter has a vector-valued function as its argument. When we study nonlinear differential equations in Section 2.4, this difference becomes crucial.

2.1.3 Inner Product Spaces

An inner product space is a special type of normed linear space in which it is possible to define geometrically appealing concepts such as orthogonality and Fourier series. An inner product space can be defined axiomatically as follows:

32 Definition *An* **inner product space** *is a linear vector space X with associated field F, together with a function $<\cdot,\cdot>: X\times X \to F$ such that the following axioms are satisfied:*

(I1) $<x, y> = <y, x>$ *if* $F = \mathbf{R}$, $<x, y> = \overline{<y, x>}$ *if* $F = C$, $\forall x, y \in X$.

(I2) $<x, y+z> = <x, y> + <x, z>$, $\forall x, y, z \in X$.

(I3) $<x, \alpha y> = \alpha <x, y>$, $\forall \alpha \in F$, $\forall x, y \in X$.

(I4) $<x, x> \geq 0$, $\forall x \in X$; $<x, x> = 0$ *if and only if* $x = 0_X$.

The quantity $<x, y>$ is an abstraction of the familiar scalar product or dot product on $\mathbf{R}^2$ or $\mathbf{R}^3$.

An inner product space can be made into a normed linear space in a natural way.

33 Theorem *Given an inner product space $(X, <\cdot,\cdot>)$, define the function $\|\cdot\|: X \to \mathbf{R}$ by*

34 $\|x\| = <x, x>^{1/2}$.

Then $\|\cdot\|$ is a norm on X, so that the pair $(X, \|\cdot\|)$ is a normed linear space.

The proof of Theorem (33) depends on the following extremely useful inequality, known as Schwarz' inequality.

35 Lemma (Schwarz' Inequality) *Let x, y belong to the inner product space $(X, <\cdot,\cdot>)$. Then*

36 $|<x, y>| \leq \|x\| \cdot \|y\|$,

and

37 $|<x, y>| = \|x\| \cdot \|y\|$

if and only if the elements x, y are linearly dependent, i.e., there exist scalars $\alpha, \beta \in F$, not both zero, such that $\alpha x + \beta y = 0_X$.

Proof of Lemma (35) The proof is only given for the case of a real linear vector space; the case where $F = C$ is quite similar and is left as an exercise.

Consider the function

38 $$f(\alpha, \beta) = \|\alpha x + \beta y\|^2 = \langle \alpha x + \beta y, \alpha x + \beta y \rangle$$

$$= \alpha^2 \|x\|^2 + 2\alpha\beta \langle x, y \rangle + \beta^2 \|y\|^2.$$

By Axiom (I4), we have that $f(\alpha, \beta) \geq 0$ for all scalars α, β. Since f is a quadratic form in these two scalars, it follows that $f(\alpha, \beta) \geq 0 \; \forall \alpha, \beta$ if and only if the discriminant of the quadratic form is nonpositive, i.e.,

39 $$\langle x, y \rangle^2 \leq \|x\|^2 \cdot \|y\|^2.$$

Taking square roots of both sides proves (36). Now suppose the vectors x and y are linearly independent, i.e., that $\alpha x + \beta y \neq 0$ whenever not both α and β are zero. Then $f(\alpha, \beta) > 0$ whenever either α or β is nonzero. This is true if and only if the discriminant of the quadratic form in (38) is negative, i.e., if

40 $$\langle x, y \rangle^2 < \|x\|^2 \cdot \|y\|^2.$$

Taking square roots of both sides proves (37).

Proof of Theorem (33) One can verify by inspection that $\|\cdot\|$ satisfies Axioms (N1) and (N2). To verify (N3), suppose $x, y \in X$. Then

41 $$\|x + y\|^2 = \langle x + y, x + y \rangle = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle$$

$$\leq \|x\|^2 + \|y\|^2 + 2\|x\| \cdot \|y\| \text{ (by Schwarz' inequality)}$$

$$= (\|x\| + \|y\|)^2.$$

Taking square roots of both sides establishes the triangle inequality. ■

Theorem (33) shows that every inner product space can be made into a normed linear space in a natural way. Hence it makes sense to ask whether an inner product space is complete (in the norm defined by the inner product).

41 Definition *An inner product space which is complete in the norm defined by the inner product is called a* **Hilbert space.**

43 Example Consider the linear vector space $\mathbf{R}^n$, together with the function $\langle \cdot, \cdot \rangle : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ defined by

44 $$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i.$$

It is routine to verify that the function in (44) satisfies all four axioms of an inner product. The norm on $\mathbf{R}^n$ defined by the inner product is

$$\|\mathbf{x}\| = [\sum_{i=1}^{n} |x_i|^2]^{1/2},$$

which is recognized as the l_2-norm defined in Example (13). Note that $\mathbf{R}^n$ together with the inner product defined in (44) is in fact a Hilbert space. ■

45 Example Let $C^n[a, b]$ be the linear space of Example (30), and define the inner product $<\cdot,\cdot>_C$ on this space as follows: If $\mathbf{x}(\cdot), \mathbf{y}(\cdot) \in C^n[a, b]$, then let

$$\mathbf{46} \quad <\mathbf{x}(\cdot), \mathbf{y}(\cdot)>_C = \int_a^b <\mathbf{x}(t), \mathbf{y}(t)> dt,$$

where the inner product inside the integral is that on $\mathbf{R}^n$ defined in Example (43). Once again the function defined in (46) satisfies all the axioms of the inner product. However, with this inner product, $C^n[a, b]$ is *not* a Hilbert space; contrast this with the fact that $C^n[a, b]$ *is* a Banach space with the norm $\|\cdot\|_C$ defined in Example (30). To see that $C^n[a, b]$ is not a Hilbert space with the inner product in (46), pick a time T such that $a < T < b$, and consider the function $\mathbf{y}(\cdot)$ defined on $[a, b]$ by

$$y_i(t) = \begin{cases} 0, & \text{if } a \le t \le T, \\ 1, & \text{if } T < t \le b. \end{cases}$$

Define the Fourier series expansion of $\mathbf{y}(t)$ in the familiar fashion, namely

$$\mathbf{y}(t) = \sum_{l=0}^{\infty} \mathbf{p}_l \sin l\omega t + \mathbf{q}_l \cos l\omega t,$$

where $\omega = 2\pi/(b-a)$. Then the Fourier series above converges to the *discontinuous* function $\mathbf{y}(\cdot)$ in the mean-squared sense, i.e., in the sense of the norm defined by the inner product of (46). Thus the partial sums of the Fourier series constitute a Cauchy sequence in the space $C^n[a, b]$ which does not converge (to an element of the space in question). Hence $C^n[a, b]$ is not a Hilbert space, even though it is an inner product space.

The completion of $C^n[a, b]$ under the norm corresponding to the inner product (46) is the space of Lebesgue-measurable, square-integrable functions mapping $[a, b]$ into $\mathbf{R}^n$, and is denoted by $L_2^n[a, b]$. The inner product on $L_2^n[a, b]$ is also defined by (46), except that the integral must now be interpreted as a Lebesgue integral. ■

This section is concluded with two useful examples of continuous functions.

47 Lemma *Let $(X, \|\cdot\|)$ be a normed linear space. Then the norm function $\|\cdot\| : X \to \mathbf{R}$ is uniformly continuous.*

Proof Use Definition (27) of uniform continuity. Given any $\varepsilon > 0$, let $\delta(\varepsilon) = \varepsilon$. To show that the definition is satisfied with this choice, suppose $x, y \in X$ and that

48 $\big|\, \|x-y\| \,\big| < \delta = \varepsilon.$

Then

49 $\big|\, \|x\| - \|y\| \,\big| < \|x-y\| < \varepsilon.$

This completes the proof. ■

50 Corollary *Suppose that* $(X, \|\cdot\|)$ *is a normed linear space, and that* $\{x_i\}$ *is a sequence in* X *converging to* $x_0 \in X$. *Then the sequence of real numbers* $\{\|x_i\|\}$ *converges to* $\|x_0\|$.

51 Lemma *Suppose* $(X, <\cdot,\cdot>)$ *is an inner product space. Then, for each* $y \in X$, *the function mapping* x *into* $<x,y>: X \to \mathbf{R}$ *is uniformly continuous.*

Proof If $y = 0$, then $<x, 0> = 0 \; \forall x \in X$, which is clearly a uniformly continuous function, so it is only necessary to study the case where $y \neq 0$. Use Definition (27) of uniform continuity, and given $\varepsilon > 0$, define $\delta(\varepsilon) = \varepsilon / \|y\|$. Now suppose

52 $x, z \in X$, and $\|x-z\| < \delta = \dfrac{\varepsilon}{\|y\|}$.

Then

53
$$|<x, y> - <z, y>| = |<x-z, y>|$$
$$\leq \|x-z\| \cdot \|y\|, \text{ by Schwarz' inequality}$$
$$< \frac{\varepsilon}{\|y\|} \cdot \|y\| = \varepsilon.$$

This completes the proof. ■

Problem 2.1 Show that the zero element of a linear vector space is unique. [Hint: Assume that the linear vector space V has two zero elements 0_1 and 0_2, and use Axiom (V3).]

Problem 2.2 Show that, in a linear vector space, the additive inverse of an element is unique.

Problem 2.3 Give an example of a set which is *not* a linear vector space.

Problem 2.4 Let S be the sequence space of Example (5), and define a subset S_r of S as the set of all sequences converging to r. For what values of r is S_r a subspace of S?

Problem 2.5 Consider the normed linear space $\mathbf{R}^2$, with the norm $\|\cdot\|_p$ defined in Example (13). Sketch the unit spheres, i.e., the sets

$$\{\mathbf{x} \in \mathbf{R}^2 : \| x \|_p = 1\}$$

for the values $p = 1, 2, 5, \infty$.

Problem 2.6 (a) Let $\| \cdot \|$ be any norm on $\mathbf{R}^n$, and let $\mathbf{x}_1, \cdots, \mathbf{x}_m$ be any collection of vectors in $\mathbf{R}^n$. Using the triangle inequality, show that

$$\| \sum_{i=1}^{m} \mathbf{x}_i \| \leq \sum_{i=1}^{m} \| \mathbf{x}_i \|.$$

(b) Let $C^n[a, b]$ be as in Example (30). Using the Riemannian approximation to the integral, show that

$$\| \int_a^b \mathbf{x}(t)\, dt \| \leq \int_a^b \| \mathbf{x}(t) \|\, dt.$$

Problem 2.7 Prove Schwarz' inequality for complex inner product spaces.

Problem 2.8 Suppose $(X, <\cdot,\cdot>)$ is an inner product space. Show that the inner product function is jointly continuous in its two arguments; i.e., show that if $\{x_i\}$, $\{y_i\}$ are two sequences in X converging respectively to x_0 and y_0, then the sequence of real numbers $\{<x_i, y_i>\}$ converges to $<x_0, y_0>$. [Hint: Write

$$<x_i, y_i> - <x_0, y_0> = <x_i, y_i> - <x_0, y_i> + <x_0, y_i> - <x_0, y_0>,$$

and use Schwarz' inequality.]

Problem 2.9 Suppose X and Y are normed linear spaces and that $f : X \to Y$ is continuous at $x_0 \in X$. Suppose $\{x_i\}$ is a sequence in X converging to x_0. Show that the sequence $\{f(x_i)\}$ in Y converges to $f(x_0)$.

2.2 INDUCED NORMS AND MATRIX MEASURES

In this section the concepts of the induced norm of a matrix and the measure of a matrix are introduced. These concepts are used in Section 2.5 to derive estimates for the solutions of nonlinear differential equations, without actually solving them.

2.2.1 Induced Norms

Let $C^{n\times n}$ (respectively, $\mathbf{R}^{n\times n}$) denote the set of all $n \times n$ matrices with complex (respectively, real) elements. Then $C^{n\times n}$ can be made into a complex linear vector space if addition and scalar multiplication are done componentwise. Moreover, for each matrix $\mathbf{A} \in C^{n\times n}$ there is a corresponding linear mapping α from C^n into itself, defined by

$$\mathbf{1} \qquad \alpha(\mathbf{x}) = \mathbf{A}\mathbf{x}, \ \forall \mathbf{x} \in C^n.$$

Conversely, for every linear mapping α from C^n into itself, there is a corresponding matrix

$\mathbf{A} \in C^{n\times n}$ such that (1) holds. Thus there is a one-to-one correspondence between matrices in $C^{n\times n}$ and linear mappings mapping C^n into itself. (Actually, this correspondence is one-to-one only after the basis on C^n has been chosen. However, in this book such subtleties of linear algebra are not explored.) We do not in general distinguish between a matrix in $C^{n\times n}$ and the corresponding linear mapping on C^n. However, this correspondence is the motivation behind the concept of the induced norm of a matrix.

2 Definition *Let $\|\cdot\|$ be a given norm on C^n. Then for each matrix $\mathbf{A} \in C^{n\times n}$, the quantity $\|\mathbf{A}\|_i$, defined by*

$$\mathbf{3} \qquad \|\mathbf{A}\|_i = \sup_{\mathbf{x}\neq\mathbf{0},\, \mathbf{x}\in C^n} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = \sup_{\|\mathbf{x}\|\leq 1} \|\mathbf{A}\mathbf{x}\|,$$

is called the **induced (matrix) norm** *of* $\mathbf{A}$ *corresponding to the vector norm* $\|\cdot\|$.

It should be noted that there are two distinct functions involved in Definition (2); one is the *norm* function $\|\cdot\|$ mapping C^n into $\mathbf{R}$, and the other is the *induced norm* function $\|\cdot\|_i$ mapping $C^{n\times n}$ into $\mathbf{R}$.

The induced norm of a matrix can be given a simple geometric interpretation. Equation (3) shows that $\|\mathbf{A}\|_i$ is the least upper bound of the ratio $\|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\|$ as $\mathbf{x}$ varies over C^n. In this sense, $\|\mathbf{A}\|_i$ can be thought of as the "gain" of the linear mapping corresponding to $\mathbf{A}$. Alternatively, let B denote the closed unit ball in C^n; i.e., let

$$\mathbf{4} \qquad B = \{\mathbf{x} \in C^n : \|\mathbf{x}\| \leq 1\}.$$

Now suppose we distort B by replacing each $\mathbf{x}$ in B by $\mathbf{A}\mathbf{x}$, i.e., its image under the mapping $\mathbf{A}$. Then what results is the image of the set B under the mapping $\mathbf{A}$. In this setting, the induced norm $\|\mathbf{A}\|_i$ of $\mathbf{A}$ can be thought of as the radius of the smallest ball in C^n that completely covers the image of B under $\mathbf{A}$.

Lemma (5) shows that the function $\|\cdot\|_i$ is a valid norm on $C^{n\times n}$.

5 Lemma *For each norm $\|\cdot\|$ on C^n, the induced norm function $\|\cdot\|_i$ maps $C^{n\times n}$ into $[0, \infty)$, satisfies Axioms (N1) through (N3), and is therefore a norm on $C^{n\times n}$.*

Proof It is clear that $\|\mathbf{A}\|_i \geq 0 \; \forall \mathbf{A} \in C^{n\times n}$, and Axioms (N1) and (N2) can be verified by inspection. To verify (N3), suppose $\mathbf{A}, \mathbf{B} \in C^{n\times n}$. Then

$$\mathbf{6} \qquad \|\mathbf{A}+\mathbf{B}\|_i = \sup_{\|\mathbf{x}\|=1} \|(\mathbf{A}+\mathbf{B})\mathbf{x}\| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{x}\|$$

$$\leq \sup_{\|\mathbf{x}\|=1} [\|\mathbf{A}\mathbf{x}\| + \|\mathbf{B}\mathbf{x}\|] \text{ by the triangle inequality on } C^n$$

$$\leq \sup_{\|\mathbf{x}\|\leq 1} \|\mathbf{A}\mathbf{x}\| + \sup_{\|\mathbf{x}\|\leq 1} \|\mathbf{B}\mathbf{x}\| = \|\mathbf{A}\|_i + \|\mathbf{B}\|_i.$$

Hence (N3) is also satisfied, and thus $\|\cdot\|_i$ is a norm on $C^{n\times n}$. ■

In view of Lemma (5), is clear that, for each norm on C^n, there is a corresponding induced norm on $C^{n\times n}$. However, the converse is not true. Consider the function $\|\cdot\|_s: C^{n\times n} \to \mathbf{R}$ defined by

$$\|\mathbf{A}\|_s = \max_{i,j} |a_{ij}|. \tag{7}$$

Then one can verify that $\|\cdot\|_s$ is a norm on $C^{n\times n}$. Indeed, $\|\mathbf{A}\|_s$ is simply the l_∞ norm of the $n^2\times 1$ vector consisting of all the components of the matrix $\mathbf{A}$. However, there is no norm on C^n such that $\|\cdot\|_s$ is the corresponding induced matrix norm. This is a consequence of the next result.

8 Lemma *Let $\|\cdot\|_i$ be an induced norm on $C^{n\times n}$. Then*

$$\|\mathbf{AB}\|_i \le \|\mathbf{A}\|_i \cdot \|\mathbf{B}\|_i, \ \forall \mathbf{A}, \mathbf{B} \in C^{n\times n}. \tag{9}$$

Proof By definition,

$$\|\mathbf{AB}\|_i = \sup_{\|\mathbf{x}\|=1} \|\mathbf{ABx}\|. \tag{10}$$

However, it follows from (3) that

$$\|\mathbf{Ay}\| \le \|\mathbf{A}\|_i \cdot \|\mathbf{y}\|, \ \forall \mathbf{y} \in C^n. \tag{11}$$

So in particular,

$$\|\mathbf{ABx}\| \le \|\mathbf{A}\|_i \cdot \|\mathbf{Bx}\|, \ \forall \mathbf{x} \in C^n. \tag{12}$$

Similarly,

$$\|\mathbf{Bx}\| \le \|\mathbf{B}\|_i \cdot \|\mathbf{x}\|, \ \forall \mathbf{x} \in C^n. \tag{13}$$

Combining (12) and (13) gives

$$\|\mathbf{ABx}\| \le \|\mathbf{A}\|_i \cdot \|\mathbf{B}\|_i \cdot \|\mathbf{x}\|, \ \forall \mathbf{x} \in C^n. \tag{14}$$

Now (9) follows immediately from (14). ■

Thus induced norms have the special feature that they are *submultiplicative;* i.e., the induced norm of the product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is less than or equal to the product of the induced norms of $\mathbf{A}$ and $\mathbf{B}$. It can be readily verified by example that the norm $\|\cdot\|_s$ of (7) does not have this property (and hence cannot be an induced norm).

In general, given a specific norm on C^n [say, for instance, the l_p-norm defined in Example (2.1.13)], it is not always easy to find an explicit expression for the corresponding induced norm on $C^{n\times n}$—the equations in (3) serve more as definitions than as computable expressions. However, the induced matrix norms corresponding to the vector norms $\|\cdot\|_\infty$, $\|\cdot\|_1$, and $\|\cdot\|_2$ [as defined in Examples (2.1.9), (2.1.11) and (2.1.13) respectively] are

known and are displayed in Table 2.1. Note that $\mathbf{A}^*$ denotes the conjugate transpose of the matrix $\mathbf{A}$, and $\lambda_{\max}(\mathbf{M})$ denotes the largest eigenvalue of the Hermitian matrix $\mathbf{M}$.

Table 2.1

Norm on C^n	Induced Norm on $C^{n\times n}$
$\|\mathbf{x}\|_\infty = \max_i \|x_i\|$	$\|\mathbf{A}\|_{i\infty} = \max_i \sum_{j=1}^{n} \|a_{ij}\|$
$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} \|x_i\|$	$\|\mathbf{A}\|_{i1} = \max_j \sum_{i=1}^{n} \|a_{ij}\|$
$\|\mathbf{x}\|_2 = (\sum_{i=1}^{n} \|x_i\|^2)^{1/2}$	$\|\mathbf{A}\|_{i2} = [\lambda_{\max}(\mathbf{A}^*\mathbf{A})]^{1/2}$

2.2.2 Matrix Measures

Let $\|\cdot\|_i$ be an induced matrix norm on $C^{n\times n}$. Then the corresponding **matrix measure** is the function $\mu(\cdot)$: $C^{n\times n} \to \mathbf{R}$ defined by

$$\mathbf{15} \quad \mu(\mathbf{A}) = \lim_{\varepsilon \to 0^+} \frac{\|I + \varepsilon \mathbf{A}\|_i - 1}{\varepsilon}.$$

Note that some authors use the term *logarithmic derivative* instead.

The measure of a matrix $\mu(\mathbf{A})$ can be thought of as the directional derivative of the induced norm function $\|\cdot\|_i$, as evaluated at the identity matrix I in the direction $\mathbf{A}$. The measure function has several useful properties, as shown next.

16 Theorem *Let $\|\cdot\|_i$ be an induced matrix norm on $C^{n\times n}$ and let $\mu(\cdot)$ be the corresponding matrix measure. Then $\mu(\cdot)$ has the following properties:*

(M1) *For each $\mathbf{A} \in C^{n\times n}$, the limit indicated in (15) exists and is well-defined.*

(M2) $-\|\mathbf{A}\|_i \le \mu(\mathbf{A}) \le \|\mathbf{A}\|_i$, $\forall \mathbf{A} \in C^{n\times n}$.

(M3) $\mu(\alpha\mathbf{A}) = \alpha\mu(\mathbf{A})$, $\forall \alpha \ge 0$, $\forall \mathbf{A} \in C^{n\times n}$.

(M4) $\max\{\mu(\mathbf{A}) - \mu(-\mathbf{B}), \mu(\mathbf{B}) - \mu(-\mathbf{A})\} \le \mu(\mathbf{A}+\mathbf{B}) \le \mu(\mathbf{A}) + \mu(\mathbf{B})$, $\forall \mathbf{A}, \mathbf{B} \in C^{n\times n}$.

(M5) $\mu(\cdot)$ *is a convex function; i.e.,*

$$\mu[\alpha\mathbf{A} + (1-\alpha)\mathbf{B}] \le \alpha\mu(\mathbf{A}) + (1-\alpha)\mu(\mathbf{B}), \quad \forall \alpha \in [0, 1], \ \forall \mathbf{A}, \mathbf{B} \in C^{n\times n}.$$

(M6) *If λ is an eigenvalue of $\mathbf{A} \in C^{n\times n}$, then*

$$-\mu(-\mathbf{A}) \le \operatorname{Re}\lambda \le \mu(\mathbf{A}).$$

Proof Since $\|\cdot\|_i$ is a convex function on $C^{n\times n}$, it can be shown to have a directional derivative at every point in $C^{n\times n}$ in every direction; see Eggleston (1966). However, a direct constructive proof is given below. Fix $\mathbf{A} \in C^{n\times n}$ and define

$$\mathbf{17} \quad f(\varepsilon) = \frac{\|I + \varepsilon\mathbf{A}\|_i - 1}{\varepsilon} = \left\|\frac{1}{\varepsilon}I + \mathbf{A}\right\|_i - \frac{1}{\varepsilon}, \ \forall \varepsilon > 0.$$

Clearly $f(\cdot)$ is continuous. It is shown that $f(\varepsilon)$ is nonincreasing as $\varepsilon \to 0^+$, and is bounded below. This shows that

$$\mathbf{18} \quad \lim_{\varepsilon \to 0^+} f(\varepsilon) =: \mu(\mathbf{A})$$

is well-defined. Towards this end, it is first shown that

$$\mathbf{19} \quad 0 < \delta < \varepsilon \Rightarrow f(\delta) \le f(\varepsilon).$$

Suppose $0 < \delta < \varepsilon$, and note that

$$\mathbf{20} \quad f(\delta) = \left\|\frac{1}{\delta}I + \mathbf{A}\right\|_i - \frac{1}{\delta}, f(\varepsilon) = \left\|\frac{1}{\varepsilon}I + \mathbf{A}\right\|_i - \frac{1}{\varepsilon}.$$

Now, using the triangle inequality and the fact that $\|I\|_i = 1$, one obtains

$$\mathbf{21} \quad \begin{aligned} \left\|\frac{1}{\delta}I + \mathbf{A}\right\|_i &= \left\|\frac{1}{\varepsilon}I + \mathbf{A} + \left(\frac{1}{\delta} - \frac{1}{\varepsilon}\right)I\right\|_i \\ &\le \left\|\frac{1}{\varepsilon}I + \mathbf{A}\right\|_i + \left\|\left(\frac{1}{\delta} - \frac{1}{\varepsilon}\right)I\right\|_i \\ &= \left\|\frac{1}{\varepsilon}I + \mathbf{A}\right\|_i + \frac{1}{\delta} - \frac{1}{\varepsilon}. \end{aligned}$$

Rearranging (21) and using (20) shows that $f(\delta) \le f(\varepsilon)$. Hence $f(\varepsilon)$ is nonincreasing as $\varepsilon \to 0^+$. Again, the triangle inequality shows that

$$\mathbf{22} \quad 1 - \varepsilon\|\mathbf{A}\|_i \le \|I + \varepsilon\mathbf{A}\|_i \le 1 + \varepsilon\|\mathbf{A}\|_i, \ \forall \varepsilon > 0,$$

$$\mathbf{23} \quad -\|\mathbf{A}\|_i \le f(\varepsilon) \le \|\mathbf{A}\|_i, \ \forall \varepsilon > 0.$$

Hence $f(\varepsilon)$ is bounded below. By previous discussion, this shows that $f(\varepsilon)$ has a well-defined limit as $\varepsilon \to 0^+$. Therefore $\mu(\mathbf{A})$ is well-defined [Property (M1)] and satisfies Property (M2). To prove (M3), observe that

$$\mathbf{24}\quad \mu(\alpha\mathbf{A}) = \lim_{\varepsilon\to 0^+} \frac{\|I+\varepsilon\alpha\mathbf{A}\|_i - 1}{\varepsilon} = \lim_{\varepsilon\alpha\to 0^+} \alpha \frac{\|I+\varepsilon\alpha\mathbf{A}\|_i - 1}{\varepsilon\alpha}.$$

To prove (M4), we begin by showing that

$$\mathbf{25}\quad \mu(\mathbf{A}+\mathbf{B}) \le \mu(\mathbf{A}) + \mu(\mathbf{B}).$$

A slight rearrangement of (15) gives

$$\mathbf{26}\quad \mu(\mathbf{A}+\mathbf{B}) = \lim_{\varepsilon\to 0^+} \|\frac{1}{\varepsilon}I + \mathbf{A} + \mathbf{B}\|_i - \frac{1}{\varepsilon}.$$

But, for each $\varepsilon > 0$, we have

$$\mathbf{27}\quad \|\frac{1}{\varepsilon}I + \mathbf{A} + \mathbf{B}\|_i - \frac{1}{\varepsilon} = \|\frac{1}{2\varepsilon}I + \mathbf{A} + \frac{1}{2\varepsilon}I + \mathbf{B}\|_i - \frac{1}{2\varepsilon} - \frac{1}{2\varepsilon}$$
$$\le \left[\|\frac{1}{2\varepsilon}I + \mathbf{A}\|_i - \frac{1}{2\varepsilon}\right] + \left[\|\frac{1}{2\varepsilon}I + \mathbf{B}\|_i - \frac{1}{2\varepsilon}\right]$$

Letting $\varepsilon \to 0^+$ in (27) proves (25). Now replace $\mathbf{A}$ by $\mathbf{A}+\mathbf{B}$ and $\mathbf{B}$ by $-\mathbf{B}$ in the right side of (25). Then in the left side of (25) $\mathbf{A}+\mathbf{B}$ is replaced by $\mathbf{A}+\mathbf{B}-\mathbf{B} = \mathbf{A}$, which gives

$$\mathbf{28}\quad \mu(\mathbf{A}) \le \mu(\mathbf{A}+\mathbf{B}) + \mu(-\mathbf{B}),$$

or

$$\mathbf{29}\quad \mu(\mathbf{A}) - \mu(-\mathbf{B}) \le \mu(\mathbf{A}+\mathbf{B}).$$

By symmetry,

$$\mathbf{30}\quad \mu(\mathbf{B}) - \mu(-\mathbf{A}) \le \mu(\mathbf{A}+\mathbf{B}).$$

This establishes (M4). Now (M5) is a ready consequence of (M3) and (25). Finally, to prove (M6), let λ be an eigenvalue of $\mathbf{A}$, and let $\mathbf{v}$ be a corresponding eigenvector. Assume without loss of generality that $\|\mathbf{v}\| = 1$, where $\|\cdot\|$ is the norm on C^n which induces the matrix norm $\|\cdot\|_i$ on $C^{n\times n}$. For each $\varepsilon > 0$, we have

$$\mathbf{31}\quad \|I+\varepsilon\mathbf{A}\|_i = \sup_{\|\mathbf{x}\|=1} \|(I+\varepsilon\mathbf{A})\mathbf{x}\|$$
$$\ge \|(I+\varepsilon\mathbf{A})\mathbf{v}\|$$
$$= |1+\varepsilon\lambda| \cdot \|\mathbf{v}\| = |1+\varepsilon\lambda|.$$

Similarly it follows that

32 $|1-\varepsilon\lambda| \le \|I-\varepsilon\mathbf{A}\|_i, \forall\varepsilon>0.$

Now, it is easy to verify that

33 $$\operatorname{Re}\lambda = \lim_{\varepsilon\to 0^+} \frac{|1+\varepsilon\lambda|-1}{\varepsilon}, \text{ and}$$

34 $$\operatorname{Re}\lambda = -\lim_{\varepsilon\to 0^+} \frac{|1-\varepsilon\lambda|-1}{\varepsilon}.$$

Combining (31) to (34) establishes (M6). ■

Comparing the properties of the matrix measure and the induced matrix norm, we see that, although both functions are convex, the similarity almost ends there. The measure can have positive as well as negative values, whereas a norm can assume only nonnegative values. The measure is "sign-sensitive" in that $\mu(-\mathbf{A})\neq\mu(\mathbf{A})$ in general, whereas $\|-\mathbf{A}\|_i = \|\mathbf{A}\|_i$. Because of these special properties, the measure function is useful in obtaining tight upper bounds on the norms of solutions of vector differential equations.

Theorem (16) lists only some of the many interesting properties of the measure function. A more complete discussion can be found in Desoer and Vidyasagar (1975) and Desoer and Haneda (1972).

In defining the measure of a matrix in $C^{n\times n}$, we have assumed that the norm used in (15) is an induced norm. It is possible, given *any* norm on $C^{n\times n}$, to define a corresponding measure function $\mu(\cdot)$ mapping $C^{n\times n}$ into $\mathbf{R}$. In this case, Properties (M1) through (M5) still hold, but (M6) does not. Such a measure function is of no use in estimating the norm of a solution to a vector differential equation; for such a purpose, only measures corresponding to *induced* matrix norms are useful.

In most applications, such as those involving differential equations, the linear vector space in question is $\mathbf{R}^n$, and the matrices of interest belong to $\mathbf{R}^{n\times n}$. Suppose $\|\cdot\|$ is a norm on $\mathbf{R}^n$, and let $\|\cdot\|_i$ denote the corresponding induced matrix norm defined on $\mathbf{R}^{n\times n}$; suppose we define the corresponding matrix measure $\mu(\cdot)$ as in (15), except that now $\mathbf{A}\in\mathbf{R}^{n\times n}$ and $\|\cdot\|_i$ is only defined on $\mathbf{R}^{n\times n}$. What properties does such a measure function have? An examination of the proofs of Properties (M1) through (M5) of Theorem (16) reveals that they carry over without modification to the case of real matrices. However, in proving Property (M6), essential use was made of the fact that the space in question is C^n and not $\mathbf{R}^n$, since in general the both the eigenvalue λ and eigenvector $\mathbf{v}$ could be complex. To get around this difficulty, one can "extend" the given norm on $\mathbf{R}^n$ to a norm on C^n. The details are not given here, but it can be shown that even Property (M6) is true for such a measure (see Problem 2.12). This can be summarized as follows:

35 Theorem *Let* $\|\cdot\|$ *be a norm defined on* $\mathbf{R}^n$, *and let* $\|\cdot\|_i: \mathbf{R}^{n\times n}\to\mathbf{R}$ *and* $\mu(\cdot): \mathbf{R}^{n\times n}\to\mathbf{R}$ *be defined in a manner analogous to (2) and (15), respectively. Then* $\mu(\cdot)$ *satisfies Properties (M1) through (M6) of Theorem (16).*

Given a particular vector norm $\|\cdot\|$ on C^n (or $\mathbf{R}^n$), it is in general a difficult task to obtain an explicit expression for the corresponding induced matrix norm (as mentioned earlier), and it is therefore still more difficult to obtain an explicit expression for the corresponding matrix measure. Nevertheless, the measure functions corresponding to the norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ can be calculated, and are displayed in Table 2.2 below.

Table 2.2

Norm on C^n	Matrix Measure on $C^{n\times n}$
$\|\mathbf{x}\|_\infty = \max_i \|x_i\|$	$\mu_\infty(\mathbf{A}) = \max_i [a_{ii} + \sum_{j\neq i} \|a_{ij}\|]$
$\|\mathbf{x}\|_1 = \sum_{i=1}^n \|x_i\|$	$\mu_1(\mathbf{A}) = \max_j [a_{jj} + \sum_{i\neq j} \|a_{ij}\|]$
$\|\mathbf{x}\|_2 = (\sum_{i=1}^n \|x_i\|^2)^{1/2}$	$\mu_2(\mathbf{A}) = \lambda_{\max}(\mathbf{A}^* + \mathbf{A})/2$

36 Example Let

$$\mathbf{A} = \begin{bmatrix} -6 & 2 & 1 \\ 0 & -1 & 2 \\ 1 & 3 & 0 \end{bmatrix}.$$

Using the formulas given in Table 2.2, one obtains by inspection that

$$\mu_1(\mathbf{A}) = 4,\ \mu_1(-\mathbf{A}) = 7;$$

$$\mu_\infty(\mathbf{A}) = 4,\ \mu_\infty(-\mathbf{A}) = 9;$$

Using Property (M6) of Theorem (35) to estimate the real parts of the eigenvalues of $\mathbf{A}$, one obtains

$$-7 \leq \operatorname{Re} \lambda_i \leq 4,$$

using the measure μ_1, and

$$-9 \leq \operatorname{Re} \lambda_i \leq 4,$$

using the measure μ_∞. The actual eigenvalues of $\mathbf{A}$ are

$$\{-6.0426, -3.1271, 2.1698\}.$$

Hence the smallest interval which contains the real parts of all eigenvalues of $\mathbf{A}$ is $[-6.0426, 2.1698]$. So the estimate obtained above, namely $[-7, 4]$ is not too bad. To complete the picture, let us compute the measure μ_2. This gives

$$\mu_2(\mathbf{A}) = 2.289, \mu_2(-\mathbf{A}) = 6.245.$$

This implies that

$$-6.245 \leq \operatorname{Re} \lambda_i \leq 2.289.$$

This estimate is almost exactly accurate. But of course it requires more work than computing either of the two measures μ_1 or μ_∞. Moreover, in another example some other measure might give a better bound. ■

Problem 2.10 Calculate the matrix norm $\|\mathbf{A}\|_i$ and the measure $\mu(\mathbf{A})$ corresponding to each of the vector norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$, for each of the matrices below:

$$\mathbf{A} = \begin{bmatrix} -4 & 1 & 1 \\ 2 & 0 & -2 \\ 1 & -3 & -6 \end{bmatrix}, \begin{bmatrix} 4 & -2 & 1 \\ 2 & -5 & -3 \\ -2 & 0 & 0 \end{bmatrix}.$$

Compute an interval in the real line containing the real parts of all the eigenvalues of $\mathbf{A}$ using Property (M6) of Theorem (16). Compare with the exact answer.

Problem 2.11 Suppose $\mathbf{M} \in \mathbf{R}^{n \times n}$ is a nonsingular matrix, and define a norm $\|\cdot\|_{M2}$ on $\mathbf{R}^n$ as follows:

$$\|\mathbf{x}\|_{M2} = \|\mathbf{M}\mathbf{x}\|_2 = (\mathbf{x}'\mathbf{M}'\mathbf{M}\mathbf{x})^{1/2}.$$

Show that the corresponding matrix measure on $\mathbf{R}^{n \times n}$ is given by

$$\mu_{M2}(\mathbf{A}) = \lambda_{\max}[(\mathbf{A}'\mathbf{P}' + \mathbf{P}\mathbf{A})/2],$$

where $\mathbf{P} = \mathbf{M}'\mathbf{M}$. Suppose we define the vector norms

$$\|\mathbf{x}\|_{M1} = \|\mathbf{M}\mathbf{x}\|_1, \quad \|\mathbf{x}\|_{M\infty} = \|\mathbf{M}\mathbf{x}\|_\infty.$$

Obtain explicit expressions for the corresponding the matrix measures.

Problem 2.12 Prove Theorem (35).

2.3 CONTRACTION MAPPING THEOREM

In this section, we state and prove a very important theorem, which is used in Section 2.4 to derive the existence and uniqueness of solutions to a class of nonlinear vector differential equations.

The theorem proved here is generally known as the *contraction mapping theorem* (or sometimes the Banach fixed point theorem), and is usually given in two forms: the global version and the local version. The local version assumes weaker hypotheses than the global version, and obtains correspondingly weaker conclusions. The global version is given first.

Note that, hereafter, the terms *mapping, function,* and *operator* are used interchangeably. Also, if T is a (possibly nonlinear) mapping, we write Tx instead of $T(x)$ in the interests of clarity.

2.3.1 Global Contractions

1 Theorem (Global Contraction Mapping) *Let* $(X, \|\cdot\|)$ *be a Banach space, and let* $T: X \to X$. *Suppose there exists a fixed constant* $\rho < 1$ *such that*

$$\mathbf{2} \qquad \|Tx - Ty\| \leq \rho \|x - y\|, \ \forall x, y \in X.$$

Under these conditions, there exists exactly one $x^* \in X$ *such that* $Tx^* = x^*$. *For each* $x_0 \in X$, *the sequence* $\{x_n\}$ *in* X *defined by*

$$\mathbf{3} \qquad x_{n+1} = Tx_n$$

converges to x^*. *Moreover,*

$$\mathbf{4} \qquad \|x^* - x_n\| \leq \frac{\rho^n}{1-\rho} \|Tx_0 - x_0\|.$$

Remarks An operator T satisfying the condition (2) is known as a **contraction**, because the images of any two elements x and y are closer together than x and y are. Moreover, T is a **global** contraction, since (2) holds for all x, y in the entire space X. An element $x \in X$ such that $Tx^* = x^*$ is called a **fixed point** of the mapping T, since x^* remains fixed when the mapping T is applied to it. Theorem (1) asserts that every contraction has exactly one fixed point in X. Moreover, this fixed point can be determined simply by taking *any* arbitrary starting point $x_0 \in X$ and repeatedly applying the mapping T to it. Finally, (4) provides an estimate of the rate of convergence of this sequence to the fixed point. Note that the bound in (4) decreases by a fixed ratio (namely ρ) at each iteration; such convergence is known as "linear convergence."

Proof Let $x_0 \in X$ be arbitrary and define the sequence $\{x_n\}$ as in (3). It is first shown that the sequence is a Cauchy sequence. For each $n \geq 0$, it follows from (2) that

$$\mathbf{5} \qquad \|x_{n+1} - x_n\| \leq \rho \|x_n - x_{n-1}\| \leq \cdots \leq \rho^n \|x_1 - x_0\| = \rho^n \|Tx_0 - x_0\|.$$

Suppose $m = n + r$, $r \geq 0$, is given. Then it follows from (5) that

$$\mathbf{6} \qquad \|x_m - x_n\| = \|x_{n+r} - x_n\|$$

$$\leq \sum_{i=0}^{r-1} \|x_{n+i+1} - x_{n+i}\|$$

$$\leq \sum_{i=0}^{r-1} \rho^{n+i} \| Tx_0 - x_0 \|$$

$$\leq \sum_{i=0}^{\infty} \rho^{n+i} \| Tx_0 - x_0 \| = \frac{\rho^n}{1-\rho} \| Tx_0 - x_0 \|.$$

Now, as $n \to \infty$, the quantity ρ^n approaches zero. Hence it is clear from (6) that $\| x_m - x_n \|$ can be made arbitrarily small by choosing n sufficiently large. Hence $\{x_n\}$ is a Cauchy sequence, and since X is assumed to be a Banach space, the sequence converges to an element of X. Let x^* denote this limit. Now, using Definition (2.1.27) of uniform continuity, one can show that T is a uniformly continuous mapping. Therefore, by Problem 2.9,

7 $$Tx^* = T(\lim_{n \to \infty} x_n) = \lim_{n \to \infty} Tx_n = \lim_{n \to \infty} x_{n+1} = x^*.$$

Hence x^* is a fixed point of T. To show that it is the *only* fixed point of T, suppose $x \in X$ is another fixed point of T, i.e., that $Tx = x$. Then, by (2),

8 $$\| x^* - x \| = \| Tx^* - Tx \| \leq \rho \| x^* - x \|.$$

Since $\rho < 1$, this inequality can be satisfied only if $\| x^* - x \| = 0$, i.e., if $x^* = x$. Finally, to prove the estimate (4), consider the inequality (6), and let $m \to \infty$. Since the norm function is continuous, it follows that

9 $$\| x^* - x_n \| = \| \lim_{m \to \infty} x_m - x_n \|$$

$$= \lim_{m \to \infty} \| x_m - x_n \| \leq \frac{\rho^n}{1-\rho} \| Tx_0 - x_0 \|,$$

where we have used the fact the right side of (6) is independent of m. ■

Note that in general it is *not* possible to replace (2) by the weaker condition

10 $\| Tx - Ty \| < \| x - y \|$, $\forall x, y \in X$, with $x \neq y$.

It is easy to show that any mapping satisfying (10) can have at most one fixed point, but quite possibly it may not have any at all. As a simple example, let $X = \mathbf{R}$, and define $f : \mathbf{R} \to \mathbf{R}$ by

11 $$f(x) = x + \frac{\pi}{2} - \tan^{-1}(x),$$

and define $Tx = f(x)$. Then

12 $$f'(x) = 1 - \frac{1}{1+x^2} < 1, \quad \forall x \in \mathbf{R}.$$

By the mean-value theorem,

13 $f(x)-f(y)=f'(z)(x-y)$ for some $z\in(x,y)$.

Hence T satisfies (10). However, it follows from (11) that $f(x)=x$ if and only if $\tan^{-1}(x)=\pi/2$. Clearly no such x exists. Hence T has no fixed point in $\mathbf{R}$.

14 Example Let $f:\mathbf{R}\to\mathbf{R}$ be a continuously differentiable function, and suppose

$$\sup_{x\in\mathbf{R}} |f'(x)| := \rho < 1.$$

Then, by the mean-value theorem, it follows as in (13) that f is a contraction on $\mathbf{R}$. Thus, by Theorem (1), there is a unique number $x^*\in\mathbf{R}$ such that $f(x^*)=x^*$. Moreover, this number can be determined as the limit of the sequence $\{x_n\}$ obtained by choosing any arbitrary $x_0\in\mathbf{R}$ and repeatedly applying the function f. The sequence of points so obtained is depicted in Figure 2.1.

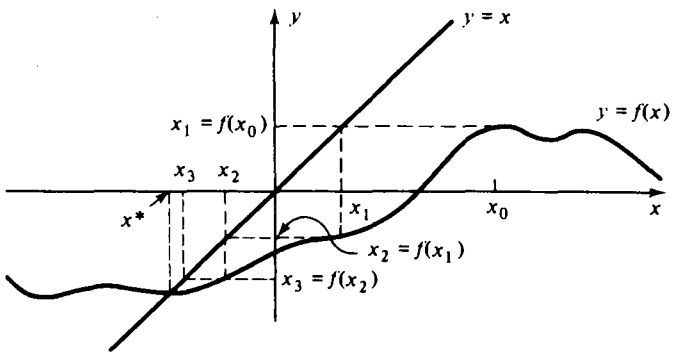


Fig. 2.1

2.3.2 Local Contractions

The applicability of Theorem (1) is limited by the fact that the operator T is required to satisfy (2) for *all* $x\in X$. In other words, T has to be a *global* contraction. In Theorem (15), we examine the case where T satisfies (2) only over some region M in X, i.e., the case where T is a *local* contraction, and derive correspondingly weaker results.

15 Theorem *Let* $(X, \|\cdot\|)$ *be a Banach space, let* M *be a subset of* X, *and let* $T: M\to X$. *Suppose there exists a constant* $\rho<1$ *such that*

16 $\|Tx-Ty\|\le\rho\|x-y\|, \forall x, y\in M,$

and suppose there exists an element $x_0\in M$ *such that the ball*

17 $$B=\{x\in X: \|x-x_0\|\le\frac{\|Tx_0-x_0\|}{1-\rho}\}$$

is contained in M. *Under these conditions,* T *has exactly one fixed point in* M. *If* x^* *denotes*

the fixed point of T in M, then the sequence $\{x_n\}$ defined by

18 $x_{n+1} = Tx_n, n \geq 0,$

converges to x^*. Moreover,

19 $\|x_n - x^*\| \leq \dfrac{\rho^n}{1-\rho} \|Tx_0 - x_0\|, \forall n \geq 0.$

Remarks

1. The significance of Theorem (15) lies in the fact that T is only required to be a contraction over the set M, not all of X. The price paid for this weaker hypothesis is that the conclusions of Theorem (15) are also weaker than those of Theorem (1).

2. Everything is contingent on finding a suitable element $x_0 \in M$ such that the ball B defined in (17) is contained in M. In effect, this means that we must be able to find an element x_0 in M such that repeated applications of T to x_0 result in a sequence that is entirely contained in M. Even if T satisfies (16), it may not be possible to find such an element x_0. For example, let $X = \mathbf{R}$, and let $T: \mathbf{R} \to \mathbf{R}$ be the function defined by

20 $$Tx = \begin{cases} 2, & \text{if } |x| \leq 1 \\ 0, & \text{if } |x| > 1. \end{cases}$$

   If M is chosen as the interval $[-1, 1]$, then T is a contraction over M. However, it is not possible to find an $x_0 \in M$ such that the ball B defined in (17) is contained in M. Accordingly, T has no fixed point in M.

3. Suppose we do succeed in finding an $x_0 \in M$ such that the hypotheses of Theorem (15) hold. Then the *particular* sequence defined in (18) converges to the unique fixed point x^* of T in M. However, if we choose *another* starting point for the iteration, there is no guarantee that the resulting sequence will converge to x^*. In contrast, if T is a global contraction, then the sequence defined in (3) converges to x^* converges for *every* starting point. There is one small consolation: If the sequence of iterations remains in M, then it must in fact converge to x^*; see Theorem (22) below.

Proof First, it is clear from (16) that T has at most one fixed point in M. If $x_0 \in M$ is chosen in such a way that the ball B defined in (17) is contained in M, then it follows that the sequence $\{x_n\}$ defined in (18) stays in B for all n; to see this, apply the inequality (6) with $n = 0$. Because the contraction condition holds in B, one can show, just as in the proof of Theorem (1), that $\{x_n\}$ is a Cauchy sequence in X and therefore converges to an element of X. Denote this limit by x^*; then a routine application of the continuity of the norm function shows that the limit must also belong to B and hence to M. The rest of the proof exactly follows that of Theorem (1). ■

21 Example Consider once again the case where $X = \mathbf{R}$, and let $f : \mathbf{R} \to \mathbf{R}$ be continuously differentiable. Suppose

$$\sup_{x \in [-1,1]} |f'(x)| := \rho < 1,$$

and that there exists an $x_0 \in [-1, 1]$ such that

$$B = \left[x_0 - \frac{f(x_0) - x_0}{1-\rho}, x_0 + \frac{f(x_0) - x_0}{1-\rho} \right] \subseteq [-1, 1].$$

Then Theorem (15) tells us that there is a unique $x^* \in [-1, 1]$ such that $f(x^*) = x^*$, and that x^* is the limit of the sequence $\{x_0, f(x_0), f[f(x_0)], \cdots\}$. The situation is depicted in Figure 2.2.

This section is concluded with another theorem whose hypotheses and conclusions lie between those of Theorems (1) and (15). This theorem is convenient for later applications.

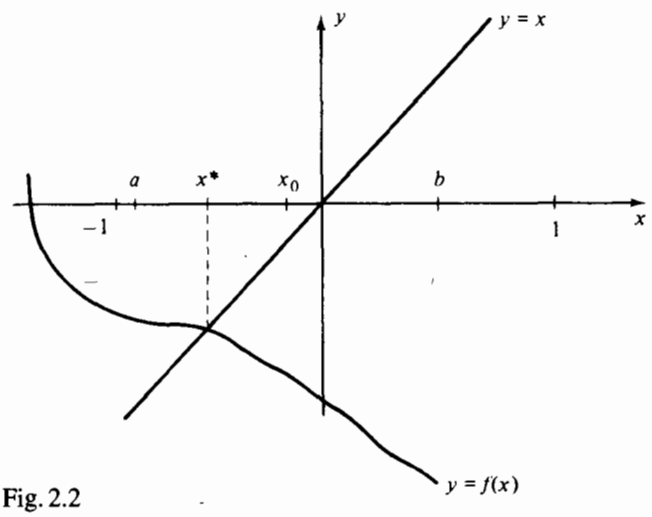


Fig. 2.2

22 Theorem *Let $(X, \|\cdot\|)$ be a Banach space, and let B be a closed ball in X, i.e., a set of the form*

23 $B = \{x : \|x - z\| \leq r\}$

for some $z \in X$ and $r \geq 0$. Let $T : X \to X$ be an operator satisfying the following conditions: (i) T maps B into itself, i.e., $Tx \in B$ whenever $x \in B$. (ii) There exists a constant $\rho < 1$ such that

24 $\|Tx - Ty\| \leq \rho \|x - y\|, \forall x, y \in B.$

Under these conditions, T has exactly one fixed point in B. If x^* denotes the fixed point of T

in B, then for each $x_0 \in B$, the sequence $\{x_n\}$ defined by

25 $x_{n+1} = Tx_n, n \geq 0,$

converges to x^*. Moreover

26 $\|x_n - x_0\| \leq \frac{\rho^n}{1-\rho} \|Tx_0 - x_0\|, \forall n \geq 0.$

The proof is obvious from Theorem (15).

The difference between Theorems (15) and (22) is that in the latter case T is assumed to map the entire ball B into itself, whereas in the former case it is only assumed that for a *particular* point $x_0 \in B$ the sequence of iterations is contained in B. As a consequence, in the latter case one can start from an *arbitrary* starting point in B to compute x^*.

Problem 2.13 Give a detailed proof of Theorem (22).

2.4 NONLINEAR DIFFERENTIAL EQUATIONS

In this section, we derive some general and very useful conditions which guarantee the existence and uniqueness of solutions to the nonlinear differential equation

1 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], t \geq 0; \mathbf{x}(0) = \mathbf{x}_0,$

where $\mathbf{x}(t) \in \mathbf{R}^n$ and $\mathbf{f}: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}^n$. As shown in Chapter 1, the existence and uniqueness of solutions to (1) is not guaranteed unless some restrictions are placed on the nature of $\mathbf{f}$. By a **solution** of (1) over an interval $[0, T]$, we mean an element $\mathbf{x}(\cdot)$ of $C^n[0, T]$ such that (i) $\mathbf{x}(\cdot)$ is differentiable everywhere, and (ii) Equation (1) holds at all t.

We first establish some conditions under which (1) has exactly one solution over every finite interval $[0, \delta]$ for sufficiently small δ, i.e., conditions for *local* existence and uniqueness. Then we present stronger results which guarantee *global* existence and uniqueness, i.e., conditions under which (1) has exactly one solution over $[0, \infty)$.

One small point is to be cleared up before we proceed to the theorems. First, if $\mathbf{x}(\cdot)$ is a solution of (1) over $[0, T]$ and $\mathbf{f}$ is continuous, then $\mathbf{x}(\cdot)$ also satisfies the integral equation

2 $\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}[\tau, \mathbf{x}(\tau)]\, d\tau, t \in [0, T].$

On the other hand, if $\mathbf{x}(\cdot) \in C^n[0, T]$ satisfies (2), then clearly $\mathbf{x}(\cdot)$ is actually differentiable everywhere and satisfies (1). Thus (1) and (2) are equivalent in the sense that every solution of (1) is also a solution of (2) and vice versa.

2.4.1 Local Existence and Uniqueness

3 Theorem (Local Existence and Uniqueness) *Suppose the function* $\mathbf{f}$ *in (1) is continuous in* t *and* $\mathbf{x}$ *and satisfies the following conditions: There exist finite constants* T, r, h, *and* k *such that*

4 $$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| \leq k \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in B, \ \forall t \in [0, T],$$

5 $$\|\mathbf{f}(t, \mathbf{x}_0)\| \leq h, \ \forall t \in [0, T],$$

where B *is a ball in* $\mathbf{R}^n$ *of the form*

6 $$B = \{\mathbf{x} \in \mathbf{R}^n : \|\mathbf{x} - \mathbf{x}_0\| \leq r\}.$$

Then (1) has exactly one solution over $[0, \delta]$ *whenever the number* δ *is sufficiently small to satisfy the inequalities*

7 $$h\delta \exp(k\delta) \leq r,$$

and

8 $$\delta \leq \min\{T, \frac{\rho}{k}, \frac{r}{h + kr}\}$$

for some constant $\rho < 1$.

9 Remarks

1. While following the proof of Theorem (3), it is important to keep in mind the distinction between $\|\cdot\|$ (which is a norm on $\mathbf{R}^n$), and $\|\cdot\|_C$, (which is a norm on $C^n[0, \delta]$). Also, it should be noted that B is a ball in $\mathbf{R}^n$, while S defined in (10) below is a ball in $C^n[0, \delta]$.

2. The condition (4) is known as a **Lipschitz condition**, and the constant k is known as a **Lipschitz constant**. Notice that we say *a* Lipschitz constant, because if k is a Lipschitz constant for the function $\mathbf{f}$, then so is any constant larger than k. Some authors reserve the term Lipschitz constant for the *smallest* number k such that (4) is satisfied. A function that satisfies a Lipschitz condition is said to be **Lipschitz-continuous.** Note that a Lipschitz-continuous function is also *absolutely continuous* [see Royden (1963)] and is therefore differentiable almost everywhere.

3. Equation (4) is known as a **local** Lipschitz condition, because it holds only for all $\mathbf{x}, \mathbf{y}$ in some ball around $\mathbf{x}_0$, for $t \in [0, T]$. Accordingly, Theorem (3) is a local existence and uniqueness theorem, because it guarantees existence and uniqueness of solutions over a sufficiently small interval $[0, \delta]$. Note that, given any finite constants k, r, T and h, (7) and (8) can always be satisfied by choosing δ sufficiently small.

Proof By a slight abuse of notation, we use $\mathbf{x}_0(\cdot)$ to denote the function in $C^n[0, \delta]$ whose value is $\mathbf{x}_0$ for all $t \in [0, \delta]$. Suppose δ satisfies (7) and (8), and let S be the ball in $C^n[0, \delta]$ defined by

$$\mathbf{10} \quad S = \{\mathbf{x}(\cdot) \in C^n[0, \delta]: \|\mathbf{x}(\cdot) - \mathbf{x}_0(\cdot)\|_C \leq r\}.$$

Let P denote the mapping of $C^n[0, \delta]$ into itself defined by

$$\mathbf{11} \quad (P\mathbf{x})(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}[\tau, \mathbf{x}(\tau)]\, d\tau, \ \forall t \in [0, \delta].$$

Clearly $\mathbf{x}(\cdot)$ is a solution of (2) over the interval $[0, \delta]$ if and only if $(P\mathbf{x})(\cdot) = \mathbf{x}(\cdot)$, i.e., $\mathbf{x}(\cdot)$ is a fixed point of the map P.

It is first shown that P is a contraction on S. Let $\mathbf{x}(\cdot)$ and $\mathbf{y}(\cdot)$ be arbitrary elements of S; then $\mathbf{x}(t)$ and $\mathbf{y}(t)$ lie in the ball B, for all $t \in [0, \delta]$. Thus

$$\mathbf{12} \quad (P\mathbf{x})(t) - (P\mathbf{y})(t) = \int_0^t \{\mathbf{f}[\tau, \mathbf{x}(\tau)] - \mathbf{f}[\tau, \mathbf{y}(\tau)]\}\, d\tau,$$

$$\mathbf{13} \quad \|(P\mathbf{x})(t) - (P\mathbf{y})(t)\| \leq \int_0^t \|\mathbf{f}[\tau, \mathbf{x}(\tau)] - \mathbf{f}[\tau, \mathbf{y}(\tau)]\|\, d\tau$$

$$\leq \int_0^t k\, \|\mathbf{x}(\tau) - \mathbf{y}(\tau)\|\, d\tau$$

$$\leq kt\, \|\mathbf{x}(\cdot) - \mathbf{y}(\cdot)\|_C$$

$$\leq \rho\, \|\mathbf{x}(\cdot) - \mathbf{y}(\cdot)\|_C,$$

where in the last line we have used the fact that $kt \leq k\delta \leq \rho$ by (8). Because the last term on the right-hand side of (13) is independent of t, it follows that

$$\mathbf{14} \quad \|(P\mathbf{x})(\cdot) - (P\mathbf{y})(\cdot)\|_C = \sup_{t \in [0, \delta]} \|(P\mathbf{x})(t) - (P\mathbf{y})(t)\| \leq \rho\, \|\mathbf{x}(\cdot) - \mathbf{y}(\cdot)\|_C.$$

This shows that P is a contraction on S.

Next it is shown that P maps S into itself. Suppose $\mathbf{x}(\cdot) \in S$. Then

Para um $t \leq$

15 $$\|(P\mathbf{x})(t) - \mathbf{x}_0\| = \left\| \int_0^t \mathbf{f}[\tau, \mathbf{x}(\tau)]\, d\tau \right\|$$

$$= \left\| \int_0^t \{\mathbf{f}[\tau, \mathbf{x}(\tau) - \mathbf{f}(\tau, \mathbf{x}_0) + \mathbf{f}(\tau, \mathbf{x}_0)\}\, d\tau \right\|$$

$$\leq \int_0^t \{\, \|\mathbf{f}[\tau, \mathbf{x}(\tau)] - \mathbf{f}(\tau, \mathbf{x}_0)\| + \|\mathbf{f}(\tau, \mathbf{x}_0)\| \,\}\, d\tau$$

$$\leq kr\delta + h\delta \leq r,$$

by (8). Hence Para todo t em $[0, \delta]$

16 $$\|(P\mathbf{x})(\cdot) - \mathbf{x}_0(\cdot)\|_C = \sup_{t \in [0, \delta]} \|(P\mathbf{x})(t) - \mathbf{x}_0\| \leq r.$$

This shows that $P\mathbf{x} \in S$, so that P maps S into itself.

Now, since P maps S into itself and is a contraction on S, it has exactly one fixed point in S, by Theorem (2.3.22). Our objective, however, is to show that P has exactly one fixed point in $C^n[0, \delta]$, not just S (the point being that S is a proper subset of $C^n[0, \delta]$). Thus the proof is completed if it can be shown that any fixed point of P in $C^n[0, \delta]$ must in fact lie in S. Accordingly, suppose $\mathbf{x}(\cdot) \in C^n[0, \delta]$ satisfies (2). Then $\mathbf{x}(0) = \mathbf{x}_0 \in B$. Also, since $\mathbf{x}(\cdot)$ is continuous, it follows that $\mathbf{x}(t) \in B$ for all sufficiently small t. Now it is shown that $\mathbf{x}(t) \in B$ for all $t \in [0, \delta]$. To show this, assume the contrary, namely that there exists a time $t_0 \in (0, \delta)$ such that $\mathbf{x}(t_0)$ does not belong to B, i.e., $\|\mathbf{x}(t_0) - \mathbf{x}_0\| > r$. Since $\|\mathbf{x}(t) - \mathbf{x}_0\|$ is a continuous function of t and since $\|\mathbf{x}(0) - \mathbf{x}_0\| = 0$, there is a unique *first time* $\alpha < t_0 < \delta$ with the property that

17 $$\|\mathbf{x}(t) - \mathbf{x}_0\| < r,\ \forall t \in [0, \alpha), \text{ and } \|\mathbf{x}(\alpha) - \mathbf{x}_0\| = r.$$

Now, since $\mathbf{x}(\cdot)$ satisfies (2), we have

18 $$\|\mathbf{x}(t) - \mathbf{x}_0\| = \int_0^t \mathbf{f}[\tau, \mathbf{x}(\tau)]\, d\tau$$

$$= \int_0^t \{\mathbf{f}[\tau, \mathbf{x}(\tau)] - \mathbf{f}(\tau, \mathbf{x}_0) + \mathbf{f}(\tau, \mathbf{x}_0)\}\, d\tau,\ \forall t \in [0, \alpha],$$

19 $$\|\mathbf{x}(t) - \mathbf{x}_0\| \leq \int_0^t k\, \|\mathbf{x}(t) - \mathbf{x}_0\|\, d\tau + ht$$

$$\leq h\alpha + \int_0^t k \| \mathbf{x}(t) - \mathbf{x}_0 \| \, d\tau, \ \forall t \in [0, \alpha].$$

Equation (19) gives an implicit bound for the quantity $\| \mathbf{x}(t) - \mathbf{x}_0 \|$. This implicit bound can be replaced by an *explicit* bound, using a result known as the Gronwall inequality [see Lemma (5.7.1)]. Applying this inequality to (19) gives

20 $\| \mathbf{x}(t) - \mathbf{x}_0 \| \leq h\alpha \exp(k\alpha), \ \forall t \in [0, \alpha].$

In particular, then,

21 $\| \mathbf{x}(\alpha) - \mathbf{x}_0 \| \leq h\alpha \exp(k\alpha) < h\delta \exp(k\delta) \leq r$, by (7).

But (21) contradicts (17). This shows that, if any function $\mathbf{x}(\cdot) \in C^n[0, \delta]$ satisfies (2), and δ is sufficiently small that (7) holds, then $\mathbf{x}(\cdot)$ must necessarily belong to S. Thus we have shown that any fixed point of P in $C^n[0, \delta]$ must in fact be in S. Since P has exactly one fixed point in S, it follows that P has exactly one fixed point in $C^n[0, \delta]$. By the manner in which P is defined, we conclude that (2) has exactly one solution over $[0, \delta]$. ■

The following result is actually a corollary to Theorem (3), but is in a form that can be readily applied.

22 Corollary *Consider the differential equation (1). Suppose that in some neighborhood of* $(0, \mathbf{x}_0)$ *the function* $\mathbf{f}(t, \mathbf{x})$ *is continuously differentiable. Then (1) has exactly one solution over* $[0, \delta]$ *provided* δ *is sufficiently small.*

Proof The differentiability properties assumed on $\mathbf{f}$ ensure that $\mathbf{f}$ satisfies (4) and (5) for some set of finite constants r, T, k and h. ■

Thus far we have studied the existence and uniqueness of solutions to (1) over *closed* intervals of the form $[0, \delta]$. The reasons for this are primarily technical. For example, $C^n[0, \delta]$ is a Banach space, but $C^n[0, \delta)$ is a much more tricky object. But now consider the following question: Suppose $\mathbf{f}(t, \mathbf{x})$ is continuously differentiable *everywhere*. What is the *largest* interval over which (1) has a unique solution? Looking back over the proof of Theorem (3), one can see that the proof is equally valid if the initial time is changed from 0 to an arbitrary time t_0, and all hypotheses are adjusted accordingly. Thus, if (1) has a unique solution over some interval $[0, \delta]$ [which it will, by Corollary (22)], then one can again apply Corollary (22) with δ as the initial time and $\mathbf{x}(\delta)$ as the initial state, and conclude that there is a unique solution to (1) over some interval $[\delta, \delta']$. This solution can be concatenated with the earlier solution over $[0, \delta]$ to construct a unique solution to (1) over the *larger* interval $[0, \delta']$. But the process can be repeated yet again with δ' as the initial time and $\mathbf{x}(\delta')$ as the initial state. Since this process can be repeated indefinitely, we see that there is no largest *closed* interval over which (1) has a unique solution. Instead, there is a number $\delta_{\max}$ (which may equal infinity) such that (1) has a unique solution over every closed interval $[0, \delta]$ in the half-open interval $[0, \delta_{\max})$; this solution is called the **maximal solution.** Now, what can happen as $t \to \delta_{\max}$? If $\delta_{\max}$ is finite and if $\mathbf{x}(t)$ remains well-behaved as $t \to \delta_{\max}$ and

approaches some *finite* vector $\mathbf{x}_{\max}$, then one can again apply Corollary (22) with $\delta_{\max}$ as the initial time and $\mathbf{x}_{\max}$ as the initial state, and thereby extend the solution still further in time, which contradicts the definition of $\delta_{\max}$. Thus, if $\delta_{\max}$ is finite, then $\|\mathbf{x}(t)\|$ *must* approach infinity as $t \to \delta_{\max}$. This discussion can be summarized as follows:

23 Corollary *Consider the differential equation (1), and suppose that* $\mathbf{f}(t, \mathbf{x})$ *is continuously differentiable everywhere. Then there exists a unique number* $\delta_{\max} = \delta_{\max}(\mathbf{x}_0)$, *which could equal infinity, such that (1) has a unique solution over* $[0, \delta_{\max})$ *and over no larger interval. If* $\delta_{\max}$ *is finite, then* $\|\mathbf{x}(t)\| \to \infty$ *as* $t \to \delta_{\max}$.

24 Example Consider the scalar differential equation

$$\dot{x}(t) = 1 + x^2, x(0) = 0.$$

Then $\delta_{\max} = \pi/2$, and the maximal solution is

$$x(t) = \tan t.$$

Predictably, $x(t) \to \infty$ as $t \to \pi/2$. ■

A solution $\mathbf{x}(t)$ with the property that $\|\mathbf{x}(t)\| \to \infty$ as t approaches some finite time is said to exhibit **finite escape time.**

Another question one can ask about the differential equation (1) is this: Is it possible to solve (1) for *negative* values of t? The answer, under the hypotheses of Theorem (3), is yes. In fact, if one is interested in solutions of (1) for both negative as well as positive values of t, then Corollary (23) should be modified to speak of a maximal interval $(-\delta_{\min}, \delta_{\max})$, rather than a half-open interval. Generally speaking, in control theory one is usually not interested in solving for the past behavior of a system, only its future. Thus the topic is not pursued further in this book. However, in the theory of dynamical systems, one is often interested in both the past as well as the future of a system. The interested reader is referred to Hirsch and Smale (1974) for further details.

2.4.2 Global Existence and Uniqueness

In this subsection, we show that (loosely speaking) if $\mathbf{f}$ satisfies a global Lipschitz condition, then (1) has a unique solution over all time.

25 Theorem (Global Existence and Uniqueness) *Suppose that for each* $T \in [0, \infty)$ *there exist finite constants* k_T *and* h_T *such that*

26 $\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| \leq k_T \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, \forall t \in [0, T],$

27 $\|\mathbf{f}(t, \mathbf{x}_0)\| \leq h_T, \forall t \in [0, T].$

Then (1) has exactly one solution over $[0, \infty)$.

Remark Recall that by a solution over $[0, \infty)$ we mean a solution over $[0, T]$ for each finite T.

Proof We give two alternate proofs.

Proof No. 1 Let $T < \infty$ be specified, and let k_T and h_T be finite constants such that (26) and (27) hold. Then the hypotheses of Theorem (3) are satisfied with $r = \infty$. In this case (7) holds for all δ. Thus, by Theorem (3), it follows that (1) has a unique solution over $[0, \delta]$ whenever δ satisfies

$$\mathbf{28} \quad \delta \leq \frac{\rho}{k_T}$$

for some constant $\rho < 1$. Suppose a positive number δ satisfying (28) is chosen. If $T \leq \delta$, then by Theorem (3) there is a unique solution over $[0, T]$, so suppose $T > \delta$. Now (1) has a unique solution over $[0, \delta]$. Denote this solution by $\mathbf{x}_1(\cdot)$, and consider the "shifted" differential equation

$$\mathbf{29} \quad \dot{\mathbf{x}}(t) = \mathbf{f}_1[t, \mathbf{x}(t)], \ \mathbf{x}(0) = \mathbf{x}_1(\delta),$$

where

$$\mathbf{30} \quad \mathbf{f}_1(t, \mathbf{x}) = \mathbf{f}(t + \delta, \mathbf{x}).$$

Then $\mathbf{f}_1$ also satisfies (26) and (27); therefore once again by Theorem (3), (29) has unique solution over $[0, \delta]$, *where δ is the same as before.* Denote this solution by $\mathbf{y}_2(\cdot)$. It is easy to verify that the function $\mathbf{x}_2(\cdot)$ defined by

$$\mathbf{31} \quad \mathbf{x}_2(t) = \begin{cases} \mathbf{x}_1(t), & 0 \leq t \leq \delta \\ \mathbf{y}_2(t - \delta), & \delta \leq t \leq 2\delta \end{cases}$$

is the unique solution of (1) over the interval $[0, 2\delta]$. Proceeding by induction, let $\mathbf{x}_m(\cdot)$ denote the unique solution of (1) over the interval $[0, m\delta]$, and consider the differential equation

$$\mathbf{32} \quad \dot{\mathbf{x}}(t) = \mathbf{f}_m[t, \mathbf{x}(t)], \ \mathbf{x}(0) = \mathbf{x}_m(m\delta),$$

where

$$\mathbf{33} \quad \mathbf{f}_m(t, \mathbf{x}) = \mathbf{f}(t + m\delta, \mathbf{x}).$$

Let $\mathbf{y}_{m+1}$ denote the unique solution of (32) over the interval $[0, \delta]$ (the same δ as before). Then the function $\mathbf{x}_{m+1}(\cdot)$ defined by

$$\mathbf{34}\quad \mathbf{x}_{m+1}(t)=\begin{cases}\mathbf{x}_m(t), & 0\le t\le m\delta\\ \mathbf{y}_{m+1}(t-m\delta), & m\delta\le t\le(m+1)\delta\end{cases}$$

is the unique solution of (1) over the interval $[0, (m+1)\delta]$. In this manner, the unique solution can be extended to all of $[0, T]$.

Proof No. 2 Let $T<\infty$ be given, let $P: C^n[0, T]\to C^n[0, T]$ be given by (11), and let $\mathbf{x}_0(\cdot)$ denote (as before) the element of $C^n[0, T]$ whose value is $\mathbf{x}_0$ for all $t\in[0, T]$. It is shown first that the sequence $\{P^m\mathbf{x}_0(\cdot)\}_{m=1}^{\infty}$ is a Cauchy sequence in $C^m[0, T]$ and that it converges to a solution of (2).

Let $\mathbf{x}_m(\cdot)=(P^m\mathbf{x}_0)(\cdot)$. Then we have, first,

$$\mathbf{35}\quad \mathbf{x}_1(t)-\mathbf{x}_0(t)=\int_0^t \mathbf{f}(\tau, \mathbf{x}_0)\,d\tau,$$

$$\mathbf{36}\quad \|\mathbf{x}_1(t)-\mathbf{x}_0(t)\|\le\int_0^t \|\mathbf{f}(\tau, \mathbf{x}_0)\|\,d\tau\le h_T t.$$

In general, for $m\ge 1$, we have

$$\mathbf{37}\quad \|\mathbf{x}_{m+1}(t)-\mathbf{x}_m(t)\|\le\int_0^t \|\mathbf{f}[\tau, \mathbf{x}_m(\tau)]-\mathbf{f}[\tau, \mathbf{x}_{m-1}(\tau)]\|\,d\tau$$

$$\le k_T\int_0^t \|\mathbf{x}_m(\tau)-\mathbf{x}_{m-1}(\tau)\|\,d\tau.$$

Substituting (36) into (37) and proceeding by induction gives

$$\mathbf{38}\quad \|\mathbf{x}_m(t)-\mathbf{x}_{m-1}(t)\|\le k_T^{m-1}h_T\frac{t^m}{m!}.$$

Thus for any integer $p\ge 0$ it follows that

$$\mathbf{39}\quad \|\mathbf{x}_{m+p}(t)-\mathbf{x}_m(t)\|\le\sum_{i=0}^{p-1}\|\mathbf{x}_{m+i+1}(t)-\mathbf{x}_{m+i}(t)\|$$

$$\le\sum_{i=0}^{p-1}h_T k_T^{m+i}\frac{t^{m+i+1}}{(m+i+1)!}$$

$$=\sum_{i=m+1}^{m+p}h_T k_T^{i-1}\frac{t^i}{i!},$$

$$\textbf{40}\quad \|\mathbf{x}_{m+p}-\mathbf{x}_m\|_C=\sup_{t\in[0,T]}\|\mathbf{x}_{m+p}(t)-\mathbf{x}_m(t)\|$$

$$\leq\sum_{i=m+1}^{m+p}h_T k_T^{i-1}\frac{T^i}{i!}\leq\sum_{i=m+1}^{\infty}h_T k_T^{i-1}\frac{T^i}{i!}.$$

Now consider the sequence of sums

$$\textbf{41}\quad \left\{\sum_{i=0}^{m}h_T k_T^{i-1}\frac{T^i}{i!}\right\}.$$

As $m\to\infty$, this sequence converges to $(h_T/k_T)\exp(k_T T)$. Moreover, the last term in (40) is the difference between this limit and the partial sum in (41) and therefore converges to zero. Thus by choosing m sufficiently large this sum can be made arbitrarily small. This shows that $\{\mathbf{x}_m(\cdot)\}$ is a Cauchy sequence in $C^n[0,T]$. Since $C^n[0,T]$ is a Banach space, the sequence converges to a limit in $C^n[0,T]$. Denote this limit by $\mathbf{x}^*(\cdot)$.

Whenever $\mathbf{z}_1(\cdot)$ and $\mathbf{z}_2(\cdot)$ are two elements in $C^n[0,T]$, we have

$$\textbf{42}\quad (P\mathbf{z}_1)(t)-(P\mathbf{z}_2)(t)=\int_0^t\{\mathbf{f}[\tau,\mathbf{z}_1(\tau)]-\mathbf{f}[\tau,\mathbf{z}_2(\tau)]\}\,d\tau,$$

$$\textbf{43}\quad \|(P\mathbf{z}_1)(t)-(P\mathbf{z}_2)(t)\|\leq\int_0^t\|\mathbf{f}[\tau,\mathbf{z}_1(\tau)-\mathbf{f}[\tau,\mathbf{z}_2(\tau)]\|\,d\tau$$

$$\leq k_T T\|\mathbf{z}_1-\mathbf{z}_2\|_C,$$

$$\textbf{44}\quad \|(P\mathbf{z}_1)(\cdot)-(P\mathbf{z}_2)(\cdot)\|_C=\sup_{t\in[0,T]}\|(P\mathbf{z}_1)(t)-(P\mathbf{z}_2)(t)\|$$

$$\leq k_T T\|\mathbf{z}_1-\mathbf{z}_2\|_C.$$

Since $k_T T$ is a finite constant, it follows that P is uniformly continuous on $C^n[0,T]$. Hence if $\{\mathbf{x}_m(\cdot)\}$ converges to $\mathbf{x}^*$, it follows that

$$\textbf{45}\quad (P\mathbf{x}^*)(\cdot)=\lim_{m\to\infty}(P\mathbf{x}_m)(\cdot)=\lim_{m\to\infty}\mathbf{x}_{m+1}(\cdot)=\mathbf{x}^*(\cdot).$$

This shows that $\mathbf{x}^*(\cdot)$ is a solution of (2).

Next, to show that $\mathbf{x}^*$ is the *only* solution to (2), suppose $\mathbf{y}(\cdot)$ also satisfies (2). Then

$$\textbf{46}\quad \mathbf{y}(t)-\mathbf{x}^*(t)=\int_0^t\{\mathbf{f}[\tau,\mathbf{y}(\tau)]-\mathbf{f}[\tau,\mathbf{x}^*(\tau)]\}\,d\tau,\ \forall t\in[0,T],$$

47 $\|\mathbf{y}(t)-\mathbf{x}^*(t)\| \le k_T \int_0^t \|\mathbf{y}(\tau)-\mathbf{x}^*(\tau)\|\, d\tau,\ \forall t \in [0, T].$

Applying Gronwall's inequality [Lemma (5.7.1)] to (47) gives

48 $\|\mathbf{y}(t)-\mathbf{x}^*(t)\| = 0,\ \forall t \in [0, T].$

Thus $\mathbf{y}(\cdot) = \mathbf{x}^*(\cdot)$, i.e., $\mathbf{x}^*(\cdot)$ is the unique solution of (2). ■

49 Remarks

1. The sequence $\{P^m \mathbf{x}_0(\cdot)\}$ that converges to the solution $\mathbf{x}^*(\cdot)$ of (2) is known as the sequence of *Picard's iterations*, and this method of generating a solution to (2) is known as *Picard's method.* Actually, it is easy to show that Picard's iterations converge starting from *any* arbitrary starting function in $C^n[0, T]$ and not just $\mathbf{x}_0(\cdot)$.
2. Note that some authors assume that $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}\ \forall t \ge 0$. This assumption, together with (4), implies (5), because then $\|\mathbf{f}(t, \mathbf{x}_0)\| \le k_T \|\mathbf{x}_0\|$. However, in "forced" nonlinear systems, it is not necessarily true that $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}\ \forall t$. The present development does not require this assumption.

We next prove two theorems regarding the solution of (2). In effect, Theorem (25) states that (2) has a unique solution corresponding to each initial condition. Theorem (50) below shows that, at any given time, there is exactly one solution trajectory of (2) passing through each point in $\mathbf{R}^n$. Theorem (57) shows that the solution of (2) depends continuously on the initial condition.

50 Theorem *Let* $\mathbf{f}$ *satisfy the hypotheses of Theorem (25). Then for each* $\mathbf{z} \in \mathbf{R}^n$ *and each* $T \in [0, \infty)$ *there exists exactly one element* $\mathbf{z}_0 \in \mathbf{R}^n$ *such that the unique solution over* $[0, T]$ *of the differential equation*

51 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)],\ \mathbf{x}(0) = \mathbf{z}_0$

satisfies

52 $\mathbf{x}(T) = \mathbf{z}.$

Proof Consider the equation

53 $\dot{\mathbf{x}}(t) = \mathbf{f}_s[t, \mathbf{x}(t)],\ \mathbf{x}(0) = \mathbf{z},$

where

54 $\mathbf{f}_s(t, \mathbf{x}) = -\mathbf{f}(T-t, \mathbf{x}),\ \forall t \in [0, T].$

Then $\mathbf{f}_s$ also satisfies the hypotheses of Theorem (25), so that (53) has a unique solution over $[0, T]$. Denote this solution by $\mathbf{y}(\cdot)$ and define $\mathbf{z}_0 = \mathbf{y}(T)$. Then one can easily verify that the

function $\mathbf{y}_s(\cdot)$ defined by

55 $\mathbf{y}_s(t) = \mathbf{y}(T-t),\ \forall t \in [0, T]$

satisfies (51) and also satisfies (52). To prove the uniqueness of the element $\mathbf{z}_0$ corresponding to a particular $\mathbf{z}$, assume by way of contradiction that there exist two functions $\mathbf{y}_1(\cdot)$ and $\mathbf{y}_2(\cdot)$ in $C^n[0, T]$ that satisfy (51) and (52). Let $\mathbf{y}_1(0) = \mathbf{z}_1$, $\mathbf{y}_2(0) = \mathbf{z}_2$. Then the functions $\mathbf{y}_a(\cdot)$ and $\mathbf{y}_b(\cdot)$ defined by

56 $\mathbf{y}_a(t) = \mathbf{y}_1(T-t),\ \mathbf{y}_b(t) = \mathbf{y}_2(T-t)$

must *both* satisfy (53). However, because the solution to (53) is unique, it follows that $\mathbf{y}_a(\cdot) = \mathbf{y}_b(\cdot)$. Hence $\mathbf{z}_1 = \mathbf{z}_2$. ■

57 Theorem *Let* $\mathbf{f}$ *satisfy the hypotheses of Theorem (25), and let* $T \in [0, \infty)$ *be specified. Then for each* $\varepsilon > 0$, *there exists a* $\delta(\varepsilon, T) > 0$ *such that the following is true: Suppose* $\mathbf{x}_0$ *and* $\mathbf{y}_0$ *are vectors in* $\mathbf{R}^n$ *that satisfy*

58 $\|\mathbf{x}_0 - \mathbf{y}_0\| < \delta(\varepsilon, T).$

Suppose $\mathbf{x}(\cdot)$ *and* $\mathbf{y}(\cdot)$ *are the corresponding solutions to the differential equations*

59 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)],\ \mathbf{x}(0) = \mathbf{x}_0,$

60 $\dot{\mathbf{y}}(t) = \mathbf{f}[t, \mathbf{y}(t)],\ \mathbf{y}(0) = \mathbf{y}_0.$

Then

61 $\|\mathbf{x}(\cdot) - \mathbf{y}(\cdot)\|_C \leq \varepsilon.$

Proof The functions $\mathbf{x}(\cdot)$ and $\mathbf{y}(\cdot)$ also satisfy

62 $$\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}[\tau, \mathbf{x}(\tau)]\, d\tau,$$

63 $$\mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}[\tau, \mathbf{y}(\tau)]\, d\tau.$$

Subtracting, we get

64 $$\mathbf{x}(t) - \mathbf{y}(t) = \mathbf{x}_0 - \mathbf{y}_0 + \int_0^t \{\mathbf{f}[\tau, \mathbf{x}(\tau)] - \mathbf{f}[\tau, \mathbf{y}(\tau)]\}\, d\tau,$$

$$\mathbf{65} \quad \|\mathbf{x}(t)-\mathbf{y}(t)\| \le \|\mathbf{x}_0-\mathbf{y}_0\| + k_T \int_0^t \|\mathbf{x}(\tau)-\mathbf{y}(\tau)\|\, d\tau.$$

Applying Gronwall's inequality [Lemma (5.7.1)] to (65) gives

$$\mathbf{66} \quad \|\mathbf{x}(t)-\mathbf{y}(t)\| \le \|\mathbf{x}_0-\mathbf{y}_0\| \exp(k_T T).$$

Hence

$$\mathbf{67} \quad \|\mathbf{x}(\cdot)-\mathbf{y}(\cdot)\|_C \le \|\mathbf{x}_0-\mathbf{y}_0\| \exp(k_T T).$$

Thus, given $\varepsilon > 0$, (61) is satisfied if we choose $\delta(\varepsilon, T) = \varepsilon/\exp(k_T T)$. ■

Remarks

1. The results contained in Theorems (50) and (57) can be given a simple interpretation in terms of certain mappings being continuous. Let $\phi: \mathbf{R}^n \to C^n[0, T]$ be the mapping that associates, with each initial condition $\mathbf{x}_0 \in \mathbf{R}^n$, the corresponding unique solution of (2). Then Theorem (57) states that ϕ is uniformly continuous on $\mathbf{R}^n$. In the same vein, let $\psi_T: \mathbf{R}^n \to \mathbf{R}^n$ be the mapping that associates, with each initial condition $\mathbf{x}_0 \in \mathbf{R}^n$, the value at time T of the corresponding unique solution of (2). Then Theorem (50) states that ψ_T is one-to-one [i.e., given $\psi_T(\mathbf{x})$, one can uniquely determine $\mathbf{x}$], and onto (i.e., the range of ψ_T is all of $\mathbf{R}^n$). Furthermore, Theorem (57) shows that both ψ_T and its inverse map ψ_T^{-1} are continuous.

2. It is important to note that Theorem (57) is strictly limited to the case where the interval $[0, T]$ is finite. Theorem (57) *does not* say that the solution *over the infinite interval* $[0, \infty)$ depends continuously on the initial condition $\mathbf{x}_0$. In fact, we shall see in Chapter 5 that one possible interpretation of so-called Lyapunov stability is precisely that the solution over the infinite interval depends continuously on the initial condition.

68 Example Consider the scalar differential equation

$$\mathbf{69} \quad \dot{x}(t) = \tanh[x(t)] =: f[x(t)], \quad x(0) = x_0.$$

Since the function $\tanh(x)$ is everywhere continuously differentiable, and since this derivative is everywhere bounded (in magnitude) by 1, it is easy to verify that $f(\cdot)$ satisfies a global Lipschitz condition of the form (26) with $k_T = 1$ for all T (see also Problem 2.15 below). Also, for every x_0, there exists a finite constant h_T such that (27) holds. Hence, by Theorem (25), it follows that (69) has a unique solution over $[0, \infty)$ corresponding to each x_0; moreover, for every *finite* number T, the map taking x_0 into the corresponding solution function in $C[0, T]$ is continuous, by Theorem (57).

70 Example Consider the *linear* vector differential equation

71 $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\,\mathbf{x}(t),\ \mathbf{x}(0) = \mathbf{x}_0,$

where $\mathbf{A}(\cdot)$ is continuous. Let $\|\cdot\|$ be a given norm on $\mathbf{R}^n$. Since $\mathbf{A}(\cdot)$ is continuous, for every finite T there exists a finite constant k_T such that

$$\|\mathbf{A}(t)\|_i \leq k_T,\ \forall t \in [0, T].$$

Hence it follows that

$$\|\mathbf{A}(t)\mathbf{x} - \mathbf{A}(t)\mathbf{y}\| \leq k_T \|\mathbf{x} - \mathbf{y}\|,\ \forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n,\ \forall t \in [0, T],$$

$$\|\mathbf{A}(t)\mathbf{x}_0\| \leq k_T \|\mathbf{x}_0\|,\ \forall t \in [0, T].$$

So (26) is satisfied with k_T as above, and (27) is satisfied with $h_T = k_T$. Therefore, by Theorem (25), (71) has a unique solution over $[0, \infty)$ corresponding to each initial condition $\mathbf{x}_0$. Moreover, over each finite interval $[0, T]$, this solution depends continuously on $\mathbf{x}_0$. ■

In conclusion, in this section we have presented some conditions that are *sufficient* to ensure that a given nonlinear vector differential equation has a unique solution over some interval, or over all intervals. It is easy to construct counterexamples to show that the conditions presented here are by no means necessary for the existence and uniqueness of solutions. For instance, consider the scalar differential equation

72 $\dot{x}(t) = -x^2,\ x(0) = 1.$

This equation has a unique solution over $[0, \infty)$, namely $x(t) = 1/(t+1)$, even though the function $f(x) = x^2$ is not globally Lipschitz-continuous.

At a first glance the condition of Lipschitz-continuity appears to be extremely restrictive, since it is known that "almost all" continuous functions are not differentiable and thus not Lipschitz-continuous. Nevertheless, it can be shown that differential equations with unique solutions are *prevalent* in the sense that "almost all" differential equations with continuous functions $\mathbf{f}$ have unique solutions. The arguments used to make this statement precise and to prove it are quite advanced; therefore, they are presented separately in Appendix A. The contents of this appendix show that it is quite reasonable to assume that a given differential equation has a unique solution. This is a useful fact to know, especially when we study the stability of differential equations in Chapter 5.

Problem 2.14 Show that Lipschitz-continuity is independent of which norm on $\mathbf{R}^n$ is used. Precisely, let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two given norms on $\mathbf{R}^n$. Show that for each finite T there exists a finite constant k_{aT} such that

$$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\|_a \leq k_{aT} \|\mathbf{x} - \mathbf{y}\|_a,\ \forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n,\ \forall t \in [0, T],$$

if and only if, for each finite T there exists a finite constant k_{bT} such that

$$\| \mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y}) \|_b \leq k_{bT} \| \mathbf{x} - \mathbf{y} \|_b, \ \forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, \ \forall t \in [0, T].$$

Problem 2.15 (a) Let $f : \mathbf{R}_+ \times \mathbf{R} \to \mathbf{R}$ be continuously differentiable in the second argument. Show that f satisfies (26) if and only if, for each finite T there exists a finite constant k_T such that

$$\left| \frac{\partial f(t, x)}{\partial x} \right| \leq k_T, \ \forall x \in \mathbf{R}, \ \forall t \in [0, T],$$

i.e., $|\partial f(t, x)/\partial x|$ is bounded independently of x over each finite interval $[0, T]$. (Hint: Use the mean-value theorem.)

(b) Let $\mathbf{f}: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}^n$ be continuously differentiable in the second argument. Show that $\mathbf{f}$ satisfies (26) if and only if, for each finite T there exists a finite constant k_T such that

$$\left| \frac{\partial f_i(t, \mathbf{x})}{\partial x_j} \right| \leq k_T, \ \forall i, j, \ \forall \mathbf{x} \in \mathbf{R}^n, \ \forall t \in [0, T].$$

(Hint: Use the results of Problem 2.14 above.)

Problem 2.16 Determine whether or not the following functions satisfy a global Lipschitz condition:

(a) $\mathbf{f}(\mathbf{x}) = [x_1^2 - x_1 x_2 \quad 2x_1 - x_2^2]'$,

(b) $\mathbf{f}(\mathbf{x}) = [x_1 \exp(-x_2^2) \quad x_2 \exp(-x_1^2)]'$.

2.5 SOLUTION ESTIMATES

In this section, we give a method for obtaining both upper and lower bounds on the norm of a solution of a given differential equation. The Gronwall inequality [Lemma (5.7.1)] does give an easily applicable upper bound on the norm of the solution of a linear differential equation, and a similar inequality known as Langenhop's inequality provides a lower bound. However, both of these bounds suffer from the deficiency of being *sign-insensitive*; i.e., they give exactly the same estimates for

1 $\dot{\mathbf{x}}(t) = \mathbf{A}\,\mathbf{x}(t)$

as for

2 $\dot{\mathbf{x}}(t) = -\mathbf{A}(t)\,\mathbf{x}(t)$.

This is because both Gronwall's inequality and Langenhop's inequality (not presented in this book) utilize $\| \mathbf{A}(t) \|$, which is of course sign-insensitive. In contrast, the method given here is based on the concept of the matrix measure, which is sign-sensitive. As a result, the bounds derived in this section are always "tighter" than (or the same as) those given by the

Gronwall and Langenhop inequalities.

3 Theorem *Consider the differential equation*

$$\mathbf{4} \qquad \dot{\mathbf{x}}(t) = \mathbf{A}(t)\,\mathbf{x}(t),\ t \geq 0,$$

where $\mathbf{x}(t) \in \mathbf{R}^n$ *and* $\mathbf{A}(t)$ *is a continuous* $n \times n$ *matrix-valued function. Let* $\|\cdot\|$ *be a norm on* $\mathbf{R}^n$, *and let* $\|\cdot\|_i$ *and* $\mu(\cdot)$ *denote respectively the corresponding induced matrix norm and the corresponding matrix measure on* $\mathbf{R}^{n\times n}$. *Then, whenever* $t \geq t_0 \geq 0$, *we have that*

$$\mathbf{5} \qquad \|\mathbf{x}(t_0)\| \exp\left\{\int_{t_0}^{t} -\mu[-\mathbf{A}(\tau)]\,d\tau\right\} \leq \|\mathbf{x}(t)\| \leq \|\mathbf{x}(t_0)\| \exp\left\{\int_{t_0}^{t} \mu[\mathbf{A}(\tau)]\,d\tau\right\}.$$

Proof From Example (2.4.70), we know that the differential equation (4) has a unique solution over $[0, \infty)$. To prove the inequalities (5), observe first that, from the integral form of (4), it follows that

$$\mathbf{6} \qquad \mathbf{x}(t+\delta) = \mathbf{x}(t) + \delta\mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{o}(\delta),\ \forall \delta > 0,$$

where $\mathbf{o}(\delta)$ denotes an error term with the property that

$$\mathbf{7} \qquad \lim_{\delta \to 0} \frac{\|\mathbf{o}(\delta)\|}{\delta} = 0.$$

Rearranging (6) gives, successively,

$$\mathbf{8} \qquad \mathbf{x}(t+\delta) = [\mathbf{I} + \delta\mathbf{A}(t)]\,\mathbf{x}(t) + \mathbf{o}(\delta),$$

$$\mathbf{9} \qquad \|\mathbf{x}(t+\delta)\| \leq \|[\mathbf{I} + \delta\mathbf{A}(t)]\|_i \cdot \|\mathbf{x}(t)\| + o(\delta),$$

$$\mathbf{10} \qquad \|\mathbf{x}(t+\delta)\| - \|\mathbf{x}(t)\| \leq (\|I + \delta\mathbf{A}(t)\|_i - 1) \cdot \|\mathbf{x}(t)\| + o(\delta),$$

$$\mathbf{11} \qquad \frac{d^+}{dt}\|\mathbf{x}(t)\| = \lim_{\delta \to 0^+} \frac{\|\mathbf{x}(t+\delta)\| - \|\mathbf{x}(t)\|}{\delta} \leq \mu[\mathbf{A}(t)]\,\|\mathbf{x}(t)\|,$$

where d^+/dt denotes the right-hand derivative. Multiplying both sides of (11) by the integrating factor

$$\mathbf{12} \qquad \exp\left\{-\int_{t_0}^{t} \mu[\mathbf{A}(\tau)]\,d\tau\right\}$$

(or, equivalently, applying the Gronwall inequality) gives the right-hand inequality in (5). The proof of the left-hand inequality in (5) is entirely similar, starting with

13 $\mathbf{x}(t-\delta) = \mathbf{x}(t) - \delta\mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{o}(\delta)$.

The completion of the proof is left as an exercise. ■

Theorem (3) provides both upper and lower bounds for the norm of the solution of the unforced linear equation (4). In applying the bounds (5), it is important to remember that the norm being used and the measure must correspond to one another. Also, using different norms in Theorem (3) will give rise to distinct bounds. This is illustrated by the following examples.

14 Example Consider the equation (4) with $n = 2$ and

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

First, let us calculate the measures μ_1, μ_2, μ_∞ of the matrix $\mathbf{A}(t)$. This gives

$$\mu_1[\mathbf{A}(t)] = \mu_\infty[\mathbf{A}(t)] = -t+1,$$

$$\mu_1[-\mathbf{A}(t)] = \mu_\infty[-\mathbf{A}(t)] = 2t + 1,$$

$$\mu_2[\mathbf{A}(t)] = -t, \ \mu_2[-\mathbf{A}(t)] = 2t.$$

Thus, applying the inequalities (5) with each of the above measures gives

$$\exp(-t-t^2) \le |x_1(t)| + |x_2(t)| \le \exp(-t-t^2/2),$$

$$\exp(-t-t^2) \le |x_1(t)|, \ |x_2(t)| \le \exp(-t-t^2/2),$$

$$\exp(-t^2) \le [|x_1(t)|^2 + |x_2(t)|^2]^{1/2} \le \exp(-t^2/2).$$

Thus the same two inequalities (5), when applied with different vector norms and corresponding matrix measures, yield different estimates for the vector $\mathbf{x}(t)$. By way of illustrating the bounds obtained above, the regions of $\mathbf{R}^2$ to which the vector $\mathbf{x}(1)$ is confined by each of the above bounds are shown in Figures 2.3, 2.4, and 2.5, respectively.

15 Example Consider the equation (4) with $n = 2$ and

$$\mathbf{A}(t) = \begin{bmatrix} -3t & t \\ 2t & -4t \end{bmatrix}, \quad \mathbf{x}(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Then the actual solution for $\mathbf{x}(t)$ is

$$\mathbf{x}(t) = [(4/3)\exp(-t^2) - (1/3)\exp(-5t^2/2) \quad (4/3)\exp(-t^2) + (2/3)\exp 9{-}5t^2/2)]'.$$

However, if we calculate the various measures of $\mathbf{A}(t)$, we get

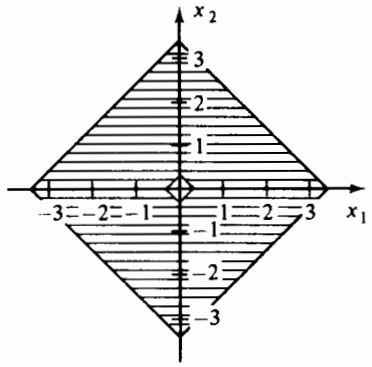


Fig. 2.3

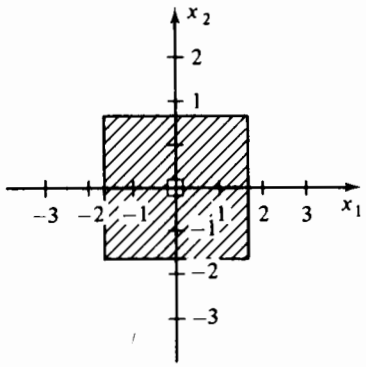


Fig. 2.4

$$\mu_1[\mathbf{A}(t)] = -t,\ \mu_1[-\mathbf{A}(t)] = 5t,$$

$$\mu_2[\mathbf{A}(t)] \approx -2.97t,\ \mu_2[-\mathbf{A}(t)] \approx 5.03t,$$

$$\mu_\infty[\mathbf{A}(t)] = -2t,\ \mu_\infty[-\mathbf{A}(t)] = 6t.$$

Thus the corresponding estimates for $\mathbf{x}(t)$ are as follows:

$$3\exp(-2.5t^2) \le |x_1(t)| + |x_2(t)| \le 3\exp(-0.5t^2),$$

$$\sqrt{5}\exp(-2.52t^2) \le [|x_1(t)|^2 + |x_2(t)|^2]^{1/2} \le \sqrt{5}\exp(-1.48t^2),$$

$$2\exp(-3t^2) \le |x_1(t)|,\ |x_2(t)| \le 2\exp(-t^2).$$

The bounds are depicted for the case $t = 0.5$ in Figures 2.6, 2.7 and 2.8, respectively.

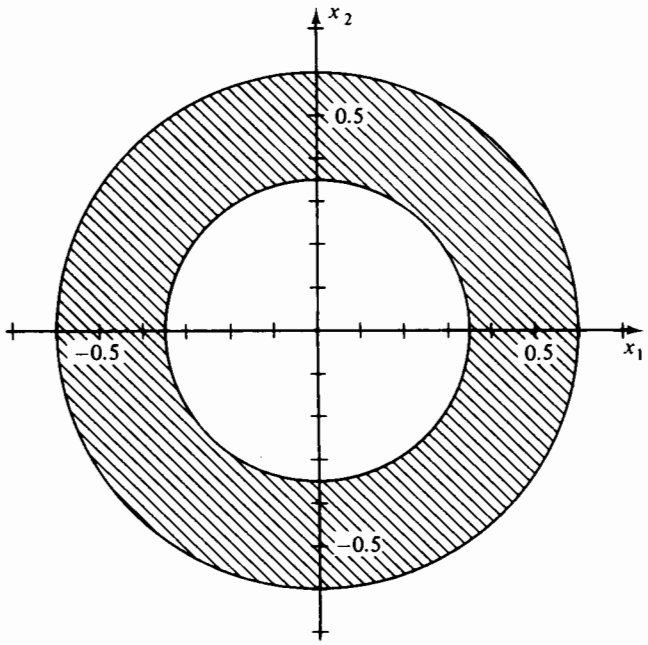


Fig. 2.5

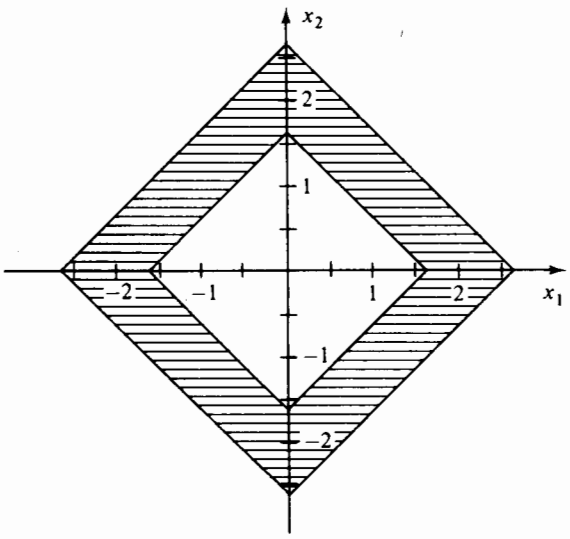


Fig. 2.6

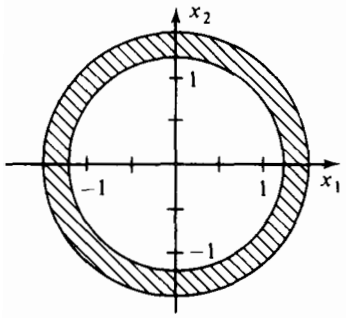


Fig. 2.7

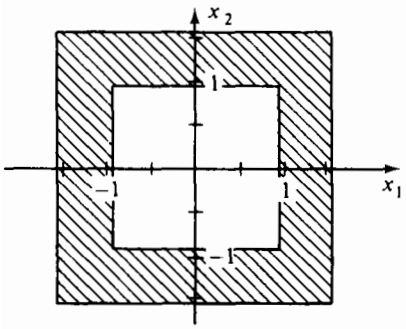


Fig. 2.8

To extend the above estimation technique to nonlinear differential equations of the form

16 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], \mathbf{x}(0) = \mathbf{x}_0,$

a preliminary result is needed.

17 Lemma *Suppose* $\mathbf{f}: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}^n$ *is continuously differentiable. Then there exists a continuous function* $\mathbf{A}: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}^{n \times n}$ *such that*

18 $\mathbf{f}(t, \mathbf{x}) = \mathbf{f}(t, \mathbf{0}) + \mathbf{A}(t, \mathbf{x})\mathbf{x}, \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$

Proof Fix t and x, and consider $\mathbf{f}(t, \lambda \mathbf{x})$ as a function of the scalar parameter λ. Then

$$\mathbf{19} \quad \mathbf{f}(t, \mathbf{x}) = \mathbf{f}(t, \mathbf{0}) + \int_0^1 \frac{d}{d\lambda} \mathbf{f}(t, \lambda \mathbf{x})\, d\lambda = \mathbf{f}(t, \mathbf{0}) + \left[\int_0^1 \nabla_{\mathbf{x}} \mathbf{f}(t, \lambda \mathbf{x})\, d\lambda \right] \cdot \mathbf{x}.$$

Hence (18) holds with

$$\mathbf{20} \quad \mathbf{A}(t, \mathbf{x}) = \int_0^1 \nabla_{\mathbf{x}} \mathbf{f}(t, \lambda \mathbf{x})\, d\lambda.$$

Note that there is nothing special about the origin in the above formula. Indeed, given *any* fixed $\mathbf{x}_0 \in \mathbf{R}^n$, we can write

$$\mathbf{21} \quad \mathbf{f}(t, \mathbf{x}) = \mathbf{f}(t, \mathbf{x}_0) + \mathbf{B}(t, \mathbf{x}, \mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

for a suitably chosen matrix-valued function $\mathbf{B}(t, \mathbf{x}, \mathbf{x}_0)$.

22 Theorem *Consider the differential equation (16), and suppose (i)* $\mathbf{f}$ *is continuously differentiable, and (ii)* $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}\ \forall t \geq 0$. *Define* $\mathbf{A}(t, \mathbf{x})$ *as in (18). Let* $\|\cdot\|$ *be a norm on* $\mathbf{R}^n$, *and let* $\|\cdot\|_i$ *and* $\mu(\cdot)$ *denote the corresponding induced norm and matrix measure on* $\mathbf{R}^{n \times n}$. *Suppose there exist continuous functions* $\alpha(\cdot)$ *and* $\beta(\cdot)$ *such that*

$$\mathbf{23} \quad \mu[\mathbf{A}(t, \mathbf{x})] \leq \alpha(t),\ \beta(t) \leq \mu[-\mathbf{A}(t, \mathbf{x})],\ \forall t \geq 0,\ \forall \mathbf{x} \in \mathbf{R}^n.$$

Then

$$\mathbf{24} \quad \|\mathbf{x}_0\| \exp\left[-\int_0^t \beta(\tau)\, d\tau\right] \leq \|\mathbf{x}(t)\| \leq \|\mathbf{x}_0\| \exp\left[\int_0^t \alpha(\tau)\, d\tau\right],\ \forall t \geq 0.$$

The proof is virtually the same as that of Theorem (1) and is left as an exercise.

Notes and References

The material in this chapter is quite standard, and can be found in most textbooks on differential equations, e.g., Hartman (1964). The matrix measure was introduced by Dahlquist (1959), while the solution estimates given in Section 2.5, based on the matrix measure are due to Coppel (1965). Appendix A contains a result due to Orlicz (1932), to the effect that "almost all" differential equations have a unique solution.

3. SECOND-ORDER SYSTEMS

3.1 PRELIMINARIES

In this chapter, we study several techniques for the analysis of autonomous second-order systems. In subsequent chapters, this restriction on the order of the system is removed and some techniques are presented for analyzing systems of any order, autonomous or otherwise. Obviously the latter techniques are also applicable to second-order systems. However, second-order systems occupy a special place in the study of nonlinear systems. The most important reason is that the solution trajectories of a second-order system can be represented by curves in the *plane*. As a result, nonlinear systems concepts such as oscillations, vector fields, etc. have simple geometric interpretations in the case of second-order systems. (All the technical terms used above will be defined shortly.) For these and other reasons, second-order systems, by themselves, have been the subject of much research, and in this chapter we present some of the simpler results that are available.

Consider a general second-order system described by the scalar differential equations

1 $\dot{x}_1(t) = f_1[t, x_1(t), x_2(t)], \dot{x}_2(t) = f_2[t, x_1(t), x_2(t)].$

A basic concept in the analysis of second-order systems is the so-called state-plane plot. The **state-plane** is the usual two-dimensional plane with the horizontal axis labeled x_1 and the vertical axis labeled x_2. Suppose $[x_1(\cdot), x_2(\cdot)]$ denotes a solution of (1). Then a plot of $x_1(t)$ versus $x_2(t)$ as t varies over $\mathbf{R}_+$ is called a **state-plane plot** or a **state-plane trajectory** of the system (1). In such a plot, the time t is a parameter that can either be explicitly displayed or omitted. In the special case where the first equation in (1) is of the form

2 $\dot{x}_1(t) = x_2(t),$

it is customary to refer to the state plane as the **phase plane.** Correspondingly, in this case one also refers to **phase-plane plots** or **phase-plane trajectories.** This special case arises quite commonly in practice. In particular, if the system under study is governed by a second-order scalar differential equation of the form

3 $\ddot{y}(t) = g[t, y(t), \dot{y}(t)],$

then a natural choice for the state variables is

4 $x_1(t) = y(t), x_2(t) = \dot{y}(t).$

In this case, the system equation (3) is equivalent to the following two first-order equations:

5 $\dot{x}_1(t)=x_2(t), \dot{x}_2(t)=g[t, x_1(t), x_2(t)]$.

In the case of autonomous systems, i.e., where the function g in (5) does not explicitly depend on the time t, phase-plane plots have another useful feature, namely: it is possible to reconstruct the implicit parameter t from the phase-plane plot. Suppose we are given a phase-plane plot denoted by **C**, and suppose it is known that a particular point (x_{10}, x_{20}) corresponds to a time t_0. Typically t_0 is the initial time and (x_{10}, x_{20}) is the initial state of the system. If (x_{1f}, x_{2f}) is another point on **C**, the value of t (say t_f) which corresponds to (x_{1f}, x_{2f}) can be determined as follows: If x_2 does not change sign along **C** between (x_{10}, x_{20}) and (x_{1f}, x_{2f}), then

6 $$t_f = t_0 + \int_{\mathbf{C}} \frac{dx_1}{x_2},$$

where the integral in (6) is taken along the curve **C** (see Figure 3.1). If x_2 changes sign along **C**, then the integral in (6) has to be evaluated as the sum of several integrals, one corresponding to each segment of **C** along which x_2 does not change sign (see Figure 3.2). Note that, as $x_{2f} \to 0$, the integral in (6) becomes an improper integral. The proof of the relationship (6) is easily obtained starting from (5) and is left as an exercise (see Problem 3.1).

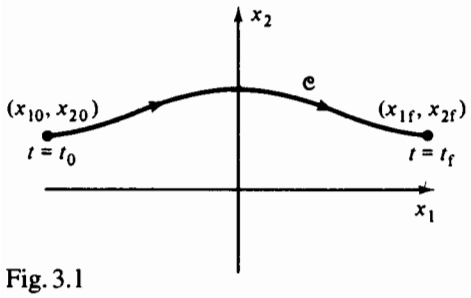


Fig. 3.1

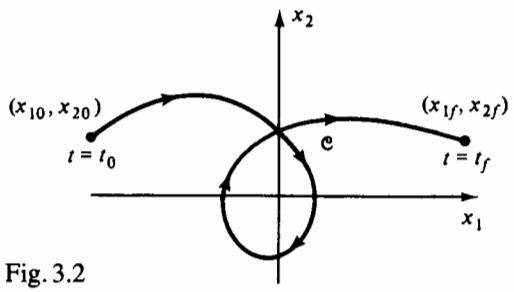


Fig. 3.2

Another very important concept is a vector field. A couple of preliminary notions are needed to introduce this concept.

A function $f: \mathbf{R}^2 \to \mathbf{R}$ is said to be **smooth** if $f(x_1, x_2)$ has continuous partial derivatives of all orders with respect to all combinations of x_1 and x_2, i.e., if the partial derivative $\partial^n f / \partial x_1^i \partial x_2^{n-i}$ is well-defined and continuous for all integers $n \geq i \geq 1$.

Suppose a, b are real numbers, not both zero. Then the two-argument arc tangent function $\text{Atan}(a, b)$ is defined as the unique number $\theta \in [0, 2\pi)$ such that

$$\mathbf{7} \qquad \cos\theta = \frac{a}{a^2 + b^2}, \sin\theta = \frac{b}{a^2 + b^2}.$$

Note that $\text{Atan}(a, b) = \text{Atan}(ra, rb)$ provided $r > 0$ (but not if $r < 0$). $\text{Atan}(0, 0)$ is undefined.

8 Definition *A function* $\mathbf{f}: \mathbf{R}^2 \to \mathbf{R}^2$ *is called a* **vector field** *if both of its components are smooth functions. A vector* $\mathbf{x} \in \mathbf{R}^2$ *is called an* **equilibrium** *of a vector field* $\mathbf{f}$ *if* $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. *If* $\mathbf{x} \in \mathbf{R}^2$ *is not an equilibrium of* $\mathbf{f}$, *then the* **direction** *of the vector field* $\mathbf{f}$ *at the point* $\mathbf{x}$ *is denoted by* $\theta_{\mathbf{f}}(\mathbf{x})$ *and is defined as*

$$\mathbf{9} \qquad \theta_{\mathbf{f}}(\mathbf{x}) = \text{Atan}[f_1(\mathbf{x}), f_2(\mathbf{x})].$$

Figure 3.3 depicts the quantity $\theta_{\mathbf{f}}(\mathbf{x})$.

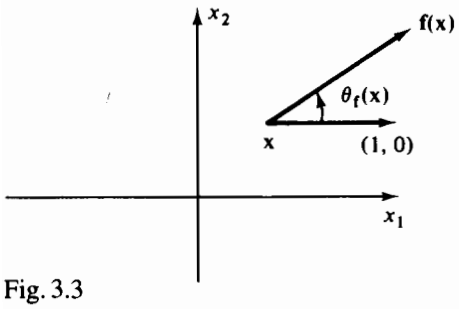


Fig. 3.3

To see the utility of these concepts, suppose $\mathbf{f}: \mathbf{R}^2 \to \mathbf{R}^2$ is a vector field, and consider the associated differential equation

$$\mathbf{10} \qquad \dot{x}_1 = f_1(x_1, x_2), \dot{x}_2 = f_2(x_1, x_2).$$

Note that here and in the remainder of the chapter we follow the standard practice of not explicitly displaying the time variable t.

Suppose $\mathbf{x} = (x_1, x_2)$ is a point in $\mathbf{R}^2$; then it is easy to see from (10) that if $\mathbf{C}$ is a solution trajectory of (10) passing through $\mathbf{x}$, then the vector $\mathbf{f}(\mathbf{x})$ is tangent to $\mathbf{C}$ at $\mathbf{x}$. Hence, in principle at least, it is possible to construct graphically the solution trajectories of (10) by plotting the vector field $\mathbf{f}(\mathbf{x})$. Actually, the concept is very deep and has many applications,

only a few of which are touched upon in this book. Furthermore, the concept of a vector field is applicable to (autonomous) systems of any order. The reader interested in a deeper knowledge of the application of the vector field concept to differential equations may consult Arnold (1973). Vector fields are encountered again in this book in Chapter 7.

Note that it is quite common to refer to $\mathbf{f}(\mathbf{x})$ as the **velocity vector field** associated with the system of equations (10).

The objective of the present chapter is to present some ways of analyzing the system (10) by either finding the state-plane trajectory of the system to a reasonably high degree of accuracy or determining some qualitative features of the state-plane trajectory without doing too much work. Throughout the chapter, the study is confined to autonomous systems, because even though the concept of a state-plane trajectory is valid for nonautonomous systems, most of the significant results are applicable only to autonomous systems. For example, the autonomous system (10) has a periodic solution $\mathbf{x}(t)$ if the corresponding solution trajectory is a closed curve in $\mathbf{R}^2$. An analogous statement for nonautonomous systems is false in general.

Finally, a word about the existence and uniqueness of solutions to the system of equations (10). Since $\mathbf{f}$ is smooth, it follows from Corollary (2.4.22) that (10) has a unique solution at least locally; that is, given the system (10) together with an initial condition

11 $x_1(0)=x_{10}, x_2(0)=x_{20},$

there exists a number δ such that (10-11) has exactly one solution over $[0, \delta)$. Additional conditions on $\mathbf{f}$ ensure that (10-11) has a unique solution over all of $[0, \infty)$; see Theorem (2.4.25).

Problem 3.1 Prove the relationship (6). Hint: Use (5) to write

$$x_1(t+\Delta t)=x_1(t)+\Delta t\, x_2(t)+o(\Delta t).$$

Problem 3.2 Show that if $\mathbf{C}$ is a solution trajectory of (10) passing through $\mathbf{x}$, then the vector field $\mathbf{f}(\mathbf{x})$ is tangent to $\mathbf{C}$ at $\mathbf{x}$. Hint: Express (10) in difference form as

$$x_1(t+\Delta t)=x_1(t)+\Delta t\, f_1[x_1(t), x_2(t)]+o(\Delta t),$$

$$x_2(t+\Delta t)=x_2(t)+\Delta t\, f_2[x_1(t), x_2(t)]+o(\Delta t),$$

and eliminate Δt as $\Delta t \to 0$.

Problem 3.3 Does the function $\mathbf{f}:\mathbf{R}^2 \to \mathbf{R}^2$ defined by

$$f_1(x_1, x_2)=x_2+[1-(x_1^2+x_2^2)^{1/2}],$$

$$f_2(x_1, x_2)=-x_1+[1-(x_1^2+x_2^2)^{1/2}]$$

constitute a vector field? Justify your answer. Hint: Consider the behavior of $\mathbf{f}$ near the

origin.

3.2 LINEARIZATION METHOD

We begin by studying linear systems, which are simpler to analyze than nonlinear systems and yet provide much insight into the behavior of nonlinear systems. The general form of a second-order autonomous linear system is

$$\mathbf{1} \qquad \dot{x}_1 = a_{11}x_1 + a_{12}x_2, \dot{x}_2 = a_{21}x_1 + a_{22}x_2,$$

together with the initial conditions

$$\mathbf{2} \qquad x_1(0) = x_{10}, x_2(0) = x_{20}.$$

In matrix notation (1) and (2) can be expressed as

$$\mathbf{3} \qquad \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \mathbf{x}(0) = \mathbf{x}_0.$$

To understand better the behavior of solutions to (3), it is helpful to make a transformation of variables. Accordingly, let

$$\mathbf{4} \qquad \mathbf{z} = \mathbf{M}\mathbf{x},$$

where $\mathbf{M}$ is a constant nonsingular 2×2 matrix with *real* coefficients. In terms of the transformed variable $\mathbf{z}$, (3) becomes

$$\mathbf{5} \qquad \dot{\mathbf{z}}(t) = \mathbf{M}\mathbf{A}\mathbf{M}^{-1}\,\mathbf{z}(t), \mathbf{z}(0) = \mathbf{M}\mathbf{x}_0.$$

It is known [see, for example, Bellman (1970)] that by appropriately choosing the matrix $\mathbf{M}$, the matrix $\mathbf{M}\mathbf{A}\mathbf{M}^{-1}$ can be made to have one of the following forms:

1. *Diagonal form*: In this case,

$$\mathbf{6} \qquad \mathbf{M}\mathbf{A}\mathbf{M}^{-1} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

where λ_1 and λ_2 are the real (and not necessarily distinct) eigenvalues of the matrix $\mathbf{A}$.

2. *Jordan form*: In this case,

$$\mathbf{7} \qquad \mathbf{M}\mathbf{A}\mathbf{M}^{-1} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix},$$

where λ is the real *repeated* eigenvalue of the matrix $\mathbf{A}$.

3. *Complex conjugate form*: In this case,

$$\mathbf{8} \qquad \mathbf{MAM}^{-1} = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix},$$

where $\alpha \pm j\beta$ are the complex conjugate eigenvalues of $\mathbf{A}$, and we choose $\beta > 0$ to be definite.

Each of these cases is studied in detail.

Case 1 *Diagonal Form*: In this case (5) assumes the form

$$\mathbf{9} \qquad \dot{z}_1(t) = \lambda_1 z_1(t),\ \dot{z}_2(t) = \lambda_2 z_2(t),$$

$$z_1(0) = z_{10},\ z_2(0) = z_{20}.$$

The solution of (9) is

$$\mathbf{10} \qquad z_1(t) = z_{10}\, e^{\lambda_1 t},\ z_2(t) = z_{20}\, e^{\lambda_2 t}.$$

At this point it can be assumed that not both λ_1 and λ_2 are zero, because if both λ_1, λ_2 are zero then $\mathbf{A} = \mathbf{0}$ and $\mathbf{z}(t) = \mathbf{z}_0$ for all t; consequently the state-plane plot consists of just a single point. Thus suppose $\lambda_1 \neq 0$. Then the parameter t can be eliminated from (10) to give

$$\mathbf{11} \qquad z_2 = z_{20} \cdot \left[\frac{z_1}{z_{10}} \right]^{\lambda_2/\lambda_1}.$$

Equation (11) describes the state-plane trajectory of (9) in the z_1–z_2 plane. If λ_1 and λ_2 are of the same sign, then the trajectories have the characteristic shape shown in Figure 3.4, but if λ_1 and λ_2 have opposite signs then the trajectories have the characteristic shape shown in Figure 3.5. The arrowheads in Figure 3.4 correspond to the case where $\lambda_2 < \lambda_1 < 0$; if λ_1 and λ_2 are both positive then the direction of the arrowheads is reversed, and the trajectories go *away* from the origin as t increases instead of going *towards* the origin as in Figure 3.4. Similarly the arrowheads in Figure 3.5 correspond to the case where $\lambda_1 < 0 < \lambda_2$. It should be emphasized that the trajectories depicted in Figures 3.4 and 3.5 are in the z_1–z_2 coordinate system; the corresponding trajectories in the x_1–x_2 coordinate system, although they will have the same general appearance as those in the z_1–z_2 coordinate system, will be a little distorted. This can be seen in Figures 3.6 and 3.7, where the trajectories in the x_1–x_2 coordinate system are illustrated for the cases where λ_1 and λ_2 are of the same sign, and where λ_1 and λ_2 are of opposite signs, respectively. If λ_1 and λ_2 are of the same sign, then the equilibrium at the origin is referred to as a **node.** It is called a **stable node** if both λ_1 and λ_2 are negative, and an **unstable node** if λ_1 and λ_2 are positive. In the case where λ_1 and λ_2 are of opposite sign, the equilibrium at the origin is called a **saddle**. The rationale for this nomenclature is that if one were to make a three-dimensional plot of $[x_1(t), x_2(t), t]$, then the resulting surface would resemble a saddle.

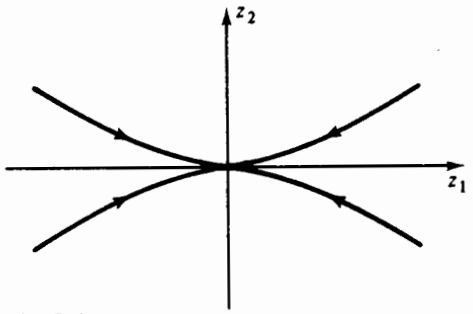


Fig. 3.4

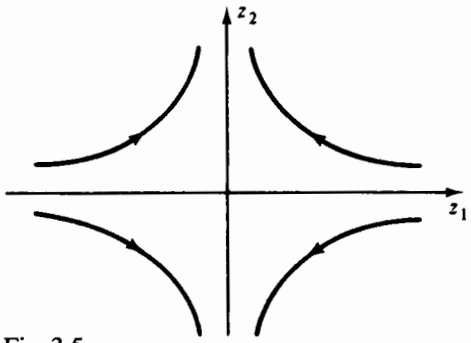


Fig. 3.5

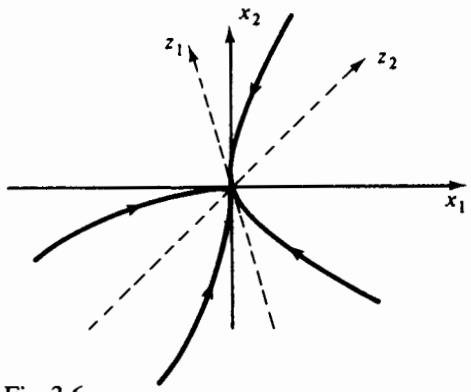


Fig. 3.6

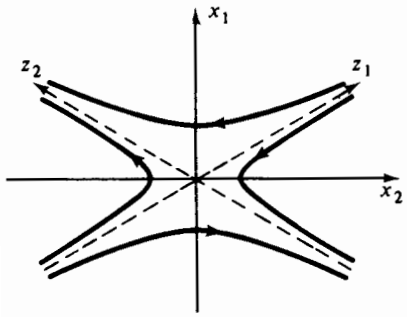


Fig. 3.7

Case 2 *Jordan Form*: In this case (5) assumes the form

12 $\dot{z}_1(t)=\lambda z_1(t)+z_2(t),\ \dot{z}_2(t)=\lambda z_2(t),$

$z_1(0)=z_{10},\ z_2(0)=z_{20}.$

The solution of (12) is

13 $z_1(t)=z_{10}\,e^{\lambda t}+z_{20}t\,e^{\lambda t},\ z_2(t)=z_{20}\,e^{\lambda t}.$

Once again, t can be eliminated from (13); the resulting expression describing the trajectory is somewhat messy and its derivation is left as a problem (see Problem 3.4). The trajectories in the z_1–z_2 coordinate system, which can be obtained from (13), are shown in Figure 3.8 for the case $\lambda<0$; if $\lambda>0$, then the direction of the arrows is reversed. The corresponding trajectories in the x_1–x_2 coordinate system are shown in Figure 3.9. In this case also, the equilibrium (0, 0) is called a **stable node** if $\lambda<0$ and an **unstable node** if $\lambda>0$.

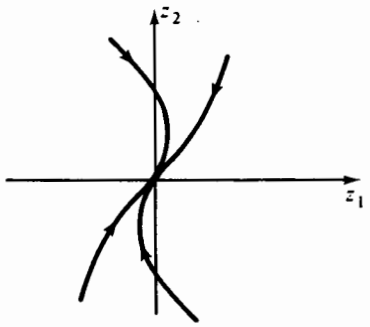


Fig. 3.8

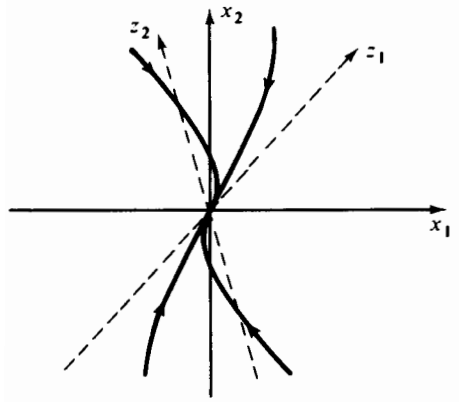


Fig. 3.9

Case 3 *Complex Conjugate Form:* In this case (5) becomes

14 $\dot{z}_1(t) = \alpha z_1(t) - \beta z_2(t),\ \dot{z}_2(t) = \beta z_1(t) + \alpha z_2(t),$

$z_1(0) = z_{10},\ z_2(0) = z_{20}.$

To simplify the equation further, introduce the polar coordinates

15 $r = (z_1^2 + z_2^2)^{1/2},\ \phi = \text{Atan}(z_1, z_2).$

Then (14) is transformed into

16 $\dot{r}(t) = \alpha r(t),\ \dot{\phi}(t) = \beta,$

which has the solution

17 $r(t) = r(0)\, e^{\alpha t},\ \phi(t) = \phi(0) + \beta t.$

In the z_1–z_2 coordinate system, (17) represents an exponential spiral. If $\alpha > 0$, then the spiral expands as t increases, whereas if $\alpha < 0$, then the spiral shrinks as t increases; and if $\alpha = 0$ the trajectory is a circle. The equilibrium (0, 0) is referred to as an **unstable focus** if $\alpha > 0$, a **stable focus** if $\alpha < 0$, and a **center** if $\alpha = 0$. The trajectories in the z_1–z_2 coordinate system corresponding to each of these cases are depicted in Figures 3.10, 3.11 and 3.12.

Table 3.1 summarizes the various kinds of equilibria for second-order linear systems. Note that λ_1, λ_2 are the eigenvalues of the matrix **A**.

Now consider an autonomous nonlinear system described by

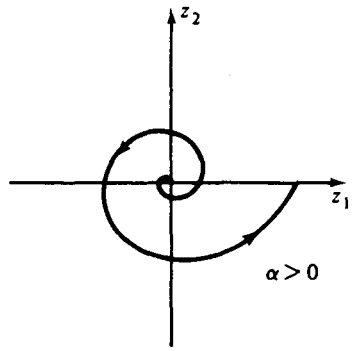


Fig. 3.10

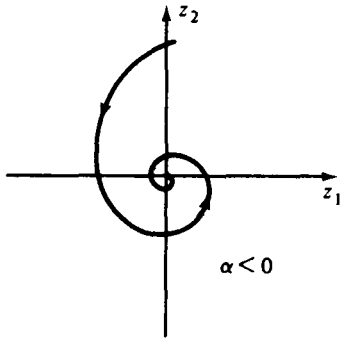


Fig. 3.11

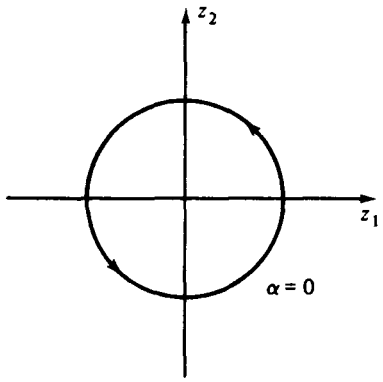


Fig. 3.12

Table 3.1

Eigenvalues of A	Type of Equilibrium
λ_1, λ_2 real, both negative	Stable node
λ_1, λ_2 real, both positive	Unstable node
λ_1, λ_2 real, $\lambda_1\lambda_2 < 0$	Saddle
λ_1, λ_2 complex, $\operatorname{Re}\lambda_i < 0$	Stable focus
λ_1, λ_2 complex, $\operatorname{Re}\lambda_i > 0$	Unstable focus
λ_1, λ_2 imaginary	Center

18 $\dot{x}_1 = f_1(x_1, x_2), \dot{x}_2 = f_2(x_1, x_2).$

The linearization method, as the name implies, consists of linearizing the given system in the neighborhood of an equilibrium and determining the behavior of the *nonlinear* system by studying the resulting *linear* system. The power of the method lies in the fact that, except for special cases to be specified later, the method yields definitive results that are valid in some neighborhood of the equilibrium.

The method can be summarized as follows: Suppose (0, 0) is an equilibrium of the system (1) and that both f_1 and f_2 are continuously differentiable in some neighborhood of (0, 0). Define

19 $$a_{ij} = \left[\frac{\partial f_i}{\partial x_j}\right]_{\mathbf{x}=\mathbf{0}}, \quad i, j = 1, 2,$$

20 $$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Then, by Taylor's theorem, it is possible to expand f_1 and f_2 in the form

21 $$f_1(x_1, x_2) = f_1(0, 0) + a_{11}x_1 + a_{12}x_2 + r_1(x_1, x_2)$$
$$= a_{11}x_1 + a_{12}x_2 + r_1(x_1, x_2),$$
$$f_2(x_1, x_2) = a_{21}x_1 + a_{22}x_2 + r_2(x_1, x_2), \quad + f_2(00)$$

where r_1 and r_2 are the remainder terms, and we have used the fact that $f_i(0, 0) = 0$ since (0, 0) is an equilibrium. If the equilibrium is not at (0, 0) but at some other point in $\mathbf{R}^2$, then one can always translate the coordinates in such a way that the equilibrium is at the origin of the new coordinate system. Now, associated with the nonlinear system (18), define the *linear* system

22 $\dot{\xi}_1 = a_{11}\xi_1 + a_{12}\xi_2, \dot{\xi}_2 = a_{21}\xi_1 + a_{22}\xi_2.$

The linearization method is based on the fact (proved in Section 5.5) that if the matrix **A** does not have any eigenvalues with zero real parts, then the trajectories of the *nonlinear* system (18) in the vicinity of the equilibrium $x_1 = 0, x_2 = 0$ have the same characteristic shape as the trajectories of the *linear* system (22) in the vicinity of the equilibrium $\xi_1 = 0, \xi_2 = 0$. Table 3.2 summarizes the situation.

Table 3.2

Equilibrium of the Linear System (22)	Equilibrium of the Nonlinear System (18)
Stable node	Stable node
Unstable node	Unstable node
Saddle	Saddle
Stable focus	Stable focus
Unstable focus	Unstable focus
Center	

The last entry in the table can be explained as follows: If the equilibrium (0, 0) of the system (22) is a center, then the linearized system exhibits perfect oscillations which neither grow nor decay with time. In such a case, the behavior of the trajectories of the original nonlinear system is determined by the remainder terms r_1 and r_2, which are neglected in the linearization. Studying the linearized system alone does not provide a definitive answer about the behavior of the nonlinear system.

23 Example Consider the following second-order equation, commonly known as Van der Pol's equation:

24 $\ddot{y} - \mu(1 - y^2)\dot{y} + y = 0,$

where $\mu > 0$ is a constant. By defining the natural state variables

$$x_1 = y, x_2 = \dot{y},$$

(24) is transformed into the pair of first-order equations

25 $\dot{x}_1 = x_2, \dot{x}_2 = -x_1 + \mu(1 - x_1^2)x_2.$

The linearization of (25) around the equilibrium (0, 0) is

26 $\dot{\xi}_1 = \xi_2, \dot{\xi}_2 = -\xi_1 + \mu\xi_2.$

The eigenvalues of the associated matrix **A** satisfy the characteristic equation

27 $\lambda^2 - \mu\lambda + 1 = 0.$

For all positive values of μ, the roots of (27) are complex with positive real parts, so that the equilibrium $\xi_1 = 0, \xi_2 = 0$ of (26) is an unstable focus. Referring to Table 3.2, we see that the equilibrium $x_1 = 0, x_2 = 0$ of the original system (8) is also an unstable focus.

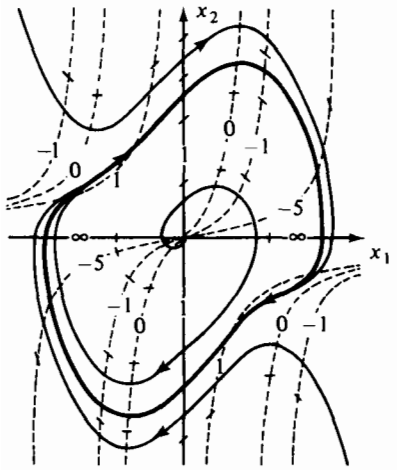


Fig. 3.13

Figure 3.13 shows the phase-plane trajectories of the Van der Pol oscillator. A notable feature of this system is that all solution trajectories starting from an initial state other than (0, 0) approach a limit cycle. This system is further analyzed in Section 3.4.

Problem 3.4 Eliminate t from (13) and obtain an expression for the state-plane trajectory involving only z_1, z_2, z_{10}, and z_{20}.

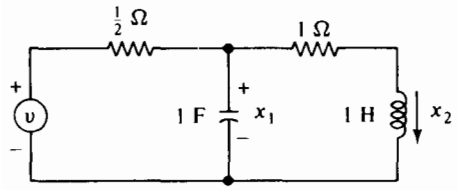


Fig. 3.14

Problem 3.5 Consider the electrical circuit shown in Figure 3.14.

(a) Select the capacitor voltage x_1 and the inductor current x_2 as the state variables, and show that the network is described by the equations

$$\dot{x}_1 = -2x_1 - x_2 + 2v,\ \dot{x}_2 = x_1 - x_2.$$

(b) Suppose $v(t) \equiv 0$. Determine the nature of the equilibrium at (0, 0) and find the matrix **M** that transforms the above equations into the appropriate canonical form.

Problem 3.6 Suppose the 1/2 Ohm resistor in Figure 3.14 is replaced by a general resistor R.

(a) Write the state equations for the network with $v(t) \equiv 0$.

(b) For what values of R is the equilibrium (0, 0) (i) a node, (ii) a focus, (iii) a saddle?

Problem 3.7 For each of the matrices **A** given below:

(a) Determine the matrix **M** that transforms **A** into the appropriate canonical form.

(b) Sketch the state-plane trajectories in both the z_1–z_2 coordinates and the x_1–x_2 coordinate system.

(c) Classify the equilibrium at (0, 0) as to its type.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 5 \\ -1 & -1 \end{bmatrix}, \begin{bmatrix} 2 & -1 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 2 & -2 \end{bmatrix}.$$

Problem 3.8 Find all equilibria of the Volterra predator-prey equations

$$\dot{x}_1 = -x_1 + x_1x_2,\ \dot{x}_2 = x_2 - x_1x_2.$$

Linearize the system around each of the equilibria and determine, if possible, the nature of the equilibrium. (Answer: One center, one saddle).

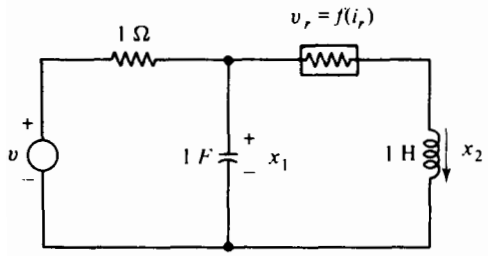


Fig. 3.15

Problem 3.9 Consider the nonlinear circuit shown in Figure 3.15. Suppose the voltage-current relationship of the nonlinear resistor is given by

$$v_r = i_r^3 - 3i_r^2 + 3i_r =: f(i_r).$$

(a) Select the capacitor voltage x_1 and the inductor x_2 as the state variables, and show that the system equations are

$$\dot{x}_1 = v - x_1 - x_2, \dot{x}_2 = x_1 - f(x_2).$$

(b) With $v = 0$, calculate the equilibria of the system.

(c) Linearize the system around each of the equilibria and determine the nature of each equilibrium.

3.3 PERIODIC SOLUTIONS

3.3.1 Introduction

Some autonomous systems exhibit periodic solutions. For example, consider a simple harmonic oscillator, which is described by the linear equations

1 $$\dot{x}_1 = x_2, \dot{x}_2 = -x_1.$$

The solution of (1) subject to the initial conditions

2 $$x_1(0) = x_{10}, x_2(0) = x_{20}$$

is given by

3 $$x_1(t) = r_0 \cos(-t + \phi_0), x_2(t) = r_0 \sin(-t + \phi_0),$$

where

4 $$r_0 = (x_{10}^2 + x_{20}^2)^{1/2}, \phi_0 = \text{Atan}(x_{10}, x_{20}).$$

Thus the solution of (1) is periodic irrespective of the initial conditions. Furthermore, the entire phase-plane is covered with periodic solutions of (1): Given any point (x_{10}, x_{20}), one can always find a periodic solution passing through it.

In contrast, consider the system of *nonlinear* equations

5 $$\dot{x}_1 = x_2 + \alpha x_1(\beta^2 - x_1^2 - x_2^2), \dot{x}_2 = -x_1 + \alpha x_2(\beta^2 - x_1^2 - x_2^2).$$

These equations can be expressed as

6 $$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x}),$$

where

$$\mathbf{7} \qquad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix}, \ \mathbf{g}(\mathbf{x}) = \alpha(\beta - x_1^2 - x_2^2)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Note that **f** is exactly the velocity vector field of the system (1), while **g** is a so-called **radial** vector field, i.e., **g**(**x**) is always aligned with the vector **x**. Now introduce the polar coordinates

$$\mathbf{8} \qquad r = (x_1^2 + x_2^2)^{1/2}, \ \phi = \text{Atan}(x_1, x_2).$$

Then the equations (5) are transformed to

$$\mathbf{9} \qquad \dot{r} = \alpha r(\beta - r^2), \ \dot{\phi} = -1.$$

It can be easily verified that the solution of (9) is

$$\mathbf{10} \qquad r(t) = \frac{\beta}{[1 + c_0 \exp(-2\beta^2 \alpha t)]^{1/2}}, \ \phi(t) = \phi(0) - t,$$

where

$$\mathbf{11} \qquad c_0 = \frac{\beta^2}{r^2(0)} - 1.$$

Thus the system (5) has only one nontrivial periodic solution, namely $r = \beta$, i.e., $x_{10}^2 + x_{20}^2 = \beta^2$. (Note that the equilibrium solution $x_1 = 0, x_2 = 0$ can also be considered a trivial periodic solution.) Furthermore, if $\alpha > 0$, any solution of (5) with $r(0) \neq 0$ approaches this periodic solution as $t \to \infty$. This example differs from the earlier example of a simple harmonic oscillator in that there is only one nontrivial periodic solution, and moreover, this periodic solution is **isolated**, i.e., there exists a neighborhood of it that does not contain any other periodic solution.

It is common to refer to a nontrivial periodic solution as a **limit cycle.** Note that some authors reserve this phrase only for an *isolated* periodic solution. By convention, an equilibrium is not regarded as a periodic solution.

In the remainder of this section, some results are presented pertaining to the existence or absence of periodic solutions in nonlinear systems.

3.3.2 Bendixson's Theorem

Bendixson's theorem presents a simple sufficient condition to guarantee that a given simply connected domain in the plane *does not* contain a periodic solution. Before stating the theorem, the terms "domain" and "simply connected" are defined. A **domain** in $\mathbf{R}^2$ is just an open set. A subset $S \subseteq \mathbf{R}^2$ is **simply connected** if it can be continuously shrunk to a single point in S, i.e., if there exists a point $\mathbf{x}_0 \in S$ and a continuous function $h : [0, 1] \times S \to S$ such that

12 $h(0, \mathbf{x}) = \mathbf{x}, h(1, \mathbf{x}) = \mathbf{x}_0, \forall \mathbf{x} \in S.$

For example, a closed disk is simply connected, whereas an annular region is not.

13 **Theorem** *Consider the second-order system*

14 $\dot{x}_1 = f_1(x_1, x_2), \dot{x}_2 = f_2(x_1, x_2).$

Suppose D is a simply connected domain in $\mathbf{R}^2$ such that the quantity $\nabla \mathbf{f}(\mathbf{x})$ defined by

15 $$\nabla \mathbf{f}(\mathbf{x}) = \frac{\partial f_1}{\partial x_1}(x_1, x_2) + \frac{\partial f_2}{\partial x_2}(x_1, x_2)$$

is not identically zero over any subdomain of D and does not change sign over D. Under these conditions, D does not contain any nontrivial periodic solutions of (14).

Proof Suppose J is a closed trajectory of (14). Then at each point $\mathbf{x} = (x_1, x_2) \in J$, the velocity vector field $\mathbf{f}(\mathbf{x})$ is tangent to J. Let $\mathbf{n}(\mathbf{x})$ denote the outward normal to J at $\mathbf{x}$. Then $\mathbf{f}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) = 0$ for all $\mathbf{x} \in J$. Therefore

16 $$\int_J \mathbf{f}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\, dl = 0.$$

But by the divergence theorem,

17 $$\int_J \mathbf{f}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\, dl = \iint_S \nabla \mathbf{f}(\mathbf{x})\, d\mathbf{x} = 0,$$

where S is the area enclosed by J. Therefore, in order for (17) to hold, either $\nabla \mathbf{f}(\mathbf{x})$ must be identically zero over S, or else $\nabla \mathbf{f}(\mathbf{x})$ must change sign over S. But if S is a subset of D, then the hypotheses of the theorem rule out both possibilities. Hence D contains no nontrivial periodic solutions of (14). ■

18 **Example** Consider the application of Theorem (13) to the *linear* system of equations

$$\dot{x}_1 = a_{11}x_1 + a_{12}x_2, \dot{x}_2 = a_{21}x_1 + a_{22}x_2.$$

From Section 3.2 we know that a necessary and sufficient condition for the system to have periodic solutions is that the matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

have two nonzero imaginary eigenvalues. Since the eigenvalues of $\mathbf{A}$ are the roots of the characteristic equation

$$\lambda^2 + (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0,$$

it is clear that the system has periodic solutions if and only if

$$a_{11} + a_{22} = 0,\ a_{11}a_{22} - a_{12}a_{21} > 0.$$

Equivalently, a necessary and sufficient condition for the absence of periodic solutions is that either of the above conditions be violated.

Applying Theorem (13) to the present case gives

$$\nabla \mathbf{f}(\mathbf{x}) = a_{11} + a_{22},\ \forall \mathbf{x} \in \mathbf{R}^2.$$

Hence Bendixson's theorem states that if $a_{11} + a_{22} \neq 0$, then the system has no periodic solutions, which is consistent with the previous discussion.

19 Example Consider the system of nonlinear equations

$$\dot{x}_1 = x_2 + x_1 x_2^2,\ \dot{x}_2 = -x_1 + x_1^2 x_2.$$

The linearization of this equation around the equilibrium at the origin is

$$\dot{x}_1 = x_2,\ \dot{x}_2 = -x_1,$$

which exhibits a continuum of periodic solutions. However, for the nonlinear system we have

$$\nabla \mathbf{f}(\mathbf{x}) = x_1^2 + x_2^2 > 0\ \forall \mathbf{x} \neq \mathbf{0}.$$

Thus $\nabla \mathbf{f}$ never changes sign, and is zero only at the origin (which is not a subdomain since it is not an open set). Hence Bendixson's theorem leads to the conclusion that the system under study has no nontrivial periodic solutions.

20 Example In applying Theorem (13), the assumption that D is a simply connected domain is crucial — it is not enough for D to be just connected. (A subset D of $\mathbf{R}^2$ is said to be **connected** if every two points in D are connected by a path lying entirely in D. Thus an annular region is connected but not simply connected.) To see this, consider the system (5), and let D be the annular region

$$D = \{(x_1, x_2): 2\beta^2/3 < x_1^2 + x_2^2 < 2\beta^2\}.$$

For this example, we have

$$\nabla \mathbf{f}(\mathbf{x}) = 2\alpha\beta^2 - 4\alpha(x_1^2 + x_2^2),$$

which is everywhere nonnegative on D. Yet D contains a periodic solution. Though the region D is connected, it is not *simply* connected. Hence Theorem (13) does not apply in the present situation.

3.3.3 Poincaré-Bendixson Theorem

The Poincaré-Bendixson theorem can be used to prove the existence of a periodic solution, provided a domain M satisfying certain conditions can be found. The strength of the theorem is its generality and simple geometric interpretation. The weakness of the theorem is the necessity of having to find the region M. A definition is introduced first.

21 Definition *Let* $\mathbf{x}(\cdot)$ *be a solution trajectory of (14). A point* $\mathbf{z} \in \mathbf{R}^2$ *is called a* **limit point** *of this trajectory if there exists a sequence* $\{t_i\}$ *in* $\mathbf{R}_+$ *such that* $t_i \to \infty$ *and* $\mathbf{x}(t_i) \to \mathbf{z}$. *The set of all limit points of a trajectory is called the* **limit set** *of the trajectory and is denoted by L.*

Remarks Basically, a limit point of a trajectory is a point to which the trajectory passes arbitrarily close infinitely many times as time progresses. We shall encounter limit points and limit sets again in Section 5.2.

22 Theorem (Poincaré-Bendixson) *Let*

23 $S = \{\mathbf{x}(t), t \geq 0\}$

denote a trajectory in $\mathbf{R}^2$ *of the system (14), and let L denote its limit set. If L is contained in a closed bounded region M in* $\mathbf{R}^2$ *and if M contains no equilibria of (14), then either*

(i) *S is a periodic solution of (14), or*

(ii) *L is a periodic solution of (14).*

The proof is omitted as it is beyond the scope of the book.

Remarks Roughly speaking, Theorem (22) states the following: Suppose we can find a closed bounded region M in $\mathbf{R}^2$ such that M does not contain any equilibria of (14) and such that all limit points of some trajectory S are contained in M. Then M contains at least one periodic solution of (14). In practice, an easy way to verify that M contains all the limit points of a trajectory S is to verify that S eventually lies entirely in M, i.e., to show that there exists a time T such that $\mathbf{x}(t) \in M \ \forall t \geq T$. Thus the theorem reduces to this: If we can find a closed bounded region M containing no equilibria such that some trajectory is eventually confined to M, then M contains at least one periodic solution. Now, a sufficient condition for a trajectory to be eventually confined to M is that, at every point along the boundary of M, the velocity vector field always points *into* M. If this is true, then any trajectory originating from within M must remain in M, and hence M contains at least one periodic solution of the system at hand. (This is depicted in Figure 3.16.)

24 Example Consider once again the system (5), and let M be the annular region defined by

$$M = \{(x_1, x_2): 0.9\beta^2 \leq x_1^2 + x_2^2 \leq 1.1\beta^2\}.$$

Then M contains no equilibria of the system. Moreover, a sketch of the velocity vector field reveals that, all along the boundary of M, the vector field always points *into* M, as depicted in

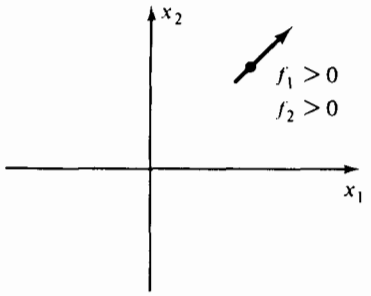


Fig. 3.16

Figure 3.16. Hence we can apply Theorem (22) and conclude that M contains a periodic solution.

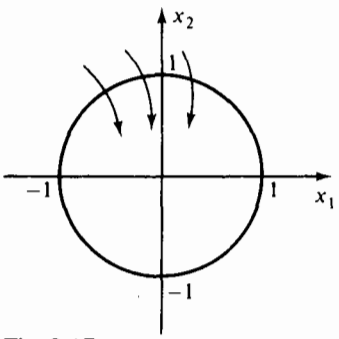


Fig. 3.17

25 Example In applying Theorem (22), the condition that M does not contain any equilibria is indispensable. To see this, consider the system

$$\dot{x}_1 = -x_1 + x_2, \dot{x}_2 = -x_1 - x_2.$$

The velocity vector field for this system is sketched in Figure 3.17. If M is chosen to be the closed unit disk centered at the origin, then all along the boundary of M the velocity vector field points into M. Hence all trajectories originating in M remain within M. The same conclusion can be reached by analytical reasoning because, in polar coordinates, the system equations become

$$\dot{r} = -r, \dot{\phi} = -1,$$

which has the solution

$$r(t) = r(0)\exp(-t),\ \phi(t) = \phi(0) - t.$$

However, even though all trajectories starting within M remain within M, M does not contain any nontrivial periodic solutions. Theorem (22) does not apply in this case because M contains the equilibrium $\mathbf{0}$.

3.3.4 Index Theorems

The concept of index is a very powerful one, and the results given below only scratch the surface of the many results that are available. Unfortunately, the arguments involved in index theory are well beyond the scope of this book. Hence almost all of the results presented in this subsection are stated without proof. For further discussion, see Nemytskii and Stepanov (1960).

The definition below introduces the concept of index. Recall that a point $\mathbf{x} \in \mathbf{R}^2$ is called an equilibrium of a vector field $\mathbf{f}$ if $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, and recall also Definition (3.1.8) of the direction of a vector field at a point (other than an equilibrium).

26 Definition *Suppose D is an open, simply connected subset of $\mathbf{R}^2$, and suppose $\mathbf{f}: \mathbf{R}^2 \to \mathbf{R}^2$ is a vector field on $\mathbf{R}^2$. Suppose D contains only isolated equilibria of the vector field $\mathbf{f}$. Let J be a simple, closed, positively oriented Jordan curve in D that does not pass through any equilibria of $\mathbf{f}$, and let $\theta_{\mathbf{f}}(\mathbf{x})$ denote the direction of the vector field at $\mathbf{x}$. Then the* **index of the curve J with respect to the vector field $\mathbf{f}$** *is denoted by $I_{\mathbf{f}}(J)$ and is defined as*

$$\mathbf{27} \quad I_{\mathbf{f}}(J) = \frac{1}{2\pi}\int_J d\theta_{\mathbf{f}}(\mathbf{x}).$$

Remarks A positively oriented curve is one which is traversed in the counterclockwise direction, i.e., a curve with the property that the area enclosed by it always lies to the left of it. Since it is assumed that J does not pass through any equilibria of $\mathbf{f}$, the direction vector $\theta_{\mathbf{f}}(\mathbf{x})$ is well-defined at all $\mathbf{x} \in J$. The index of J with respect to $\mathbf{f}$ is just the net change in the direction of $\mathbf{f}(\mathbf{x})$ as $\mathbf{x}$ traverses around J, divided by 2π. Clearly $I_{\mathbf{f}}(J)$ is always an integer.

28 Definition *Let $\mathbf{p}$ be an isolated equilibrium of the vector field $\mathbf{f}$. Then the* **index of $\mathbf{p}$** *is denoted by $I_{\mathbf{f}}(\mathbf{p})$ and is defined as $I_{\mathbf{f}}(J)$ where J is any suitable Jordan curve such that (i) $\mathbf{p}$ is enclosed by J, and (ii) J does not enclose any other equilibria of $\mathbf{f}$.*

Note that the same symbol $I_{\mathbf{f}}(\cdot)$ is used for both the index of a closed curve and of an equilibrium.

Now some facts are stated without proof.

29 Fact *Suppose J does not enclose any equilibria of $\mathbf{f}$. Then $I_{\mathbf{f}}(J) = 0$.*

30 Fact *The indices of a center, focus, and node are each equal to* 1, *while the index of a saddle is* -1.

This fact can be verified by sketching the vector field near each of the above types of equilibria.

31 Fact *Suppose J encloses a finite number of equilibria of* $\mathbf{f}$, *say* $\mathbf{p}_1, \cdots, \mathbf{p}_n$. *Then*

$$\textbf{32} \quad I_{\mathbf{f}}(J) = \sum_{i=1}^{n} I_{\mathbf{f}}(\mathbf{p}_i).$$

33 Fact *Let* $\mathbf{f}$ *and* $\mathbf{g}$ *be two vector fields on* $\mathbf{R}^2$. *Let J be a simple, closed, positively oriented Jordan curve, and suppose that* $\mathbf{f}$ *and* $\mathbf{g}$ *are never in opposition along the boundary of J; i.e., suppose that* $|\theta_{\mathbf{f}}(\mathbf{x}) - \theta_{\mathbf{g}}(\mathbf{x})| < \pi$ *at all* $\mathbf{x}$ *along the boundary of J. Suppose in addition that J does not pass through any equilibria of either* $\mathbf{f}$ *or* $\mathbf{g}$. *Under these conditions,*

$$\textbf{34} \quad I_{\mathbf{f}}(J) = I_{\mathbf{g}}(J).$$

This fact follows from Definition (28) and the fact that both $I_{\mathbf{f}}(J)$ and $I_{\mathbf{g}}(J)$ are integers.

35 Fact *Let J be a simple, closed, positively oriented trajectory of the system*

$$\textbf{36} \quad \dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t)].$$

Then

$$\textbf{37} \quad I_{\mathbf{f}}(J) = 1.$$

This can be seen from the fact that the vector field $\mathbf{f}$ is always tangent to J.

On the basis of these facts, we can state the following general theorem.

38 Theorem *Suppose the system (36) has only isolated equilibria. Then every closed trajectory of (36) (if any) encloses at least one equilibrium. Moreover, the sum of the indices of the equilibria enclosed by the closed trajectory is equal to* 1.

39 Example As an illustration of Theorem (38), consider the Volterra predator-prey equations (introduced earlier in Problem 3.8)

$$\textbf{40} \quad \dot{x}_1 = -x_1 + x_1 x_2, \dot{x}_2 = x_2 - x_1 x_2.$$

Let us digress briefly to discuss the rationale behind the above model. Let x_1 denote the number of predators (foxes, let us say), and let x_2 denote the number of prey (rabbits). If $x_2 = 0$, then the first equation in (40) reduces to $\dot{x}_1 = -x_1$, which states that in the absence of prey the number of predators will dwindle exponentially to zero. If $x_2 \neq 0$, then the same equation shows that $\dot{x}_1$ contains an exponential growth term proportional to x_2. The situation in the case of x_2 is just the opposite. If $x_1 = 0$, then x_2 will grow exponentially, while if

$x_1 \neq 0$, then $\dot{x}_2$ contains an exponential decay term proportional to x_1.

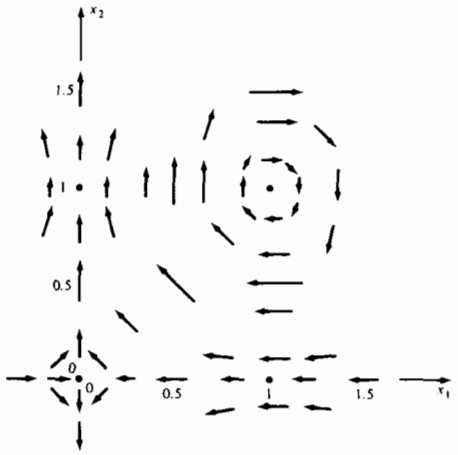


Fig. 3.18

The velocity vector field for the predator-prey system is shown in Figure 3.18. Clearly there are two equilibria, namely (0, 0) and (1, 1). By linearizing the system (40) around each of these equilibria one can readily determine that (0, 0) is a saddle while (1, 1) is a center. Hence the index of (0, 0) is -1 while the index of (1, 1) is 1. Now, by Theorem (38), any closed trajectory of the system (40) must enclose (1, 1), and it must *not* enclose (0, 0). Thus, by examining the index alone, one can derive a great deal of qualitative information about the possible closed trajectories of a system.

3.3.5 An Analytical Method

In this subsection, a technique is presented for obtaining analytical expressions for the closed trajectories of some nonlinear systems that exhibit a continuum of periodic trajectories. Rather than presenting a general theorem, which would have to be rather weak because of all the possible pathological cases, we illustrate the method by means of a few examples.

The basic idea of the method is as follows: Given the system (14) and a continuously differentiable function $V: \mathbf{R}^2 \to \mathbf{R}$, define the function $\dot{V}: \mathbf{R}^2 \to \mathbf{R}$ by

$$\mathbf{41} \quad \dot{V}(x_1, x_2) = \frac{\partial V}{\partial x_1} f_1(x_1, x_2) + \frac{\partial V}{\partial x_2} f_2(x_1, x_2).$$

The function $\dot{V}$ is known as the **derivative of** V along the trajectories of the system (14), because if $[x_1(.), x_2(.)]$ is a trajectory of the system (14), then the derivative with respect to t of the function $V[x_1(t), x_2(t)]$ is precisely $\dot{V}[x_1(t), x_2(t)]$. We shall encounter this concept again in Section 5.2. Now suppose we are able to find a domain D in $\mathbf{R}^2$ such that $\dot{V}(x_1, x_2) \equiv 0$ for all $\mathbf{x} \in D$. Let $(x_{10}, x_{20}) \in D$, and let $\mathbf{C}$ denote the solution trajectory of (14)

originating from (x_{10}, x_{20}). The hypothesis that $\dot{V}$ is identically zero implies that $V(x_1, x_2)$ is constant along **C**. In other words,

42 $V[x_1(t), x_2(t)] = V(x_{10}, x_{20}), \forall t \geq 0.$

Let us now consider the set

43 $S = \{(x_1, x_2): V(x_1, x_2) = V(x_{10}, x_{20})\}.$

Then **C** is a subset of S. In particular, if S is itself a closed curve, then we can conclude (under reasonably mild additional assumptions) that **C** is itself a closed trajectory and is equal to S.

Of course a great deal depends on the choice of the function V. If we choose $V(x_1, x_2) = 1$ for all (x_1, x_2), then naturally $\dot{V} \equiv 0$; but the set S in (43) is all of $\mathbf{R}^2$, and as a result no insight has been gained into the nature of the trajectories. However, in some cases, by properly choosing V, we can show that the family of sets

44 $\{(x_1, x_2): V(x_1, x_2) = c\}$

as c varies over an appropriate subset of real numbers, defines a continuum of closed trajectories of the system (14).

45 Example A very simple application which illustrates this procedure is the harmonic oscillator

$$\dot{x}_1 = x_2, \dot{x}_2 = -x_1.$$

Let us choose

$$V(x_1, x_2) = x_1^2 + x_2^2.$$

Then

$$\dot{V} = 2x_1(x_2) + 2x_2(-x_1) = 0.$$

Since the equation $V(x_1, x_2) = c$ defines a closed curve for each $c > 0$, we conclude that the system above exhibits a continuum of closed trajectories, described by

$$x_1^2 + x_2^2 = x_{10}^2 + x_{20}^2.$$

46 Example Consider again the predator-prey equations (40), and try a function V of the form

47 $V(x_1, x_2) = h_1(x_1) + h_2(x_2).$

where h_1 and h_2 are to be selected so that $\dot{V}$ is identically zero. Now

$$\dot{V}(x_1, x_2) = h_1'(x_1)(-x_1 + x_1x_2) + h_2'(x_2)(x_2 - x_1x_2).$$

In order that $\dot{V}$ be identically zero, it is necessary that

$$h_1'(x_1)x_1(x_2 - 1) = h_2'(x_2)x_2(1 - x_1) = 0,$$

which can be rearranged as

$$h_1'(x_1)\frac{x_1}{1-x_1} = h_2'(x_2)\frac{x_2}{1-x_2}.$$

The left side of this equation is independent of x_2, while the right side is independent of x_1. Hence, in fact both must equal a constant, say c. In other words,

$$\textbf{48} \quad h_1'(x_1)\frac{x_1}{1-x_1} = c,\ h_2'(x_2)\frac{x_2}{1-x_2} = c.$$

The solution of (48) is

$$h_1(x_1) = c(\ln x_1 - x_1),\ h_2(x_2) = c(\ln x_2 - x_2).$$

Hence an appropriate choice in this example is

$$V(x_1, x_2) = (\ln x_1 - x_1 + \ln x_2 - x_2),$$

where the arbitrary constant c has been dropped without loss of generality. For the above choice of V, any set of the form (43) is actually a closed curve. Hence the family of curves defined by

$$\ln x_1 - x_1 + \ln x_2 - x_2 = \text{const.}$$

constitutes a set of closed trajectories for the predator-prey system. Note that V is defined only in the first quadrant, i.e., if $x_1 > 0$, $x_2 > 0$.

49 Example Consider the pendulum equation

$$\ddot{\theta} + \frac{g}{l}\sin\theta = 0,$$

where θ denotes the angle of the pendulum from the vertical axis, g is the acceleration due to gravity, and l is the length of the pendulum. With the natural choice of state variables

$$x_1 = \theta,\ x_2 = \dot{\theta},$$

the pendulum equation can be rewritten as

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

Let us once again choose V to be of the form (47). Then, in order for $\dot{V}$ to be identically zero, we must have

$$h_1'(x_1)x_2 - \frac{g}{l}h_2'(x_2)\sin x_1 = 0,$$

which implies, as in Example (46), that

$$\frac{h_1'(x_1)}{\sin x_1} = \frac{g\,h_2'(x_2)}{l\,x_2} = \text{const.} = c.$$

The solution of these equations is

$$h_1(x_1) = -c\cos x_1, h_2(x_2) = \frac{cg}{l}\frac{x_2^2}{2}.$$

Hence the family of curves

$$\frac{x_2^2}{2} - \frac{g}{l}\cos x_1 = \text{const.}$$

constitute a set of closed trajectories of the pendulum equation.

Remarks Examples (46) and (49) illustrate how the method presented here can sometimes yield good results. However, it should be clear that (i) a function V of the form (47) does not always work, and (ii) even if it does, there is no guarantee that all closed trajectories are of the form (44). Despite these limitations, however, the method nevertheless has some value, as indicated by these two examples.

Problem 3.10 Consider a mechanical system consisting of a unit mass, a nonlinear spring, and a nonlinear damper. Such a system can be modelled by the set of equations

$$\dot{x}_1 = x_2, \dot{x}_2 = -g(x_1) - h(x_2),$$

where x_1 is the position of the mass, $g(\cdot)$ is the restoring force of the spring, and $h(\cdot)$ is the damping force exerted by the damper. Assume that both $g(\cdot)$ and $h(\cdot)$ are continuously differentiable. Using Bendixson's theorem, show that this system has no periodic solutions if $h'(\xi) \neq 0$ for all $\xi \neq 0$, i.e., there is always some amount of damping when the mass is in motion.

Problem 3.11 Sketch a vector field with exactly one node and one saddle. Show that it is not possible to deform this vector field continuously in such a way that there is a periodic solution enclosing both the node and the saddle.

Problem 3.12 Using the method of Section 3.3.5, show that the *undamped* unit mass-nonlinear spring system described by

$$\dot{x}_1 = x_2, \dot{x}_2 = -g(x_1)$$

always has a continuum of periodic solutions if

$$x_1 g(x_1) > 0 \; \forall x_1 \neq 0.$$

Derive an expression for these closed trajectories.

3.4 TWO ANALYTICAL APPROXIMATION METHODS

In this section, we describe two techniques for obtaining analytical expressions that approximate the periodic solution of second-order nonlinear differential equations. In contrast with the method presented in Section 3.3.5, which gives exact expressions if it works, the two methods presented here are only approximate. However, they have the advantage of having a wide range of applicability and of enabling one to study the so-called "slowly varying" oscillations. It should be emphasized that, depending on the particular problem to which they are applied, one technique might work better than the other. Moreover, the two methods presented here are only a small part of the numerous techniques that are available for analyzing slowly varying oscillations.

3.4.1 Krylov-Boguliubov Method

The Krylov-Boguliubov method is an example of a so-called "averaging" method. It is applicable to differential equations of the form

$$\ddot{y} + y = \mu f(y, \dot{y}), \quad (1)$$

where μ is a "small" parameter. The class of equations of the form (1) include several commonly encountered ones, such as the Van der Pol equation and the pendulum equation. Note that, in (1), the angular velocity of the oscillations corresponding to $\mu = 0$ has been normalized to 1. This presents no loss of generality, and can always be achieved by scaling the time variable t.

If $\mu = 0$, the solution of (1) is of the form

$$y(t) = a \sin(t + \phi), \quad (2)$$

where a and ϕ are constants determined by the initial conditions. With this in mind, let us assume that the solution of (1) when $\mu \neq 0$ is of the form

$$y(t) = a(t) \sin[t + \phi(t)], \quad (3)$$

4 $\dot{y}(t) = a(t)\cos[t + \phi(t)]$,

where $a(\cdot)$ and $\phi(\cdot)$ are "slowly varying," i.e., $\dot{a}(t)$ and $\dot{\phi}(t)$ are "small." Actually, if $y(\cdot)$ is given by (3), then

5 $\dot{y}(t) = \dot{a}(t)\sin[t + \phi(t)] + a(t)\cos[t + \phi(t)][1 + \dot{\phi}(t)]$.

Hence, in order for (4) to be valid, we must have

6 $\dot{a}\sin(t + \phi) + a\dot{\phi}\cos(t + \phi) = 0$,

where the dependence of a and ϕ on t has been suppressed in the interests of clarity. Substituting for y and $\dot{y}$ from (3) and (4) into (1) gives

7 $\dot{a}\cos(t + \phi) - a\dot{\phi}\sin(t + \phi) = \mu f[a\sin(t + \phi), a\cos(t + \phi)]$.

Equations (6) and (7) represent two linear equations in the two unknowns $\dot{a}$ and $\dot{\phi}$. Solving for these quantities gives

8 $\dot{a} = \mu\cos(t + \phi) f[a\sin(t + \phi), a\cos(t + \phi)]$,

9 $\dot{\phi} = -\dfrac{\mu}{a}\sin(t + \phi) f[a\sin(t + \phi), a\cos(t + \phi)]$.

To find solutions to (1) of the form (3) where $a(\cdot)$ is periodic, an extra condition is imposed, namely

10 $\dfrac{a(T) - a(0)}{T} = 0$,

or, equivalently,

11 $\dfrac{1}{T}\displaystyle\int_0^T \dot{a}(t)\,dt = 0$,

where T is the period of the function $a(\cdot)$. Unfortunately, (11) cannot be applied directly, since the period T is in general dependent on μ and hence unknown. To get around this difficulty, we observe that $a(\cdot)$ goes through one complete period as the phase $\theta = t + \phi(t)$ goes from 0 to 2π. Thus the variable of integration in (11) can be changed from t to θ. Then the limits of the integration become 0 and 2π, and the integrand $\dot{a}(t)$ becomes, in view of (8),

12 $\mu\cos\theta f(a\sin\theta, a\cos\theta)$.

Finally, we make the approximation

$$\textbf{13} \quad \frac{d\theta}{2\pi} = \frac{dt}{T}.$$

Equation (13) expresses the fact that as t varies over one period, θ varies over 2π. Thus (11) becomes

$$\textbf{14} \quad \frac{1}{2\pi}\int_0^{2\pi} \mu \cos\theta f(a\sin\theta, a\cos\theta)\, d\theta = 0.$$

Similarly, if ϕ is also required to be periodic with period T, this leads to the relationship

$$\textbf{15} \quad \frac{1}{2\pi}\int_0^{2\pi} \frac{\mu}{a} \sin\theta f(a\sin\theta, a\cos\theta)\, d\theta = 0.$$

Equations (14) and (15) can be used in the following way: Suppose we are interested in approximating the periodic solutions of (1) by functions of the form

$$\textbf{16} \quad y(t) = a\sin[(1+\delta)t],$$

where a and δ are now unknown *constants*. In this case, (14) and (15) simplify to

$$\textbf{17} \quad \int_0^{2\pi} \cos\theta f(a\sin\theta, a\cos\theta)\, d\theta = 0,$$

$$\textbf{18} \quad \int_0^{2\pi} \sin\theta f(a\sin\theta, a\cos\theta)\, d\theta = 0.$$

Since a is a constant, the function $f(a\sin\theta, a\cos\theta)$ is periodic in θ with period 2π and hence can be expanded in a Fourier series. Now (17) and (18) state that in order for (16) to approximate (to the first order in μ) a periodic solution of (1), it is necessary for the first harmonic of the periodic function $f(a\sin\theta, a\cos\theta)$ to be zero. This requirement is sometimes called the "principle of harmonic balance." We shall encounter the same reasoning in Chapter 4 in connection with the so-called describing function method.

Note that the parameter μ does not appear in (17) and (18), because when we study the periodic solutions of (1), we are in effect examining the *steady-state* oscillations of (1), and μ does not affect the steady-state solutions. However, μ is prominently present when the so-called "slowly varying" or transient solutions of (1) are studied. For this purpose, we make the approximations

$$\textbf{19} \quad \dot{a}(t) \approx \frac{a(T) - a(0)}{T},$$

20 $$\dot{\phi}(t) \approx \frac{\phi(T) - \phi(0)}{T},$$

where T is the period of the steady-state oscillations. However, as in studying the steady-state oscillations, we have

21 $$\frac{a(T) - a(0)}{T} = \frac{1}{2\pi} \int_0^{2\pi} \mu \cos\theta f(a \sin\theta, a \cos\theta)\, d\theta,$$

22 $$\frac{\phi(T) - \phi(0)}{T} = -\frac{1}{2\pi} \int_0^{2\pi} \frac{\mu}{a} \sin\theta f(a \sin\theta, a \cos\theta)\, d\theta.$$

Hence the approximate equations describing the slowly varying oscillations of (1) are

23 $$\dot{a} = \frac{1}{2\pi} \int_0^{2\pi} \mu \cos\theta f(a \sin\theta, a \cos\theta)\, d\theta,$$

24 $$\dot{\phi} = -\frac{1}{2\pi} \int_0^{2\pi} \frac{\mu}{a} \sin\theta f(a \sin\theta, a \cos\theta)\, d\theta.$$

25 Example Let us apply the Krylov-Boguliubov method to the Van der Pol equation, which can be rewritten in the form

$$\ddot{y} + y = \mu\dot{y}(1 - y^2).$$

This is of the form (1) with

$$f(y, \dot{y}) = \dot{y}(1 - y^2).$$

Hence

$$f(a \sin\theta, a \cos\theta) = a \cos\theta (1 - a^2 \sin^2\theta)$$

$$= \left(a - \frac{a^3}{4}\right) \cos\theta + \frac{a^3}{4} \cos 3\theta.$$

The integrals in (23) and (24) are just the Fourier coefficients of this function, multiplied by some constants. Thus the approximate equations (23) and (24) governing the slowly varying oscillations of (1) are given by

26 $$\dot{a} = \frac{\mu}{2}(a - a^3/4),$$

27 $\dot{\phi} = 0.$

To find a steady-state periodic solution of Van der Pol's equation, we set $\dot{a} = 0$ and $\dot{\phi} = 0$, which gives $a = 2$. Hence, to first order in μ, the limit cycle of the Van der Pol oscillator is described by

$$y(t) = 2 \sin(t + \phi_0).$$

To get the slowly varying solution, we solve (26) and (27) which results in

$$a(t) = 2\left[\frac{1}{1 + c \exp(-\mu t)}\right]^{1/2}, \phi(t) = \phi_0,$$

where c is a constant determined by the initial conditions. Hence the slowly varying solution of Van der Pol's equation is

$$y(t) = 2\left[\frac{1}{1 + c \exp(-\mu t)}\right]^{1/2} \sin(t + \phi_0).$$

Thus we see that, even though the parameter μ does not affect the steady-state solution, it does affect the rate at which the transient solution approaches the steady-state solution.

3.4.2 Power Series Method

The power series method is applicable to autonomous second-order differential equations containing a "small" parameter μ and consists of attempting to expand the solution of the given equation as a power series in μ. Mathematically the method is full of pitfalls, but it sometimes works reasonably well. The method is illustrated by an example.

Consider the differential equation

28 $\ddot{y} + y + \mu y^3 = 0,$

together with the special initial condition

29 $y(0) = a, \dot{y}(0) = 0.$

This equation can represent, for example, the motion of a unit mass constrained by a nonlinear spring. If $\mu > 0$, the spring is said to be "hard," whereas if $\mu < 0$, the spring is said to be "soft."

Clearly, if $\mu = 0$, the solution of (28) satisfying the initial condition (29) is

30 $y_0(t) = a \cos t.$

If $\mu \neq 0$ but is "small," then we can attempt to express the solution of (28) – (29) as a power series in μ, in the form

31 $y(t) = y_0(t) + \mu y_1(t) + \mu^2 y_2(t) + \cdots$

The idea is to substitute (31) into (28) and equate the coefficients of all powers of μ to zero. However, if this is done blindly, some of the $y_i(\cdot)$ may contain **secular** terms, i.e., functions which are unbounded. To see this phenomenon, let us substitute (31) into (28) and set the coefficients of all powers of μ equal to zero. This gives

32 $\ddot{y}_0 + y_0 = 0; y_0(0) = a, \dot{y}_0(0) = 0.$

33 $\ddot{y}_1 + y_1 + y_0^3 = 0; y_1(0) = 0, \dot{y}_1(0) = 0.$

Solving first for $y_0(\cdot)$ gives

34 $y_0(t) = a \cos t.$

This is as expected, since $y_0(\cdot)$ is the solution of (28) corresponding to $\mu = 0$. Now the equation for $y_1(\cdot)$ becomes

35 $$\ddot{y}_1 + y_1 = -y_0^3 = -\cos^3 t = -\frac{3a^3}{4}\cos t - \frac{a^3}{4}\cos 3t.$$

The solution of this equation is

36 $$y_1(t) = -\frac{3a^3}{8} t \sin t - \frac{a^3}{32}\cos t + \frac{a^3}{32}\cos 3t.$$

The $t \sin t$ term on the right side is the secular term, which arises because the forcing function of the nonhomogeneous equation (35) for y_1 contains a component of angular frequency 1, which is also the resonance frequency of the unforced system corresponding to (35). Combining the above expressions for y_0 and y_1 gives an approximate solution to (28) which is good to the first order in μ:

37 $$y(t) \approx y_0(t) + \mu y_1(t)$$

$$= (1 - \mu a^3/32)\cos t - \frac{3\mu a^3}{8} t \sin t + \frac{\mu a^3}{32}\cos 3t.$$

It is clear that the above approximation is unacceptable because it is an unbounded function of t.

The presence of the secular terms can be rationalized as follows: If $\mu = 0$, the solution of (28) is periodic with period 2π. However, if $\mu \neq 0$, the period of the resulting solution need not necessarily equal 2π, though it will be close. On the other hand, since $y_0(\cdot)$, the so-called "generating solution" of the sequence of functions $y_1(\cdot), y_2(\cdot), \cdots$, has period 2π, so will all functions $y_i(\cdot)$. This attempt to express a function whose period is not 2π as a series using functions whose period is 2π leads to secular terms. As an example, suppose δ is "small," and let us expand $\cos[(1+\delta)t]$ as a power series in δ. This leads to

38 $$\cos[(1+\delta)t] = \cos t - \delta t \sin t - \frac{\delta^2 t^2}{2}\cos t \cdots$$

This power series converges uniformly in t over any finite interval, and can therefore be considered as a valid expression. However, if the series is truncated after a finite number of terms, the resulting finite summation contains secular terms. Moreover, the periodicity and boundedness properties of the function $\cos[(1+\delta)t]$ are not at all apparent from the above power series expansion.

To alleviate this difficulty, suppose that the solution $y(\cdot)$ of (28) – (29) is periodic with angular frequency ω, which is itself expressed as a power series in μ. In other words, suppose

39 $$\omega^2 = 1 + \mu\xi_1(a) + \mu^2\xi_2(a) + \cdots$$

This can be rewritten as

40 $$1 = \omega^2 - \mu\xi_1(a) - \mu^2\xi_2(a) - \cdots$$

Note that in (39) and (40) the dependence of the frequency on the initial condition a is explicitly identified. This is a purely nonlinear phenomenon which has no analog in linear systems. Substituting (40) and (31) into (28) and displaying only the constant and the first order terms in μ yields

41 $$\ddot{y}_0 + \mu\ddot{y}_1 + \omega^2 y_0 - \mu\xi_1 y_0 + \mu\omega^2 y_1 + \mu y_0^3 + \cdots = 0.$$

Collecting terms gives

42 $$\ddot{y}_0 + \omega^2 y_0 = 0; y_0(0) = a,\ \dot{y}_0(0) = 0,$$

43 $$\ddot{y}_1 + \omega^2 y_1 = -y_0^3 + \xi_1 y_0,\ y_1(0) = 0,\ \dot{y}_1(0) = 0,$$

and so on. Solving these equations gives

44 $$y_0(t) = a\cos\omega t,$$

45 $$\ddot{y}_1(t) + \omega y_1(t) = -a^3\cos^3\omega t + \xi_1 a\cos\omega t$$

$$= -\frac{3}{4}a^3\cos\omega t - \frac{1}{4}a^3\cos 3\omega t + \xi_1 a\cos\omega t.$$

Now, in order that the solution for $y_1(\cdot)$ does not contain any secular term, it is necessary (and sufficient) that the coefficient of $\cos\omega t$ on the right side of (45) be equal to zero. Thus

46 $\xi_1 = \frac{3}{4}a^3.$

With this condition, the solution for $y_1(\cdot)$ is obtained as

47 $y_1(t) = -\frac{a^3}{32\omega^2}\cos\omega t + \frac{a^3}{32\omega^2}\cos 3\omega t,$

where

48 $\omega^2 = 1 + \frac{3}{4}\mu a^3.$

Hence the overall solution of (28)-(29), accurate to first order in μ, is given by

49 $y(t) = a\cos\omega t - \frac{a^3}{32\omega^2}\cos\omega t + \frac{a^3}{32\omega^2}\cos 3\omega t.$

50 **Example** Consider the simple pendulum equation

51 $\ddot{y} + \sin y = 0.$

Equation (51) can be approximated by

$$\ddot{y} + y - \frac{y^3}{6} = 0.$$

This equation is of the form (28) within $\mu = -1/6$. Using the foregoing analysis, we conclude that the frequency of oscillation of the simple pendulum is related to the initial amplitude by

52 $\omega^2 = 1 - \frac{a^3}{8}.$

This is a refinement of the analysis based on the linearization of (51), which states that the frequency of oscillation is independent of the initial amplitude. That conclusion is indeed valid to first order in a, as can be seen from (52).

Problem 3.13 Apply the Krylov-Boguliubov method to Rayleigh's equation

$$\ddot{y} + y = \mu\left[\dot{y} - \frac{\dot{y}^3}{3}\right].$$

Solve the same equation using the perturbation method, and show that both methods give the same solution to the first order in μ.

Problem 3.14 Apply the perturbation method to the Van der Pol equation

$$\ddot{y} + y = \mu\dot{y}(1 - y^2).$$

Problem 3.15 Apply the Krylov-Boguliubov method to the pendulum equation

$$\ddot{y} + y - \frac{y^3}{6} = 0.$$

Show that the expression derived for the frequency of oscillation is the same as (52).

Problem 3.16 Consider the second-order equation

$$\ddot{y} + y = \mu f(y, \dot{y}),\ y(0) = 0,\ \dot{y}(0) = b.$$

Assuming that the function f is continuously differentiable with respect to both of its arguments, show that both the Krylov-Boguliubov method and the perturbation method give the same results.

Notes and References

Most of the material in this section is historic, and much of it can be generalized to higher-order systems. Discussions of nonlinear oscillations can be found in many classical texts, including Nemytskii and Stepanov (1960). The method of averaging, briefly introduced in Section 3.5, can be made rigorous; see Boguliuboff and Mitropolsky (1961) or Sanders and Verhulst (1985).

4. APPROXIMATE ANALYSIS METHODS

In this chapter, we present two methods for *approximately* analyzing a given nonlinear system. Since a closed-form analytical solution of a nonlinear differential equation is usually impossible to obtain (except in some special examples, which are often contrived), it is useful in practice to have some methods for carrying out an approximate analysis. Two methods are presented here. The **Describing Function Method** consists of replacing a nonlinear element within a system by an "equivalent" linear time-invariant system which is in some sense the best possible linear approximation of the given nonlinear system. This method is often used to predict the existence of periodic solutions in feedback systems. **Singular Perturbation Methods**, just touched upon here, are well-suited for the analysis of systems where the inclusion or exclusion of a particular component changes the *order* of the differential equation describing the system. It should be emphasized that these are just two of the many methods that are available for approximate analysis. Moreover, even with regard to these methods, the presentation here is merely an introduction, especially in the case of singular perturbations.

4.1 DESCRIBING FUNCTIONS

In this section, the concept of describing functions is introduced, and it is demonstrated that they can be used to predict the existence of periodic solutions in feedback systems.

4.1.1 Optimal Quasi-Linearization

The problem studied in this subsection is that of approximating a given nonlinear system by a linear time-invariant system. Let $C[0, \infty)$ denote the set of continuous real-valued functions over $[0, \infty)$, and suppose N is a given operator mapping $C[0, \infty)$ into itself. In other words, given any continuous function $x \in C[0, \infty)$, the operator N associates with it another function $Nx \in C[0, \infty)$. One can think of Nx as the output of a nonlinear system in response to the input x. By a slight abuse of notation, the nonlinear system is also denoted by the symbol N.

The problem at hand is to approximate the given nonlinear system N by a linear time-invariant system H in an optimal fashion. More precisely, suppose a function $r \in C[0, \infty)$, called the **reference input**, is specified. If H is a linear time-invariant system with the impulse response $h(\cdot)$,[1] then the output of H in response to the input r is given by

[1] The concept of the impulse response is formalized in Section 6.4.

$$\mathbf{1} \qquad (Hr)(t) = \int_0^t h(t-\tau)\, r(\tau)\, d\tau.$$

A measure of how well the linear system H approximates the nonlinear system N is provided by the error criterion

$$\mathbf{2} \qquad E(H) = \lim_{T\to\infty} \frac{1}{T} \int_0^T [(Nr)(t) - (Hr)(t)]^2\, dt,$$

assuming of course that the indicated limit exists.[2] Thus the objective is to choose the linear system H in such a way that the error criterion $E(H)$ is minimized. A linear time-invariant system H that minimizes the criterion $E(H)$ is called an **optimal quasi-linearization** of the nonlinear system N, and the problem of finding such an H is called the optimal quasi-linearization problem.

The solution to this problem is provided in Theorem (12) below. But first a couple of technical questions are laid to rest.

A function $x \in C[0, \infty)$ is said to **have finite average power** if

$$\mathbf{3} \qquad \lim_{T\to\infty} \frac{1}{T} \int_0^T x^2(t)\, dt =: P(x) < \infty.$$

Note that "$< \infty$" means that the indicated quantity exists and is finite. In such a case the quantity $P(x)$ is called the **average power** of the function x. Thus the quantity $E(H)$ in (2) is just the average power of the *error function* $e = Nr - Hr$, i.e., the difference between the true system output Nr and the output of the approximate system Hr. It is natural to ask when $E(H)$ is well-defined.

4 Lemma *Suppose $x, y \in C[0, \infty)$ have finite power. Then so does $x + y$.*

Proof The proof is analogous to that of Theorem (2.1.33). Let $\mathbf{F}$ denote the subset of $C[0, \infty)$ consisting of all functions with finite average power, that is,

$$\mathbf{5} \qquad \mathbf{F} = \{x \in C[0, \infty) : P(x) < \infty\}.$$

The claim is that $\mathbf{F}$ is a linear vector space. To show this, suppose T is any *finite* number, that $f, g \in C[0, T]$, and define

$$\mathbf{6} \qquad \langle f, g \rangle_{TP} = \frac{1}{T} \int_0^T f(t) g(t)\, dt.$$

It is straight-forward to verify that $\langle \cdot, \cdot \rangle_{TP}$ satisfies all the axioms of an inner product space

[2] This issue is cleared up later; see Lemma (4).

(see Section 2.1.3). Thus, if we define

$$\mathbf{7} \qquad \|f\|_{TP} = \left[\frac{1}{T}\int_0^T f^2(t)\,dt\right]^{1/2} = \langle f, f\rangle^{1/2},$$

then it follows from Schwarz' inequality [Lemma (2.1.38)] that

$$\mathbf{8} \qquad |\langle f, g\rangle_{TP}| \leq \|f\|_{TP} \cdot \|g\|_{TP}.$$

Next, observe that a function $f \in C[0, \infty)$ also belongs to $\mathbf{F}$ if and only if

$$\mathbf{9} \qquad \lim_{T\to\infty} \|f\|_{TP} < \infty.$$

Now suppose $x, y \in \mathbf{F}$. Then

$$\mathbf{10} \qquad \|x+y\|_{TP}^2 = \|x\|_{TP}^2 + \|y\|_{TP}^2 + 2\langle x, y\rangle_{TP}, \text{ from (4) and (7)}$$

$$\leq [\|x\|_{TP} + \|y\|_{TP}]^2, \text{ from (8).}$$

Letting $T \to \infty$ shows that $x + y \in \mathbf{F}$. ■

Theorem (12) below characterizes solutions to the problem of optimal quasi-linearization in terms of the so-called cross-correlation function. Suppose $x, y \in \mathbf{F}$. Then their **cross-correlation function** $\phi_{x,y}(\cdot)$ is defined by

$$\mathbf{11} \qquad \phi_{x,y}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T x(t)\,y(t+\tau)\,dt.$$

Note that $\phi_{x,y}$ is well-defined since x and y both have finite average power.

12 Theorem *Suppose the reference input r and the corresponding output Nr of the nonlinear system both have finite average power. Suppose H is a linear time-invariant operator of the form (1) such that Hr* $\in \mathbf{F}$. *Then H minimizes the error criterion E of (2) if and only if*

$$\mathbf{13} \qquad \phi_{r,Hr}(\tau) = \phi_{r,Nr}(\tau), \ \forall \tau \geq 0.$$

Proof First, since both Nr and Hr belong to $\mathbf{F}$ by assumption, it follows from Lemma (4) that the quantity $E(H)$ is well-defined and finite. Now suppose G is another linear time-invariant system of the form

$$\mathbf{14} \qquad (Gx)(t) = \int_0^t g(t-\tau)\,x(\tau)\,d\tau$$

such that $Gr \in \mathbf{F}$, and define

15 $d(t) = g(t) - h(t)$,

16 $(Dx)(t) = (Gx)(t) - (Hx)(t) = \int_0^t d(t-\tau)\, x(\tau)\, d\tau = \int_0^t d(\tau)\, x(t-\tau)\, d\tau.$

Since $Gr \in \mathbf{F}$, the quantity $E(G)$ is also well-defined. Now from (2) we get

17 $$E(G) - E(H) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \{[(Nr)(t) - (Gr)(t)]^2 - [(Nr)(t) - (Hr)(t)]^2\}\, dt$$

$$= \lim_{T\to\infty} \int_0^T \frac{1}{T} \{[(Dr)(t)]^2 + 2(Dr)(t)(Hr - Nr)(t)\}\, dt.$$

Clearly, H minimizes the error criterion E if and only if $E(G) \geq E(H)$ for all suitable G. Now the right side of (17) is nonnegative for all suitable operators D if and only if the linear term in Dr is identically zero, i.e.,

18 $\lim_{T\to\infty} \frac{1}{T} \int_0^T (Dr)(t)\, (Hr - Nr)(t)\, dt = 0, \ \forall D.$

For brevity let e denote the function $Hr - Nr$. Substituting for Dr from (16) and interchanging the order of integration gives

19 $$0 = \lim_{T\to\infty} \frac{1}{T} \int_0^T \int_0^t d(\tau)\, r(t-\tau)\, e(t)\, d\tau\, dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T \left[\int_\tau^T r(t-\tau)\, e(t)\, dt \right] d(\tau)\, d\tau.$$

However, since $d(\cdot)$ is an *arbitrary* impulse response, subject only to the condition that $Gr \in \mathbf{F}$, the coefficient of $d(\cdot)$ must be identically zero. Thus, after interchanging the order of integration with respect to τ and taking the limit with respect to T, the optimality condition becomes

20 $$0 = \lim_{T\to\infty} \frac{1}{T} \int_\tau^T r(t-\tau)\, e(t)\, dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T r(t)\, e(t+\tau)\, dt = \phi_{r,e}(\tau), \ \forall \tau \geq 0.$$

Finally, note that the cross-correlation function is *bilinear*; hence (20) becomes

21 $0 = \phi_{r,Hr-Nr}(\tau) = \phi_{r,Hr}(\tau) - \phi_{r,Nr}(\tau), \ \forall \tau \geq 0.$

Equation (21) is the same as (13). ■

Theorem (12) represents an important principle, namely: H is a best possible linear approximation to the nonlinear system N if and only if the linear system faithfully reproduces the input-output cross-correlation function of the nonlinear system. But it must be emphasized that the approximation is best *for the given reference input.* If the reference

input is altered, the corresponding "best" approximation is also altered in general. Moreover, even for a given reference input, there is in general more than one optimal quasi-linearization.

22 Example Suppose $N: C[0, \infty) \to C[0, \infty)$ is a memoryless time-invariant nonlinearity of the form

$$(Nx)(t) = n[x(t)], \ \forall t \geq 0,$$

where $n: \mathbf{R} \to \mathbf{R}$ is continuous, and suppose the reference input r is a nonzero constant, i.e.,

$$r(t) = k, \ \forall t \geq 0.$$

Thus $r(\cdot)$ is a d.c. signal. It is easy to see that both r and Nr have finite average power. Now an easy calculation shows that

$$\phi_{r,Nr}(\tau) = kn(k), \ \forall \tau \geq 0.$$

Hence the optimality condition (13) is satisfied if

$$h(t) = \frac{n(k)}{k}\,\delta(t), \ (Hx)(t) = \frac{n(k)}{k}\,x(t), \ \forall t \geq 0,$$

where $\delta(\cdot)$ denotes the unit impulse distribution. Thus an optimal quasi-linearization of N with respect to the chosen reference input is a constant gain of $n(k)/k$, sometimes called the "d.c. gain" of N.

The above quasi-linearization is not unique. In fact one can show that if H is a stable linear time-invariant system with the transfer function $\hat{h}(s)$, then H is an optimal quasi-linearization of N if and only if $\hat{h}(0) = n(k)/k$. (See Problem 4.1.) ■

In the preceding discussion it is assumed that the reference input r is deterministic. However, it is possible to define the notion of an optimal quasi-linearization of a nonlinear operator with respect to a *random* input. Also, the development can be extended with no essential changes to *multi-input, multi-output* systems. For a more detailed discussion, see Gelb and Vander Velde (1968).

4.1.2 Describing Functions

By far the most commonly used reference input in optimal quasi-linearization is the sinusoidal function

23 $r(t) = a \sin \omega t.$

It is easily verified that $r \in \mathbf{F}$, i.e., r has finite average power. If the operator N is bounded-input, bounded-output (BIBO) stable in the sense defined in Chapter 6, the output Nr is the sum of two functions: (i) the **steady-state response** z_{ss} which is periodic, and the (ii) the **transient response** z_{tr} which decays to zero after some time. The same situation prevails if

r is applied to a BIBO stable linear time-invariant system of the form (1). The determination of an optimal linear approximation to N is greatly facilitated by the next two results.

24 Lemma *Suppose* $r \in \mathbf{F}$, *and* $f \in C[0, \infty)$ *satisfies*

$$\int_0^\infty f^2(t)\,dt < \infty. \tag{25}$$

Then

$$\phi_{r,f} \equiv 0. \tag{26}$$

Proof Note that (25) implies that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f^2(t)\,dt = 0. \tag{27}$$

By definition [cf. (11)],

$$\phi_{r,f}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T r(t) f(t+\tau)\,dt. \tag{28}$$

By analogy with the inequality (8), we have

$$|\phi_{r,f}(\tau)|^2 \le \left[\frac{1}{T}\int_0^T r^2(t)\,dt\right] \cdot \left[\frac{1}{T}\int_0^T f^2(t+\tau)\,dt\right]. \tag{29}$$

For each fixed τ, as $T \to \infty$ the first term on the right side remains bounded while the second term approaches zero. ■

Lemma (24) does not depend on r being a pure sinusoid of the form (23). The next result is more specialized.

30 Lemma *Let r be as in (23). Suppose* $f \in C[0, \infty)$ *is periodic with period* $2\pi/\omega$, *and expand f in a Fourier series of the form*

$$f(t) = f_0 + \sum_{k=1}^{\infty} [f_{rk} \sin k\omega t + f_{ik} \cos k\omega t]. \tag{31}$$

Define

$$f_1(t) = f_{r1} \sin \omega t + f_{i1} \cos \omega t. \tag{32}$$

Then

33 $\phi_{r,f} = \phi_{r,f_1}$.

Remarks The lemma asserts that the cross-correlation between a periodic function f and the pure sinusoid r depends only on the *first harmonic* of the function f.

Proof Let $T = 2\pi l/\omega$ where l is an integer. In view of the well-known orthogonality property of trigonometric functions, it follows that

$$\textbf{34} \quad \int_0^{2\pi l/\omega} r(t) \sin k\omega(t+\tau)\, dt = 0, \quad \int_0^{2\pi l/\omega} r(t) \cos k\omega(t+\tau)\, dt = 0, \quad \forall \tau \geq 0, \forall l \geq 1, \forall k \neq 1.$$

Hence

$$\textbf{35} \quad \int_0^{2\pi l/\omega} r(t) f(t+\tau)\, dt = \int_0^{2\pi l/\omega} r(t) f_1(t+\tau)\, dt, \quad \forall l \geq 1.$$

The desired conclusion follows upon dividing both sides of (35) by $T = 2\pi l/\omega$ and letting $l \to \infty$. ■

Now we come to the main result.

36 Theorem *Let r be as in (23). Suppose $Nr =: z$ is of the form $z = z_{ss} + z_{tr}$, where z_{ss} is continuous and periodic with period $2\pi/\omega$, and*

$$\textbf{37} \quad \int_0^{\infty} z_{tr}^2(t)\, dt < \infty.$$

Finally, suppose H is an operator of the form (1), and suppose $\hat{h}(s)$, the Laplace transform of $h(\cdot)$, is a proper rational function whose poles all have negative real parts. Under these conditions, H is an optimal quasi-linearization of N with respect to the input r if and only if

$$\textbf{38} \quad \hat{h}(j\omega) = \frac{g_{re} + j g_{im}}{a},$$

where

$$\textbf{39} \quad z_1(t) = g_{re} \sin \omega t + g_{im} \cos \omega t$$

is the first harmonic of z_{ss}.

Remarks Theorem (36) presents a condition sometimes called the **principle of harmonic balance.** Suppose H is a linear time-invariant operator of the form (1). Then, as is the case with the nonlinear output Nr, the function $Hr =: y$ is a sum of the steady-state response y_{ss} which is periodic with the same period as r, and the transient response y_{tr} which eventually decays to zero in view of the assumption about the pole locations of the transfer function $\hat{h}(s)$. Moreover, y_{ss} is a pure sinusoid unlike z_{ss}. If $\hat{h}(j\omega) = h_{re} + j h_{im}$, then

40 $y_{ss}(t) = h_{re}\, a \sin \omega t + h_{im}\, a \cos \omega t.$

Thus Theorem (36) states that the optimal quasi-linearizations of N with respect to the sinusoidal reference input r are precisely those whose steady-state outputs (in response to r) precisely match the first harmonic of the steady-state part of Nr. Since the condition (38) specifies the value of $\hat{h}$ at only one frequency, it is clear that there are infinitely many optimal quasi-linearizations of N.

Proof Condition (37) and Lemmas (24) and (30) together imply that

41 $\phi_{r,Nr} = \phi_{r,z_1}.$

Theorem (12) states that H is an optimal quasi-linearization of N if and only if $\phi_{r,Nr} = \phi_{r,Hr}$. But, as discussed in the remarks above, $y := Hr$ is the sum of the steady-state response y_{ss} and the transient response y_{tr}. The assumptions on $\hat{h}$ ensure that the transient response y_{tr} is a finite sum of decaying (and possibly oscillating) exponentials; hence y_{tr} satisfies a condition analogous to (37). Thus

42 $\phi_{r,Hr} = \phi_{r,y_{ss}},$

and H is an optimal quasi-linearization if and only if

43 $\phi_{r,z_1} = \phi_{r,y_{ss}}.$

It is left as an exercise to show that (43) holds if and only if $z_1 = y_{ss}$. From (40), it follows that $z_1 = y_{ss}$ if and only if (38) holds. ■

44 Definition *Let r be as in (23), and suppose $z = Nr$ satisfies (37). Define z_1 as in (39). Then the* **describing function** $\eta(\cdot,\cdot)$ *of the operator N is the complex-valued function defined by*

45 $$\eta(a, \omega) = \frac{g_{re} + j g_{im}}{a}.$$

As has been observed above, for a given reference input r of the form (23) there can be infinitely many optimal quasi-linearizations of N. However, once a and ω are fixed, the describing function defined in (45) is unique.

46 Lemma *Suppose N is a memoryless time-invariant nonlinear operator of the form*

47 $(Nx)(t) = n[x(t)], \forall t \geq 0,$

where $n : \mathbf{R} \to \mathbf{R}$ is continuous. Then $\eta(a, \omega)$ is independent of ω.

Proof Since r is given by (23), it follows that

48 $z(t) = (Nr)(t) = n(a \sin \omega t)$

is also periodic with period $2\pi/\omega$, i.e., $z_{tr} \equiv 0$. Let z_1 be the first harmonic of z. Then $\eta(a, \omega)$ is given by (45). Now let the reference input be

49 $x(t) = a \sin \bar{\omega} t.$

Thus r and x have the same amplitude, but different frequencies. Then

50 $x(t) = r(\bar{\omega} t/\omega),$

i.e., x can be obtained from r by *time-scaling*. Now, *since N is memoryless*, it follows from (47) that

51 $(Nx)(t) = (Nr)(\bar{\omega} t/\omega).$

Hence the first harmonics of Nx and Nr are the same, allowing for the time scaling. Therefore $\eta(a, \omega) = \eta(a, \bar{\omega})$. ■

52 Lemma *Suppose N is of the form (47), and in addition, $n(\cdot)$ is an odd function. Then $\eta(a)$ is real for all a.*

Proof Observe first that one can write $\eta(a)$ instead of $\eta(a, \omega)$ since η is independent of ω by Lemma (46). If $n(\cdot)$ is odd, then $(Nr)(t)$ is an odd function of t, and there are no cosine terms in the Fourier series expansion of nr. Hence g_{im} in (45) is zero and $\eta(a)$ is real. ■

53 Lemma *Suppose N is a memoryless time-invariant operator of the form (47), and suppose in addition that $n(\cdot)$ is odd. Finally, suppose there exist constants k_1 and k_2 such that*

54 $k_1\sigma^2 \leq \sigma n(\sigma) \leq k_2\sigma^2, \forall \sigma \in \mathbf{R}.$

Then the describing function η of N satisfies

55 $k_1 \leq \eta(a) \leq k_2, \forall a.$

Remarks A nonlinearity $n(\cdot)$ satisfying the bounds (59) is said to **lie in the sector** $[k_1, k_2]$, since the graph of the function $n(\cdot)$ lies between two straight lines of slope k_1 and k_2 passing through the origin (see Figure 4.1). Note that $k_1 \leq k_2$, but one or both of the constants could be negative.

Proof By Lemma (52), $\eta(a)$ is real for all a. Moreover,

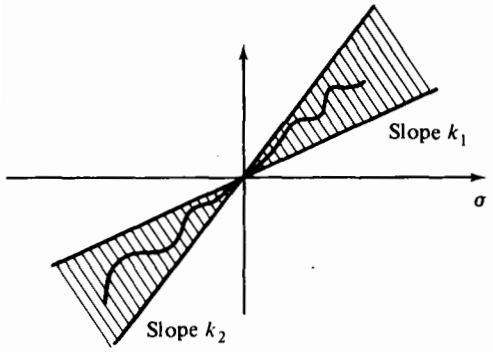


Fig. 4.1

56 $$\eta(a) = \frac{\omega}{\pi a} \int_0^{2\pi/\omega} n(a \sin \omega t) \sin \omega t \, dt$$

$$= \frac{1}{\pi a} \int_0^{2\pi} n(a \sin \theta) \sin \theta \, d\theta, \text{ letting } \theta = \omega t$$

$$\geq \frac{1}{\pi a^2} \int_0^{2\pi} k_1 (a \sin \theta)^2 \, d\theta, \text{ by (54)}$$

$$= k_1.$$

The proof that $\eta(a) \leq k_2$ is similar. ■

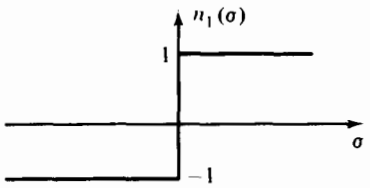


Fig. 4.2

57 Example Consider the nonlinearity shown in Figure 4.2, usually referred to as the "sign" nonlinearity. If an input r of the form (23) is applied to this system, the resulting output is a square wave of amplitude 1, irrespective of what a is (so long as $a \neq 0$). The first harmonic of a square wave has amplitude $4/\pi$, so the describing function of this nonlinearity is

$$\eta(a) = \frac{4}{\pi a}.$$

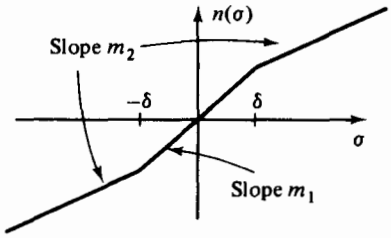


Fig. 4.3

58 Example Consider an element $n(\cdot)$ which is piecewise linear, as shown in Figure 4.3. For $|\sigma| \leq \delta$, the nonlinear element acts like a gain of value m_1, but the gain reduces to m_2 if the input value exceeds δ in magnitude. If a sinusoidal input r of the form (23) is applied to this element and if $|a| \leq \delta$, the output is another sinusoid of amplitude $a\delta$ and is in phase with the input. Therefore

$$\eta(a) = \delta \text{ if } 0 < a \leq \delta.$$

However, if $|a| > \delta$, the output of the nonlinearity is a "clipped" sine wave as shown in Figure 4.4. In this case, it can be verified through laborious but routine calculations that

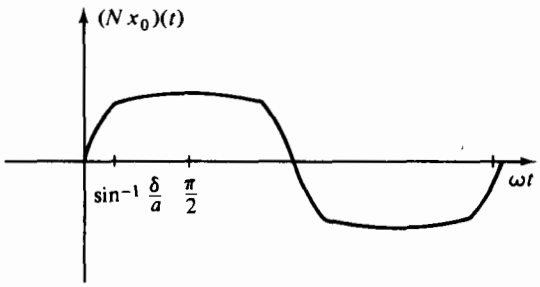


Fig. 4.4

$$\eta(a) = \frac{2(m_1 - m_2)}{\pi}\left[\sin^{-1}\frac{\delta}{a} + \frac{\delta}{a}\left[1 - \frac{\delta^2}{a^2}\right]^{1/2}\right] + m_2, \text{ if } |a| > \delta.$$

This can be expressed more compactly. Define the function

$$\textbf{59} \quad f(x) := \begin{cases} 1, & \text{for } x \geq 1 \\ \dfrac{2}{\pi}[\sin^{-1}x + x(1-x^2)^{1/2}], & \text{for } 0 \leq x \leq 1 \end{cases}$$

A sketch of the graph of the function f is shown in Figure 4.5. It is easy to verify that

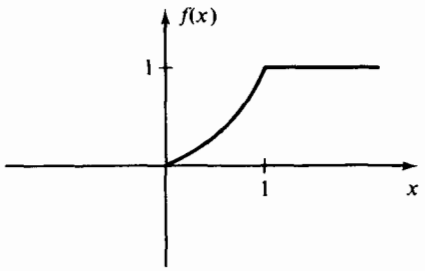


Fig. 4.5

$$\eta(a) = (m_1 - m_2) f(\delta/a) + m_2.$$

Though Figure 4.3 depicts the case where $m_1 > m_2$, the above expression is valid for any choice of m_1 and m_2. By choosing various values for these two constants, one can obtain the describing functions of several common nonlinear characteristics. For example, if $m_1 = 0$, then the nonlinearity becomes the "dead zone" characteristic shown in Figure 4.6. If $m_1 > 0$ but $m_2 = 0$, then we get the "limiter" characteristic of Figure 4.7.

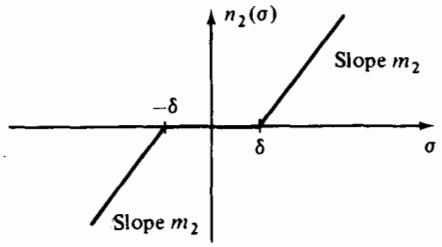


Fig. 4.6

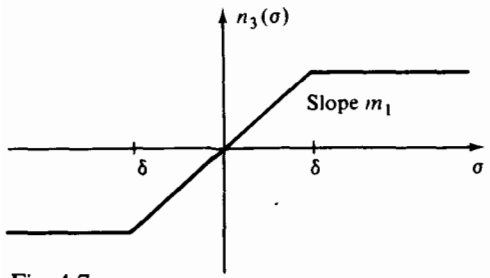


Fig. 4.7

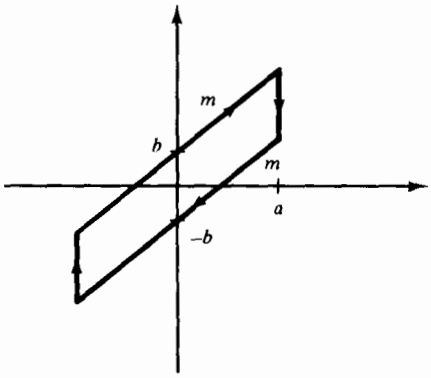


Fig. 4.8

Thus far all the examples have involved only memoryless nonlinearities. The next two examples discuss nonlinearities with memory.

60 Example Consider the hysteresis nonlinear operator N shown in Figure 4.8. In the steady-state, the output $(Nx)(t)$ follows the upper straight line when the input is increasing [i.e., $\dot{x}(t) > 0$] and the lower straight line when the input is decreasing. The number a in Figure 4.8 depends on the amplitude of the input and is not a characteristic of N itself. Thus

$$(Nx)(t) = \begin{cases} mx(t) + b \text{ if } \dot{x}(t) > 0 \\ mx(t) - b \text{ if } \dot{x}(t) < 0 \end{cases}$$

and "jumps" when $\dot{x}(t)$ goes through zero.

Suppose a sinusoidal input r of the form (23) is applied to N. The resulting *steady-state* output is shown in Figure 4.9. One can express $Nr(\cdot)$ as the sum of two signals as shown in Figure 4.10. From this it is clear that the first harmonic of the steady-state part of Nr is

$$z_1(t) = ma \sin \omega t + \frac{4b}{\pi} \cos \omega t.$$

Hence

$$\eta(a, \omega) = m + j\frac{4b}{\pi a}.$$

Note that η is once again independent of ω, because time scaling does not affect the output of N. Hence, even though N is not memoryless, the arguments in the proof of Lemma (46) apply.

61 Example Consider the hysteresis nonlinear operator N shown in Figure 4.11. In this case, if the input amplitude is less than δ, the output is simply equal to zero.

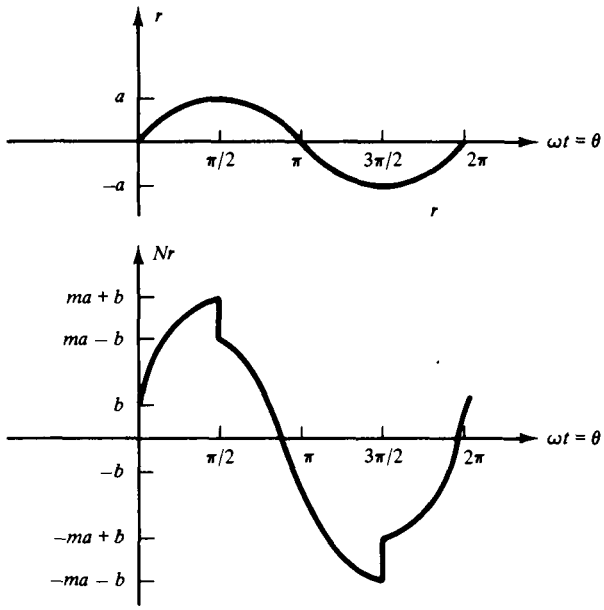


Fig. 4.9

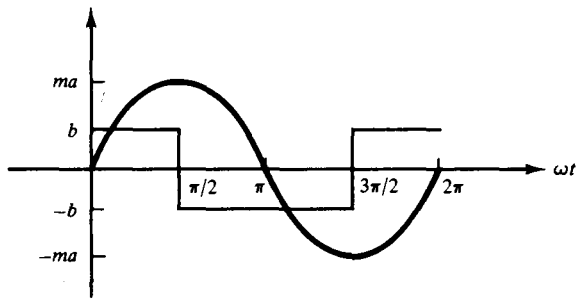


Fig. 4.10

$$\eta(a) = 0 \text{ if } a \leq \delta.$$

If the input amplitude a exceeds δ, then, *in the steady-state*, the output signal z_{ss} is as shown in Figure 4.12. Let $\beta \in (0, \pi/2)$ be the unique number such that

$$\sin\beta = 1 - \frac{2\delta}{a}.$$

For convenience let θ denote ωt. Then the steady-state output is described by

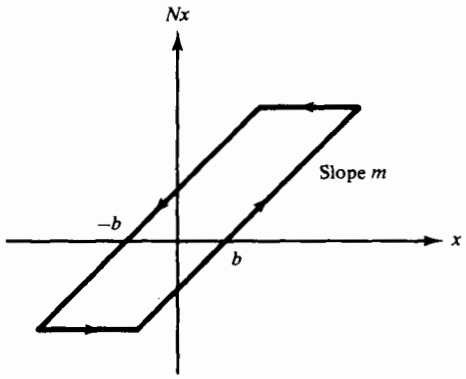


Fig. 4.11

$$z_{ss}(t)=\begin{cases} ma\sin\theta - m\delta, & \text{for } 0\le\theta\le\pi/2 \\ m(a-\delta), & \text{for } \pi/2\le\theta\le\pi-\beta \\ ma\sin\theta + m\delta, & \text{for } \pi-\beta\le\theta\le 3\pi/2 \\ -m(a-\delta), & \text{for } 3\pi/2\le\beta\le 2\pi-\beta \\ ma\sin\theta - m\delta, & \text{for } 2\pi-\beta\le\theta\le 2\pi \end{cases}$$

After some character-building computations, one finds that if $a \ge \delta$, then

$$\eta(a)=\eta_{re}+j\eta_{im},$$

where

$$\eta_{re}(a)=\frac{ma}{2}[1-f(1-2\delta/a)],\ \eta_{im}(a)=\frac{4m\delta}{\pi}\left(\frac{\delta}{a}-1\right),$$

and $f(\cdot)$ is the function defined in (59). Of course, if $|a| < \delta$, then $\eta(a)=0$.

4.1.3 Periodic Solutions: Informal Arguments

In this section, we discuss the application of the describing function method to the problem of predicting the existence of periodic solutions. The arguments given here are informal, and can only be used to predict the *likelihood* of periodic solutions. The next section presents some precise results under which one can *guarantee* that periodic solutions exist, or that they do not exist.

Consider the nonlinear feedback system shown in Figure 4.13, where G is linear and time-invariant, and N is nonlinear. Specifically, it is assumed that the operator $\mathcal{G}$ is of the form

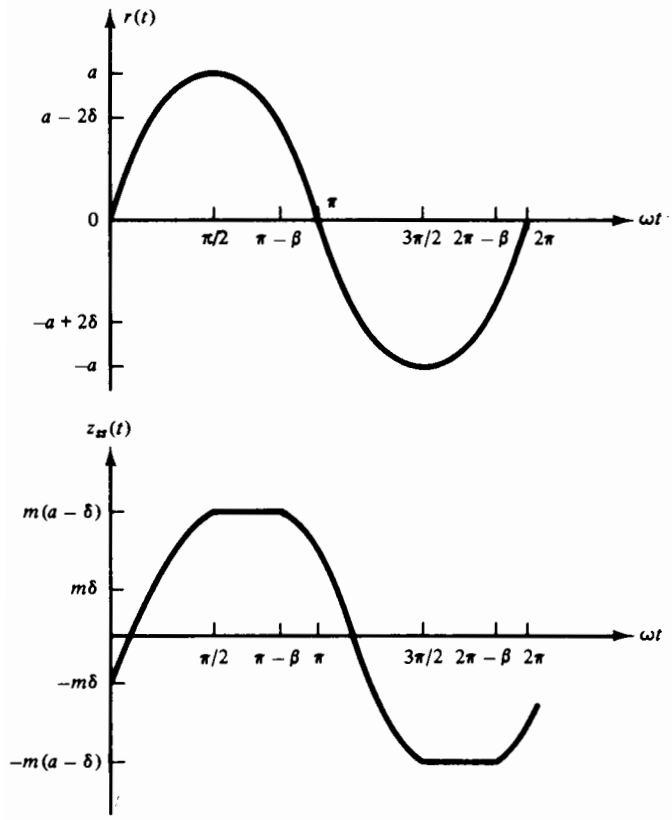


Fig. 4.12

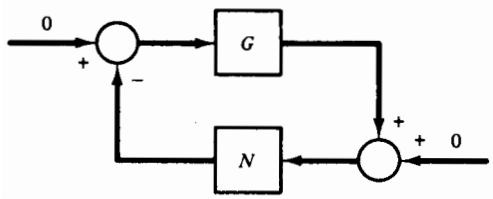


Fig. 4.13

62 $$(Gr)(t) = \int_0^t g(t-\tau)\, r(\tau)\, d\tau,$$

where $g(\cdot)$ satisfies the condition

63 $\int_0^\infty |g(t)|\, dt < \infty.$

As shown in Section 6.4, this hypothesis implies that the operator G is BIBO stable. Moreover, by minor variations of the arguments used therein, one can show that G maps $\mathbf{F}$ into $\mathbf{F}$; in other words, Gr has finite average power whenever r does. The nonlinear element N is also assumed to map $\mathbf{F}$ into itself. Moreover, it is assumed that if r is a pure sinusoid of the form (23), then $Nr =: z = z_{ss} + z_{tr}$, where z_{ss} is continuous and periodic with period $2\pi/\omega$, and z_{tr} satisfies (37). In effect, the assumption is that N has a describing function $\eta(a, \omega)$ for all $a,\ \omega \geq 0$. Now, in the absence of an input, the system of Figure 4.13 is described by

64 $x = -GNx.$

The problem is to determine whether there exists a *nonzero* periodic function $x(\cdot)$ satisfying (64).

The common approach to solve this problem is to *assume* that (64) has a periodic solution of the form

65 $x(t) = a \sin \omega_0 t,$

where a and ω_0 are to be determined. Now, by assumption, the steady-state part z_{ss} of Nx is also periodic with a period of $2\pi/\omega$. Moreover, by definition, the first harmonic z_1 of Nx is given by

66 $z_1(t) = \mu a \sin(\omega_0 t + \phi),$

where

67 $\mu \exp(j\phi) := \eta(a, \omega_0)$

is the describing function of the nonlinearity N. Now define the transfer function $\hat{g}(s)$ of the linear time-invariant operator in the familiar way, namely

68 $\hat{g}(s) = \int_0^\infty g(t) \exp(-st)\, dt.$

If the input $z = Nx$ is applied to G, then $Gz =: y$ also consists of a sum $y_{ss} + y_{tr}$, where y_{tr} satisfies a condition analogous to (37), and y_{ss} is periodic with period $2\pi/\omega_0$. Moreover, if

69 $\hat{g}(j\omega_0) = \gamma \exp(j\theta),$

then the first harmonic of y_{ss} equals

70 $y_1(t) = -\gamma\mu a \sin(\omega_0 t + \phi + \theta).$

Now the essence of the method is to *equate* y_1 *to* x. Strictly speaking, this is not correct, because (64) is the equation we want to solve, not

71 $x =$ first harmonic of $-GNx$.

But the rationale is that if G is a low-pass filter, then $\hat{g}(j\omega) \to 0$ very rapidly as ω increases, so that the higher harmonics of Nx are attenuated by G, and GNx is virtually a pure sinusoid. Thus (71) is a good approximation to (64).

By substituting from (70) into (71), one observes that (71) is equivalent to

72 $1 + \hat{g}(j\omega_0)\eta(a, \omega_0) = 0.$

Some authors refer to (72) as the *principle of harmonic balance.* Note that (72) is quite unrelated to Theorem (36). If (72) is satisfied for some choice of a and ω_0, then informally one believes that there is a periodic solution of (64) which is "close" to (65).

In the special case where η is independent of ω, (72) is particularly easy to solve. As the preceding examples show, η is indeed independent of ω for a wide class of nonlinearities, so this is a very useful and important special case. In this case (72) can be rewritten as

73 $$\hat{g}(j\omega_0) = -\frac{1}{\eta(a)}.$$

This equation can be solved graphically by plotting $\hat{g}(j\omega)$ as a function of ω and $-1/\eta(a)$ as a function of a on the same plane. Every intersection of these curves corresponds to a solution of (72) and thus represents a potential periodic solution.

74 **Example** Consider the feedback system of Figure 4.13, where

$$\hat{g}(s) = \frac{50}{(s+1)(s+2)(s+3)},$$

and N is the hysteresis nonlinearity of Example (60) with

$$m = 0.3, b = 0.01.$$

Then, as derived in Example (60), we have

$$\eta(a) = 0.3 + j\frac{0.04}{\pi a}.$$

Figure 4.14 shows the plots of $\hat{g}(j\omega)$ and of $-1/\eta(a)$. One can see that the two plots intersect at $-1.23 + j1.61$, which corresponds roughly to

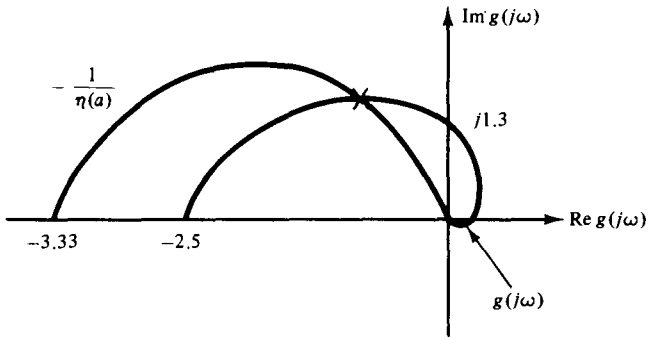


Fig. 4.14

$$\omega = 0.62, a = 0.0325.$$

Hence one predicts a limit cycle with an amplitude of 0.0325 and an angular frequency of 0.62 rad/sec, i.e., a period of roughly ten seconds. ■

An even simpler special case occurs when N is an odd memoryless operator of the form studied in Lemma (53). Suppose to be specific that the function $n(\cdot)$ lies in the sector $[k_1, k_2]$ where $k_1 \geq 0$. Then $\eta(a)$ is always real and the plot of $-1/\eta(a)$ is contained within the interval $[-1/k_1, -1/k_2]$. If $k_1 = 0$, then the plot of $-1/\eta(a)$ is contained within the half-line $(-\infty, -1/k_2]$. Hence the frequencies of potential periodic solutions are precisely those frequencies at which the plot of $\hat{g}(j\omega)$ intersects the negative real axis.

75 Example Consider the feedback system of Figure 4.13, where

$$\hat{g}(s) = \frac{(s+20)^2}{(s+1)(s+2)(s+3)}.$$

Suppose N is the limiter nonlinearity of Figure 4.7, with

$$m_1 = 4, m_2 = 0, \delta = 1.$$

The Nyquist plot of $\hat{g}(j\omega)$ is shown in Figure 4.15. The plot intersects the negative real axis twice; roughly

$$\hat{g}(j5.4197) = -2.1771, \hat{g}(j11.9007) = -0.3062.$$

The corresponding values of η satisfying (73) are given by

$$\eta_1 = \frac{1}{2.1771} = 0.4593, \eta_2 = \frac{1}{0.3062} = 3.2657.$$

Since the slope m_1 equals 4, using the results of Example (58) shows that we must have

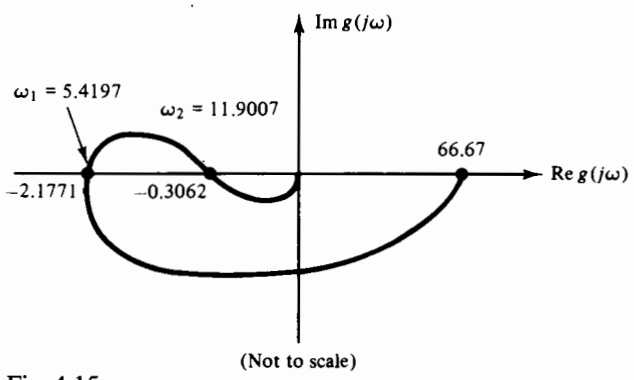


Fig. 4.15

$$f(1/a_1) = \frac{\eta_1}{4} = 0.1148, f(1/a_2) = \frac{\eta_2}{4} = 0.8164.$$

Now one can compute that

$$a_1 = 11.0742, a_2 = 1.4184.$$

Hence for this system we predict two limit cycles: one at an angular frequency of 5.4197 rad/sec and an amplitude of 11.0742, and another at an angular frequency of 11.9007 rad/sec and an amplitude of 1.4184.

Problem 4.1 Suppose N is defined as in Example (22), and let the reference input be the constant signal k. Show that a linear time-invariant system with the transfer function $\hat{h}(s)$ is an optimal quasi-linearization of N if and only if $\hat{h}(0) = n(k)/k$.

Problem 4.2 Given two operators N_1 and N_2 of the type studied in Section 4.1.1, define their sum $N_1 + N_2$ to be the operator defined by

$$[(N_1 + N_2)x](t) = (N_1 x)(t) + (N_2 x)(t).$$

Let $r \in \mathbf{F}$ be a given reference input. Show that if H_i is an optimal quasi-linearization of N_i for $i = 1, 2$, then $H_1 + H_2$ is an optimal quasi-linearization of $N_1 + N_2$.

Problem 4.3 Using the results of Problem 4.2, show that the describing function of the operator $N_1 + N_2$ is given by $\eta_1(a, \omega) + \eta_2(a, \omega)$.

Problem 4.4 Verify that the nonlinearity of Figure 4.3 is the sum of the nonlinearities of Figures 4.6 and 4.7 respectively. Verify that the describing function of the sum function is the sum of the describing functions of the individual nonlinearities.

Problem 4.5 Using the results of Problem 4.3, show that the describing function of the dead-zone limiter shown in Figure 4.16 is

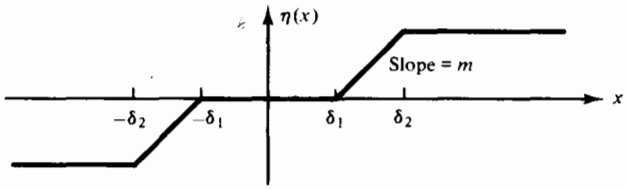


Fig. 4.16

$$\eta(a) = m[f(\delta_2/a) - f(\delta_1/a)].$$

Problem 4.6 Using the results of Problem 4.3, show that the describing function of the piecewise-linear nonlinear element shown in Figure 4.17 is

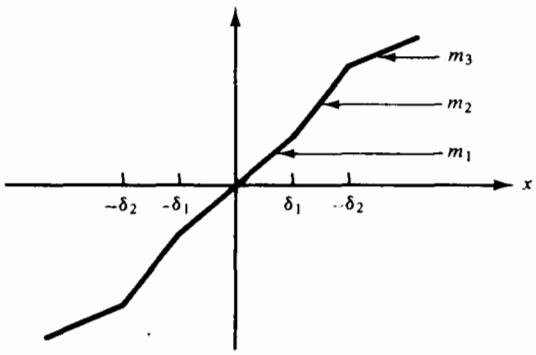


Fig. 4.17

$$\eta(a) = (m_1 - m_2)f(\delta_1/a) + (m_2 - m_3)f(\delta_2/a) + m_3.$$

Problem 4.7 Let $n(\cdot)$, $n_1(\cdot)$, and $n_2(\cdot)$ be odd functions such that

$$n_1(\sigma) \le n(\sigma) \le n_2(\sigma), \ \forall \sigma \ge 0.$$

Let $\eta(\cdot)$, $\eta_1(\cdot)$, and $\eta_2(\cdot)$ denote respectively the describing functions of these nonlinearities. Show that

$$\eta_1(a) \le \eta(a) \le \eta_2(a), \ \forall a > 0.$$

Problem 4.8 Consider the feedback system of Figure 4.13 with

$$\hat{g}(s) = \frac{(s+20)(s+30)}{(s+1)(s+2)(s+4)}.$$

Analyze the possible existence of periodic solutions when

(a) N is the sign nonlinearity of Figure 4.2;

(b) N is the limiter nonlinearity of Figure 4.7 with slope = 4 and $\delta = 2$;

(c) N is the dead-zone nonlinearity of Figure 4.6 with slope = 2 and $\delta = 0.1$.

4.2 PERIODIC SOLUTIONS: RIGOROUS ARGUMENTS

In this section, we study the existence of periodic solutions in the system of Figure 4.13. But in contrast to Section 4.1.3, the arguments presented here are mathematically rigorous.

Though the contents of the present section logically belong just after Section 4.1 for pedogogical reasons, they make use of several concepts that are only introduced in Chapter 6. Thus the reader is advised to read Chapter 6, at least through Section 6.5, before tackling the present section.

Throughout the section, the object of attention is the single-input, single-output feedback system shown in Figure 4.13, where G is linear and time-invariant, and N is a memoryless nonlinear element. Specifically, it is assumed that the operator G is of the form

$$\mathbf{1} \qquad (Gx)(t) = \int_0^t g(t-\tau)\, x(\tau)\, d\tau,$$

where

$$\mathbf{2} \qquad \int_0^\infty |g(t)|\, dt < \infty.$$

The operator N is of the form

$$\mathbf{3} \qquad (Nx)(t) = n[x(t)],$$

where $n(\cdot)$ is an odd continuous function. The describing function of N is denoted by $\eta(\cdot)$. In the absence of an input, the system of Figure 4.13 is described by

$$\mathbf{4} \qquad x = -GNx.$$

Note that since $n(0) = 0$, $x \equiv 0$ is always a solution of (4). Thus the objective is to determine whether the equation (4) has any *nontrivial* periodic solution.

Two types of results are presented here. The first type of result gives conditions under which one can conclude that there *does not* exist a nontrivial periodic solution to (4), while the second type of result gives conditions under which one can conclude that there *does* exist a nontrivial periodic solution to (4).

The hypotheses on $\hat{g}$ and N are stated next. As stated above, G is of the form (1)-(2), while N is assumed to be of the form (3) where $n(\cdot)$ is an odd continuous function. Moreover, it is assumed that the function $n(\cdot)$ is continuously differentiable, and that its derivative n' satisfies

5 $\quad n'(x) \in [k_1, k_2], \forall x \in \mathbf{R}.$

A function satisfying (5) is said to belong to the *incremental sector* $[k_1, k_2]$. Note that (5) is a more restrictive condition than (4.1.54).

It is now shown that, using a technique known as *loop transformation*, one can assume that $k_1 = -k_2$ without loss of generality. Suppose $c \neq 0$ is a real number, and suppose the plot of $\hat{g}(j\omega)$ neither intersects nor encircles the point $-1/c$. Then, by the Nyquist criterion [see e.g., Theorem (6.5.35)], it follows that the operator $(I + cG)^{-1}G$ is also of the form (1)-(2). Now a periodic function x satisfies (4) if and only if

6 $\quad x + cGx = -GNx + cGx = -G(N - cI)x$, or, equivalently, $x = (I + cG)^{-1}G(N - cI)x$,

where I denotes the identity operator. Now the operator $N - cI$ is also memoryless, and $n - cI$ is also odd. Moreover, it is easy to see that it belongs to the incremental sector $[k_1 - c, k_2 - c]$. In particular, if one chooses

7 $\quad c = \dfrac{k_2 + k_1}{2}, r = \dfrac{k_2 - k_1}{2},$

then $N - cI$ belongs to the incremental sector $[-r, r]$. Hence, without loss of generality, one can assume that in (4) the nonlinear element N belongs to the *symmetric* incremental sector $[-r, r]$.

There is another notational change to be introduced. We are seeking periodic functions $x(\cdot)$ satisfying (4), of whatever period. Now, the set of all periodic functions is quite an unwieldy object. In fact, it is not even a linear vector space. So, in order to impart some structure to the problem, let us *normalize* the period to 2π by scaling the time variable. Thus, if $x(t)$ is periodic with a period of $2\pi/\omega$ and angular frequency ω, then $x(t/\omega)$ is also periodic, but with a period of 2π. By scaling time in this manner, it can be assumed that the unknown quantity x in (4) belongs to the set $L_2[0, 2\pi]$ of real-valued, measurable, square-integrable functions over $[0, 2\pi]$. (See Section 6.1 for more details about the set L_2.) But this time-scaling has an important side effect: Let x_ω denote the function mapping t into $x(t/\omega)$. Since N is memoryless, it is easy to see that N is not affected by time-scaling; i.e., $Nx_\omega = (Nx)_\omega$. However, the operator G *is* affected. In fact, it is necessary to replace G by a *family* of operators G_ω. Though it is intuitively clear how this is done, the procedure is explained below so as to leave no misunderstanding.

Suppose $s \in L_2[0, 2\pi]$. Then s has a Fourier series in the familiar form

$$\mathbf{8} \quad s(t) = \beta_0 + \sum_{m=1}^{\infty} (\alpha_m \sin mt + \beta_m \cos mt) = \sum_{m=0}^{\infty} \text{Im}\,[\gamma_m \exp(jmt + \theta_m)],$$

where

$$\mathbf{9} \quad \gamma_m \exp(j\theta_m) = \alpha_m + j\beta_m.$$

Now suppose x is periodic with period $2\pi/\omega$. Then $s(t) = x(t/\omega)$ is periodic with period 2π. In the steady state $(Gx)(t)$ is periodic with period $2\pi/\omega$, while $(Gx)(t/\omega)$ is periodic with period 2π and hence belongs to $L_2[0, 2\pi]$. To compute its Fourier series, define the sequence of complex numbers

$$\mathbf{10} \quad \xi_m(\omega) = \hat{g}(j\omega m),$$

and suppose $s(\cdot)$ is of the form (8). Then it is easy to see that

$$\mathbf{11} \quad (Gx)(t/\omega) = \sum_{m=0}^{\infty} \text{Im}\,[\hat{g}(j\omega m)\,\gamma_m \exp(jmt + \theta_m)].$$

Let us now *define* the operator $G_\omega: L_2[0, 2\pi] \to L_2[0, 2\pi]$ by (11). With this definition, it is true that $(Gx)_\omega = G_\omega x_\omega$; that is, if $x_\omega \in L_2[0, 2\pi]$ is the time-scaled version of x, then $G_\omega x_\omega$ is the time-scaled version of (the steady-state part of) Gx.

With this notation as well as the loop transformation described earlier, the problem under study can be restated as follows: Suppose G is an operator of the form (1)-(2), and let $\{G_\omega\}$ denote the corresponding *family* of operators defined by (11). Suppose N is a memoryless, time-invariant, odd nonlinearity of the form (3), belonging to the incremental sector $[-r, r]$. It is desired to know whether there exist (i) a function $x \in L_2[0, 2\pi]$, and (ii) a number $\omega > 0$, such that

$$\mathbf{12} \quad x = -G_\omega N x.$$

If (12) is satisfied for some $x \in L_2[0, 2\pi]$ and some $\omega > 0$, then ω gives the frequency of oscillation while x gives the waveform.

To simplify the problem further, we restrict attention only to those functions in $L_2[0, 2\pi]$ that possess no d.c. bias, i.e., those functions $x(\cdot)$ with the property that

$$\mathbf{13} \quad \int_0^{2\pi} x(t)\,dt = 0.$$

Let $L_{20}[0, 2\pi]$ denote the set of all such functions. One can see that $L_{20}[0, 2\pi]$ is precisely the subspace of those functions in $L_2[0, 2\pi]$ whose Fourier series do not contain a constant term. If we define the norm of a function x in $L_{20}[0, 2\pi]$ by

$$\textbf{14} \quad \|x\| = \left[\frac{1}{2\pi} \int_0^{2\pi} x^2(t)\, dt \right]^{1/2},$$

then $L_{20}[0, 2\pi]$ is a closed subspace of $L_2[0, 2\pi]$ and is therefore a Banach space in its own right.

Theorem (19) below gives conditions under which (12) does *not* have a solution. To make the theorem statement concise, some notation is first introduced. Given $\hat{g}$, N, define the sets $\Omega_0 \subseteq \Omega \subseteq \mathbf{R}$ as follows:

$$\textbf{15} \quad \Omega_0 = \{\omega \in \mathbf{R}: \sup_{k \geq 1} |\hat{g}(j\omega k)| < r^{-1}\},$$

$$\textbf{16} \quad \Omega = \{\omega \in \mathbf{R}: \sup_{k \geq 2} |\hat{g}(j\omega k)| < r^{-1}\}.$$

For each $\omega \in \mathbf{R}$, define

$$\textbf{17} \quad \lambda(\omega) = \sup_{k \geq 2} |\hat{g}(j\omega k)|,$$

and for each $\omega \in \Omega$ define

$$\textbf{18} \quad \sigma(\omega) = \frac{\lambda(\omega) r^2}{1 - \lambda(\omega) r}.$$

Finally, for each $\omega \in \Omega - \Omega_0$, define $D(\omega)$ to be the disk in the complex plane centered at $-1/\hat{g}(j\omega)$ and of radius $\sigma(\omega)$.

19 Theorem (Nonexistence of Periodic Solutions) *(i) For all $\omega \in \Omega_0$, (12) has no nontrivial solution in $L_{20}[0, 2\pi]$. (ii) Suppose $\omega \in \Omega - \Omega_0$ and that the disk $D(\omega)$ does not intersect the plot of $\eta(a)$ as a varies over $\mathbf{R}$; then (12) has no nontrivial solution in $L_{20}[0, 2\pi]$.*

Proof For each $x \in L_{20}[0, 2\pi]$, let $Px := x_1$ denote the first harmonic of x. Then P is a projection operator on $L_{20}[0, 2\pi]$ in the sense that $P^2 = P$. Let $Q = I - P$ where I is the identity operator, and denote Qx by x_h. Thus x_h denotes the higher harmonics of x. Clearly

$$\textbf{20} \quad \|Px\| \leq \|x\|, \ \|Qx\| \leq \|x\|, \ \forall x \in L_{20}[0, 2\pi].$$

It is easy to see from (11) that $G_\omega P = PG_\omega$, and that $G_\omega Q = QG_\omega$. Now it is clear from (11) that

$$\textbf{21} \quad \|G_\omega Px\| = |\hat{g}(j\omega)| \cdot \|Px\|, \ \forall \omega > 0, \ \forall x \in L_{20}[0, 2\pi],$$

$$\textbf{22} \quad \|G_\omega Qx\| \leq \sup_{k \geq 2} |\hat{g}(j\omega k)| \cdot \|Qx\| = \lambda(\omega) \|Qx\|, \ \forall \omega > 0, \ \forall x \in L_{20}[0, 2\pi],$$

23 $\| G_\omega x \| \leq \sup_{k \geq 1} | \hat{g}(j\omega k) | \cdot \| x \|, \forall \omega > 0, \forall x \in L_{20}[0, 2\pi].$

Finally, observe that since $n(\cdot)$ belongs to the incremental sector $[-r, r]$, we have

24 $\| Nx - Ny \| \leq r \| x - y \|, \forall x, y \in L_{20}[0, 2\pi].$

To prove (i), suppose $\omega \in \Omega_0$. Then (23) and (24) show that the operator $-G_\omega N$ is a contraction on $L_{20}[0, 2\pi]$. Hence, by the contraction mapping theorem [Theorem (2.3.1)], it follows that (12) has a unique solution for x. Since $x = 0$ is already a solution, we conclude that it is the *only* solution.

To prove (ii), suppose $\omega \in \Omega - \Omega_0$, which means that $\lambda(\omega) < r^{-1}$. Let $\mathbf{H}$ denote the range of the projection Q. Thus $\mathbf{H}$ consists of those functions in $L_{20}[0, 2\pi]$ whose first harmonic is identically zero. It is a closed subspace of $L_{20}[0, 2\pi]$ and is thus a Banach space in its own right. Similarly, let $\mathbf{F} \subseteq L_{20}[0, 2\pi]$ denote the range of the projection P. (Note that the present usage of $\mathbf{F}$ is different from what it is in Section 4.1.) Thus $\mathbf{F}$ is a two-dimensional subspace spanned by the two functions $\sin t$ and $\cos t$. Now, by successively applying the operators P and Q to (12), we see that (12) is equivalent to the following *two* equations:

25 $x_1 = -PG_\omega Nx = -PG_\omega N(x_1 + x_h),$

26 $x_h = -QG_\omega Nx = -QG_\omega N(x_1 + x_h).$

Let us first study (26), treating x_1 as a given quantity and x_h as the unknown. Since $\lambda(\omega)r < 1$, it follows from (22) and (24) that the map $x_h \mapsto -QG_\omega N(x_1 + x_h)$ is a contraction. Hence, once again by the contraction mapping theorem, it follows that (26) has a unique solution for x_h corresponding to each x_1. Let us denote this solution by $\alpha(x_1)$. Next, one can obtain an estimate for $\| \alpha(x_1) \|$ using (2.3.1) with $\rho = \lambda(\omega)r$ and the initial guess $x_h = 0$. Then (2.3.1) implies that

27 $$\| \alpha(x_1) \| \leq \frac{\| QG_\omega N(x_1) \|}{1 - \lambda(\omega)r} = \frac{\| G_\omega QN(x_1) \|}{1 - \lambda(\omega)r}.$$

Since $N(0) = 0$, (24) shows that

28 $\| Nx_1 \| \leq r \| x_1 \|.$

Also, it follows from (17) that $\| G_\omega Q \| \leq \lambda(\omega)$. Combining (27) and (28) shows that

29 $$\| \alpha(x_1) \| \leq \frac{\lambda(\omega)r}{1 - \lambda(\omega)r} \| x_1 \|.$$

Now, since (26) has a unique solution for x_h corresponding to each $x_1 \in \mathbf{F}$, it follows that (25) and (26) together are satisfied if and only if there exists an $x_1 \in \mathbf{F}$ such that

30 $x_1 = -PG_\omega N[x_1 + \alpha(x_1)] = -G_\omega PN[x_1 + \alpha(x_1)],$

or equivalently,

31 $x_1 + G_\omega PNx_1 = -G_\omega P\{N[x_1 + \alpha(x_1)] - N(x_1)\}.$

At this stage, it is convenient to identify the set **F** with the complex plane, as follows: Suppose

32 $x_1(t) = a \sin t + b \cos t \in \mathbf{F}.$

Then we define $\phi(x_1)$ as the complex number $a + jb$; note that $\phi(x_1)$ is sometimes called the "phasor" representing x_1. One advantage of this representation is that

33 $\phi(G_\omega x_1) = \hat{g}(j\omega)\,\phi(x_1),$

a fact well-known to undergraduates. Now let ϕ be the phasor representing the function

34 $y := P\{N[x_1 + \alpha(x_1)] - N(x_1)\}.$

Then from (24) and (20), it follows that

35 $$\|y\| \le r\|\alpha(x_1)\| \le \frac{\lambda(\omega)r^2}{1-\lambda(\omega)r}\|x_1\| = \sigma(\omega)\|x_1\|,$$

where the second step follows from (29). Now look at the phasor representation of (31). By the definition of the describing function, we have

36 $\phi(PNx_1) = \eta(\|x_1\|)\,\phi(x_1).$

In other words, the (phasor representing the) first harmonic of Nx_1 is just the describing function of N times the (phasor representing the) signal x_1. Hence (31) becomes

37 $[1 + \hat{g}(j\omega)\eta(\|x_1\|)]\,\phi(x_1) = -\hat{g}(j\omega)\,\phi(y).$

If $\hat{g}(j\omega) \neq 0$, this is equivalent to

38 $$\left[\frac{1}{\hat{g}(j\omega)} + \eta(\|x_1\|)\right]\phi(x_1) = -\phi(y).$$

One advantage of the phasor representation is that it is norm-preserving, i.e.,

39 $|\phi(x_1)| = \|x_1\|, \forall x_1 \in \mathbf{F}.$

Hence, in order for (38) to hold, we must have

$$\textbf{40} \quad \left| \frac{1}{\hat{g}(j\omega)} + \eta(\|x_1\|) \right| \cdot |\phi(x_1)| = |\phi(y)| \leq \sigma(\omega)|\phi(x_1)|,$$

where the last inequality follows from (35). If $\phi(x_1) \neq 0$, then it is possible to divide both sides of (40) by $\phi(x_1)$ to yield

$$\textbf{41} \quad \left| \frac{1}{\hat{g}(j\omega)} + \eta(\|x_1\|) \right| \leq \sigma(\omega).$$

From the definition of the disk $D(\omega)$, (41) implies that $\eta(\|x_1\|) \in D(\omega)$. However, by hypothesis, the plot of $\eta(a)$ as a varies over $\mathbf{R}$ does not intersect the disk $D(\omega)$. Hence (41) cannot be satisfied. Therefore (40) can be satisfied only if $x_1 = 0$, which implies that (12) has no nontrivial solution. ■

The next result gives conditions under which (12) does have a nontrivial solution. Suppose there exist a real number a_0 and a frequency ω_0 such that the harmonic balance condition

$$\textbf{42} \quad N(a_0) + \frac{1}{\hat{g}(j\omega_0)} = 0$$

is satisfied. Then, under certain relatively mild conditions, one can *guarantee* that (12) has a nontrivial solution with a frequency "near" ω_0 and a first harmonic amplitude "near" a_0. Again, some preliminary definitions and notation make the theorem statement more concise.

Let us plot $\eta(a)$ as a varies over $(0, \infty)$, and denote the resulting plot by E. If (42) holds, then clearly the disk $D(\omega_0)$ intersects the plot E, since $-1/\hat{g}(j\omega_0)$ belongs to $D(\omega_0)$. Now, if ω is either increased or decreased from ω_0, the disk $D(\omega_0)$ might eventually not intersect the plot E. Let $[\omega_l, \omega_u]$ be the interval containing ω_0 such that $D(\omega)$ intersects the plot E for all $\omega \in [\omega_l, \omega_u]$, but not if ω is increased beyond either limit. Of course, if (42) is satisfied for several values of ω_0, then there is one such interval corresponding to each ω_0. The definition of ω_l and ω_u makes it clear that the disks $D(\omega_l)$ and $D(\omega_u)$ are tangent to the plot E, and hence to the real axis. The situation can be depicted as in Figure 4.18. Now define

$$\textbf{43} \quad S = \bigcup_{\omega \in [\omega_l, \omega_u]} D(\omega).$$

Then it is easy to show that S is a connected set. Moreover, every point on the boundary ∂S of S is a boundary point of $D(\omega)$ for some $\omega \in [\omega_l, \omega_u]$. Let $\eta(a_l)$, $\eta(a_u)$ denote the two points on the real axis where the boundary of S intersects the real axis. (Again, refer to Figure 4.18.) Now we make a couple of assumptions.

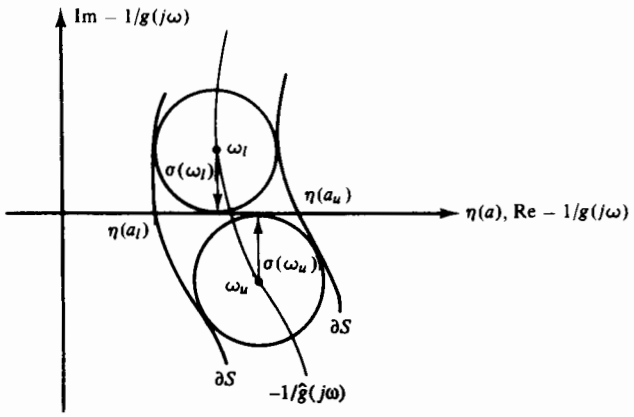


Fig. 4.18

A1. The map η: $[a_l, a_u] \mapsto [\eta(a_l), \eta(a_u)]$ is continuous, one-to-one, and has a continuous inverse.

A2. On the interval $[\omega_l, \omega_u]$, the map taking ω into $\operatorname{Im} \hat{g}(j\omega)$ is continuous, one-to-one, and has a continuous inverse.

Now the main result can be stated.

44 Theorem *Suppose Assumptions (A1) and (A2) are satisfied, and that* $[\omega_l, \omega_u] \subseteq \Omega$. *Then there exist an* $a \in [a_l, a_u]$ *and an* $\omega \in [\omega_l, \omega_u]$ *such that (12) has a nontrivial solution for x, and moreover* $x_1(t) = a \sin \omega t$.

The proof of Theorem (44) is based on the following result, known as the **Leray-Schauder fixed point theorem.** Actually, what is given here is a simplified version suitable for the present situation. The general result can be found in Lloyd (1978).

45 Theorem *Suppose* $M \subseteq \mathbf{R}^2$ *is closed, bounded, simply connected, and contains* 0. *Suppose* $H : M \to \mathbf{R}^2$ *is continuous. Finally, suppose that for every point* $p \in \partial M$,

46 $H(p) \neq \lambda p, \forall \lambda > 1.$

Then H has a fixed point in M; i.e., there exists a $p \in M$ *such that* $H(p) = p$.

Remarks Comparing the Leray-Schauder fixed point theorem with the contraction mapping theorem of Section 2.3, one can see several important differences. Unlike the contraction mapping theorem, the Leray-Schauder theorem does not state that there is *only one* fixed point; moreover, it does not give any systematic procedure for finding a fixed point. Thus the Leray-Schauder theorem is purely an existence theorem. On the other hand, its hypotheses are quite different from those of the contraction mapping theorem.

Proof of Theorem (45) Define the rectangle

47 $R = [a_l, a_u] \times [\omega_l, \omega_u]$,

and define a map $z: \mathbf{R} \to C$ by

48 $z(a, \omega) = \eta(a) + \dfrac{1}{\hat{g}(j\omega)}$.

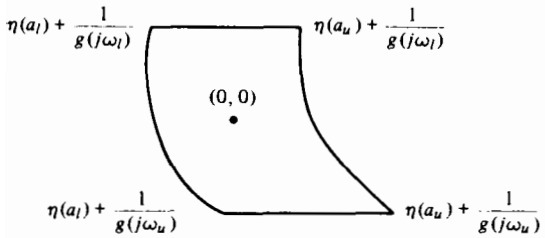


Fig. 4.19

Then the image of R under the map z is the set M shown in Figure 4.19. By identifying the complex plane C with $\mathbf{R}^2$ in the obvious fashion, one can also think of M as a subset of $\mathbf{R}^2$. Clearly $0 \in M$, because of (42). Hence M satisfies all the hypotheses of Theorem (45).

The map $z: R \to M$ is continuous and onto. Now it is claimed that z is one-to-one and that z^{-1} is continuous. To establish this claim, suppose z is given. Since $\eta(a)$ is real, it follows that

49 $\operatorname{Im} \dfrac{1}{\hat{g}(j\omega)} = \operatorname{Im} z.$

By Assumption (A2), this determines ω uniquely. Once ω is known, we have

50 $\eta(a) = \operatorname{Re} z - \operatorname{Re} \dfrac{1}{\hat{g}(j\omega)}$.

By Assumption (A1), this determines a uniquely. Thus z is one-to-one. It is clear that z^{-1} is also continuous in view of Assumptions (A1) and (A2).

Now define a map $H: M \to \mathbf{R}^2$ as follows: For each $p \in M$, let $(a, \omega) = z^{-1}(p)$. Let

51 $x_1 = a \sin \omega t$.

Let $\alpha(x_1)$ denote the corresponding unique solution of (26), and let y be as in (34). Finally, define

52 $H(p) = -\dfrac{\phi(y)}{a}$,

where $\phi(y)$ is the phasor representation of y. Here again we identify the complex number $\phi(y)$ with an ordered pair of real numbers in the obvious fashion. Note that H is the composition of continuous maps and is therefore itself continuous.

Now it is claimed that H satisfies the condition (46). To show this, suppose $p \in \partial M$. It is clear from Figure 4.19 that if $p \in \partial M$ and $(a, \omega) = z^{-1}(p)$, then either $a = a_l$ or a_u, or else $\omega = \omega_l$ or ω_u, or both. Now, since $\omega \in [\omega_l, \omega_u] \subseteq \Omega$, all the arguments used in the proof of Theorem (19) apply. In particular, (35) holds; i.e.,

53 $\|y\| \leq \sigma(\omega)\|x_1\| = \sigma(\omega)a.$

Hence

54 $|H(p)| \leq \sigma(\omega).$

However, it is claimed that, whenever $p \in \partial M$,

55 $|p| \geq \sigma(\omega).$

To establish (55), we consider four cases, namely: (i) $a = a_l$, (ii) $a = a_u$, (iii) $\omega = \omega_l$, and (iv) $\omega = \omega_u$. At least one of these must be true whenever $p \in \partial M$. Suppose first that $a = a_l$. Then, from Figure 4.20, one sees that

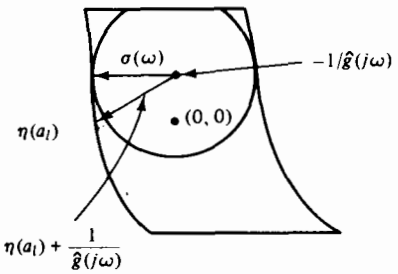


Fig. 4.20

56 $\left|\eta(a_l) + \dfrac{1}{\hat{g}(j\omega)}\right| \geq \sigma(\omega), \ \forall \omega \in [\omega_l, \omega_u].$

Similar reasoning applies if $a = a_u$. Now suppose $\omega = \omega_l$. Since the disk $D(\omega_l)$ is tangent to the real axis and $\eta(a)$ is always real, it follows that

$$\mathbf{57} \quad \left| \eta(a) + \frac{1}{\hat{g}(j\omega_l)} \right| \geq \sigma(\omega_l), \ \forall a \in [a_l, a_u].$$

This can also be seen from Figure 4.18. Similar reasoning applies if $\omega = \omega_u$. This establishes (55). Finally, (54) and (55) together prove (46).

Since all the hypotheses of Theorem (45) are satisfied, we conclude that there exists a $p \in M$ such that $H(p) = p$. Equivalently, there exist an $a \in [a_l, a_u]$ and an $\omega \in [\omega_l, \omega_u]$ such that

$$\mathbf{58} \quad \eta(a) + \frac{1}{\hat{g}(j\omega)} = -\frac{\phi(y)}{a},$$

where y is given by (34). Multiplying both sides of (58) by $\hat{g}(j\omega)a$ leads to

$$\mathbf{59} \quad a + \hat{g}(j\omega)\,\eta(a)\,a = -\hat{g}(j\omega)\,\phi(y).$$

If we define x_1 by (51), then the left side of (59) is precisely the phasor representation of $x_1 + G_\omega PNx_1$, while the right side of (59) is the phasor representation of

$$\mathbf{60} \quad -G_\omega y = -G_\omega P\{N[x_1 + \alpha(x_1)] - N(x_1)\}.$$

Hence x_1 satisfies (31). By the discussion preceding (31), we conclude that

$$\mathbf{61} \quad x = x_1 + \alpha(x_1) \in L_{20}[0, 2\pi]$$

is a nontrivial solution of (12). ■

Theorems (19) and (44) apply to the case where the nonlinear element $n(\cdot)$ belongs to the incremental sector $[-r, r]$. As mentioned above, this represents no loss of generality, as one can always carry out a loop transformation; see the paragraph containing Equations (6) and (7). Nevertheless, it is desirable to recast Theorems (19) and (44) to cover the case of a general nonlinearity. This is done next.

62 Corollary *Consider (12), where the operator N satisfies (3) and (5). Define c and r as in (7), and define the sets* $\Omega_0 \subseteq \Omega \subseteq \mathbf{R}$ *as follows:*

$$\mathbf{63} \quad \Omega_0 = \{\omega \in \mathbf{R}: \sup_{k \geq 1} \left| \frac{\hat{g}(j\omega k)}{1 + c\hat{g}(j\omega k)} \right| < r^{-1}\},$$

$$\mathbf{64} \quad \Omega = \{\omega \in \mathbf{R}: \sup_{k \geq 2} \left| \frac{\hat{g}(j\omega k)}{1 + c\hat{g}(j\omega k)} \right| < r^{-1}\}.$$

For each $\omega \in \Omega$, *define*

$$\mathbf{65}\quad \lambda(\omega)=\sup_{k\geq 2}\left|\frac{\hat{g}(j\omega k)}{1+c\hat{g}(j\omega k)}\right|,$$

$$\mathbf{66}\quad \sigma(\omega)=\frac{\lambda(\omega)r^2}{1-\lambda(\omega)r}.$$

For each $\omega\in\Omega-\Omega_0$, define $D(\omega)$ to be the disk in the complex plane centered at $-1/\hat{g}(j\omega)$ and of radius $\sigma(\omega)$. Under these conditions, we have the following: (i) For all $\omega\in\Omega_0$, (12) has no nontrivial solution in $L_{20}[0, 2\pi]$. (ii) Suppose $\omega\in\Omega-\Omega_0$, and the disk $D(\omega)$ does not intersect the plot of $\eta(a)$ as a varies over $\mathbf{R}$; then (12) has no nontrivial solution in $L_{20}[0, 2\pi]$. (iii) Suppose there exist an ω_0 and an $a_0\in\mathbf{R}$ such that

$$\mathbf{67}\quad \eta(a_0)+\frac{1}{\hat{g}(\omega_0)}=0.$$

Define a_l, a_u, ω_l, and ω_u as in Figure 4.18. Then there exist an $a\in[a_l, a_u]$ and an $\omega\in[\omega_l, \omega_u]$ such that (12) has a nontrivial solution for x, and moreover $x_1=a\sin\omega t$.

Proof As shown in (6) and (7), one can apply Theorems (19) and (44) after replacing N by $N_t=N-cI$ and $\hat{g}(j\omega)$ by

$$\mathbf{68}\quad \hat{g}_t(j\omega)=\frac{\hat{g}(j\omega)}{1+c\hat{g}(j\omega)}.$$

Clearly the describing function η_t of the operator N_t is given by

$$\mathbf{69}\quad \eta_t(a)=\eta(a)-c,$$

while

$$\mathbf{70}\quad \frac{-1}{\hat{g}_t(j\omega)}=\frac{-1}{\hat{g}(j\omega)}-c.$$

Hence, by loop transformation, all that happens is that the plots of $\eta(a)$ and $-1/\hat{g}(j\omega)$ get shifted by c. ■

One rather disadvantageous feature of Corollary (62) is the need to plot $-1/\hat{g}(j\omega)$ and $\eta(a)$, contrary to the usual practice of plotting $-1/\eta(a)$ and $\hat{g}(j\omega)$. One way to surmount this difficulty is to plot $\eta(a)$ as a varies over $\mathbf{R}$ and to plot the *reciprocal disk*

$$\mathbf{71}\quad B(\omega)=\{z\in C:-z^{-1}\in D(\omega)\}.$$

The only potential difficulty with doing so is that if $0\in D(\omega)$, then $R(\omega)$ becomes the *complement* of a closed disk instead of being itself a closed disk, making for a rather awkward situation. Similarly, the plot of $-1/\eta(a)$ becomes disconnected if $\eta(a)$ changes sign as a varies. One can of course add an assumption to the effect that $0\notin D(\omega)$ and that $\eta(a)$ always

has the same sign; but this leads to a result which is slightly less general than Corollary (62). It is stated next.

72 Corollary *Consider (12), where the operator N satisfies (3) and (5), and* $k_1 \geq 0$. *Define the sets* Ω, Ω_0 *and the constants* $\lambda(\omega)$ *and* $\sigma(\omega)$ *by (63) to (66). Define the set* $\Omega' \subseteq \Omega$ *by*

$$\textbf{73} \quad \Omega' = \{\omega \in \Omega: \sigma(\omega) | \hat{g}(\omega)| < 1\}.$$

For each $\omega \in \Omega'$, *define* $B(\omega)$ *to be the disk in the complex plane centered at*

$$\textbf{74} \quad \zeta(\omega) = \frac{\hat{g}(j\omega)}{1 - \sigma^2(\omega)|\hat{g}(j\omega)|^2},$$

and with radius

$$\textbf{75} \quad \rho(\omega) = \frac{\sigma(\omega)|\hat{g}(j\omega)|^2}{1 - \sigma^2(\omega)|\hat{g}(j\omega)|^2}.$$

Under these conditions, we have the following: (i) For all $\omega \in \Omega_0$, *(12) has no nontrivial solution in* $L_2[0, 2\pi]$. *(ii) Suppose* $\omega \in \Omega'$ *and disk* $B(\omega)$ *does not intersect the plot of* $-1/\eta(a)$; *then (12) has no nontrivial solution in* $L_2[0, 2\pi]$. *(iii) Suppose there exist an* $a_0 \in \mathbf{R}$ *and an* $\omega_0 \in \mathbf{R}$ *such that*

$$\textbf{76} \quad \hat{g}(j\omega_0) + \frac{1}{\eta(a_0)} = 0.$$

Choose ω_l, ω_u *such that (i)* $\omega_0 \in [\omega_l, \omega_u]$, *(ii) the disk* $B(\omega)$ *intersects the plot of* $-1/\eta(a)$ *for all* $\omega \in [\omega_l, \omega_u]$, *and (iii) the disks* $B(\omega_l)$, $B(\omega_u)$ *are tangent to the plot of* $-1/\eta(a)$. *Define a subset* $T \subseteq C$ *by*

$$\textbf{77} \quad T = \bigcup_{\omega \in [\omega_l, \omega_u]} B(\omega).$$

Define $-1/\eta(a_l)$, $-1/\eta(a_u)$ *to be the two points at which the boundary of T intersects the plot of* $-1/\eta(a)$. *(See Figure 4.21.) Suppose Assumptions (A1) and (A2) [stated just before Theorem (44)] are satisfied. Under these conditions, there exist an* $a \in [a_l, a_u]$ *and an* $\omega \in [\omega_l, \omega_u]$ *such that (12) has a nontrivial solution, and moreover* $x_1 = a \sin \omega t$.

The proof follows directly from Corollary (62) upon noting that the disks $B(\omega)$ and $D(\omega)$ are related by (71).

Actually, Corollary (72) is only *slightly* less general than Corollary (62). By the Riemann-Lebesgue lemma, (2) implies that $\hat{g}(j\omega) \to 0$ as $\omega \to \infty$. Thus $\sigma(\omega) \to 0$ as $\omega \to \infty$. and as a consequence all sufficiently large ω belong to Ω'. Note also that the point $\hat{g}(j\omega)$ lies in the disk $B(\omega)$ for all $\omega \in \Omega'$, but it is *not the center* of the disk. Hence the locus of the disks $B(\omega)$ contains the Nyquist plot of $\hat{g}(j\omega)$, and can be thought of as a "broadening" of the

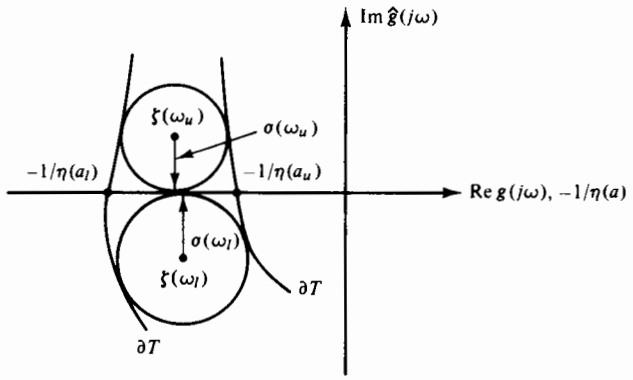


Fig. 4.21

Nyquist plot to reflect the high frequency behavior of $\hat{g}$, i.e., the effects of neglecting the higher harmonics of $G_\omega Nx$.

78 Example Consider the feedback system of Figure 4.13, where

$$\hat{g}(s) = \frac{100}{(s+2)(s+5)(s+10)},$$

and N is the odd piecewise-linear characteristic shown in Figure 4.22. This nonlinearity is of the form studied in Example (4.1.58) with

$$m_1 = 15,\ m_2 = 10,\ \delta = 1.$$

Hence its describing function is given by

$$\eta(a) = 10 + 5f(1/a),$$

where $f(\cdot)$ is the function defined in (4.1.59).

Let us first use the heuristic arguments of Section 4.1.3. The Nyquist plot of $\hat{g}(j\omega)$ is shown in Figure 4.23. The plot intersects the negative real axis when $\omega_0 = \sqrt{80} = 8.9443$ rad/sec, and

$$\hat{g}(j\omega_0) = -0.0794.$$

Thus

$$\eta(a_0) = -\frac{1}{\hat{g}(j\omega_0)} = 12.6,$$

which corresponds to

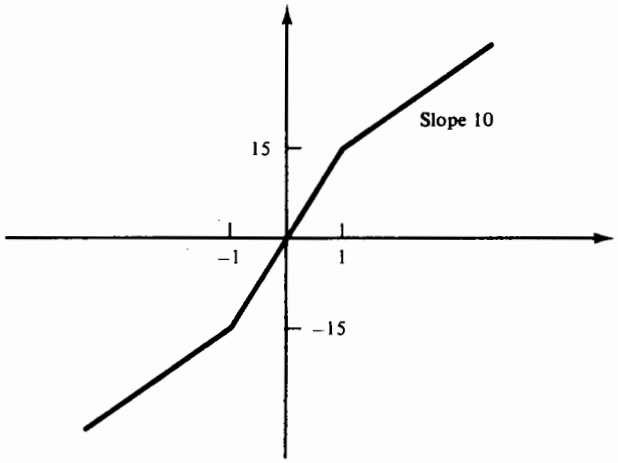


Fig. 4.22

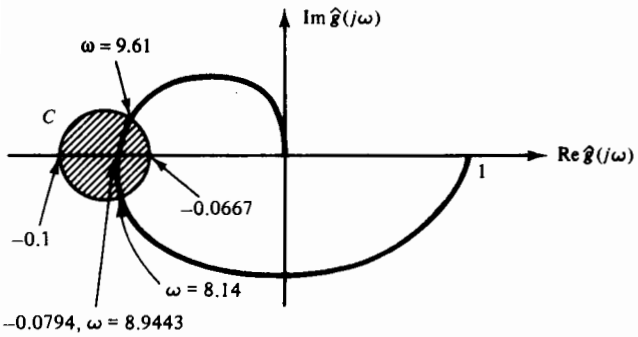


Fig. 4.23

$$a_0 = 2.3742.$$

Hence we "predict" a periodic solution with an angular frequency of 8.9443 rad/sec and an amplitude of 2.3742.

Now let us apply Corollary (72). It is easy to see that N belongs to the incremental sector [10, 15], so one can choose

$$c = 12.5,\ r = 2.5.$$

First let us determine the sets Ω and Ω_0 of (63) and (64) respectively. The shaded region shown in Figure 4.23 is called the *critical disk*, and passes through the points

$$-\frac{1}{m_1} = -0.0667, -\frac{1}{m_2} = -0.01.$$

The critical disk is denoted by $\mathbf{C}$. It is an easy exercise in algebra to show that, if z is any complex number, then

$$\left| \frac{z}{1+cz} \right| < r^{-1} \text{ iff } z \notin \mathbf{C}.$$

In any case, a derivation of the above relationship is found in the proof of Theorem (6.6.40). Now one can see from Figure 4.23 that the Nyquist plot enters the critical disk when $\omega = 8.14$ and leaves it when $\omega = 9.61$. Thus

$$\Omega_0 = \{\omega: k\omega \notin [8.14, 9.61] \; \forall k \geq 1\},$$

$$\Omega = \{\omega: k\omega \notin [8.14, 9.61] \; \forall k \geq 2\},$$

Hence, for simplicity, one can take

$$\Omega_0 = (9.61, \infty), \Omega = (4.805, \infty).$$

So by (i) of Corollary (72), we can state our first *precise* conclusion:

The system has no periodic solution with a frequency > 9.61 rad/sec.

Next, let us determine the frequencies ω_l and ω_u. These are the frequencies at which the disk $B(\omega)$ is tangent to the plot of $-1/\eta(a)$, which in this case is just the interval $(-0.1, -0.667]$. Now it is easy to see that $B(\omega)$ is tangent to the real axis if and only if

$$\text{Im } \zeta(\omega) = \pm \rho(\omega),$$

where $\zeta(\omega)$ and $\rho(\omega)$ are given by (74) and (75) respectively. This allows one to determine that

$$\omega_l = 8.873, \omega_u = 9.011.$$

If $\omega < \omega_l$ or $\omega > \omega_u$, then $B(\omega)$ does not intersect the real axis. Based on (ii) of Corollary (72), we can now state our second precise conclusion:

The system has no periodic solution with frequency in the intervals (4.805, 8.873) and (9.011, 9.61].

Finally, the disk $B(\omega_l)$ is tangent to real axis at

$$\text{Re } \zeta(\omega_l) = -0.0808.$$

So

$$\eta(a_u) = -\frac{1}{\operatorname{Re}\zeta(\omega_l)} = 12.3841, \text{ or } a_u = 2.6028.$$

Similarly,

$$\eta(a_l) = -\frac{1}{\operatorname{Re}\zeta(\omega_u)} = -\frac{1}{0.0781} = 12.8037, \text{ or } a_l = 2.2821.$$

So by (iii) of Corollary (72) we have our third precise conclusion:

There is a periodic solution with angular frequency in the interval [8.873, 9.011] and first harmonic amplitude in the interval [2.1821, 2.6028].

79 Example Consider again the feedback system of Figure 4.13, with

$$\hat{g}(s) = \frac{10(10-s)}{(s+2)(s+3)(s+5)},$$

and N is the limiter nonlinearity shown in Figure 4.24. From Example (4.1.58), it follows that the describing function of N is

$$\eta(a) = 2f(10/a),$$

where f is defined in (4.1.59).

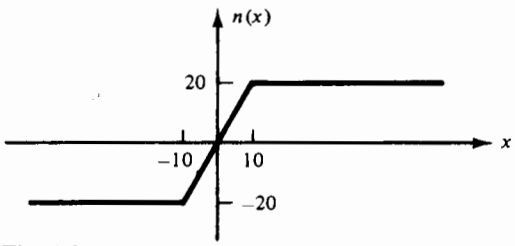


Fig. 4.24

Since the arguments here closely parallel those in Example (78), most of the details are omitted, and only the final conclusions are given.

The Nyquist plot of $\hat{g}(j\omega)$ is shown in Figure 4.25. It crosses the negative real axis at $-1/1.4 \approx 0.71283$, corresponding to $\omega_0 = \sqrt{17} \approx 4.1231$ rad/sec. Now

$$\eta(a_0) = -\frac{1}{\hat{g}(j\omega_0)} = 1.4, \text{ or } a_0 = 17.0911.$$

Thus the classical describing function analysis predicts a periodic solution of amplitude 17.0911 and an angular frequency of 4.1231 rad/sec.

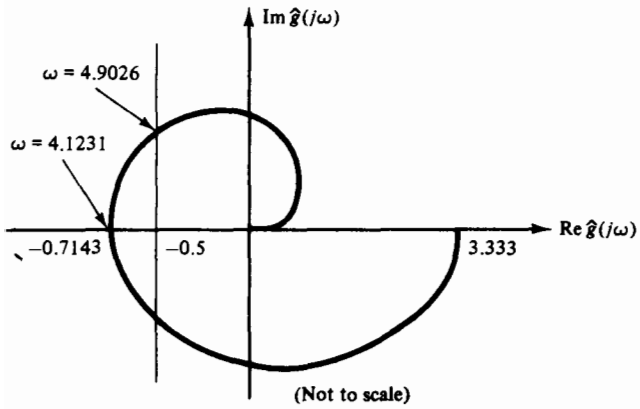


Fig. 4.25

Now let us apply Corollary (72). The nonlinear element N belongs to the incremental sector [0, 2], so we can take

$$c = 1, r = 1.$$

Now, if z is any complex number, then

$$\left| \frac{z}{1+z} \right| < 1 \text{ iff } \operatorname{Re} z > -0.5.$$

From the Nyquist plot, one can see that

$$\operatorname{Re} \hat{g}(j\omega) > -0.5 \ \forall \omega > 4.9026.$$

Hence we can choose

$$\Omega_0 = (4.9026, \infty), \Omega = (2.4513, \infty).$$

Next, the tangency conditions are satisfied when

$$\omega_l = 3.41, \omega_u = 4.42.$$

Now

$$\zeta(\omega_l) = -1.0260 - j0.5296, \zeta(\omega_u) = -0.6366 + j0.0783,$$

$$\eta(a_l) = -\frac{1}{\operatorname{Re} \zeta(\omega_l)} = 1.5707, \text{ or } a_l = 14.894,$$

$$\eta(a_u) = -\frac{1}{\mathrm{Re}\,\zeta(\omega_u)} = 0.9747,\ a_u = 25.439.$$

So we can draw the following conclusions, based on Corollary (72):

(i) *There is no periodic solution with a frequency* > 4.9026.

(ii) *There is no periodic solution with a frequency in* $(2.4513, 3.41)$ *or* $(4.42, 4.9026]$.

(iii) *There is a periodic solution with a frequency* $\omega \in [3.41, 4.42]$ *and a first harmonic amplitude* $a \in [14.894, 25.439]$.

Compared to Example (78), we see that the "spread" in both ω and a is considerably larger in the present instance. This is because the describing function of the nonlinearity in the present example varies over a much larger range than in Example (78). That nonlinearity is much "closer" to being linear. One would naturally expect that there is much less uncertainty in the results obtained using quasi-linearization methods when the nonlinearity is close to being linear. This is reflected in the results of Examples (78) and (79).

Problem 4.9 Consider again the transfer function $\hat{g}(s)$ of Example (78), and suppose the nonlinear element N is the form shown in Figure 4.3 (or 4.22), where the initial slope m_1 for small values of the input equals 13, the final slope m_2 equals 12, and the width δ equals 1. Using Corollary (72), find upper and lower bounds on the amplitude and frequency of the periodic solution. Show that, in this case, the "spreads" between the upper and lower bounds for both the amplitude and the frequency of the periodic solution are less than they are in Example (78). How do you explain this?

Problem 4.10 Consider again the transfer function $\hat{g}(s)$ of Example (79), and suppose the nonlinear element N is a dead-zone limiter of the form shown in Figure 4.7, with slope $m_2 = 20$ and a dead-zone width $\delta = 0.01$. Using Corollary (72), find upper and lower bounds on the amplitude and frequency of the periodic solution, if any. Compare with the results of Example (79).

4.3 SINGULAR PERTURBATIONS

In this section, a brief introduction is given to the method of singular perturbations. This method is valuable in analyzing systems whose dynamic order changes as a result of neglecting some elements, or making some simplifying assumptions [see Example (40) below].

Consider a system of nonlinear differential equations

$$\mathbf{1} \qquad \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{y}),$$

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

where $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{y} \in \mathbf{R}^m$, $\mathbf{f}: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$, and $\mathbf{g}: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^m$. Note that for any value of ε other than zero, the system (1) consists of $n + m$ differential equations. However, if $\varepsilon = 0$, the

system (1) becomes a set of *n differential* equations, and *m algebraic* equations, namely

$$\mathbf{2} \qquad \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},\mathbf{y}),$$

$$\mathbf{0} = \mathbf{g}(\mathbf{x},\mathbf{y}).$$

Suppose it is possible to solve the m algebraic equations above to obtain an explicit expression for $\mathbf{y}$ in terms of $\mathbf{x}$, of the form

$$\mathbf{3}' \qquad \mathbf{y} = \mathbf{h}(\mathbf{x}),$$

where $\mathbf{h}: \mathbf{R}^n \to \mathbf{R}^m$. Then one can substitute (3) into the first equation of (2) to obtain the set of differential equations

$$\mathbf{4} \qquad \dot{\mathbf{x}} = \mathbf{f}[\mathbf{x}, \mathbf{h}(\mathbf{x})].$$

Setting $\varepsilon = 0$ in (1) is called a **singular perturbation** since it changes the *order* of the system. This is to be contrasted with a so-called "regular" perturbation, described next. Suppose we are given a system of n differential equations

$$\mathbf{5} \qquad \dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t), \mathbf{p}],$$

where $\mathbf{x}(t) \in \mathbf{R}^n$, $\mathbf{p} \in \mathbf{R}^m$, and $\mathbf{f}: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ is *continuous*. One can think of $\mathbf{p}$ as a vector of physical parameters appearing in the system description. If the vector $\mathbf{p}$ is perturbed slightly, then the *order* of the system is not affected. In contrast, suppose (1) is rewritten as

$$\mathbf{6} \qquad \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{f}(\mathbf{x},\mathbf{y}) \\ \dfrac{1}{\varepsilon}\mathbf{g}(\mathbf{x},\mathbf{y}) \end{bmatrix}.$$

Then obviously the right side of (6) is *not* continuous with respect to ε at $\varepsilon = 0$. This is why setting $\varepsilon = 0$ is a "singular" perturbation.

The system (1) is called the **full-order, unsimplified,** or **original** system, while (4) is called the **reduced-order** or **simplified** system. In broad terms, the basic objective of singular perturbation theory is to draw conclusions about the behavior of the original system (1) based upon a study of the simplified system (4).

At the moment we do not have the tools to study the stability of nonlinear systems of the form (1) or (4); these are presented in Chapter 5. So in the present section the scope of the study is limited to *linear* singularly perturbed systems, of the form

$$\mathbf{7} \qquad \begin{bmatrix} \dot{\mathbf{x}} \\ \varepsilon\dot{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

where $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{y} \in \mathbf{R}^m$, and the matrices $\mathbf{A}_{ij}$ have compatible dimensions. If ε is set equal to

zero in (7), the second equation becomes

8 $\mathbf{0} = \mathbf{A}_{21}\mathbf{x} + \mathbf{A}_{22}\mathbf{y}$.

If $\mathbf{A}_{22}$ is nonsingular, (8) can be solved to yield

9 $\mathbf{y} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{x}$.

Substituting from (9) into (7) gives the simplified system

10 $\dot{\mathbf{x}} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})\mathbf{x} =: \mathbf{A}_0\mathbf{x}$.

Theorem (12) below presents the main result of this section. To make the theorem statement concise, a little notation is introduced first. Define

11 $\mathbf{A}_0 = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$.

Let $\Lambda = \{\lambda_1, \cdots, \lambda_n\}$ denote the **spectrum** of the matrix $\mathbf{A}_0$, i.e., the set of eigenvalues of $\mathbf{A}_0$, where repeated eigenvalues are listed as many times as their multiplicity. Similarly, let $\Gamma = \{\gamma_1, \cdots, \gamma_m\}$ denote the spectrum of $\mathbf{A}_{22}$.

12 Theorem *Consider the system (7). Suppose $\mathbf{A}_{22}$ is nonsingular, and define $\mathbf{A}_0$ as in (11). Then given any $\delta > 0$, there exists an $\varepsilon_0 > 0$ such that, whenever $0 < |\varepsilon| < \varepsilon_0$, the $n+m$ eigenvalues $\{\alpha_1, \cdots, \alpha_{n+m}\}$ of the matrix*

13 $$\mathbf{A}_\varepsilon = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21}/\varepsilon & \mathbf{A}_{22}/\varepsilon \end{bmatrix}$$

satisfy the bounds

14 $|\lambda_i - \alpha_i| < \delta$, for $i = 1, \cdots, n$,

15 $|\gamma_{i-n} - \varepsilon\alpha_i| < \delta$, for $i = n+1, \cdots, n+m$.

Remarks Clearly the eigenvalues of the matrix $\mathbf{A}_\varepsilon$ are the natural modes of the unsimplified system (7). The inequalities (14) and (15) imply that, as $\varepsilon \to 0$, exactly n of the eigenvalues of $\mathbf{A}_\varepsilon$ converge to the eigenvalues of $\mathbf{A}_0$, while the remaining m eigenvalues "approach" infinity, asymptotically like γ_i/ε. Moreover, if λ_i is an r_i-times repeated eigenvalue of $\mathbf{A}_0$, then exactly r_i eigenvalues of $\mathbf{A}_\varepsilon$ converge to λ_i; similarly for γ_i. If all eigenvalues of $\mathbf{A}_\varepsilon$ have negative real parts, then the solution of (7) approaches $\mathbf{0}$ as $t \to \infty$ for each initial condition; in this case, the system (7) is said to be **asymptotically stable.** On the other hand, if some eigenvalue of $\mathbf{A}_\varepsilon$ has a positive real part, then the norm of the solution of (7) approaches infinity as $t \to \infty$ for almost all initial conditions; in this case the system (7) is said to be **unstable.** See Chapter 5 for precise definitions of these concepts.

Proof To compute the eigenvalues of $\mathbf{A}_\varepsilon$, let us carry out a similarity transformation of $\mathbf{A}_\varepsilon$ such that the resulting matrix is in a block-triangular form. More precisely, let us find a matrix $\mathbf{M}_\varepsilon \in \mathbf{R}^{n\times m}$ such that

$$\mathbf{16}\quad \begin{bmatrix} I & -\mathbf{M}_\varepsilon \\ \mathbf{0}_{m\times n} & I \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21}/\varepsilon & \mathbf{A}_{22}/\varepsilon \end{bmatrix} \begin{bmatrix} I & \mathbf{M}_\varepsilon \\ \mathbf{0}_{m\times n} & I \end{bmatrix} = \begin{bmatrix} \mathbf{F}_\varepsilon & \mathbf{0}_{n\times m} \\ \mathbf{G}_\varepsilon & \mathbf{H}_\varepsilon \end{bmatrix}.$$

Expanding the triple matrix product on the left side of (16) shows that, in order for the 1,2 block of the product to equal zero, the matrix $\mathbf{M}_\varepsilon$ must satisfy the following equation:

$$\mathbf{17}\quad \mathbf{A}_{11}\mathbf{M}_\varepsilon + \mathbf{A}_{12} - \mathbf{M}_\varepsilon \frac{\mathbf{A}_{21}}{\varepsilon}\mathbf{M}_\varepsilon - \mathbf{M}_\varepsilon \frac{\mathbf{A}_{22}}{\varepsilon} = \mathbf{0}.$$

As $\varepsilon \to 0$, some matrices in (17) approach infinity. To get around this difficulty, suppose $\mathbf{M}_\varepsilon$ has the form

$$\mathbf{18}\quad \mathbf{M}_\varepsilon = \varepsilon \mathbf{P}_\varepsilon.$$

Substituting for $\mathbf{M}_\varepsilon$ in (17) and clearing fractions gives the following equation for $\mathbf{P}_\varepsilon$:

$$\mathbf{19}\quad \varepsilon\mathbf{A}_{11}\mathbf{P}_\varepsilon + \mathbf{A}_{12} - \varepsilon\mathbf{P}_\varepsilon\mathbf{A}_{21}\mathbf{P}_\varepsilon - \mathbf{P}_\varepsilon\mathbf{A}_{22} = \mathbf{0}.$$

This equation is quite well-behaved as $\varepsilon \to 0$. In fact, substituting $\varepsilon = 0$ in (19) gives

$$\mathbf{20}\quad \mathbf{A}_{12} - \mathbf{P}_0\mathbf{A}_{22} = \mathbf{0},$$

which has the unique solution

$$\mathbf{21}\quad \mathbf{P}_0 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}.$$

If $\varepsilon \neq 0$, then (19) is a quadratic equation in $\mathbf{P}_\varepsilon$. Nevertheless, since the coefficients in (19) are continuous in ε, one can conclude that for sufficiently small ε *there exists* a solution $\mathbf{P}_\varepsilon$ of (19) which is close to $\mathbf{P}_0$; but this need not be the unique solution of (19). Choose such a solution $\mathbf{P}_\varepsilon$, and define $\mathbf{M}_\varepsilon$ as in (18). Now expanding (16) gives

$$\mathbf{22}\quad \begin{bmatrix} \mathbf{F}_\varepsilon & \mathbf{0}_{n\times m} \\ \mathbf{G}_\varepsilon & \mathbf{H}_\varepsilon \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} - \mathbf{M}_\varepsilon\mathbf{A}_{21}/\varepsilon & \mathbf{0}_{n\times m} \\ \mathbf{A}_{21}/\varepsilon & \mathbf{A}_{21}\mathbf{M}_\varepsilon/\varepsilon + \mathbf{A}_{22}/\varepsilon \end{bmatrix}.$$

Now we know that

$$\mathbf{23}\quad \mathbf{P}_\varepsilon = \mathbf{A}_{12}\mathbf{A}_{22}^{-1} + \mathbf{O}(\varepsilon).$$

Hence, from (22),

$$\mathbf{24}\quad \mathbf{F}_\varepsilon = \mathbf{A}_{11} - \mathbf{P}_\varepsilon \mathbf{A}_{21} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{O}(\varepsilon) = \mathbf{A}_0 + \mathbf{O}(\varepsilon),$$

$$\mathbf{25}\quad \mathbf{H}_\varepsilon = \frac{\mathbf{A}_{21}\mathbf{M}_\varepsilon + \mathbf{A}_{22}}{\varepsilon} = \frac{\mathbf{A}_{22}}{\varepsilon} + \mathbf{A}_{21}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} + \mathbf{O}(\varepsilon).$$

Since the matrix in (22) is block-triangular, it follows that the spectrum of $\mathbf{A}_\varepsilon$ is the union of the spectrum of $\mathbf{F}_\varepsilon$ and the spectrum of $\mathbf{H}_\varepsilon$. As $\varepsilon \to 0$, it is clear from (24) that $\mathbf{F}_\varepsilon \to \mathbf{A}_0$; hence the spectrum of $\mathbf{F}_\varepsilon$ approaches that of $\mathbf{A}_0$, i.e., (14) is satisfied. Finally, as $\varepsilon \to 0$, the $\mathbf{A}_{22}/\varepsilon$ term on the right side of (25) swamps the other terms (recall that $\mathbf{A}_{22}$ is nonsingular), and (15) is satisfied. ■

To state the next result concisely, a couple of terms are first introduced. A square matrix is said to be **hyperbolic** if it has no eigenvalues with zero real part, and it is said to be **Hurwitz** if all of its eigenvalues have negative real parts.

26 Corollary *Consider the matrix $\mathbf{A}_\varepsilon$ of (12), and suppose $\mathbf{A}_{22}$ is hyperbolic. Then the following two statements are equivalent:*

1. *There exists an $\varepsilon_0 > 0$ such that $\mathbf{A}_\varepsilon$ is Hurwitz whenever $0 < \varepsilon < \varepsilon_0$.*
2. *The matrices $\mathbf{A}_0$ and $\mathbf{A}_{22}$ are both Hurwitz.*

In the parlance of singular perturbation theory, the matrix $\mathbf{A}_{22}$ is said to represent the **fast dynamics**, while $\mathbf{A}_0$ is said to represent the **slow dynamics.** Without getting too deeply into the technicalities of the subject (and there are plenty of them), the reasoning behind this terminology can be briefly explained as follows: Suppose both $\mathbf{A}_0$ and $\mathbf{A}_{22}$ are Hurwitz. For each ε sufficiently small and positive, define $\mathbf{M}_\varepsilon$ as the solution of (17) such that $\varepsilon\mathbf{M}_\varepsilon = \mathbf{P}_\varepsilon \to \mathbf{A}_{12}\mathbf{A}_{22}^{-1}$ as $\varepsilon \to 0$. Now define a new state variable vector by

$$\mathbf{27}\quad \begin{bmatrix} \mathbf{z}_\varepsilon \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} I & -\mathbf{M}_\varepsilon \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \text{ i.e., } \mathbf{z}_\varepsilon = \mathbf{x} - \mathbf{M}_\varepsilon \mathbf{y}.$$

Now (16) makes it clear that the dynamics of the new state vector are governed by

$$\mathbf{28}\quad \begin{bmatrix} \dot{\mathbf{z}}_\varepsilon \\ \dot{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_\varepsilon & \mathbf{0} \\ \mathbf{G}_\varepsilon & \mathbf{H}_\varepsilon \end{bmatrix} \begin{bmatrix} \mathbf{z}_\varepsilon \\ \mathbf{y} \end{bmatrix},$$

or, in expanded form,

$$\mathbf{29}\quad \dot{\mathbf{z}}_\varepsilon = \mathbf{F}_\varepsilon \mathbf{z}_\varepsilon,$$

$$\dot{\mathbf{y}} = \mathbf{G}_\varepsilon \mathbf{z}_\varepsilon + \mathbf{H}_\varepsilon \mathbf{y}.$$

Now define the "fast" time variable

30 $\tau = t/\varepsilon,$

and define the functions $\bar{\mathbf{y}}, \bar{\mathbf{z}}_\varepsilon$ by

31 $\bar{\mathbf{y}}(\tau) = \mathbf{y}(\varepsilon\tau), \bar{\mathbf{z}}_\varepsilon(\tau) = \mathbf{z}_\varepsilon(\varepsilon\tau).$

Thus $\bar{\mathbf{y}}(\cdot)$ is just the function $\mathbf{y}(\cdot)$ with the time scaled by the factor ε; similarly for $\bar{\mathbf{z}}_\varepsilon$. It is easy to see that

32 $$\frac{d\bar{\mathbf{y}}(\tau)}{d\tau} = \varepsilon\left[\frac{d\mathbf{y}(t)}{dt}\right]_{t=\tau/\varepsilon}$$

With this change of independent variable, the system equations (29) can be rewritten as

33 $$\frac{d\mathbf{z}_\varepsilon(t)}{dt} = \mathbf{F}_\varepsilon \mathbf{z}_\varepsilon(t),$$

34 $$\frac{d\bar{\mathbf{y}}(\tau)}{d\tau} = \varepsilon\mathbf{G}_\varepsilon \bar{\mathbf{z}}_\varepsilon(\tau) + \varepsilon\mathbf{H}_\varepsilon \bar{\mathbf{y}}(\tau).$$

These equations enable us to understand more clearly the time behavior of the functions $\mathbf{z}_\varepsilon$ and $\bar{\mathbf{y}}$. First, (33) shows that the time response of $\mathbf{z}_\varepsilon$ is independent of the initial condition $\mathbf{y}(0)$ and depends only on the initial condition $\mathbf{z}_\varepsilon(0)$. Since $\mathbf{M}_\varepsilon \to \mathbf{0}$ as $\varepsilon \to 0$, we see that $\mathbf{z}_\varepsilon(0)$ becomes closer and closer to $\mathbf{x}(0)$ as $\varepsilon \to 0$ [cf. (27)]. Second, (34) shows that $\bar{\mathbf{z}}_\varepsilon(\cdot)$ acts like a forcing function to $\bar{\mathbf{y}}(\cdot)$. If $\mathbf{z}_\varepsilon(0) = \mathbf{0}$, then $\bar{\mathbf{z}}_\varepsilon \equiv \mathbf{0}$, and (34) reduces to

35 $$\frac{d\bar{\mathbf{y}}(\tau)}{d\tau} = \varepsilon\mathbf{H}_\varepsilon \bar{\mathbf{y}}(\tau).$$

Now note from (25) that $\varepsilon\mathbf{H}_\varepsilon \to \mathbf{A}_{22}$ as $\varepsilon \to 0$. Hence, as $\varepsilon \to 0$, $\bar{\mathbf{y}}(\cdot)$ looks approximately like

36 $\bar{\mathbf{y}}(\tau) \simeq \exp(\mathbf{A}_{22}\tau)\bar{\mathbf{y}}(0).$

If $\mathbf{z}_\varepsilon(0) \neq \mathbf{0}$, then from (33),

37 $\mathbf{z}_\varepsilon(t) = \exp(\mathbf{F}_\varepsilon t)\mathbf{z}_\varepsilon(0) \simeq \exp(\mathbf{A}_0 t)\mathbf{z}_\varepsilon(0),$

since $\mathbf{F}_\varepsilon \to \mathbf{A}_0$ as $\varepsilon \to 0$. Now, if ε is very small, then for the purposes of analyzing (34) one can treat $\bar{\mathbf{z}}_\varepsilon(\tau)$ as a *constant* vector $\mathbf{z}_\varepsilon(0)$, and replace $\varepsilon\mathbf{G}_\varepsilon$ by $\mathbf{A}_{22}$ [cf. (20)]. Then the approximate solution of (34) is

38 $\bar{\mathbf{y}}(\tau) \simeq \exp(\mathbf{A}_{22}\tau)[\bar{\mathbf{y}}(0) + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{z}_\varepsilon(0)],$

or

39 $\mathbf{y}(t) \simeq \exp(\mathbf{A}_{22}t/\varepsilon)\,[\mathbf{y}(0) + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{z}_\varepsilon(0)].$

Equation (39) shows why $\mathbf{y}$ is often referred to as the **fast state variable**, and also why the fast dynamics are determined by the matrix $\mathbf{A}_{22}$. Now, as $\varepsilon \to 0$, the matrix $\mathbf{M}_\varepsilon \to \mathbf{0}$, and the vector $\mathbf{z}_\varepsilon \to \mathbf{x}$ [cf. (27)]. Thus the vector $\mathbf{x}$ is referred to as the **slow state variable**, and its time evolution is governed by the matrix $\mathbf{A}_0$ as demonstrated by (37).

40 Example In practice, singularly perturbed differential equations arise when some dynamical elements (often called "parasitics") are neglected during the modelling process. This is illustrated in this example.

Consider the circuit shown in Figure 4.26, and suppose the operating point of the tunnel diode is so selected that its small-signal resistance is negative. If the stray capacitance of the diode is included in the network, then one obtains the linearized model shown in Figure 4.27. Following the common practice in network theory, let us choose the capacitor voltages and the inductor current as the state variables. Let us suppose also that the diode capacitance ε is very small. Then the dynamics of the network are described by the third-order equation

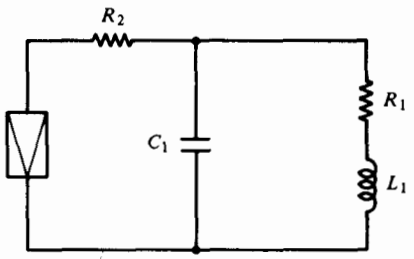


Fig. 4.26

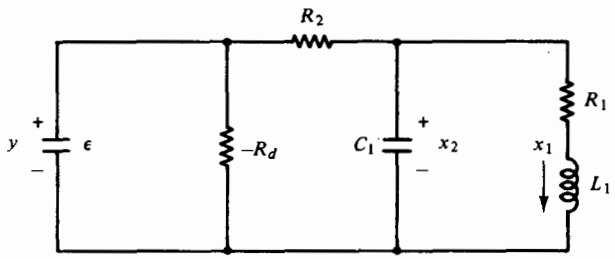


Fig. 4.27

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \varepsilon\dot{y} \end{bmatrix} = \begin{bmatrix} -\frac{R_1}{L_1} & \frac{1}{L_1} & 0 \\ -\frac{1}{C_1} & -\frac{1}{R_1 C_1} & \frac{1}{R_2 C_1} \\ 0 & \frac{1}{R_2} & \frac{R_2 - R_d}{R_2 R_d} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ y \end{bmatrix}.$$

Setting $\varepsilon = 0$ leads to

$$y = -\frac{R_d}{R_2 - R_d} x_2,$$

and to the simplified model

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -\frac{R_1}{L_1} & \frac{1}{L_1} \\ -\frac{1}{C_1} & -\frac{1}{(R_2 - R_d)C_1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The coefficient matrix above is $\mathbf{A}_0$. In this instance the matrix $\mathbf{A}_{22}$ is just a scalar and equals

$$\mathbf{A}_{22} = \frac{R_2 - R_d}{R_2 R_d}.$$

If $R_2 > R_d$, then $\mathbf{A}_{22} > 0$, which means that $\mathbf{A}_{22}$ is *not* Hurwitz. On the other hand, one can easily verify that $\mathbf{A}_0$ is Hurwitz if $R_2 > R_d$. Using Corollary (26), we conclude that if $0 < R_d < R_2$, then the simplified system is asymptotically stable, but the original system is in fact unstable whenever ε is sufficiently small and positive. Thus neglecting the stray capacitance in this instance gives a highly misleading conclusion. ■

It is possible to generalize the preceding development to systems of the form

41 $\dot{\mathbf{x}} = \mathbf{A}(1/\varepsilon)\mathbf{x},$

where each entry of the matrix $\mathbf{A}$ is a polynomial in $1/\varepsilon$. But the analysis of such systems requires much more advanced methods.

Notes and References

A good reference for the standard material on describing functions is Gelb and Vander Velde (1968). The rigorous treatment of periodic solutions in Section 4.2 is adapted from Mees and Bergen (1975).

5. LYAPUNOV STABILITY

In this chapter we study the concept of Lyapunov stability, which plays an important role in control and system theory. We have seen in Chapter 1 that if a system is initially in an equilibrium, it remains in the same state thereafter. Lyapunov stability is concerned with the behavior of the trajectories of a system when its initial state is *near* an equilibrium. From a practical viewpoint, this issue is very important because external disturbances such as noise, wind, and component errors are always present in a real system to knock it out of equilibrium.

Stability theory is a very old subject, dating back almost to the advent of the theory of differential equations. The object of stability theory is to draw conclusions about the behavior of a system without actually computing its solution trajectories. Perhaps the first person to study stability in the "modern" sense was Lagrange (1788), who analyzed mechanical systems using what we now refer to (naturally enough) as Lagrangian mechanics. One of his conclusions was that, in the absence of external forces, an equilibrium of a conservative mechanical system is stable (in a sense to be defined shortly) provided it corresponds to a minimum of the potential energy. Several researchers followed up Lagrange's methods, but for the most part their work was restricted to conservative mechanical systems described by Lagrangian equations of motion. The quantum advance in stability theory that allowed one to analyze *arbitrary* differential equations is due to the Russian mathematician A. M. Lyapunov (1892). He not only introduced the basic definitions of stability that are in use today, but also proved many of the fundamental theorems. Lyapunov's work was largely unknown in the West until about 1960, and almost all the advances in Lyapunov stability theory until that time are due to Russian mathematicians. Today the foundations of the theory are well-established, and the theory is an indispensable tool in the analysis and synthesis of nonlinear systems.

Lyapunov theory abounds in a variety of notions of stability, and one can easily list nearly two dozen definitions of stability. In this book, however, we focus on only a few of these, namely: stability (and its absence, instability), asymptotic stability, and exponential stability.

5.1 STABILITY DEFINITIONS

In this section various types of stability are defined, and the definitions are illustrated by examples.

Throughout the chapter, the object of study is the vector differential equation

1 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], t \geq 0,$

where $\mathbf{x}(t) \in \mathbf{R}^n$, and $\mathbf{f}: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}^n$ is continuous. It is further assumed that the equation (1) has a unique solution corresponding to each initial condition. This is the case, for example, if $\mathbf{f}$ satisfies a global Lipschitz condition [see Theorem (2.4.25)]. It is shown in Appendix A that, roughly speaking, the preceding assumption is true for almost all continuous functions $\mathbf{f}$. Let $\mathbf{s}(t, t_0, \mathbf{x}_0)$ denote the solution of (1) corresponding to the initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$, evaluated at time t. In other words, $\mathbf{s}$ satisfies the equation

2 $$\frac{d}{dt}\mathbf{s}(t, t_0, \mathbf{x}_0) = \mathbf{f}[t, \mathbf{s}(t, t_0, \mathbf{x}_0)], \forall t \geq t_0; \mathbf{s}(t_0, t_0, \mathbf{x}_0) = \mathbf{x}_0.$$

The solution map $\mathbf{s}$ maps $\mathbf{R}_+ \times \mathbf{R}^n$ into $\mathbf{R}^n$, and satisfies the following properties:

3 $\mathbf{s}(t_0, t_0, \mathbf{x}_0) = \mathbf{x}_0, \forall \mathbf{x}_0 \in \mathbf{R}^n,$

4 $\mathbf{s}[t, t_1, \mathbf{s}(t_1, t_0, \mathbf{x}_0)] = \mathbf{s}(t, t_0, \mathbf{x}_0), \forall t \geq t_1 \geq t_0 \geq 0, \forall \mathbf{x}_0 \in \mathbf{R}^n.$

Recall that a vector $\mathbf{x}_0 \in \mathbf{R}^n$ is an **equilibrium** of the system (1) if

5 $\mathbf{f}(t, \mathbf{x}_0) = \mathbf{0}, \forall t \geq 0.$

Clearly, if (5) is true, then

6 $\mathbf{s}(t, t_0, \mathbf{x}_0) = \mathbf{x}_0, \forall t \geq t_0 \geq 0.$

In other words, if the system starts at an equilibrium, it stays there. The converse is also true, as is easily shown. Throughout this chapter it is assumed that $\mathbf{0}$ is an equilibrium of the system (1). If the equilibrium under study is not the origin, one can always redefine the coordinates on $\mathbf{R}^n$ in such a way that the equilibrium of interest becomes the new origin. Thus, without loss of generality, it is assumed that

7 $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}, \forall t \geq 0.$

This is equivalent to the statement

8 $\mathbf{s}(t, t_0, \mathbf{0}) = \mathbf{0}, \forall t \geq t_0.$

Lyapunov theory is concerned with the behavior of the function $\mathbf{s}(t, t_0, \mathbf{x}_0)$ when $\mathbf{x}_0 \neq 0$ but is "close" to it. Occasionally, however, the case where $\mathbf{x}_0$ is "far" from $\mathbf{0}$ is also of interest.

9 **Definition** *The equilibrium* $\mathbf{0}$ *is* **stable** *if, for each* $\varepsilon > 0$ *and each* $t_0 \in \mathbf{R}_+$, *there exists a* $\delta = \delta(\varepsilon, t_0)$ *such that*

10 $\|\mathbf{x}_0\| < \delta(\varepsilon, t_0) \Rightarrow \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| < \varepsilon, \forall t \geq t_0.$

It is **uniformly stable** *if, for each* $\varepsilon > 0$, *there exists a* $\delta = \delta(\varepsilon)$ *such that*

11 $\|\mathbf{x}_0\| < \delta(\varepsilon), t_0 \geq 0 \Rightarrow \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| < \varepsilon, \forall t \geq t_0.$

The equilibrium is **unstable** *if it is not stable.*

According to Definition (9), the equilibrium $\mathbf{0}$ is stable if, given that we do not want the norm $\|\mathbf{x}(t)\|$ of the solution of (1) to exceed a prespecified positive number ε, we are able to determine an *a priori* bound $\delta(t_0, \varepsilon)$ on the norm of the initial condition $\|\mathbf{x}(t_0)\|$ in such a way that any solution trajectory of (1) starting at time t_0 from an initial state inside the ball of radius $\delta(t_0, \varepsilon)$ always stays inside the ball of radius ε at all future times $t \geq t_0$. In other words: arbitrarily small perturbations of the initial state $\mathbf{x}(t_0)$ about the initial state $\mathbf{0}$ result in arbitrarily small perturbations in the corresponding solution trajectories of (1).

It is also possible to interpret stability as a form of continuity of the solution trajectories with respect to the initial conditions. We have seen [Theorem (2.4.57)] that, under reasonable hypotheses such as Lipschitz continuity of $\mathbf{f}$, the solution of (1) is a continuous function of the initial condition. This means that, given any $t_0 \geq 0$ and any *finite* T, the map $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ which takes the initial condition $\mathbf{x}_0$ into the corresponding solution trajectory in $C^n[t_0, T]$ is continuous. This property is true whether or not the equilibrium $\mathbf{0}$ is stable. However, stability requires something more. To state what it is, let $C^n[t_0, \infty)$ denote the linear space of continuous n-vector valued functions on $[t_0, \infty)$, and let $BC^n[t_0, \infty)$ denote the subset of $C^n[t_0, \infty)$ consisting of *bounded* continuous functions. If we define the norm

12 $$\|\mathbf{x}(\cdot)\|_s = \sup_{t \in [t_0, \infty)} \|\mathbf{x}(t)\|,$$

then $BC^n[t_0, \infty)$ is a Banach space. Now stability is equivalent to the following statements:

1) For each $t_0 \geq 0$, there is a number $d(t_0)$ such that $\mathbf{s}(\cdot, t_0, \mathbf{x}_0) \in BC^n[t_0, \infty)$ whenever $\mathbf{x}_0 \in B_{d(t_0)}$, where B_d is the ball

13 $B_d = \{\mathbf{x} \in \mathbf{R}^n : \|\mathbf{x}\| < d\}.$

2) The map $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ which maps an initial condition $\mathbf{x}_0 \in B_{d(t_0)}$ into the corresponding solution trajectory in $BC^n[t_0, \infty)$ is continuous at $\mathbf{x}_0 = 0$ for each $t_0 \geq 0$.

In other words, stability is approximately the same as continuous dependence of the solution on the initial condition over an *infinite* interval.

Another small point needs to be cleared up. In (10), $\|\cdot\|$ is *any* norm on $\mathbf{R}^n$. Because all norms on $\mathbf{R}^n$ are topologically equivalent [see Example (2.1.13)], it follows that the stability status of an equilibrium does not depend on the particular norm used to verify (10).

Once the notion of stability is understood, it is easy to understand what uniform stability means. According to Definition (9), the equilibrium $\mathbf{0}$ is stable if, for each $\varepsilon \geq 0$ and each $t_0 \geq 0$, a corresponding δ can be found such that (10) holds. In general, this δ depends on both ε and t_0. However, if a δ can be found that depends only on ε and not on t_0, then the equilibrium $\mathbf{0}$ is uniformly stable. If the system (1) is autonomous ($\mathbf{f}$ does not depend explicitly on t), then there is no distinction between stability and uniform stability, since changing the initial time merely translates the resulting solution trajectories in time by a like amount. In terms of the map $\mathbf{s}$, uniform stability is roughly equivalent to uniform continuity with respect to t_0. More precisely, uniform stability is equivalent to the following two statements:

1') There is a number $d > 0$ such that $\mathbf{s}(\cdot, t_0, \mathbf{x}_0) \in BC^n[t_0, \infty)$ whenever $\mathbf{x}_0 \in B_d$, $t_0 \in \mathbf{R}_+$.

2') The map $\mathbf{s}(\cdot, t_0, \mathbf{x}_0): B_d \to BC^n[t_0, \infty)$ is uniformly continuous in $\mathbf{x}_0$ at $\mathbf{0}$ with respect to t_0.

14 Example Consider the motion of a simple pendulum. If l is the length of the pendulum, θ is the angle of the pendulum measured from a vertical line, and g is the acceleration due to gravity, then the motion of the pendulum is governed by

$$\mathbf{15} \quad \ddot{\theta} + (g/l) \sin \theta = 0.$$

By introducing the standard state variables $x_1 = \theta$, $x_2 = \dot{\theta}$, (15) becomes

$$\mathbf{16} \quad \dot{x}_1 = x_2, \dot{x}_2 = -\frac{g}{l} \sin x_1.$$

As shown in Example (3.4.49), the trajectories of this system are described by

$$\mathbf{17} \quad \frac{x_2^2}{2} - \frac{g}{l} \cos x_1 = \frac{x_{20}^2}{2} - \frac{g}{l} \cos x_{10} =: a_0.$$

One can now show that $\mathbf{0}$ is a stable equilibrium by verifying the condition of Definition (9) directly. Suppose $\varepsilon > 0$ is given; then it is possible to choose a number $a_0 > 0$ such that the curve described by (17) lies entirely within the ball B_ε. Now choose a $\delta > 0$ such that the ball B_δ lies entirely within this curve (see Figure 5.1 for the construction). Then (10) is satisfied. Since this procedure can be carried out for any $\varepsilon > 0$, $\mathbf{0}$ is a stable equilibrium. ■

As mentioned above, there is no distinction between stability and uniform stability for autonomous systems. The next example illustrates that for nonautonomous systems the two concepts are indeed distinct.

18 Example (Massera 1949) Consider the scalar differential equation

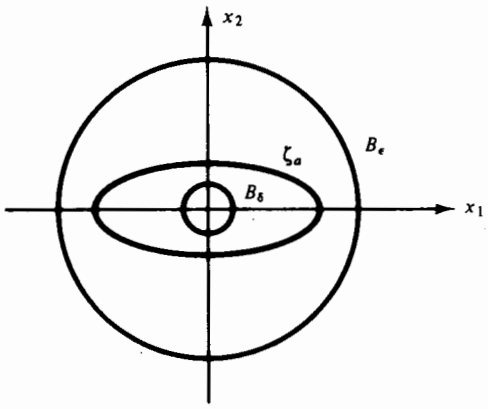


Fig. 5.1

19 $\dot{x}(t) = (6t \sin t - 2t)\, x(t).$

The solution to (19) is given by

20 $x(t) = x(t_0) \exp\{6 \sin t - 6t \cos t - t^2 - 6 \sin t_0 + 6 t_0 \cos t_0 + t_0^2\}.$

To show that the origin is a stable equilibrium, let $t_0 \geq 0$ be any fixed initial time. Then

21 $$\left| \frac{x(t)}{x(t_0)} \right| = \exp\{6 \sin t - 6t \cos t - t^2 - 6 \sin t_0 + 6 t_0 \cos t_0 + t_0^2\}.$$

Now, if $t - t_0 > 6$, then the function on the right side of (21) is bounded above by $\exp[12 + T(6 - T)]$, where $T = t - t_0$. Since this function is continuous in t, it is bounded over $[t_0, T]$ as well. Hence if we define

22 $$c(t_0) = \sup_{t \geq t_0} \exp\{6 \sin t - 6t \cos t - t^2 - 6 \sin t_0 + 6 t_0 \cos t_0 + t_0^2\},$$

then $c(t_0)$ is finite for each fixed t_0. Thus, given any $\varepsilon > 0$, the condition (10) is satisfied if we choose $\delta = \varepsilon / c(t_0)$. This shows that $\mathbf{0}$ is a stable equilibrium. On the other hand, if $t_0 = 2n\pi$, then it follows from (20) that

23 $\mathbf{x}[(2n+1)\pi] = \mathbf{x}(2n\pi) \exp[(4n+1)(6 - \pi)\pi].$

This shows that

24 $c(2n\pi) \geq \exp[(4n+1)(6 - \pi)\pi].$

Hence $c(t_0)$ is unbounded as a function of t_0. Thus, given $\varepsilon > 0$, it is *not* possible to find a single $\delta(\varepsilon)$, independent of t_0, such that (11) holds. Therefore the equilibrium $\mathbf{0}$ is *not* uniformly stable. ■

There is nothing particularly special about the preceding example. Problem 5.1 shows how one may go about constructing a class of systems for which **0** is a stable but not uniformly stable equilibrium.

Finally, let us turn to a discussion of instability. According to Definition (9), instability is merely the absence of stability. It is unfortunate that the term "instability" leads some to visualize a situation where some trajectory of the system "blows up" in the sense that $\|\mathbf{x}(t)\| \to \infty$ as $t \to \infty$. While this is one way in which instability can occur, it is by no means the only way. Stability of the equilibrium **0** means that, given *any* $\varepsilon > 0$, one can find a corresponding $\delta > 0$ such that (10) holds. Therefore, **0** is an unstable equilibrium if, for *some* $\varepsilon > 0$, *no* $\delta > 0$ can be found such that (10) holds; equivalently, there is a ball B_ε such that for *every* $\delta > 0$, no matter how small, there is a nonzero initial state $x(t_0)$ in B_δ such that the corresponding trajectory eventually leaves B_ε. This, and only this, is the definition of instability. It may happen that some trajectories starting in B_δ actually "blow up," but this is not necessary for instability. This is distinction illustrated next.

25 Example Consider the Van der Pol oscillator, described by

$$\dot{x}_1 = x_2, \dot{x}_2 = x_1 + (1 - x_1^2)x_2. \tag{26}$$

The origin is an equilibrium of this system. However, solution trajectories starting from every nonzero initial state, no matter how close to the origin, will eventually approach the limit cycle as shown in Figure 5.2. Now let us study the stability of the equilibrium **0** using Definition (9). By choosing $\varepsilon > 0$ sufficiently small, we can ensure that the ball B_ε is contained entirely within the limit cycle (see Figure 5.2). Therefore all trajectories starting from a nonzero initial state within B_ε will eventually leave B_ε, and so no $\delta > 0$ can be found such that (10) is satisfied. Accordingly, the origin is an unstable equilibrium. Note that all trajectories of the system are bounded, and none blows up. So the system is well-behaved in this sense.

27 Definition *The equilibrium* **0** *is* **attractive** *if, for each* $t_0 \in \mathbf{R}_+$, *there is an* $\eta(t_0) > 0$ *such that*

$$\|\mathbf{x}_0\| < \eta(t_0) \Rightarrow \mathbf{s}(t_0 + t, t_0, \mathbf{x}_0) \to \mathbf{0} \text{ as } t \to \infty. \tag{28}$$

The equilibrium **0** *is* **uniformly attractive** *if there is a number* $\eta > 0$ *such that*

$$\|\mathbf{x}_0\| < \eta, t_0 \geq 0 \Rightarrow \mathbf{s}(t_0 + t, t_0, \mathbf{x}_0) \to \mathbf{0} \text{ as } t \to \infty, \text{ uniformly in } \mathbf{x}_0, t_0. \tag{29}$$

Thus attractivity simply means that, at each initial time $t_0 \in \mathbf{R}_+$, every solution trajectory starting sufficiently close to **0** actually approaches **0** as $t_0 + t \to \infty$. Note that there is no requirement of uniformity at all, in two ways: First, the size of the "ball of attraction" $\eta(t_0)$ can depend on t_0. Second, even for a fixed t_0, the solution trajectories starting inside the ball $B_{\eta(t_0)}$ but at different initial states can approach **0** at different rates. In contrast, uniform attractivity requires first that there be a ball of attraction B_η whose size is independent of t_0, and second that the solution trajectories starting inside B_η all approach **0** at a uniform rate.

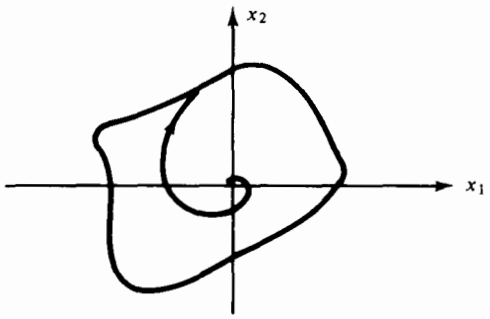


Fig. 5.2

Note that (29) is equivalent to the following statement: For each $\varepsilon > 0$ there exists a $T = T(\varepsilon)$ such that

30 $\|\mathbf{x}_0\| < \eta,\ t_0 \geq 0 \Rightarrow \|\mathbf{s}(t_0 + t,\ t_0,\ \mathbf{x}_0)\| < \varepsilon,\ \forall t \geq T(\varepsilon).$

Note that, in order for an equilibrium to be attractive, a necessary (though not sufficient) condition is that it be *isolated*, i.e., that there exist a neighborhood of the equilibrium that does not contain any other equilibria. This is in contrast to the property of stability, which can apply even to equilibria that are not isolated.

It is possible to define a property called *equi-attractivity* which is intermediate between attractivity and uniform attractivity. This corresponds to η in (30) being allowed to depend on t_0; in other words, the size of the ball of attraction $\eta(t_0)$ may depend on t_0, but all trajectories starting inside $B_{\eta(t_0)}$ must approach **0** at a uniform rate. This concept is not discussed further in this book; the interested reader is referred to Rouche, Habets and Laloy (1977), Section I.2, or to Hahn (1967), Section 36.

31 Definition *The equilibrium* **0** *is* **asymptotically stable** *if it is stable and attractive. It is* **uniformly asymptotically stable (u.a.s.)** *if it is uniformly stable and uniformly attractive.*

At this stage one can ask whether attractivity and stability are really independent properties, i.e., whether an equilibrium can be attractive without being stable. The answer is yes as shown by the following example, due originally to Vinograd (1957) and reproduced in Hahn (1967), Section 40.

32 Example Consider the second order system

33
$$\dot{x}_1 = \frac{x_1^2(x_2 - x_1) + x_2^5}{(x_1^2 + x_2^2)[1 + (x_1^2 + x_2^2)^2]},$$

$$\dot{x}_2 = \frac{x_2^2(x_2 - 2x_1)}{(x_1^2 + x_2^2)[1 + (x_1^2 + x_2^2)^2]}.$$

The right sides of both equations are defined to be 0 at $x_1 = x_2 = 0$. If we introduce polar

coordinates by defining

34 $r = (x_1^2 + x_2^2)^{1/2}, \phi = \text{Atan}(x_1, x_2),$

and denote $\tan \phi$ by u, then the system is described by

35 $$\dot{r} = \frac{r}{(1+r^4)(1+u^2)^2}\,(u^4 - 2u^3 + u - 1 + u^3 r^2 \sin^2\phi).$$

The reader is referred to Hahn (1967), pp. 191 – 194 for a detailed analysis of this system. But the situation can be summarized as shown in Figure 5.3. First, note that the trajectories of the system are symmetric about the origin. Next, in the first quadrant (and of course the third quadrant, by symmetry), there is a curve S such that if the initial state is inside the curve S, then so is the resulting trajectory; if the initial state is outside S, then so is the resulting trajectory. Thus the origin is attractive, since all trajectories approach the origin as $t \to \infty$. However, this does not mean that the origin is *uniformly* attractive, since the closer the initial state is to $\mathbf{0}$, the more slowly the resulting trajectory converges to $\mathbf{0}$. Now the origin is an *unstable* equilibrium. This can be established, as in Example (25), by choosing ε so small that S does not lie inside B_ε. This shows that it is possible for an equilibrium to be attractive yet unstable.

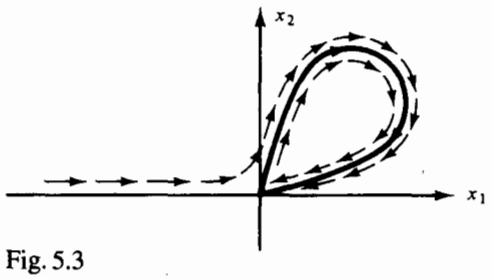


Fig. 5.3

It can be shown [see Theorem (49)] that if the origin of an autonomous system is stable and attractive, then it is also uniformly attractive. To the best of the author's knowledge, it has not been settled whether it is possible for an equilibrium to be uniformly attractive yet unstable.

36 Definition *The equilibrium* $\mathbf{0}$ *is* **exponentially stable** *if there exist constants* $r, a, b > 0$ *such that*

37 $$\|\mathbf{s}(t_0 + t, t_0, \mathbf{x}_0)\| \leq a \|\mathbf{x}_0\| \exp(-bt), \forall t, t_0 \geq 0, \ \forall \mathbf{x}_0 \in B_r.$$

Clearly exponential stability is a stronger property than uniform asymptotic stability.

All of the concepts of stability introduced thus far are *local* in nature, in the sense that they pertain only to the behavior of solution trajectories starting from initial states near the equilibrium. The final definition pertains, in contrast, to the *global* behavior of solution trajectories.

38 Definition *The equilibrium* $\mathbf{0}$ *is* **globally uniformly asymptotically stable (g.u.a.s.)** *if (i) it is uniformly stable, and (ii) for each pair of positive numbers* M, ε *with* M *arbitrarily large and* ε *arbitrarily small, there exists a finite number* $T = T(M, \varepsilon)$ *such that*

$$\textbf{39} \quad \|\mathbf{x}_0\| < M,\ t_0 \geq 0 \Rightarrow \|\mathbf{s}(t_0 + t,\ t_0,\ \mathbf{x}_0)\| < \varepsilon,\ \forall t \geq T(M, \varepsilon).$$

The equilibrium $\mathbf{0}$ *is* **globally exponentially stable (g.e.s.)** *if there exist constants* $a,\ b > 0$ *such that*

$$\textbf{40} \quad \|\mathbf{s}(t_0 + t,\ t_0,\ \mathbf{x}_0)\| \leq a\,\|\mathbf{x}_0\|\exp(-bt),\ \forall t,\ t_0 \geq 0,\ \forall \mathbf{x}_0 \in \mathbf{R}^n.$$

Note that in order for an equilibrium to be either globally uniformly asymptotically stable or globally exponentially stable, a necessary condition is that it be the *only* equilibrium.

Next we discuss the special case of autonomous and periodic systems. The system (1) is *periodic* with period T if

$$\textbf{41} \quad \mathbf{f}(t + T,\ \mathbf{x}) = \mathbf{f}(t,\ \mathbf{x}),\ \forall t \geq 0,\ \forall \mathbf{x} \in \mathbf{R}^n.$$

For a periodic system it is clear that

$$\textbf{42} \quad \mathbf{s}(t + T,\ t_0 + T,\ \mathbf{x}_0) = \mathbf{s}(t,\ t_0,\ \mathbf{x}_0),\ \forall t \geq t_0 \geq 0,,\ \forall \mathbf{x}_0 \in \mathbf{R}^n.$$

If the system (1) is autonomous, i.e., if $\mathbf{f}$ does not depend explicitly on t, then we can think of it as a periodic system with an arbitrary period. Hence all the results presented below for periodic systems apply equally well to autonomous systems.

43 Theorem *Suppose the system (1) is periodic. Then the equilibrium* $\mathbf{0}$ *is uniformly stable if and only if it is stable.*

Remarks If the system (1) is autonomous, the theorem is obvious, since (42) holds for *every* $T > 0$. But the proof for periodic systems requires a bit of work.

Proof Clearly uniform stability implies stability, and only the reverse implication needs to be proved. Suppose $t_0 \in [0,\ T]$ where T is the period, and define

$$\textbf{44} \quad \mu(\mathbf{x}_0,\ t_0) = \sup_{t \geq 0} \|\mathbf{s}(t_0 + t,\ t_0,\ \mathbf{x}_0)\|.$$

Since $\mathbf{0}$ is a stable equilibrium, there is a number $d(t_0)$ such that $\mu(\mathbf{x}_0,\ t_0)$ is finite for all $\mathbf{x}_0 \in B_{d(t_0)}$ [cf. the remark preceding (13)]. In general, it may happen that as t_0 varies over $[0,\ \infty)$, the number $d(t_0)$, though nonzero for each fixed t_0, cannot be bounded away from 0. However, due to the periodicity of the system, one can safely assume that $d(t_0 + T) = d(t_0)$, and it is only necessary to consider the number $d(t_0)$ as t_0 varies over $[0,\ T]$. Thus we can find a number $d > 0$, independent of t_0, such that

45 $\mu(\mathbf{x}_0, t_0) < \infty, \forall \mathbf{x}_0 \in B_d, \forall t_0 \in [0, T]$.

Since $\mu(\mathbf{x}_0, t_0)$ is clearly a continuous function of t_0, the function

46 $$\eta(\mathbf{x}_0) = \sup_{t_0 \in [0,T]} \mu(\mathbf{x}_0, t_0)$$

is finite for all $\mathbf{x}_0 \in B_d$, and is continuous at $\mathbf{x}_0 = 0$.

To show that $\mathbf{0}$ is a uniformly stable equilibrium, suppose $\varepsilon > 0$ is given; we must find $\delta > 0$ such that (11) holds. Due to periodicity, it is only necessary to show that the same δ works for all $t_0 \in [0, T]$ instead of for all $t_0 \geq 0$. Since $\eta(\cdot)$ is continuous, we can find a $\delta > 0$ such that

47 $\|\mathbf{x}_0\| < \delta \Rightarrow \eta(\mathbf{x}_0) < \varepsilon$.

But by the definition of η, (47) is equivalent to

48 $\|\mathbf{x}_0\| < \delta, t_0 \in [0, T] \Rightarrow \|\mathbf{s}(t_0 + t, t_0, \mathbf{x}_0)\| \leq \varepsilon, \forall t \geq 0$.

This completes the proof. ■

49 Theorem *Suppose the system (1) is periodic. Then the equilibrium* $\mathbf{0}$ *is uniformly asymptotically stable if and only if it is asymptotically stable.*

The proof of this theorem is quite involved, even for autonomous systems, in contrast to Theorem (43), which is quite obvious for autonomous systems. The source of difficulty can be explained as follows: Suppose $\mathbf{0}$ is asymptotically stable, and that the system is autonomous. Then, by Definitions (31) and (27), there is a number $r = r(0)$ such that

50 $\|\mathbf{x}_0\| < r \Rightarrow \mathbf{s}(t, 0, \mathbf{x}_0) \to \mathbf{0}$ as $t \to \infty$.

By autonomy, it follows that

51 $\|\mathbf{x}_0\| < r, t_0 \geq 0 \Rightarrow \mathbf{s}(t_0 + t, t_0, \mathbf{x}_0) \to \mathbf{0}$ as $t \to \infty$, uniformly in t_0.

But uniform attractivity requires something more, namely that the convergence in (51) be uniform with respect to $\mathbf{x}_0$ as well. Proving this properly requires some work. As can be imagined, the periodic case is still more complex. For this reason, the proof is omitted, and the reader is referred to Hahn (1967); the relevant theorems are 38.3 and 38.5.

Finally, let us recast the various stability definitions in terms of so-called functions of class K and class L, so as to simplify considerably the proofs of subsequent stability theorems.

52 Definition *A function* $\phi: \mathbf{R}_+ \to \mathbf{R}_+$ *is of* **class K** *if it is continuous, strictly increasing, and* $\phi(0) = 0$*; it is of* **class L** *if it is continuous on* $[0, \infty)$*, strictly decreasing,* $\phi(0) < \infty$*, and* $\phi(r) \to 0$ *as* $r \to \infty$.

53 Theorem *The equilibrium* $\mathbf{0}$ *of the system (1) is stable if and only if, for each* $t_0 \in \mathbf{R}_+$, *there exist a number* $d(t_0) > 0$ *and a function* ϕ_{t_0} *of class K such that*

$$\mathbf{54} \quad \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| \leq \phi_{t_0}(\|\mathbf{x}_0\|), \ \forall t \geq t_0, \ \forall \mathbf{x}_0 \in B_{d(t_0)}.$$

The equilibrium is uniformly stable if and only if there exist a number $d > 0$ *and a function* ϕ *of class K such that*

$$\mathbf{55} \quad \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| \leq \phi(\|\mathbf{x}_0\|), \ \forall t \geq t_0 \geq 0, \ \forall \mathbf{x}_0 \in B_d.$$

Remarks We can thus use (54) and (55) as the *definitions* of stability and uniform stability, if we wish.

Proof First, if (54) holds, then $\mathbf{0}$ is a stable equilibrium. To see this, given any $\varepsilon > 0$ and any $t_0 \in \mathbf{R}_+$, choose

$$\mathbf{56} \quad \delta(\varepsilon, t_0) = \min\{d(t_0), \phi_{t_0}^{-1}(\varepsilon)\}.$$

To prove the converse, fix t_0, and suppose $\varepsilon > 0$ is given. Then by definition there exists a number $\delta > 0$ such that (10) holds. Define $\delta_m(\varepsilon, t_0)$ to be the supremum of all possible choices of δ such that (10) holds. The function $\psi_{t_0}: \varepsilon \mapsto \delta_m(\varepsilon, t_0)$ is nondecreasing, satisfies $\psi_{t_0}(0) = 0$, and $\psi_{t_0}(\varepsilon) > 0 \ \forall \varepsilon > 0$; of course, it need not be continuous nor strictly increasing. However it is possible to find a function θ_{t_0} of class K such that $\theta_{t_0}(\varepsilon) \leq \psi_{t_0}(\varepsilon) \ \forall \varepsilon \geq 0$. [See Lemma (5.2.1).] Now let $\phi_{t_0} = \theta_{t_0}^{-1}$.

The proof of the second assertion regarding uniform stability is entirely similar and is left as an exercise (see Problem 5.3). ■

57 Lemma *The equilibrium* $\mathbf{0}$ *of the system (1) is attractive if and only if, for each* $t_0 \geq 0$, *there exist a number* $r(t_0) > 0$, *and for each* $\mathbf{x}_0 \in B_{r(t_0)}$ *a function* $\sigma_{t_0, \mathbf{x}_0}$ *of class L such that*

$$\mathbf{58} \quad \|\mathbf{s}(t_0 + t, t_0, \mathbf{x}_0)\| \leq \sigma_{t_0, \mathbf{x}_0}(t), \ \forall t \geq 0, \ \forall \mathbf{x}_0 \in B_{r(t_0)}.$$

The equilibrium $\mathbf{0}$ *is uniformly attractive if and only if there exist a number* $r > 0$ *and a function* σ *of class L such that*

$$\mathbf{59} \quad \|\mathbf{s}(t_0 + t, t_0, \mathbf{x}_0)\| \leq \sigma(t), \ \forall t, \ t_0 \geq 0, \ \forall \mathbf{x}_0 \in B_r.$$

Proof If (58) holds then $\mathbf{0}$ is attractive, since $\sigma_{t_0, \mathbf{x}_0}(t) \to 0$ as $t \to \infty$. Conversely, suppose $\mathbf{0}$ is attractive. For each $\varepsilon > 0$, define $T(\varepsilon)$ to be the smallest number T with the property that

$$\mathbf{60} \quad \|\mathbf{s}(t_0 + t, t_0, \mathbf{x}_0)\| < \varepsilon, \ \forall t \geq T.$$

Such a number exists, in view of (28). Define a function $\psi = \psi_{t_0, \mathbf{x}_0}: \varepsilon \mapsto T(\varepsilon)$, and note that $\psi(\varepsilon) = 0$ for ε sufficiently large, say $\psi(\varepsilon) = 0$ for $\varepsilon \geq m$. Now ψ is nonincreasing as a function

of ε since $T(\varepsilon)$ gets larger as ε gets smaller, and $\psi(\varepsilon) \to \infty$ as $\varepsilon \to 0$. Hence it is possible to find a function σ of class L such that $\psi(\varepsilon) \leq \sigma^{-1}(\varepsilon)$. Define this function as $\sigma_{t_0, \mathbf{x}_0}$, and repeat the procedure for each t_0, $\mathbf{x}_0$.

The proof of the second assertion is entirely parallel and is left as an exercise (see Problem 5.4). ■

61 Theorem *The equilibrium* $\mathbf{0}$ *of the system (1) is uniformly asymptotically stable if and only if there exist a number* $r > 0$, *a function* ϕ *of class K, and a function* σ *of class L, such that*

$$\mathbf{62} \quad \|\mathbf{s}(t_0 + t, t_0, \mathbf{x}_0)\| \leq \phi(\|\mathbf{x}_0\|)\,\sigma(t), \forall t, t_0 \geq 0, \forall \mathbf{x}_0 \in B_r.$$

The proof can be found in Hahn (1967), Chap. V. But notice the similarity between (62) and Equation (37) defining exponential stability.

Problem 5.1 The purpose of this problem is to generalize Example (18) by developing an entire class of linear systems with an equilibrium at $t = 0$ which is stable but not uniformly stable. Consider the linear scalar differential equation

$$\dot{x}(t) = a(t)\,x(t), \forall t \geq 0,$$

where $a(\cdot)$ is a continuous function.

(a) Verify that the general solution of this equation is

$$x(t) = x(t_0) \exp\left[\int_{t_0}^{t} a(\tau)\,d\tau\right].$$

(b) Show, using Definition (9), that the equilibrium 0 is stable if and only if

$$\sup_{t \geq t_0} \exp\left[\int_{t_0}^{t} a(\tau)\,d\tau\right] =: m(t_0) < \infty.$$

[Hint: In this case, (10) is satisfied with $\delta(t_0, \varepsilon) = \varepsilon / m(t_0)$.]

(c) Show, using Definition (9), that the equilibrium 0 is uniformly stable if and only if $m(t_0)$ is bounded as a function of t_0.

(d) Construct several functions $a(\cdot)$ with the property that $m(t_0)$ is finite for each finite t_0, but is unbounded as a function of t_0.

Problem 5.2 Construct other systems similar to (33) which have the property that the origin is attractive but not stable.

Problem 5.3 Complete the proof of Theorem (53).

Problem 5.4 Complete the proof of Lemma (57).

5.2 SOME PRELIMINARIES

In this section we present several concepts that are used in the next section to prove the fundamental results of Lyapunov stability theory. These include various types of definiteness, invariance, and the domain of attraction.

Let us begin with a simple but useful result.

1 Lemma *Suppose* $\phi: \mathbf{R}_+ \to \mathbf{R}_+$ *is continuous, that* $\phi(0) = 0$, ϕ *is nondecreasing, and that* $\phi(r) > 0\ \forall r > 0$. *Then there exists a function* α *of class K such that* $\alpha(r) \le \phi(r)\ \forall r$. *Moreover, if* $\phi(r) \to \infty$ *as* $r \to \infty$, *then* α *can be chosen to have the same property.*

Proof Pick a strictly increasing sequence $\{q_i\}$ of positive numbers approaching infinity, and a strictly increasing sequence $\{k_i\}$ of positive numbers approaching 1. Define

$$\textbf{2}\qquad \alpha(r) = \begin{cases} \dfrac{r}{q_1} k_1 \phi(r), & \text{if } 0 \le r \le q_1, \\ k_i \phi(q_i) + \dfrac{r - q_i}{q_{i+1} - q_i} [k_{i+1}\phi(r) - k_i \phi(q_i)], & \text{if } q_i < r \le q_{i+1}. \end{cases}$$

A pictorial interpretation of α is shown in Figure 5.4. ■

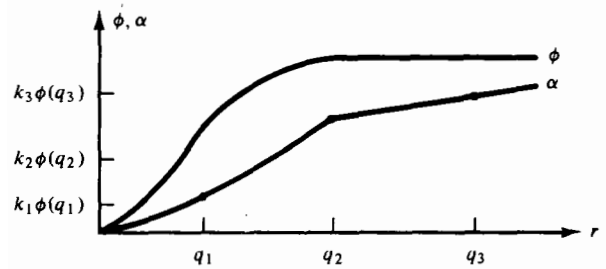


Fig. 5.4

3 Definition *A function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *is said to be a* **locally positive definite function (lpdf)** *if (i) it is continuous, (ii)* $V(t, \mathbf{0}) = 0\ \forall t \ge 0$, *and (iii) there exist a constant* $r > 0$ *and a function* α *of class K such that*

$$\textbf{4}\qquad \alpha(\|\mathbf{x}\|) \le V(t, \mathbf{x}),\ \forall t \ge 0,\ \forall \mathbf{x} \in B_r.$$

V is **decrescent** *if there exist a constant* $r > 0$ *and a function* β *of class K such that*

5 $V(t, \mathbf{x}) \leq \beta(\|\mathbf{x}\|), \forall t \geq 0, \forall \mathbf{x} \in B_r$.

V is a **positive definite function (pdf)** *if (4) holds for all $\mathbf{x} \in \mathbf{R}^n$ (i.e., if $r = \infty$). V is* **radially unbounded** *if (4) is satisfied for all $\mathbf{x} \in \mathbf{R}^n$ and for some continuous function α (not necessarily of class K) with the additional property that $\alpha(r) \to \infty$ as $r \to \infty$. V is a* **locally negative definite function** *if $-V$ is an lpdf, and is a* **negative definite function** *if $-V$ is a pdf.*

Given a continuous function $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$, it is rather difficult to determine whether or not V is a pdf or an lpdf using Definition (3). The main source of difficulty is the need to exhibit the function $\alpha(\cdot)$. Lemmas (6) and (9) give equivalent characterizations of lpdf's and pdf's, and have the advantage that the conditions given therein are more readily verifiable than those in Definition (3).

6 Lemma A continuous function $W: \mathbf{R}^n \to \mathbf{R}$ is an lpdf if and only if it satisfies the following two conditions:

(i) $W(\mathbf{0}) = 0$,

(ii) there exists a constant $r > 0$ such that

$W(\mathbf{x}) > 0, \forall \mathbf{x} \in B_r - \{0\}$.

W is a pdf only if it satisfies the following three conditions:

(iii) $W(\mathbf{0}) = 0$,

(iv) $W(\mathbf{x}) > 0, \forall \mathbf{x} \in \mathbf{R}^n - \{0\}$.

(v) There exists a constant $r > 0$ such that

$$\inf_{\|\mathbf{x}\| \geq r} W(\mathbf{x}) > 0.$$

W is radially unbounded if and only if

(vi) $W(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$, uniformly in $\mathbf{x}$.

Proof Consider first the case of lpdf's. Suppose W is an lpdf in the sense of Definition (3); then clearly (i) and (ii) above hold. To prove the converse, suppose (i) and (ii) above are true, and define

7 $$\phi(p) = \inf_{p \leq \|\mathbf{x}\| \leq r} W(\mathbf{x}).$$

Then $\phi(0) = 0$, ϕ is continuous, and ϕ is nondecreasing because as p increases, the infimum is taken over a smaller region. Further, $\phi(p) > 0$ whenever $p > 0$; to see this, note that the annular region over which the infimum in (7) is taken is compact. Hence, if $\phi(p) = 0$ for some positive p, then there would exist a nonzero $\mathbf{x}$ such that $W(\mathbf{x}) = 0$, which contradicts (ii). Now by Lemma (1), there exists an α of class K such that $\alpha(p) \leq \phi(p)\ \forall p \in [0, r]$. By the definition of ϕ, it now follows that

8 $\alpha(\|\mathbf{x}\|) \leq \phi(\|\mathbf{x}\|) \leq W(\mathbf{x}), \forall \mathbf{x} \in B_r$.

Hence W is an lpdf in the sense of Definition (3).

In the case of pdf's, the necessity of conditions (iii) to (v) is immediate from Definition (3). The remainder of the proof is left as an exercise. ■

Remark Note that conditions (iii) and (iv) alone, without condition (v), are not sufficient for W to be a pdf; consider the function $W: \mathbf{R} \to \mathbf{R}$ defined by $W(x) = x^2/(1+x^4)$.

9 Lemma *A continuous function $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ is an lpdf if and only if (i) $V(t, \mathbf{0}) = 0 \; \forall t$, and (ii) there exists an lpdf $W: \mathbf{R}^n \to \mathbf{R}$ and a constant $r > 0$ such that*

10 $V(t, \mathbf{x}) \geq W(\mathbf{x}), \forall t \geq 0, \forall \mathbf{x} \in B_r$.

V is a pdf if and only if (i) $V(t, \mathbf{0}) = 0 \; \forall t$, and (ii) there exists a pdf $W: \mathbf{R}^n \to \mathbf{R}$ such that

11 $V(t, \mathbf{x}) \geq W(\mathbf{x}), \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n$.

V is radially unbounded if and only if there exists a radially unbounded function $W: \mathbf{R}^n \to \mathbf{R}$ such that (11) is satisfied.

Proof The proof is given only for lpdf's, since the other proofs are entirely similar. Suppose W is an lpdf and that (10) holds; then it is easy to verify that V is an lpdf in the sense of Definition (3). Conversely, suppose V is an lpdf in the sense of Definition (3), and let $\alpha(\cdot)$ be the function of class K such that (4) holds; then $W(\mathbf{x}) = \alpha(\|\mathbf{x}\|)$ is an lpdf such that (10) holds.

The completion of the proof is left as an exercise (see Problem 5.8). ■

Remarks

1. The conditions given in Lemma (6) are easier to verify than those in Definition (3).
2. If $W(\mathbf{x}) = \mathbf{x}'\mathbf{M}\mathbf{x}$, where $\mathbf{M}$ is a real symmetric $n \times n$ matrix, then it is easy to show that W is a positive definite function if and only if $\mathbf{M}$ is a positive definite matrix (see Problem 5.6). Thus the two common usages of the term "positive definite" are consistent.
3. If $W(\mathbf{x})$ is a polynomial in the components of $\mathbf{x}$, then one can systematically check, in a finite number of operations, whether or not W is positive definite; see Bose (1982) for details.
4. Lemma (9) shows that a continuous function of t and $\mathbf{x}$ is an lpdf if and only if it dominates, at each instant of time and over some ball in $\mathbf{R}^n$, an lpdf of $\mathbf{x}$ alone. Similarly, a continuous function of t and $\mathbf{x}$ is a pdf if and only if it dominates, for all t and $\mathbf{x}$, a pdf of $\mathbf{x}$ alone.

5. A function $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ is decrescent if and only if, for each p in some interval $(0, r)$, we have

12 $$\sup_{\|\mathbf{x}\| \le p} \sup_{t \ge 0} V(t, \mathbf{x}) < \infty.$$

13 **Examples** The function

14 $$W_1(x_1, x_2) = x_1^2 + x_2^2$$

is a simple example of a radially unbounded pdf. Clearly $W_1(\mathbf{0}) = 0$ and $W_1(\mathbf{x}) > 0 \ \forall \mathbf{x} \neq \mathbf{0}$. Also $W_1(\mathbf{x}) = \|\mathbf{x}\|^2$ if we take $\|\cdot\|$ to be the Euclidean norm on $\mathbf{R}^n$; hence W_1 is radially unbounded.

The function

15 $$V_1(t, x_1, x_2) = (t+1)(x_1^2 + x_2^2)$$

is a pdf because it dominates the time-invariant pdf W_1. For the same reason, V_1 is radially unbounded. However, it is not decrescent, because for each $\mathbf{x} \neq \mathbf{0}$, the function $V_1(t, \mathbf{x})$ is unbounded as a function of t.

The function

16 $$V_2(t, x_1, x_2) = \exp(-t)(x_1^2 + x_2^2)$$

is not a pdf because no pdf $W: \mathbf{R}^n \to \mathbf{R}$ exists such that (11) holds. This can be seen from the fact that, for each $\mathbf{x}$, $V_2(t, \mathbf{x}) \to 0$ as $t \to \infty$. This example shows that it is *not* possible to weaken the condition (11) to the statement

17 $$V(t, \mathbf{x}) > 0, \forall t \ge 0, \forall \mathbf{x} \neq \mathbf{0}.$$

The present function V_2 is decrescent.

The function

18 $$W_2(x_1, x_2) = x_1^2 + \sin^2 x_2$$

is an lpdf but is not a pdf. Note that $W(\mathbf{0}) = 0$, and that $W(\mathbf{x}) > 0$ whenever $\mathbf{x} \neq \mathbf{0}$ and $|x_2| < \pi$. This is enough to ensure that W_2 is an lpdf. However, W_2 is not a pdf, since it vanishes at points other than $\mathbf{0}$, for example, at $(0, \pi)$.

The function

19 $$W_3(x_1, x_2) = x_1^2 + \tanh^2 x_2$$

is a pdf since $W(\mathbf{0}) = 0$ and $W(\mathbf{x}) > 0 \ \forall \mathbf{x} \neq \mathbf{0}$. However, it is not radially unbounded, since $\tanh^2 x_2 \to 1$ as $|x_2| \to \infty$. ■

Next we introduce the concept of the derivative of a function along the trajectories of a differential equation. Suppose $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ has continuous partial derivatives, and suppose $\mathbf{x}(\cdot)$ satisfies the differential equation

20 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], \forall t \geq 0.$

Then the function $V[t, \mathbf{x}(t)]$ is differentiable with respect to t, and

21 $$\frac{d}{dt} V[t, \mathbf{x}(t)] = \frac{\partial V}{\partial t}[t, \mathbf{x}(t)] + \nabla V[t, \mathbf{x}(t)]\, \mathbf{f}[t, \mathbf{x}(t)].$$

We use the symbol $\dot{V}[t, \mathbf{x}(t)]$ to denote the right-hand side of (21). This choice of symbols is motivated by the fact that

22 $$V[t, \mathbf{x}(t)] = V[t_0, \mathbf{x}(t_0)] + \int_{t_0}^{t} \dot{V}[\tau, \mathbf{x}(\tau)]\, d\tau$$

whenever $\mathbf{x}(\cdot)$ is a solution of (20). This leads to the following definition.

23 Definition *Let $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable with respect to all of its arguments, and let ∇V denote the gradient of V with respect to $\mathbf{x}$ (written as a row vector). Then the function $\dot{V}: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ is defined by*

24 $$\dot{V}(t, \mathbf{x}) = \frac{\partial V}{\partial t}(t, \mathbf{x}) + \nabla V(t, \mathbf{x})\, \mathbf{f}(t, \mathbf{x}),$$

and is called the **derivative of V along the trajectories of (20).**

Remarks

1. Note that $\dot{V}$ depends not only on the function V but also on the system (20). If we keep the same V but change the system (20), the resulting $\dot{V}$ will in general be different.
2. The quantity $\dot{V}(t, \mathbf{x})$ can be interpreted as follows: Suppose a solution trajectory of (20) passes through $\mathbf{x}_0$ at time t_0. Then, at the instant t_0, the rate of change of the quantity $V[t, \mathbf{x}(t)]$ is $\dot{V}(t_0, \mathbf{x}_0)$.
3. Note that if V is independent of t and the system (20) is autonomous, then $\dot{V}$ is also independent of t.

This section concludes with a discussion of invariance and of domains of attraction.

25 Definition *A set $M \subseteq \mathbf{R}^n$ is called an* **invariant** *set of the differential equation (20) if for each $\mathbf{x}_0 \in M$ there exists a $t_0 \in \mathbf{R}_+$ such that*

26 $\mathbf{s}(t, t_0, \mathbf{x}_0) \in M, \forall t \geq t_0.$

In other words, a set is invariant if, for every initial state in the set, a suitable initial time can be found such that the resulting trajectory stays in the set at all future times. Note that, in the dynamical systems literature, one often views a differential equation as being defined for all real t, rather than just all nonnegative t; in such a case, a set M satisfying (26) would be called **positively invariant.**

A few simple examples of invariant sets can be given. First, let $\mathbf{x}_0 \in \mathbf{R}^n$, $t_0 \in \mathbf{R}_+$ be arbitrary, and define $S(t_0, \mathbf{x}_0)$ to be the resulting trajectory viewed as a subset of $\mathbf{R}^n$; in other words, let $S(t_0, \mathbf{x}_0) = \bigcup_{t \geq t_0} \mathbf{s}(t, t_0, \mathbf{x}_0)$. Then $S(t_0, \mathbf{x}_0)$ is invariant (Problem 5.9). An equilibrium is an invariant set; more generally, so is any periodic solution.

27 Definition *Suppose* $\mathbf{x}_0 \in \mathbf{R}^n$, $t_0 \in \mathbf{R}_+$. *Then a point* $\mathbf{p} \in \mathbf{R}^n$ *is called a* **limit point** *of the trajectory* $\mathbf{s}(t, t_0, \mathbf{x}_0)$ *if there exists a sequence* $\{t_i\}$ *of real numbers in* $[t_0, \infty)$ *such that* $t_i \to \infty$ *and*

28 $$\lim_{i \to \infty} \| \mathbf{p} - \mathbf{s}(t_i, t_0, \mathbf{x}_0) \| = 0.$$

The set of all limit points of $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ *is called the* **limit set** *of* $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$, *and is denoted by* $\Omega(t_0, \mathbf{x}_0)$.

An equivalent definition is as follows: $\mathbf{p}$ is a limit point of the trajectory $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ if, given any $\varepsilon > 0$ and $T < \infty$, there exists a $t \geq T$ such that

29 $$\| \mathbf{p} - \mathbf{s}(t, t_0, \mathbf{x}_0) \| < \varepsilon.$$

Again, if one thinks of a trajectory as being defined for all $t \in \mathbf{R}$, then the above set $\Omega(t_0, \mathbf{x}_0)$ would be called the **positive limit set**; the **negative limit set** is obtained by requiring that $t_i \to -\infty$ as $i \to \infty$. Sometimes the negative limit set is called the α-limit set and the positive limit set is called the ω-limit set, on the basis that α and ω are respectively the first and last letters of the Greek alphabet. Considering the ethnicity of the present author, perhaps the negative limit set should be referred to as the $\exists\overline{\exists}$-limit set. Fortunately, this concept is not used in the book.

30 Lemma *Let* $\mathbf{x}_0 \in \mathbf{R}^n$, $t_0 \in \mathbf{R}_+$, *and suppose* $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ *is bounded. Then* $\Omega(t_0, \mathbf{x}_0)$ *is nonempty, closed, and bounded.*

Proof Clearly $\Omega(t_0, \mathbf{x}_0)$ is nonempty and bounded if $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ is bounded. To show that it is closed, let $\{\mathbf{p}_i\}$ be a sequence in $\Omega(t_0, \mathbf{x}_0)$ converging to $\mathbf{p} \in \mathbf{R}^n$; it must be shown that $\mathbf{p} \in \Omega(t_0, \mathbf{x}_0)$. Let $\varepsilon > 0$ and $T < \infty$ be arbitrary; we must then find a $t \geq T$ such that (29) holds. First, choose i such that

31 $$\| \mathbf{p} - \mathbf{p}_i \| < \varepsilon/2.$$

Such an i exists because $\mathbf{p}_i \to \mathbf{p}$. Next, choose $t \geq T$ such that

32 $\| \mathbf{p}_i - \mathbf{s}(t, t_0, \mathbf{x}_0) \| < \varepsilon/2.$

Such a t exists because $\mathbf{p}_i \in \Omega(t_0, \mathbf{x}_0)$. Combining (31) and (32) gives (29). ■

Let us define the distance $d(\mathbf{x}, \Omega)$ between a point $\mathbf{x}$ and a nonempty closed set Ω as

33 $$d(\mathbf{x}, \Omega) = \min_{\mathbf{y} \in \Omega} \| \mathbf{x} - \mathbf{y} \|.$$

Then we have a further result.

34 **Lemma** *Let $\mathbf{x}_0 \in \mathbf{R}^n$, $t_0 \in \mathbf{R}_+$, and suppose $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ is bounded. Then*

35 $d[\mathbf{s}(t, t_0, \mathbf{x}_0), \Omega(t_0, \mathbf{x}_0)] \to 0$ as $t \to \infty$.

Proof If (35) is false, then there exists an $\varepsilon > 0$ and a sequence of times $\{t_i\}$ approaching ∞ such that

36 $d[\mathbf{s}(t_i, t_0, \mathbf{x}_0), \Omega(t_0, \mathbf{x}_0)] \geq \varepsilon, \forall i.$

However, the sequence $\{\mathbf{s}(t_i, t_0, \mathbf{x}_0)\}$ is bounded. Hence it contains a convergent subsequence. By the definition of $\Omega(t_0, \mathbf{x}_0)$, the limit of this convergent subsequence must belong to $\Omega(t_0, \mathbf{x}_0)$, which contradicts (36). Hence (35) is true. ■

The results stated thus far apply to arbitrary systems. The next lemma states a property that is special to periodic (and hence also to autonomous) systems.

37 **Lemma** *Suppose the system (20) is periodic, and let $\mathbf{x}_0 \in \mathbf{R}^n$, $t_0 \in \mathbf{R}_+$. If $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ is bounded, then $\Omega(t_0, \mathbf{x}_0)$ is an invariant set of (20).*

Proof Let T be the period of (20), so that

38 $\mathbf{s}(t, t_0, \mathbf{x}_0) = \mathbf{s}(t + kT, t_0 + kT, \mathbf{x}_0)$, for all integers $k \geq 0$.

Let $\mathbf{p} \in \Omega(t_0, \mathbf{x}_0)$; it must be shown that there exists an initial time $\tau \in \mathbf{R}_+$ such that

39 $\mathbf{s}(t, \tau, \mathbf{p}) \in \Omega(t_0, \mathbf{x}_0), \forall t \geq \tau.$

Since $\mathbf{p} \in \Omega(t_0, \mathbf{x}_0)$, there exists a sequence $\{t_i\}$ approaching infinity such that (28) holds. Now, for each i, find an integer k_i such that $t_i - k_i T \in [0, T)$. Then the sequence $\{t_i - k_i T\}$ is bounded, and therefore contains a convergent subsequence. Choose such a subsequence, renumber it once again as $\{t_i\}$, let $\tau \in [0, T]$ denote its limit, and note that (28) continues to hold. Now, since solutions depend in a continuous fashion on the initial conditions and on the time, we have

40 $$\mathbf{s}(t, \tau, \mathbf{p}) = \lim_{i \to \infty} \mathbf{s}[t, \tau, \mathbf{s}(t_i, t_0, \mathbf{x}_0)]$$

$$= \lim_{i \to \infty} \mathbf{s}[t + k_i T, \tau + k_i T, \mathbf{s}(t_i, t_0, \mathbf{x}_0)], \text{ by (38)}$$

$$= \lim_{i \to \infty} \mathbf{s}[t + k_i T, t_i, \mathbf{s}(t_i, t_0, \mathbf{x}_0)]$$

$$= \lim_{i \to \infty} \mathbf{s}(t + k_i T, t_0, \mathbf{x}_0),$$

where we have used the fact that $\tau = \lim_{i \to \infty} (t_i - k_i T)$. This shows that (39) holds. ■

Now let us restrict attention to autonomous systems of the form

41 $\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t)]$.

Note that, for such systems,

42 $\mathbf{s}(t, \tau, \mathbf{x}_0) = \mathbf{s}(t + \tau, 0, \mathbf{x}_0), \forall t \geq \tau \geq 0, \forall \mathbf{x}_0 \in \mathbf{R}^n$.

Suppose $\mathbf{0}$ is an attractive equilibrium of the system (41). By Definition (5.1.27), this implies that there exists a ball B_r such that every trajectory starting inside B_r approaches $\mathbf{0}$ as $t \to \infty$.

43 Definition *Suppose* $\mathbf{0}$ *is an attractive equilibrium of the system (41). The* **domain of attraction** $D(\mathbf{0})$ *is defined as*

44 $D(\mathbf{0}) = \{\mathbf{x}_0 \in \mathbf{R}^n : \mathbf{s}(t, 0, \mathbf{x}_0) \to \mathbf{0} \text{ as } t \to \infty\}$.

Definition (5.1.27) implies that $\mathbf{0}$ is an interior point of $D(\mathbf{0})$. Note that the terms "region of attraction" and "basin" are also used by some authors instead of "domain of attraction." Also, we identify the region $D(\mathbf{0})$ with the equilibrium $\mathbf{0}$, since (41) may have more than one attractive equilibrium, in which case each equilibrium will have its own domain of attraction.

Lemma (45) below states, among other things, that $D(\mathbf{0})$ is a connected set. As a prelude to this lemma, the notion of connectedness is defined. Let $S \subseteq \mathbf{R}^n$ be a given set. Two points $\mathbf{x}$ and $\mathbf{y}$ are said to be **connected** in S if there is a path between $\mathbf{x}$ and $\mathbf{y}$ lying entirely in S; more precisely, $\mathbf{x}$ and $\mathbf{y}$ are connected in S if there is a continuous function $h : [0, 1] \to S$ such that $h(0) = \mathbf{x}$, $h(1) = \mathbf{y}$. Obviously, the property of being connected in S is symmetric and transitive; thus, if $\mathbf{x}$, $\mathbf{y}$ are connected in S and $\mathbf{y}$, $\mathbf{z}$ are connected in S, then so are $\mathbf{x}$ and $\mathbf{z}$. The entire set S is said to be **connected** if every pair of points in S is connected in S. In $\mathbf{R}^n$, a set S is connected if and only if it cannot be contained in the union of two disjoint open sets.

45 Lemma *Suppose* $\mathbf{0}$ *is an attractive equilibrium of the system (41). Then* $D(\mathbf{0})$ *is open, connected, and invariant.*

Proof To show that $D(\mathbf{0})$ is invariant, suppose $\mathbf{x}_0 \in D(\mathbf{0})$; it is enough to show that

46 $\mathbf{s}(\tau, 0, \mathbf{x}_0) \in D(\mathbf{0}), \forall \tau \geq 0.$

By definition, $\mathbf{x}_0 \in D(\mathbf{0})$ implies that

47 $\mathbf{s}(t, 0, \mathbf{x}_0) \to \mathbf{0}$ as $t \to \infty$.

Now fix $\tau \geq 0$ and note that

48 $\mathbf{s}[t, 0, \mathbf{s}(\tau, 0, \mathbf{x}_0)] = \mathbf{s}(t + \tau, 0, \mathbf{x}_0) \to 0$ as $t \to \infty$, by (47).

Hence (46) follows.

To show that $D(\mathbf{0})$ is open, observe first that $\mathbf{0}$ is an interior point of $D(\mathbf{0})$, and choose $r > 0$ such that B_r is contained in $D(\mathbf{0})$. Now let $\mathbf{x}_0 \in D(\mathbf{0})$ be arbitrary; it must be shown that there exists a ball

49 $B_{\mathbf{x}_0, r} = \{\mathbf{y}_0 \in \mathbf{R}^n : \| \mathbf{x}_0 - \mathbf{y}_0 \| < r\}$

which is contained in $D(\mathbf{0})$. For this purpose, first select a number $T < \infty$ such that

50 $\| \mathbf{s}(T, 0, \mathbf{x}_0) \| < r/2.$

Such a T exists since $\mathbf{s}(t, 0, \mathbf{x}_0) \to \mathbf{0}$. Next, select a $d > 0$ such that

51 $\| \mathbf{s}(t, 0, \mathbf{x}_0) - \mathbf{s}(t, 0, \mathbf{y}_0) \| < r/2, \forall t \in [0, T], \forall \mathbf{y}_0 \in B_{\mathbf{x}_0, d}.$

Such a d exists since solutions of (41) depend continuously on the initial conditions. Now (50) and (51) together imply that

52 $\| \mathbf{s}(T, 0, \mathbf{y}_0) \| < r, \forall \mathbf{y}_0 \in B_{\mathbf{x}_0, d},$

or, equivalently,

53 $\mathbf{s}(T, 0, \mathbf{y}_0) \in B_r, \forall \mathbf{y}_0 \in B_{\mathbf{x}_0, d}.$

But since B_r is contained in $D(\mathbf{0})$, (53) implies that

54 $\lim_{t \to \infty} \mathbf{s}(t, 0, \mathbf{y}_0) = \lim_{t \to \infty} \mathbf{s}[t - T, 0, \mathbf{s}(T, 0, \mathbf{y}_0)] = \mathbf{0}, \forall \mathbf{y}_0 \in B_{\mathbf{x}_0, d}.$

This shows that $B_{\mathbf{x}_0, d}$ is contained in $D(\mathbf{0})$. Hence $D(\mathbf{0})$ is open.

Finally, to show that $D(\mathbf{0})$ is connected, again choose $r > 0$ as in the preceding paragraph, i.e., such that $B_r \subseteq D(\mathbf{0})$. Let $\mathbf{x}_0, \mathbf{y}_0 \in D(\mathbf{0})$ be arbitrary, and select times $T_{\mathbf{x}_0}, T_{\mathbf{y}_0}$ such that

55 $\mathbf{s}(T_{\mathbf{x}_0}, 0, \mathbf{x}_0) \in B_r$, $\mathbf{s}(T_{\mathbf{y}_0}, 0, \mathbf{y}_0) \in B_r$.

By the invariance of $D(\mathbf{0})$, which has already been established, we know that

56 $\mathbf{s}(t, 0, \mathbf{x}_0) \in D(\mathbf{0})$, $\forall t \in [0, T_{\mathbf{x}_0}]$, and $\mathbf{s}(t, 0, \mathbf{y}_0) \in D(\mathbf{0})$, $\forall t \in [0, T_{\mathbf{y}_0}]$.

Hence, $\mathbf{x}_0$ is connected to $\mathbf{s}(T_{\mathbf{x}_0}, 0, \mathbf{x}_0) =: \mathbf{p}_0$ in $D(\mathbf{0})$, and $\mathbf{y}_0$ is connected to $\mathbf{s}(T_{\mathbf{y}_0}, 0, \mathbf{y}_0) =: \mathbf{q}_0$ in $D(\mathbf{0})$. Also, since B_r is convex and contained in $D(\mathbf{0})$, $\mathbf{p}_0$ and $\mathbf{q}_0$ are connected in $D(\mathbf{0})$ (see Figure 5.5). Using the transitivity of connectedness, we can finally conclude that $\mathbf{x}_0$ and $\mathbf{y}_0$ are connected in $D(\mathbf{0})$. ■

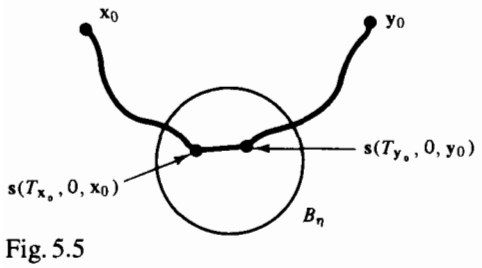


Fig. 5.5

For an extension of Lemma (45) to nonautonomous systems, see Problem 5.10.

Problem 5.5 Determine whether or not each of the following functions is (i) locally positive definite, (ii) positive definite, (iii) decrescent, and (iv) radially unbonded:

(a) $x_1^4 + x_2^4$.

(b) $x_1^2 + x_2^4$.

(c) $(x_1 + x_2^2)^2$.

(d) $t(x_1^2 + x_2^2)$.

(e) $(x_1^2 + x_2^2)/(t+1)$.

(f) $\sin^2(x_1 + x_2) + \sin^2(x_1 - x_2)$.

Problem 5.6 Suppose $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ is defined by

$$V(t, \mathbf{x}) = \mathbf{x}'\mathbf{M}(t)\mathbf{x},$$

where $\mathbf{M}$ is a continuous function of t, and $\mathbf{M}(t)$ is real and symmetric for each t. Determine necessary and sufficient conditions for V to be (i) positive definite, and (ii) decrescent.

Problem 5.7 Complete the proof of Lemma (6).

Problem 5.8 Complete the proof of Lemma (9).

Problem 5.9 Suppose $\mathbf{x}_0 \in \mathbf{R}^n$, $t_0 \in \mathbf{R}_+$, and let $S(t_0, \mathbf{x}_0)$ denote the resulting trajectory viewed as a subset of $\mathbf{R}^n$ [see the paragraph after Definition (25)]. Show that $S(t_0, \mathbf{x}_0)$ is an invariant set.

Problem 5.10 Consider the *nonautonomous* system (20), and suppose the origin is uniformly attractive in the sense of Definition (5.1.27). Define the domain of attraction $D(\mathbf{0})$ as

$$D(\mathbf{0}) = \{\mathbf{x}_0 \in \mathbf{R}^n : \exists\, t_0 \in \mathbf{R}_+ \text{ s.t. } \mathbf{s}(t, t_0, \mathbf{x}_0) \to \mathbf{0} \text{ as } t \to \infty\}.$$

Is Lemma (45) still true? Justify your answer.

5.3 LYAPUNOV'S DIRECT METHOD

The idea behind the various Lyapunov theorems on stability, asymptotic stability, and instability is as follows: Consider a system which is "isolated" in the sense that there are no external forces acting on the system. Equation (5.1.1) is a suitable model for such a system because no input is explicitly identified on the right side of this equation. Suppose that one can identify the various equilibrium states of the system, and that $\mathbf{0}$ is one of the equilibria (possibly the only equilibrium). Now suppose it is possible to define, in some sense, the **total energy** of the system, which is a function having the property that it is zero at the origin and positive everywhere else. (In other words, the energy function has either a global or a local minimum at $\mathbf{0}$.) If the system, which was originally in the equilibrium state $\mathbf{0}$, is perturbed to a new nonzero initial state (where the energy level is positive, by assumption), then there are several possibilities. If the system dynamics are such that the energy of the system is nonincreasing with time, then the energy level of the system never increases beyond the initial positive value. Depending on the nature of the energy function, this may be sufficient to conclude that the equilibrium $\mathbf{0}$ is stable. If the dynamics are such that the energy of the system is monotonically decreasing with time and the energy eventually reduces to zero, this may be sufficient to conclude that the equilibrium $\mathbf{0}$ is asymptotically stable. Finally, if the energy function continues to increase beyond its initial value, then one may be able to conclude that the equilibrium $\mathbf{0}$ is unstable. Such an approach to analyzing the qualitative behavior of mechanical systems was pioneered by Lagrange, who showed that an equilibrium of a conservative mechanical system is stable if it corresponds to a local minimum of the potential energy function, and that it is unstable if it corresponds to a local maximum of the potential energy function. The genius of Lyapunov lay in his ability to extract from this type of reasoning a general theory that is applicable to *any* differential equation. This theory requires one to search for a function which satisfies some prespecified properties. This function is now commonly known as a Lyapunov function, and is a generalization of the energy of a mechanical system. Subsequent researchers have of course refined the theory considerably.

This section deals with the so-called direct method of Lyapunov, sometimes also called the second method. The first, or indirect, method is based on power series expansions and does not find much favor today. Three basic types of theorems are presented in this section, namely: stability theorems, asymptotic stability theorems, and instability theorems. The various theorems are illustrated by several examples.

Throughout this section, the following three abbreviations are employed to make the theorem statements more compact:

C^1: continuously differentiable

lpdf: locally positive definite function

pdf: positive definite function

5.3.1 Theorems on Stability

Theorem (1) is the basic stability theorem of Lyapunov's direct method.

1 Theorem *The equilibrium* $\mathbf{0}$ *of the system (5.1.1) is stable if there exist a* C^1 *lpdf* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *and a constant* $r > 0$ *such that*

$$\mathbf{2} \qquad \dot{V}(t, \mathbf{x}) \leq 0, \ \forall t \geq t_0, \ \forall \mathbf{x} \in B_r,$$

where $\dot{V}$ *is evaluated along the trajectories of (5.1.1).*

Proof Since V is an lpdf, there exist a function α of class K and a constant $s > 0$ such that

$$\mathbf{3} \qquad \alpha(\|\mathbf{x}\|) \leq V(t, \mathbf{x}), \ \forall t \geq 0, \ \forall \mathbf{x} \in B_s.$$

To show that $\mathbf{0}$ is a stable equilibrium, we must show that given any $\varepsilon > 0$ and any $t_0 \geq 0$, we can find a $\delta = \delta(\varepsilon, t_0)$ such that (5.1.10) is satisfied. Accordingly, given ε and t_0, let $\varepsilon_1 = \min\{\varepsilon, r, s\}$, and pick $\delta > 0$ such that

$$\mathbf{4} \qquad \sup_{\|\mathbf{x}\| \leq \delta} V(t_0, \mathbf{x}) =: \beta(t_0, \delta) < \alpha(\varepsilon_1).$$

Such a δ can always be found because $\alpha(\varepsilon_1) > 0$ and $\beta(t_0, \delta) \to 0$ as $\delta \to 0$. To show that the above choice of δ satisfies (5.1.10), suppose $\|\mathbf{x}_0\| < \delta$. Then $V(t_0, \mathbf{x}_0) \leq \beta(t_0, \delta) < \alpha(\varepsilon_1)$. But since $\dot{V}(t, \mathbf{x}) \leq 0$ whenever $\|\mathbf{x}\| < \delta$ (note that $\delta \leq \varepsilon_1 \leq r$), it follows that

$$\mathbf{5} \qquad V[t, \mathbf{s}(t, t_0, \mathbf{x}_0)] \leq V(t_0, \mathbf{x}_0) < \alpha(\varepsilon_1), \ \forall t \geq t_0.$$

Now, since

$$\mathbf{6} \qquad V[t, \mathbf{s}(t, t_0, \mathbf{x}_0)] \geq \alpha[\|\mathbf{s}(t, t_0, \mathbf{x}_0)\|],$$

(5) and (6) together imply that

$$\mathbf{7} \qquad \alpha[\|\mathbf{s}(t, t_0, \mathbf{x}_0)\|] < \alpha(\varepsilon_1), \ \forall t \geq t_0.$$

Since $\alpha(\cdot)$ is strictly increasing, (7) in turn implies that

8 $\|\mathbf{s}(t, t_0, \mathbf{x}_0)\| < \varepsilon_1 \leq \varepsilon, \forall t \geq t_0.$

Hence (5.1.10) is satisfied, and $\mathbf{0}$ is a stable equilibrium. ■

Remarks Strictly speaking, the preceding proof is circular and therefore incorrect. The circularity in the argument comes in (5), where it is blandly asserted that $V[t, \mathbf{s}(t, t_0, \mathbf{x}_0)] \leq V(t_0, \mathbf{x}_0)$, because $\dot{V}(t, \mathbf{x}) \leq 0\ \forall t \geq 0$ and $\forall \mathbf{x} \in B_r$. But this reasoning presupposes that the trajectory $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ stays inside B_r for all $t \geq t_0$, which is one of the things that we are trying to prove! To get around this circular argument, it is possible to reason as follows: Define δ by (4), and suppose by way of contradiction that (8) is violated. Let T be the smallest time t at which $\|\mathbf{s}(t, t_0, \mathbf{x}_0)\| \geq \varepsilon_1$. This quantity is well-defined, since $\mathbf{s}(\cdot, t_0, \mathbf{x}_0)$ is a continuous function. Now, by definition, we have

9 $\|\mathbf{s}(t, t_0, \mathbf{x}_0)\| < \varepsilon_1, \forall t \in [t_0, T),$

10 $\|\mathbf{s}(T, t_0, \mathbf{x}_0)\| = \varepsilon_1.$

But now, since $\varepsilon_1 \leq r$, it follows from (2) and (9) that

11 $$\frac{d}{dt} V[t, \mathbf{s}(t, t_0, \mathbf{x}_0)] = \dot{V}[t, \mathbf{s}(t, t_0, \mathbf{x}_0)] \leq 0, \forall t \in [t_0, T).$$

Hence, from (4),

12 $V[T, \mathbf{s}(T, t_0, \mathbf{x}_0)] \leq V(t_0, \mathbf{x}_0) < \alpha(\varepsilon_1).$

But, from (10) and (3), and again noting that $\varepsilon_1 \leq s$, we have

13 $V[T, \mathbf{s}(T, t_0, \mathbf{x}_0)] \geq \alpha[\|\mathbf{s}(T, t_0, \mathbf{x}_0)\|] = \alpha(\varepsilon_1).$

Clearly (12) and (13) are in contradiction. This shows that the original assumption is false; no such T can exist, and (8) is true.

Note that the rigorous argument given here is in some ways much less intuitive than the circular "proof" given earlier. For this reason, we shall give the same type of circular "proofs" for all the other theorems in this section, since they generally bring out the logic behind the theorem better than a more rigorous proof would. The reader is assured, however, that all the proofs given in this section can be fixed up in the above manner, and is invited to do so.

A simple modification of the hypotheses of Theorem (1) leads to a criterion for uniform stability.

14 Theorem *The equilibrium* $\mathbf{0}$ *of the system (5.1.1) is uniformly stable if there exist a* C^1, *decrescent, lpdf* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *and a constant* $r > 0$ *such that*

15 $\dot{V}(t, \mathbf{x}) \leq 0, \forall t \geq 0, \forall \mathbf{x} \in B_r$.

Proof Since V is decrescent, the function

16 $$\beta(\delta) := \sup_{\|\mathbf{x}\| \leq \delta} \sup_{t \geq 0} V(t, \mathbf{x})$$

is finite for all sufficiently small δ, and is nondecreasing in δ. Now let $\varepsilon_1 = \min\{\varepsilon, r, s\}$ and pick $\delta > 0$ such that

17 $\beta(\delta) < \alpha(\varepsilon_1)$.

Now proceed exactly as in the proof of Theorem (1) to show that (5.1.11) holds with this choice of δ. The details are left as an exercise. ■

Remarks

1. Theorem (1) states that if we can find a C^1 lpdf V such that its derivative along the trajectories of (5.1.1) is always nonpositive, then the equilibrium **0** is stable. Theorem (14) shows that in order to conclude the *uniform* stability of the equilibrium **0**, it is enough to add the assumption that V is also decrescent. It should be noted that Theorems (1) and (14) provide only sufficient conditions for stability and uniform stability respectively. [But the converses of these theorems are also true; see Hahn (1967).]

2. The definitions of stability given in Section 5.1 are qualitative, in the sense that given an $\varepsilon > 0$, one is only required to find *some* $\delta > 0$ satisfying (5.1.10); to put it another way, one is only required to demonstrate the *existence* of a suitable δ. In the same way, Theorems (1) and (14) are also qualitative in the sense that they provide conditions under which the existence of a suitable δ can be concluded. However, in principle at least, the conditions (4) if Theorem (1) is being applied, or (17) if Theorem (14) is being applied, can be used to determine a suitable δ. But in practice this procedure is rather messy, and often gives too conservative an estimate for δ.

3. The function V is commonly known as a *Lyapunov function* or a *Lyapunov function candidate*. The term *Lyapunov function* is a source of great confusion. In an attempt to avoid confusion, the following convention is adopted: Suppose, for example, that we are attempting to show that **0** is a stable equilibrium by applying Theorem (1). Then a function V is referred to as a **Lyapunov function candidate** if it satisfies the requirements imposed on V in the hypotheses of Theorem (1), i.e., if V is C^1 and is an lpdf. If, for a particular system (5.1.1), the conditions imposed on $\dot{V}$ are also satisfied, then V is referred to as a **Lyapunov function.** The rationale behind this convention is as follows: Theorems (1) and (14) are sufficient conditions for certain stability properties. To apply them to a particular system, it is a fairly simple matter to find a function V satisfying the requirements on V. At this stage, V is a *Lyapunov function candidate.* Now, for the particular

system under study and for the particular choice of V, the conditions on $\dot{V}$ may or may not be met. If the requirements on $\dot{V}$ are also met, then definite conclusions can be drawn, and V then becomes a *Lyapunov function.* On the other hand, if the requirements on $\dot{V}$ are not met, since these theorems are only sufficient conditions, no definite conclusions can be drawn, and one has to start again with another Lyapunov function candidate. The examples that follow illustrate this usage.

18 Example Consider again the simple pendulum, which is described by

$$\ddot{\theta} + \sin\theta = 0$$

after suitable normalization. The state variable representation of this system is

$$\dot{x}_1 = x_2, \dot{x}_2 = -\sin x_1.$$

Now the total energy of the pendulum is the sum of the potential and kinetic energies, which is

$$V(x_1, x_2) = (1 - \cos x_1) + \frac{1}{2}x_2^2,$$

where the first term represents the potential energy and the second term represents the kinetic energy. One can readily verify that V is C^1 and an lpdf, so that V is a suitable Lyapunov function candidate for applying Theorem (1). Computing $\dot{V}$ gives

$$\dot{V}(x_1, x_2) = \sin x_1 \dot{x}_1 + x_2 \dot{x}_2 = \sin x_1 (x_2) + x_2(-\sin x_1) = 0.$$

Therefore $\dot{V}$ also satisfies the requirements of Theorem (1). Hence V is actually a Lyapunov function, and the equilibrium $\mathbf{0}$ is stable by Theorem (1). Further, because the system is autonomous, $\mathbf{0}$ is a uniformly stable equilibrium; see Theorem (5.1.43).

19 Example Consider the rotational motion of a rigid body in three-dimensional space. If $\boldsymbol{\omega}$ denotes the angular velocity of the body and $\mathbf{I}$ the 3×3 inertia matrix of the body (both measured in a coordinate frame rigidly attached to the body), then in the absence of external torques the motion is described by

$$\mathbf{I}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega} \times \mathbf{I}\boldsymbol{\omega} = 0, \tag{20}$$

where $\times$ denotes the vector cross product. Equation (20) can be simplified considerably if the coordinate axes are chosen to be the principal axes of the body, i.e., a set of axes with respect to which $\mathbf{I}$ is a diagonal matrix. Accordingly, let

$$\boldsymbol{\omega} = [\omega_x \ \omega_y \ \omega_z]', \ \mathbf{I} = \text{Diag}\{I_x, I_y, I_z\}.$$

Then (20) reduces to

21 $I_x \dot{\omega}_x = -(I_z - I_y)\,\omega_y\,\omega_z,$

$I_y \dot{\omega}_y = -(I_x - I_z)\,\omega_x\,\omega_z,$

$I_z \dot{\omega}_z = -(I_y - I_x)\,\omega_x\,\omega_y.$

Again, suppose without loss of generality that $I_x \geq I_y \geq I_z > 0$. For notational simplicity, let us replace ω_x, ω_y, ω_z by x, y, z, respectively, and define

$$a = \frac{I_y - I_z}{I_x},\ b = \frac{I_x - I_z}{I_y},\ c = \frac{I_x - I_y}{I_z}.$$

Note that $a, b, c \geq 0$. Then (21) finally assumes the form

22 $\dot{x} = ayz,\ \dot{y} = -bxz,\ \dot{z} = cxy.$

At this stage, assume for simplicity that the principal axes are unique; this is equivalent to assuming that $I_x > I_y > I_z$, or that $a, b, c > 0$. This assumption excludes bodies with some symmetry, such as a spinning top for example. Then the system (22) is in equilibrium if and only if *at least two* of the quantities x, y, z are equal to zero. Hence the set of equilibria consists of the union of the x, y, and the z axes. Physically this corresponds to rotation around one of the principal axes at a constant angular velocity. Note that none of the equilibria is isolated.

Consider first the equilibrium at the origin, and try the obvious Lyapunov function candidate

$$V(x, y, z) = px^2 + qy^2 + rz^2,$$

where $p, q, r > 0$. Then V is an lpdf. (Actually, V is a pdf and is radially unbounded, but this fact is not needed.) Computing $\dot{V}$ gives

$$\dot{V} = 2(px\dot{x} + qy\dot{y} + rz\dot{z}) = 2xyz\,(ap - bq + cr).$$

Clearly it is possible to choose $p, q, r > 0$ such that

$$ap - bq + cr = 0.$$

For such a choice, $\dot{V} \equiv 0$, which satisfies the hypotheses of Theorem (1). Hence $\mathbf{0}$ is a (uniformly) stable equilibrium.

Next, consider an equilibrium of the form $(x_0, 0, 0)$ where $x_0 \neq 0$. At this stage we can do one of two things: (i) We can translate the coordinates such that $(x_0, 0, 0)$ becomes the origin of the new coordinate system. This would enable us to apply Theorem (1) directly, but would have the effect of making the system equations (22) more complicated. (ii) Alternatively, we can construct a Lyapunov function candidate V such that V is C^1, $V(x_0, 0, 0) = 0$, and $V(x, y, z) > 0$ for all $(x, y, z) \neq (x_0, 0, 0)$ and sufficiently near $(x_0, 0, 0)$.

For the sake of variety, the second approach is followed here; the first approach is discussed further in Problem 5.11. For the new equilibrium, let us try the Lyapunov function candidate

$$W(x, y, z) = cy^2 + bz^2 + [2acy^2 + abz^2 + bc(x^2 - x_0^2)]^2.$$

Then $W(x_0, 0, 0) = 0$, and

$$W(x, y, z) > 0 \text{ if } (x, y, z) \neq (\pm x_0, 0, 0).$$

Hence W is an lpdf with respect to the equilibrium $(x_0, 0, 0)$, but is not a pdf since it vanishes at another point as well. Now routine computations show that $\dot{W} \equiv 0$. Hence $(x_0, 0, 0)$ is a stable equilibrium.

By entirely analogous reasoning, one can show that *every* equilibrium of the form $(0, 0, z_0)$ is also stable (Problem 5.11). However, it turns out that *every* equilibrium of the form $(0, y_0, 0)$ with $y_0 \neq 0$ is unstable. This is shown later in this section [see Example (105)]. Physically, this means that an object can be made to spin about its major axis and its minor axis, but not about its "intermediate" axis.

23 Example Consider the system described by

$$\ddot{y}(t) + \dot{y}(t) + (2 + \sin t)\, y(t) = 0,$$

or, in state variable form,

$$\dot{x}_1 = x_2, \dot{x}_2 = -x_2 - (2 + \sin t)\, x_1.$$

Note that the system is linear and periodic. This is an example of a so-called damped Mathieu equation. (If the $\dot{y}$ term were not there it would be an undamped Mathieu equation.) In this case there is no physical intuition readily available to guide us in the choice of V. Thus (after possibly a great deal of trial and error), we might be led to try the Lyapunov function candidate

$$V(t, x_1, x_2) = x_1^2 + \frac{x_2^2}{2 + \sin t}.$$

Note that V is periodic with the same period as the system. Now V is C^1, it dominates the time-invariant pdf

$$W_1(x_1, x_2) = x_1^2 + x_2^2/3,$$

and is dominated by the time-invariant function

$$W_2(x_1, x_2) = x_1^2 + x_2^2.$$

Hence, by Lemma (5.2.9), V is a pdf and decrescent, and is thus a suitable Lyapunov

function candidate for applying Theorem (14). Now

$$\begin{aligned}\dot{V}(t, x_1, x_2) &= -x_2^2 \frac{\cos t}{(2+\sin t)^2} + 2x_1\dot{x}_1 + \frac{2x_2}{2+\sin t}\dot{x}_2 \\ &= -x_2^2 \frac{\cos t}{(2+\sin t)^2} + 2x_1 x_2 + \frac{2x_2}{2+\sin t}[-x_2 - 2(2+\sin t)x_1] \\ &= -\frac{\cos t + 2(2+\sin t)}{(2+\sin t)^2} x_2^2 \\ &= -\frac{4+2\sin t+\cos t}{(2+\sin t)^2} x_2^2 \\ &\leq 0, \ \forall t \geq 0, \ \forall x_1, x_2.\end{aligned}$$

Thus the requirements on $\dot{V}$ in Theorem (14) are also met. Hence V is a Lyapunov function for this system, and $\mathbf{0}$ is a uniformly stable equilibrium.

24 Example One of the main applications of Lyapunov theory is in obtaining stability conditions involving the parameters of the system under study. As an illustration, consider the system

$$\ddot{y}(t) + p(t)\dot{y}(t) + e^{-t}y(t) = 0,$$

which can be represented in state variable form as

$$\dot{x}_1 = x_2, \dot{x}_2 = -p(t)x_2 - e^{-t}x_1.$$

The objective is to find some conditions on the function $p(\cdot)$ that ensure the stability of the equilibrium $\mathbf{0}$. For this purpose, let us choose

$$V(t, x_1, x_2) = x_1^2 + e^t x_2^2.$$

Since V is C^1 and dominates the pdf

$$W(x_1, x_2) = x_1^2 + x_2^2,$$

V is a suitable Lyapunov function candidate for applying Theorem (1). However, since V is not decrescent, it is *not* a suitable Lyapunov function candidate for applying Theorem (14). Hence, using this particular V-function, we cannot hope to establish uniform stability by applying Theorem (14).

Differentiation of V gives

$$\dot{V}(t, x_1, x_2) = e^t x_2^2 + 2x_1(x_2) + 2e^t x_2[-p(t)x_2 - e^{-t}x_1]$$

$$= e^t x_2^2[-2p(t) + 1].$$

Hence we see that $\dot{V}$ is always nonpositive provided

$$p(t) \geq \frac{1}{2}, \forall t \geq 0.$$

Thus the equilibrium **0** is stable provided the above condition holds.

It should be emphasized that by employing a different Lyapunov function candidate, we might be able to obtain entirely different stability conditions involving $p(\cdot)$.

5.3.2 Theorems on Asymptotic Stability

In this section we present some theorems that give sufficient conditions for uniform asymptotic stability, exponential stability, and global versions of the same.

25 Theorem *The equilibrium* **0** *of (5.1.1) is uniformly asymptotically stable if there exists a* C^1 *decrescent lpdf* V *such that* $-\dot{V}$ *is an lpdf.*

Proof If $-\dot{V}$ is an lpdf, then clearly $\dot{V}$ satisfies the hypothesis of Theorem (14), so that **0** is a uniformly stable equilibrium. Thus, according to Definition (5.1.31), it is only necessary to prove that **0** is uniformly attractive. Precisely, it is necessary to show the existence of a $\delta_1 > 0$ such that, for each $\varepsilon > 0$ there exists a $T(\varepsilon) < \infty$ such that

26 $\|\mathbf{x}_0\| < \delta_1, t_0 \geq 0 \Rightarrow \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| < \varepsilon, \forall t \geq T(\varepsilon).$

The hypotheses on V and $\dot{V}$ imply that there are functions $\alpha(\cdot), \beta(\cdot), \gamma(\cdot)$ of class K and a constant $r > 0$ such that

27 $\alpha(\|\mathbf{x}\|) \leq V(t, \mathbf{x}) \leq \beta(\|\mathbf{x}\|), \forall t \geq t_0, \forall \mathbf{x} \in B_r,$

28 $\dot{V}(t, \mathbf{x}) \leq -\gamma(\|\mathbf{x}\|), \forall t \geq t_0, \forall \mathbf{x} \in B_r.$

Now, given $\varepsilon > 0$, define positive constants δ_1, δ_2, and T by

29 $\delta_1 < \beta^{-1}[\alpha(r)],$

30 $\delta_2 < \min\{\beta^{-1}[\alpha(\varepsilon)], \delta_1\},$

31 $T = \dfrac{\beta(\delta_1)}{\gamma(\delta_2)}.$

We now show that these are the required constants. First, one can show, following the reasoning of Equations (9) – (13), that every trajectory that starts inside the ball B_δ stays inside the ball B_r; hence the inequalities (27) and (28) apply over the course of the trajectory. Next,

it is shown that

32 $\|\mathbf{x}_0\| < \delta_1 \Rightarrow \|\mathbf{s}(t_1, t_0, \mathbf{x}_0)\| < \delta_2$ for some $t_1 \in [t_0, t_0 + T]$.

To prove (32), suppose by way of contradiction that (32) is false, so that

33 $\|\mathbf{s}(t, t_0, \mathbf{x}_0)\| \geq \delta_2, \forall t \in [t_0, t_0 + T]$.

Then

34 $0 < \alpha(\delta_2) \leq V[t_0 + T, \mathbf{s}(t_0 + T, t_0, \mathbf{x}_0)]$ by (27)

$$= V(t_0, \mathbf{x}_0) + \int_{t_0}^{t_0+T} \dot{V}[\tau, \mathbf{s}(\tau, t_0, \mathbf{x}_0)]\, d\tau$$

$\leq \beta(\delta_1) - T\gamma(\delta_2)$ by (27), (28), and (33)

$= 0$ by (31).

This contradiction shows that (32) is true.

To complete the proof, suppose $t \geq t_0 + T$. Then, with t_1 defined in (32), we have

35 $\alpha[\|\mathbf{s}(t, t_0, \mathbf{x}_0)\|] \leq V[t, \mathbf{s}(t, t_0, \mathbf{x}_0)]$ by (27) $\leq V[t_1, \mathbf{s}(t_1, t_0, \mathbf{x}_0)]$

by the nonpositivity of $\dot{V}$. Finally,

36 $V[t_1, \mathbf{s}(t_1, t_0, \mathbf{x}_0)] \leq \beta[\|\mathbf{s}(t_1, t_0, \mathbf{x}_0)\|] \leq \beta(\delta_2)$ by (32).

Now (35) and (36) together imply that

37 $\alpha[\|\mathbf{s}(t, t_0, \mathbf{x}_0)\|] \leq \beta(\delta_2) < \alpha(\varepsilon)$.

The inequality (37) establishes (26). ■

In the above theorem, it is worth noting that only V is required to be decrescent, and that $-\dot{V}$ need not be decrescent.

Theorem (25) not only gives a sufficient condition for asymptotic stability, but also provides a way of estimating the domain of attraction. Suppose the system under study is autonomous, and that the Lyapunov function candidate V is independent of time; then $\dot{V}$ is also independent of time. Now suppose we have succeeded in finding a domain (i.e., an open connected set) S in $\mathbf{R}^n$ containing $\mathbf{0}$ with the property that

38 $V(\mathbf{x}) > 0$, $\dot{V}(\mathbf{x}) < 0$, $\forall \mathbf{x} \neq \mathbf{0}$ in S.

Then Theorem (25) applies and $\mathbf{0}$ is an asymptotically stable equilibrium. Now one can ask: Does (38) imply that S is contained in the domain of attraction $D(\mathbf{0})$ defined in (5.2.43)? In other words, does (38) imply that $\mathbf{s}(t, 0, \mathbf{x}_0) \to \mathbf{0}$ as $t \to \infty$ whenever $\mathbf{x}_0 \in S$? One might be tempted to think that the answer is yes because (38) implies that $V[\mathbf{s}(t, 0, \mathbf{x})]$ gradually decays to 0 as $t \to \infty$, but this reasoning is false. If $\mathbf{x}_0 \in S$ *and if* $\mathbf{s}(t, 0, \mathbf{x}_0) \in S$ $\forall t \geq 0$, then (38) would imply that $V[\mathbf{s}(t, 0, \mathbf{x}_0)] \to 0$, and hence that $\mathbf{s}(t, 0, \mathbf{x}_0) \to \mathbf{0}$. However, (38) alone does not guarantee that every solution that starts in S stays in S. The valid conclusion is this: If (38) holds, then *every invariant set* of (5.1.1) contained in S is also contained in $D(\mathbf{0})$, but S itself need not be contained in $D(\mathbf{0})$. But how does one go about finding such invariant sets? As easy way is to use so-called level sets of the Lyapunov function V. Let $c \in \mathbf{R}_+$, and consider the set

39 $M_V(c) = \{\mathbf{x} \in \mathbf{R}^n : V(\mathbf{x}) \leq c\}$.

Note that, depending on the nature of the function V, the set $M_V(c)$ need not be connected (see Figure 5.6). However, $\mathbf{0}$ always belongs to $M_V(c)$. Now the **level set** $L_V(c)$ is defined as the *connected component of* $M_V(c)$ containing $\mathbf{0}$. Another equivalent definition is the following: $L_V(c)$ is the set of all $\mathbf{x} \in \mathbf{R}^n$ with the property that there exists a continuous function $\mathbf{h}: [0, 1] \to \mathbf{R}^n$ such that $\mathbf{h}(0) = \mathbf{x}$, $\mathbf{h}(1) = \mathbf{0}$, and $V[\mathbf{h}(r)] \leq c$, $\forall r \in [0, 1]$. This definition is illustrated in Figure 5.6.

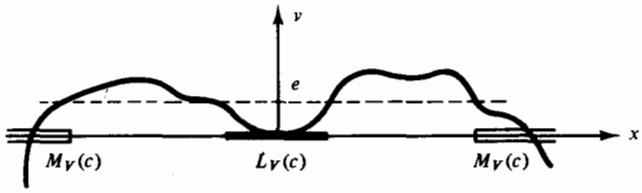


Fig. 5.6

The usefulness of level sets is brought out next.

40 Lemma *Consider the autonomous system*

41 $\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t)]$.

Suppose there exist a C^1 *function* $V: \mathbf{R}^n \to \mathbf{R}$ *and a domain* S *containing* $\mathbf{0}$ *such that* $V(\mathbf{0}) = 0$ *and (38) is satisfied. Let* c *be any positive constant such that the level set* $L_V(c)$ *is contained in* S *and is bounded. Then* $L_V(c)$ *is a subset of* $D(\mathbf{0})$.

Proof First it is shown that the level set $L_V(c)$ is invariant for the system (41). Suppose $\mathbf{x}_0 \in L_V(c)$. Then

42 $V[\mathbf{s}(t, 0, \mathbf{x}_0)] \leq V(\mathbf{x}_0)$ since $\dot{V}(\mathbf{x}) \leq 0 \; \forall \mathbf{x} \in L_V(c) \subseteq S$,

$\leq c$,

which implies that $\mathbf{s}(t, 0, \mathbf{x}_0) \in L_V(c) \; \forall t \geq 0$. Since $L_V(c)$ is bounded, the solution trajectory does not escape to infinity. Next, since $-\dot{V}$ is an lpdf in $L_V(c)$, one can now proceed as in the proof of Theorem (25) to show that $\mathbf{s}(t, 0, \mathbf{x}_0) \to \mathbf{0}$ as $t \to \infty$. ■

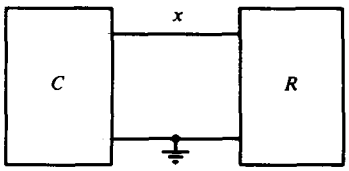


Fig. 5.7

43 Example As an illustration of Theorem (25) and Lemma (40), consider the nonlinear *RC* network shown in Figure 5.7, where R is a nonlinear resistive network, terminated in a bank of n linear capacitors. Denote these capacitances by $C_1, \cdots, C_n$, and assume that all capacitances are positive. Let x_i denote the voltage across the i-th capacitor, and let $\mathbf{x}$ denote the vector $[x_1 \cdots x_n]'$. Then the current vector through the capacitors is just $\mathbf{C}\dot{\mathbf{x}}$, where

$$\mathbf{C} = \text{Diag}\,\{C_1, \cdots, C_n\}.$$

Let $\mathbf{i}(\mathbf{x})$ denote the current vector that results when a voltage vector $\mathbf{x}$ is applied across the terminals of the resistive network. Suppose the network is unbiased, in the sense that $\mathbf{i}(\mathbf{0}) = \mathbf{0}$. If $\mathbf{i}$ is a C^1 function of $\mathbf{x}$, then there exists a continuous function $\mathbf{G}: \mathbf{R}^n \to \mathbf{R}^{n \times n}$ such that [see Lemma (2.5.17)]

$$\mathbf{i}(\mathbf{x}) = \mathbf{G}(\mathbf{x})\,\mathbf{x}.$$

One can think of $\mathbf{G}(\cdot)$ as the nonlinear version of the conductance matrix. Hence the overall system is described by

$$\mathbf{C}\dot{\mathbf{x}} + \mathbf{G}(\mathbf{x})\,\mathbf{x} = 0, \text{ or } \dot{\mathbf{x}} = -\mathbf{C}^{-1}\mathbf{G}(\mathbf{x})\,\mathbf{x}.$$

Now $\mathbf{0}$ is an equilibrium of this network, by assumption.

To study the stability of this equilibrium, let us try the obvious Lyapunov function candidate, namely the total energy stored in the capacitors. This equals

$$V(\mathbf{x}) = \frac{1}{2}\mathbf{x}'\mathbf{C}\mathbf{x}.$$

Therefore

$$\dot{V}(\mathbf{x}) = \frac{1}{2}(\dot{\mathbf{x}}'\mathbf{C}\mathbf{x} + \mathbf{x}'\mathbf{C}\dot{\mathbf{x}}) = -\frac{1}{2}\mathbf{x}'[\mathbf{G}'(\mathbf{x}) + \mathbf{G}(\mathbf{x})]\mathbf{x},$$

where $\mathbf{G}'(\mathbf{x})$ is a shorthand for $[\mathbf{G}(\mathbf{x})]'$. Define

$$\mathbf{M}(\mathbf{x}) = \mathbf{G}'(\mathbf{x}) + \mathbf{G}(\mathbf{x}).$$

If there is a constant $r > 0$ such that $\mathbf{M}(\mathbf{x})$ is a positive definite matrix for all $\mathbf{x} \in B_r$, then $-\dot{V}$ is an lpdf, and the equilibrium $\mathbf{0}$ is asymptotically stable. Actually, the condition can be simplified further: Since $\mathbf{M}(\cdot)$ is continuous, the equilibrium $\mathbf{0}$ is asymptotically stable if $\mathbf{M}(\mathbf{0})$ is positive definite.

Next, we focus attention on the particular network shown in Figure 5.8, and show how Lemma (40) can be used to estimate the domain of attraction $D(\mathbf{0})$. The element ϕ_1 in Figure 5.8 is a conventional diode with the i–v characteristic shown in Figure 5.9, while the element ϕ_2 is a tunnel diode with the i–v characteristic shown in Figure 5.10; the element denoted by g_3 is a linear resistor with positive conductance g_3. It is easy to see that

$$\mathbf{i}(\mathbf{x}) = \begin{bmatrix} \phi_1(x_1) + g_3(x_1 - x_2) \\ \phi_2(x_2) + g_3(x_2 - x_1) \end{bmatrix}.$$

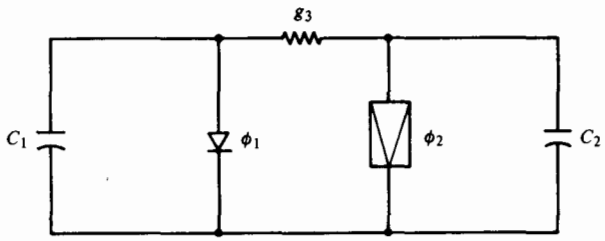


Fig. 5.8

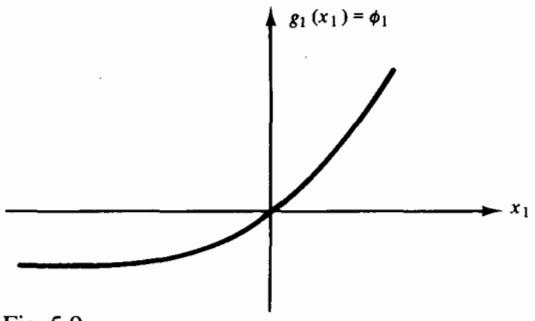


Fig. 5.9

If we define

$$g_1(x_1) = \frac{\phi_1}{x_1}, \; g_2(x_2) = \frac{\phi_2}{x_2},$$

then the nonlinear conductance matrix $G(\mathbf{x})$ can be written as

$$G(\mathbf{x}) = \begin{bmatrix} g_1 + g_3 & -g_3 \\ -g_3 & g_2 + g_3 \end{bmatrix},$$

where the arguments of g_1 and g_2 have been omitted for clarity. Clearly G is positive definite if and only if

$$g_1 + g_3 > 0, \det G(\mathbf{x}) = g_1 g_2 + g_3(g_1 + g_2) > 0.$$

Since $g_1(x_1) > 0 \; \forall x_1$ and $g_3 > 0$, the first condition is always satisfied, and only the second condition needs to be verified. Rewrite this as

$$\mathbf{44} \quad g_2 > -\frac{g_1 g_3}{g_1 + g_3} =: -g_{eq}.$$

The right side of (44), without the minus sign, is the equivalent conductance of g_1 and g_3 connected in series. At $x_1 = x_2 = 0$, (44) is satisfied since $g_2 > 0$. Hence $G(\mathbf{0})$ is positive definite, and by earlier reasoning, this is enough to show that $\mathbf{0}$ is an asymptotically stable equilibrium.

Next, let us determine a region S in the x_1–x_2 plane such that (44) is satisfied. Note that wherever (44) is satisfied, $G(\mathbf{x})$ is a positive definite matrix and thus $\dot{V}(\mathbf{x}) < 0$ (except at the origin of course). However, even if $G(\mathbf{x})$ fails to be a positive definite matrix at a particular point $\mathbf{x}$, it is nevertheless possible that $\dot{V}(\mathbf{x})$ is negative. The reason is that the positive definiteness of $G(\mathbf{x})$ implies that $\mathbf{y}' G(\mathbf{x})\mathbf{y} > 0 \; \forall \mathbf{y} \neq 0$, whereas all we really need is that $\mathbf{x}' G(\mathbf{x})\mathbf{x} > 0$. If $x_2 < 0$, then $g_2(x_2) > 0$, so that (44) is automatically satisfied. Now suppose $x_2 \geq 0$. If x_1 is large and positive, then $g_1 \to \infty$ and $g_{eq} \to g_3$. Hence $g_2(x_2) > -g_3$ provided x_2 does not belong to $[b, c]$, where b and c are identified in Figure 5.10. If x_1 is large and negative, then $g_1 \to 0$, and $g_{eq} \to 0$. In this case $g_2(x_2) > -g_{eq}$ provided x_2 does not belong to $[a, d]$, where a and d are identified in Figure 5.10. In summary, the region where $G(\mathbf{x})$ is not positive definite is the shaded region shown in Figure 5.11. Hence, if we define S to be the complement of the shaded region, then (38) holds. Now we know that every bounded invariant set contained in S is also contained in the domain of attraction $D(\mathbf{0})$. Lemma (40) tells us that every bounded level set of V contained in S is invariant, and is thus a subset of $D(\mathbf{0})$. In the present example, the level sets

$$L_V(d) = \{(x_1, x_2): C_1 x_1^2 + C_2 x_2^2 \leq d\}$$

are ellipses centered at the origin. Hence an estimate for $D(\mathbf{0})$ based on Lemma (40) is given by the ellipse shown in Figure 5.11.

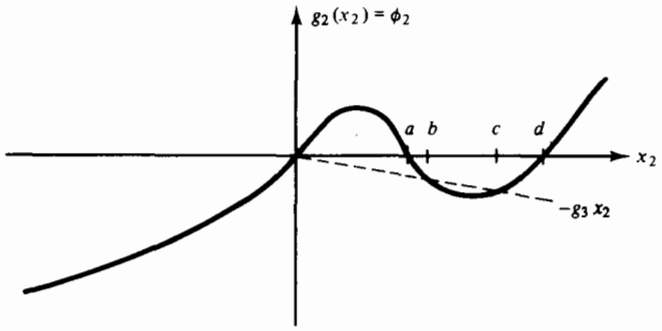


Fig. 5.10

Further analysis of this network is suggested in Problem 5.12. ■

The next theorem gives a sufficient condition for exponential stability.

45 Theorem *Suppose there exist constants* $a, b, c, r > 0$, $p \geq 1$, *and a* C^1 *function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *such that*

46 $a \|\mathbf{x}\|^p \leq V(t, \mathbf{x}) \leq b \|\mathbf{x}\|^p$, $\forall t \geq 0$, $\forall x \in B_r$,

47 $\dot{V}(t, \mathbf{x}) \leq -c \|\mathbf{x}\|^p$, $\forall t \geq 0$, $\forall \mathbf{x} \in B_r$.

Then the equilibrium $\mathbf{0}$ *is exponentially stable.*

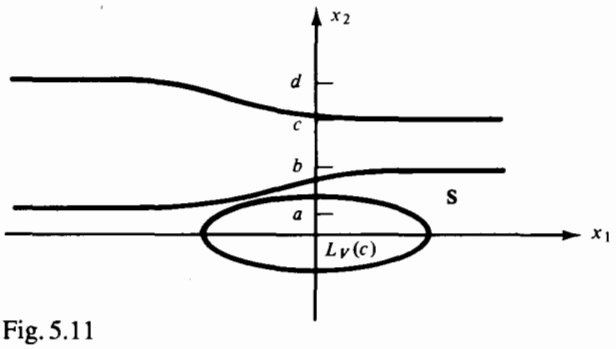


Fig. 5.11

Proof Define

48 $$\eta = r \left[\frac{a}{b} \right]^{1/p} \leq r,$$

and suppose $\mathbf{x}_0 \in B_\eta$, $t_0 \geq 0$. Then, letting $\mathbf{x}(t)$ denote the solution $\mathbf{s}(t, t_0, \mathbf{x}_0)$, we have

49 $$\frac{d}{dt}\dot{V}[t, \mathbf{x}(t)] \leq -c \|\mathbf{x}(t)\|^p \leq -\frac{c}{b} V[t, \mathbf{x}(t)].$$

Hence

50 $$V[t_0 + t, \mathbf{x}(t_0 + t)] \leq V[t_0, \mathbf{x}_0] \exp[-(c/b)t], \forall t \geq 0.$$

But, since

51 $$V(t_0, \mathbf{x}_0) \leq b \|\mathbf{x}_0\|^p, \text{ and}$$

52 $$a \|\mathbf{x}(t_0 + t)\|^p \leq V[t_0 + t, \mathbf{x}(t_0 + t)],$$

it follows that

53 $$a \|\mathbf{x}(t_0 + t)\|^p \leq b \|\mathbf{x}_0\|^p \exp[-(c/b)t], \forall t \geq 0.$$

Finally,

54 $$\|\mathbf{x}(t_0 + t)\| \leq \left[\frac{b}{a}\right]^{1/p} \|\mathbf{x}_0\| \exp[-(c/bp)t], \forall t \geq 0.$$

Hence (5.1.37) is satisfied with $(b/a)^{1/p}$ playing the role of a and c/bp playing the role of b. Thus $\mathbf{0}$ is an exponentially stable equilibrium. ■

55 **Example** Consider again the nonlinear circuit of Example (43). Let

$$C_m = \min_i C_i, \; C_M = \max_i C_i.$$

Then

$$\frac{1}{2} C_m \|\mathbf{x}\|^2 \leq V(\mathbf{x}) \leq \frac{1}{2} C_M \|\mathbf{x}\|^2, \forall \mathbf{x} \in \mathbf{R}^n.$$

If the matrix $\mathbf{M}(\mathbf{0})$ is positive definite, then by continuity $\mathbf{M}(\mathbf{x})$ is also positive definite for each $\mathbf{x}$ belonging to some ball B_r. Let

$$d = \inf_{\mathbf{x} \in B_r} \lambda_{\min}[\mathbf{M}(\mathbf{x})],$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a symmetric matrix, and choose $r > 0$ sufficiently small that $d > 0$. Then

$$\dot{V}(\mathbf{x}) = -\mathbf{x}'\mathbf{M}(\mathbf{x})\mathbf{x} \leq -d \|\mathbf{x}\|^2, \forall \mathbf{x} \in B_r.$$

Thus all hypotheses of Theorem (45) are satisfied, and we conclude that the equilibrium $\mathbf{0}$ is in fact exponentially stable. ■

The theorems for global uniform asymptotic stability and global exponential stability are straight-forward generalizations of Theorems (25) and (45) respectively.

56 Theorem *The equilibrium* $\mathbf{0}$ *of (5.1.1) is globally uniformly asymptotically stable if there exists a* C^1 *function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *such that (i) V is a pdf, decrescent and radially unbounded, and (ii)* $-\dot{V}$ *is a pdf.*

Proof The hypotheses include those of Theorem (25); hence $\mathbf{0}$ is uniformly asymptotically stable. Thus, to prove the theorem, it only remains to prove that $\mathbf{0}$ is globally uniformly attractive, i.e. that, given any $M < \infty$ and any $\varepsilon > 0$, there exists a $T = T(M, \varepsilon)$ such that

$$\mathbf{57} \quad \|\mathbf{x}_0\| < M, t_0 \geq 0 \Rightarrow \|\mathbf{s}(t_0 + t, t_0, \mathbf{x}_0)\| < \varepsilon, \forall t \geq T.$$

For this purpose, select functions α, β and γ of class K, with α radially unbounded, such that

$$\mathbf{58} \quad \alpha(\|\mathbf{x}\|) \leq V(t, \mathbf{x}) \leq \beta(\|\mathbf{x}\|), \dot{V}(t, \mathbf{x}) \leq -\gamma(\|\mathbf{x}\|), \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$$

These conditions are the same as (27) and (28) with B_r replaced by $\mathbf{R}^n$. Now select a constant $r > 0$ such that

$$\mathbf{59} \quad \beta(M) < \alpha(r).$$

This is possible since $\alpha(r) \to \infty$ as $r \to \infty$. Then, following the reasoning of Equations (9) – (13), it can be shown that every trajectory that starts in the ball B_M stays inside the ball B_r, i.e. all trajectories of the system are bounded. Now choose

$$\mathbf{60} \quad \delta_2 < \beta^{-1}[\alpha(\varepsilon)],$$

and define

$$\mathbf{61} \quad T = \frac{\beta(r)}{\gamma(\delta_2)}.$$

From this point onwards the proof is the same as that of Theorem (25). ■

62 Theorem *The equilibrium* $\mathbf{0}$ *is globally exponentially stable if there exist constants* $a, b, c > 0, p \geq 1$, *and a* C^1 *function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *such that*

$$\mathbf{63} \quad a\|\mathbf{x}\|^p \leq V(t, \mathbf{x}) \leq b\|\mathbf{x}\|^p, \dot{V}(t, \mathbf{x}) \leq -c\|\mathbf{x}\|^p, \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$$

Proof Entirely analogous to that of Theorem (45). ■

Remarks Note that, in Theorem (56), V is required to be radially unbounded, but $-\dot{V}$ is not. The hypothesis on $\dot{V}$ in Theorem (62) implies that $-\dot{V}$ is also radially unbounded.

64 Example Consider again the nonlinear circuit of Example (43). If

$$0 < \inf_{\mathbf{x} \in \mathbb{R}^n} \lambda_{\min}[M(\mathbf{x})] =: \lambda,$$

then

$$\dot{V}(\mathbf{x}) \leq -\lambda \|\mathbf{x}\|^2, \ \forall \mathbf{x} \in \mathbf{R}^n,$$

and the equilibrium $\mathbf{0}$ is globally exponentially stable by Theorem (62). ■

65 Example The purpose of this example is to illustrate that in Theorem (56), the assumption that V is radially unbounded is indispensable. Without this assumption, the theorem is not valid. In fact, even if one can find a function V that satisfies all the hypotheses of Theorem (56) *except* for radial unboundedness, it is still possible that the solution trajectories of the system exhibit finite escape time.

Consider the second-order system

66
$$\dot{x}_1 = (x_2 - 1)x_1^3,$$

$$\dot{x}_2 = -\frac{x_1^4}{(1+x_1^2)^2} - \frac{x_2}{1+x_2^2}.$$

Let

$$V(x_1, x_2) = \frac{x_1^2}{1+x_1^2} + x_2^2.$$

Note that V is a pdf, but is *not* radially unbounded. In fact, the level set

$$L_V(c) = \{(x_1, x_2): V(x_1, x_2) \leq c\}$$

is bounded if $c < 1$, but is unbounded if $c \geq 1$ (see Figure 5.12). Now

$$\dot{V}(x_1, x_2) = -\frac{2x_1^4}{(1+x_1^2)^2} - \frac{2x_2^2}{1+x_2^2}.$$

Hence $-\dot{V}$ is also a pdf. Thus V satisfies all the hypotheses of Theorem (56) except for radial unboundedness. In spite of this, the origin $\mathbf{0}$ is *not* a globally attractive equilibrium. In fact, solution trajectories starting from initial conditions sufficiently far from the origin exhibit finite escape time.

It takes a bit of work to establish this fact. First, consider the scalar differential equation

$$\dot{r} = 2\alpha r^2,$$

where $\alpha > 0$ is a constant. The solution of this equation is

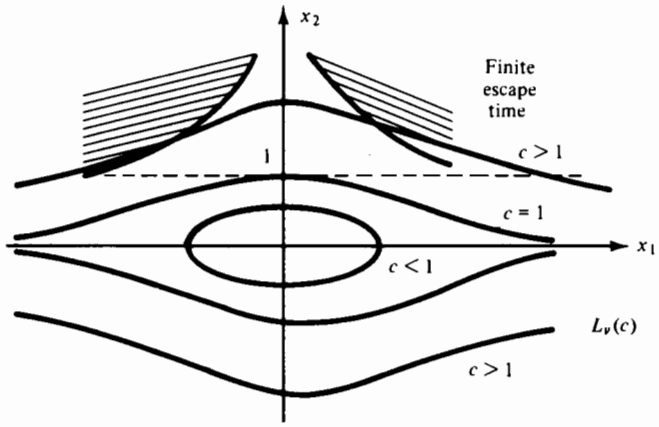


Fig. 5.12

$$r(t) = \frac{r(0)}{[1 - 2\alpha r(0)t]^{1/2}}.$$

Hence, if $r(0) > 0$, it follows that

$$r(t) \to \infty \text{ as } t \to T = \frac{1}{2\alpha r(0)}.$$

In other words, the trajectory exhibits finite escape time. Now consider the analogous equation

$$\dot{r}(t) = 2\beta(t)\, r^2(t),$$

where

$$\beta(t) \geq \alpha > 0, \ \forall t.$$

If $r(0) > 0$, then the solution of this equation increases at least as rapidly as the solution of the previous equation, since the present $\dot{r}(t)$ is at least as large as it is in the previous equation. Therefore the solution of this equation also exhibits finite escape time, and the escape time is *no larger than*

$$T = \frac{1}{2\alpha r(0)}.$$

Returning now to the system at hand, observe that

$$\left|\frac{x_1^4}{1+x_1^4}\right| < 1, \quad \left|\frac{x_2}{1+x_2^2}\right| < 0.5, \ \forall x_1, x_2.$$

Hence, from (66),

$$\dot{x}_2 \geq -1.5, \text{ or } x_2(t) \geq x_2(0) - 1.5t, \ \forall t \geq 0.$$

Now let $r = x_1^2$. Then from (66),

67 $\dot{r} = 2(x_2 - 1)\, r^2.$

Now it is claimed that, whenever the initial condition $\mathbf{x}_0 = (x_{10}, x_{20})$ of (66) lies in the first or second quadrant and satisfies

68 $(x_{20} - 1) \cdot |\dot{x}_{10}| > \sqrt{3},$

the resulting solution trajectory exhibits finite escape time. (Note: It is *not* claimed that these are the *only* initial conditions that lead to finite escape time.) Suppose (68) holds. Then

$$x_2(t) \geq x_{20} - 1.5t, \ \forall t \geq 0.$$

Define

$$T_m = \frac{x_{20} - 1}{3}.$$

Then

$$x_2(t) - 1 \geq x_{20} - 1.5t - 1 \geq \frac{x_{20} - 1}{2}, \ \forall t \in [0, T_m].$$

In view of the earlier discussion, the differential equation (67) exhibits finite escape time which is no larger than

$$T_e = \frac{1}{(x_{20} - 1)\, x_{10}^2} = \frac{1}{[(x_{20} - 1)^2\, x_{10}^2]/(x_{20} - 1)} < \frac{1}{3/(x_{20} - 1)} = \frac{x_{20} - 1}{3} = T_m.$$

This "closes the loop" on the circular reasoning and establishes the claim. Figure 5.12 shows the initial conditions which are guaranteed to lead to trajectories having finite escape time. ■

The final two theorems on asymptotic stability are applicable only to autonomous or periodic systems. This is in contrast to all of the preceding theorems in this section, which can be applied to arbitrary nonautonomous systems. The main feature of these theorems is that they enable one to claim (uniform) asymptotic stability if there exists a C^1 lpdf V whose derivative $\dot{V}$ is nonpositive along trajectories, even if $\dot{V}$ is not locally negative definite.

However, it should be noted that these theorems do not allow one to conclude exponential stability.

These theorems were first proved in the Soviet Union by Barbashin and Krasovskii (1952) in a special case, and by Krasovskii (1959) in the general case. Later they were independently rediscovered in the West by LaSalle (1960). In Western literature these theorems are often referred to as LaSalle's theorems, but it is more accurate to call them Krasovskii-LaSalle theorems.

As a prelude to stating these theorems, let us extend the notion of level sets to functions of both t and $\mathbf{x}$. Suppose $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ is continuous, and suppose $V(t, \mathbf{0}) = 0 \; \forall t \geq 0$. Define

69 $M_V(c) = \{x \in \mathbf{R}^n : \exists t \geq 0$ such that $V(t, \mathbf{x}) \leq c\}$.

Note that if V is independent of t, then (69) reduces to the earlier definition (39). Now $\mathbf{0} \in M_V(c)$ whenever $c \geq 0$. The **level set** $L_V(c)$ is now defined as before, namely, the connected component of $M_V(c)$ containing $\mathbf{0}$. Next we define another set, namely

70 $A_V(c) = \{x \in L_V(c) : V(t, \mathbf{x}) \leq c, \; \forall t \geq 0\}$.

Note the difference between the quantifiers in (69) and (70); also note that if V is independent of t, then $A_V(c)$ is the same as $L_V(c)$.

The following lemma is of independent interest, even if $\mathbf{0}$ is not asymptotically stable.

71 Lemma *Suppose the system (5.1.1) is periodic. Suppose there exists a C^1 function $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ such that (i) V is periodic with the same period as the system, (ii) V is an lpdf, (iii) there exists an open neighborhood N of $\mathbf{0}$ such that*

72 $\dot{V}(t, \mathbf{x}) \leq 0, \; \forall t \geq 0, \; \forall \mathbf{x} \in N$.

Choose a constant $c > 0$ such that the level set $L_V(c)$ is bounded and contained in N. Finally, let

73 $S = \{\mathbf{x} \in L_V(c) : \exists t \geq 0$ such that $\dot{V}(t, \mathbf{x}) = 0\}$,

and let M denote the largest invariant set of (5.1.1) contained in S. Then

74 $\mathbf{x}_0 \in A_V(c), \; t_0 \geq 0 \Rightarrow d[\mathbf{s}(t, t_0, \mathbf{x}_0), M] \to 0$ as $t \to \infty$,

where $d(\mathbf{y}, M)$ denotes the distance from the point $\mathbf{y}$ to the set M [cf. (5.2.33)].

Proof Since V is periodic, it is decrescent, so that $\mathbf{0}$ is an interior point of $A_V(c)$. Using the methods of Equations (9) – (13), it is easy to show that if $\mathbf{x}_0 \in A_V(c)$ then $\mathbf{s}(t, t_0, \mathbf{x}_0) \in L_V(c) \; \forall t \geq t_0$. Since $L_V(c)$ is bounded, the limit set $\Omega(t_0, \mathbf{x}_0)$ [see Definition (5.2.27)] is nonempty. Further, by Lemma (5.2.34),

75 $d[\mathbf{x}(t), \Omega(t_0, \mathbf{x}_0)] \to 0$ as $t \to \infty$,

where $\mathbf{x}(t)$ is a shorthand for $\mathbf{s}(t, t_0, \mathbf{x}_0)$. Further, by Lemma (5.2.37), $\Omega(t_0, \mathbf{x}_0)$ is an invariant set.

Consider now what happens to the function $V[t, \mathbf{x}(t)]$. Since $\dot{V}[t, \mathbf{x}(t)] \leq 0 \; \forall t \geq t_0$, $V[t, \mathbf{x}(t)]$ is monotonic and has a definite limit as $t \to \infty$; also $\dot{V}[t, \mathbf{x}(t)] \to 0$ as $t \to \infty$. Suppose $\mathbf{y} \in \Omega(t_0, \mathbf{x}_0)$. Then, by definition, there exists a sequence $\{t_i\}$ approaching ∞ such that $\mathbf{x}(t_i) \to \mathbf{y}$. Select integers k_i such that $t_i - k_i T \in [0, T)$ where T is the period. Then the sequence $\{t_i - k_i T\}$ is bounded and therefore contains a convergent subsequence. Renumber the subsequences again as $\{t_i\}$ and $\{k_i\}$, and let τ denote the limit of the sequence $\{t_i - k_i T\}$. Then

76
$$\begin{aligned} \dot{V}(\tau, \mathbf{y}) &= \lim_{i \to \infty} \dot{V}(t_i - k_i T, \mathbf{y}) \\ &= \lim_{i \to \infty} \dot{V}(t_i, \mathbf{y}) \quad \text{since } \dot{V} \text{ is periodic} \\ &= \lim_{i \to \infty} \dot{V}[t_i, \mathbf{x}(t_i)] \\ &= 0. \end{aligned}$$

In other words, $\mathbf{y} \in S$. Since this is true of every $\mathbf{y} \in S$, it follows that $\Omega(t_0, \mathbf{x}_0) \subseteq S$, and since $\Omega(t_0, \mathbf{x}_0)$ is an invariant set, it follows that $\Omega(t_0, \mathbf{x}_0) \subseteq M$. The desired conclusion (74) now follows from (75). ■

77 Theorem (Krasovskii-LaSalle) *Suppose the system (5.1.1) is periodic. Suppose there exists a C^1 lpdf $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ having the same period as the system, and an open neighborhood N of $\mathbf{0}$ such that (72) holds. Choose a constant $c > 0$ such that the level set $L_V(c)$ is bounded and contained in N, and define S as in (73). Under these conditions, if S contains no trajectories of the system other than the trivial trajectory $\mathbf{x}(t) \equiv 0 \; \forall t_0 \geq 0$, then the equilibrium $\mathbf{0}$ is uniformly asymptotically stable.*

Proof This theorem is essentially a corollary of Lemma (71). Let M be the largest invariant set contained in S. It is claimed that $M = \{\mathbf{0}\}$. To see this, let $\mathbf{y} \in M$. Then, by the definition of invariance, there exists a $t_0 \geq 0$ such that the corresponding trajectory $\mathbf{s}(t, t_0, \mathbf{y}) \in M \; \forall t \geq t_0$. However, by assumption, S does not contain any trajectories other than the trivial trajectory, and M is a subset of S. Hence $\mathbf{y} = \mathbf{0}$, i.e., $M = \{\mathbf{0}\}$.

Next, note that since $M = \{\mathbf{0}\}$, the distance $d(\mathbf{z}, M)$ is just $\|\mathbf{z}\|$. Hence, by Lemma (71), and in particular (74),

78 $\mathbf{x}_0 \in A_V(c) \Rightarrow \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| \to 0$, as $t \to \infty$.

Hence the origin is attractive. It is also stable, by Theorem (1), and is thus asymptotically stable. Finally, by Theorem (5.1.49), $\mathbf{0}$ is a uniformly asymptotically stable equilibrium. ■

Remark The proof of Theorem (77) makes it clear that the set $A_V(c)$ defined in (70) is contained in the domain of attraction, in the sense that every trajectory starting in $A_V(c)$, at whatever initial time, approaches $\mathbf{0}$ as $t \to \infty$.

79 Theorem (Krasovskii-LaSalle) *Suppose the system (5.1.1) is periodic. Suppose there exists a C^1 function $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ having the same period as the system such that (i) V is a pdf and is radially unbounded, and (ii)*

80 $\dot{V}(t, \mathbf{x}) \leq 0, \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n$.

Define

81 $R = \{\mathbf{x} \in \mathbf{R}^n : \exists t \geq 0 \text{ such that } \dot{V}(t, \mathbf{x}) = 0\}$,

and suppose R does not contain any trajectories of the system other than the trivial trajectory. Then the equilibrium $\mathbf{0}$ is globally uniformly asymptotically stable.

Proof Since V is radially unbounded, each set $M_V(c)$ defined in (69) is bounded, whence $L_V(c)$ is also bounded for each $c > 0$. Thus, proceeding as in the proof of Theorem (77), one can show that (78) holds for every $c > 0$. Thus the equilibrium $\mathbf{0}$ is globally attractive; it is also uniformly stable, by Theorem (1). Thus it only remains to show that the attraction to $\mathbf{0}$ is uniform with respect t_0 and $\| \mathbf{x}_0 \|$. This part of the proof is omitted, and the reader is referred to Hahn (1967), Theorems 38.3 and 38.5, or Krasovskii (1959), Theorem 14.1. ■

The application of Theorems (77) and (79) is illustrated through several examples.

82 Example Consider a unit mass constrained by a nonlinear spring and nonlinear friction. Such a system can be represented in state variable form by

$$\dot{x}_1 = x_2, \dot{x}_2 = -g(x_1) - f(x_2),$$

where $g(\cdot)$ is the restoring force of the spring and $f(\cdot)$ is the force due to friction. Suppose that $f(\cdot)$, $g(\cdot)$ are continuous, and that $f(0) = g(0) = 0$; then $\mathbf{0}$ is an equilibrium. In addition, suppose there exists a constant $\mu > 0$ such that

$$rf(r) > 0, \forall r \neq 0, r \in [-\mu, \mu],$$

$$rg(r) > 0, \forall r \neq 0, r \in [-\mu, \mu],$$

Finally, define the function

$$\Phi(r) = \int_0^r g(\sigma)\, d\sigma,$$

and suppose there is a constant $c > 0$ such that the level set $L_\Phi(c)$ is bounded. Under these conditions, it is claimed that $\mathbf{0}$ is an asymptotically stable equilibrium.

Before establishing the claim, let us reflect on what it means. The claim, simply, is that if the restoring spring force and the friction force are first and third quadrant functions in some neighborhood of the origin, then the origin is asymptotically stable.

To prove the claim, select the total energy of the system as an obvious Lyapunov function candidate. This is the sum of the potential energy stored in the spring and the kinetic energy of the mass. Thus

$$V(x_1, x_2) = \Phi(x_1) + \frac{1}{2}x_2^2 = \int_0^{x_1} g(\sigma)\, d\sigma + \frac{1}{2}x_2^2.$$

Then

$$\dot{V}(x_1, x_2) = g(x_1)\dot{x}_1 + x_2\dot{x}_2 = -x_2 f(x_2).$$

Now various properties of V and $\dot{V}$ are demonstrated, and finally asymptotic stability is concluded on the basis of Theorem (77).

1) V is an lpdf. To show this, suppose that $\mathbf{x} \neq \mathbf{0}$ and that $|x_1|, |x_2| \leq \mu$. Then $V(\mathbf{x}) > 0$ by virtue of the conditions on $f(\cdot)$ and $g(\cdot)$.

(2) $\dot{V} \leq 0$ whenever $|bx_1|, |bx_2| \leq \mu$. This too follows from the condition on $f(\cdot)$.

(3) The level set $L_V(c)$ is bounded. To see this, note that

$$V(\mathbf{x}) \leq c \Rightarrow \Phi(x_1) \leq c \text{ and } |x_2| \leq \sqrt{2c} =: d.$$

Hence $L_V(c)$ is contained in the bounded set $L_\Phi(c) \times [-d, d]$.

To apply Theorem (77), it is necessary to determine the set S of (73). Suppose $\mathbf{x} \in L_V(c)$ and $\dot{V}(\mathbf{x}) = 0$. Then $x_2 f(x_2) = 0$, which implies that $x_2 = 0$. Hence

$$\Rightarrow S = \{(x_1, x_2): x_2 = 0\}.$$

To apply Theorem (77), it only remains to verify that S contains no nontrivial system trajectories. Suppose $\mathbf{x}(t)$, $t \geq 0$ is a trajectory that lies entirely in S. Then

$$x_2(t) \equiv 0, \forall t \geq 0.$$

But this in turn implies that $f[x_2(t)] \equiv 0 \; \forall t \geq 0$. Also, since $\dot{x}_2 = -g(x_1) - f(x_2)$, it follows that

$$\dot{x}_2(t) \equiv 0 \Rightarrow g[x_1(t)] \equiv 0 \Rightarrow x_1(t) \equiv 0.$$

In other words, $\mathbf{x}(t)$ is the trivial trajectory. Thus, by Theorem (77), the origin is asymptotically stable.

If the conditions on f and g are strengthened to

$$rf(r) > 0,\ rg(r) > 0,\ \forall r \neq 0,$$

$$\Phi(r) \to \infty \text{ as } |r| \to \infty,$$

then Theorem (79) would apply and we can conclude that the origin is globally uniformly asymptotically stable.

83 Example A phase-locked loop in communication networks can be described by the equation

$$\ddot{y}(t) + [a + b(t)\cos y(t)]\,\dot{y}(t) + c(t)\sin y(t) = 0,$$

where $b(\cdot)$, $c(\cdot)$ are periodic functions with the same period. In this example, the stability of this system is analyzed. One of the objectives of this example is to illustrate the difference between the sets L_V and A_V in (70).

We begin by rewriting the system equation in the form

$$\dot{x}_1(t) = x_2(t),\ \dot{x}_2(t) = -[a + b(t)\cos x_1(t)]\,x_2(t) - c(t)\sin x_1(t).$$

Suppose $a > 0$ and that the following conditions hold.

$b(\cdot)$ is continuous and $c(\cdot)$ is C^1,

$$|b(t)| \leq b_M < a,\ \forall t,$$

$$0 < c_m := \min_t c(t),\ \max_t c(t) =: c_M < \infty.$$

$$|\dot{c}(t)| < 2(a - b_M)\,c_m,\ \forall t.$$

Then it is claimed that the origin is uniformly asymptotically stable.

It is worthwhile to reflect on what the above conditions mean. One can think of the system as a standard second order system with nonlinear damping and restorative force. The damping is always in the interval $[a - b_M, a + b_M]$, and is bounded away from zero by assumption. The restoring force is always positive. The last condition takes into account the time-varying nature of the system, and basically means that the function $c(\cdot)$ varies sufficiently slowly.

One might be tempted to try the "natural" Lyapunov function candidate

$$W(t, x_1, x_2) = c(t)(1 - \cos x_1) + \frac{1}{2}x_2^2,$$

but this function does not work nearly so well as

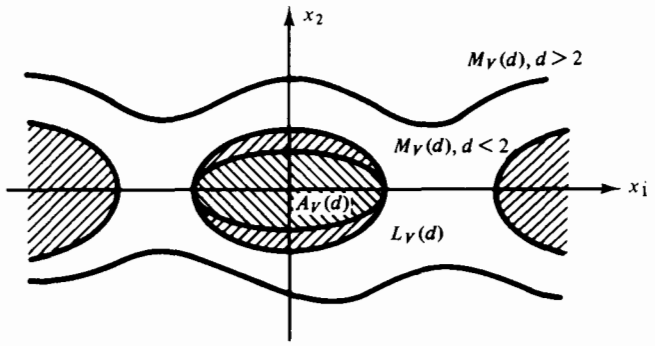


Fig. 5.13

$$V(t, x_1, x_2) = 1 - \cos x_1 + \frac{1}{2c(t)} x_2^2.$$

Note that V is periodic with the same period as the system. Now suppose $d \geq 0$ is a real number. If $d \geq 2$, then the set $M_V(d)$ defined in (69) is connected and unbounded (see Figure 5.13). If $d < 2$, the set $M_V(d)$ splits into an infinite number of components. The level set $L_V(d)$ consists of the component containing the origin, and is described by

$$L_V(d) = \{\mathbf{x}: 1 - \cos x_1 + (1/2c_M) x_2^2 \leq d, \ |x_1| \leq \cos^{-1} d\}.$$

Every point $\mathbf{x}$ in $L_V(d)$ has the property that $V(t, \mathbf{x}) \leq d$ for *some* t [e.g. the time t at which $c(\cdot)$ attains its maximum value c_M]. In contrast, the set $A_V(d)$ consists of those $\mathbf{x}$ in $L_V(d)$ such that $V(t, \mathbf{x}) \leq d$ for *all* t. Clearly

$$A_V(d) = \{\mathbf{x} \in L_V(d): 1 - \cos x_1 + (1/2c_m) x_2^2 \leq d\}.$$

This set is also shown in Figure 5.13.

Now the equilibrium $\mathbf{0}$ is shown to be asymptotically stable using Theorem (77). Since

$$(1 - \cos x_1) + \frac{1}{2c_M} x_2^2 \leq V(t, \mathbf{x}) \leq (1 - \cos x_1) + \frac{1}{2c_m} x_2^2,$$

V is an lpdf and is decrescent. Next,

$$\dot{V} = \dot{x}_1 \sin x_1 + \frac{\dot{x}_2 x_2}{c(t)} - \frac{x_2^2 \dot{c}(t)}{2c^2(t)}$$

$$= -\frac{x_2^2}{2c^2(t)} \{2c(t)[a + b(t) \cos x_1] - \dot{c}(t)\}.$$

Now consider the function inside the braces. Clearly

$$2c(t)[a(t)+b(t)\cos x_1]-\dot{c}(t)\geq 2c_m(a-b_M)-\dot{c}(t)=:d(t)>0,\ \forall t,$$

by assumption. Hence $\dot{V}(t, \mathbf{x})\leq 0\ \forall t, \mathbf{x}$. To apply Theorem (77), the only remaining condition is that the set S defined in (73) does not contain any nontrivial trajectories. Since the function $d(\cdot)$ defined above is continuous, periodic, and positive-valued, it follows that $d(t)$ is bounded away from zero. Hence

$$\exists t\geq 0 \text{ such that } \dot{V}(t, \mathbf{x})=0 \text{ iff } x_2=0,$$

In other words,

$$S=\{\mathbf{x}\in L_V(d): x_2=0\}=\{(x_1, 0): |x_1|\leq \cos^{-1} d\}.$$

Suppose now that $\mathbf{x}(\cdot)$ is a trajectory of the system lying entirely in S. Then

$$x_2(t)\equiv 0\Rightarrow \dot{x}_2(t)\equiv 0\Rightarrow c(t)\sin x_1(t)\equiv 0\Rightarrow x_1(t)\equiv 0.$$

Hence $\mathbf{x}(\cdot)$ is the trivial trajectory. Therefore, by Theorem (77), $\mathbf{0}$ is a uniformly asymptotically stable equilibrium. Moreover, by the remark following the proof of Theorem (77), the set $A_V(d)$ for each $d<2$ is in the domain of attraction; but the same is not necessarily true of $L_V(d)$.

84 Example (Stabilization of a Rigid Robot) As an application of the Krasovskii-LaSalle theorem, consider the problem of stabilizing a rigid robot which is operating in a gravity-free environment. The absence of gravity can come about because the robot is operating in outer space. Even in more "down to earth" applications, this assumption is valid if the robot is constrained to operate in a plane which is perpendicular to gravity, for example, a table-top robot operating on an air cushion. The assumption of rigidity ensures that the number of degrees of freedom equals the number of control actuators.

Let $\mathbf{q}=[q_1\cdots q_n]'$ denote the vector of generalized coordinates of the robot, and let $\mathbf{u}=[u_1\cdots u_n]'$ denote the vector of generalized forces. The assumption of rigidity means, in effect, that the vector $\mathbf{u}$ can be chosen arbitrarily and is thus a suitable control input. Now the dynamics of the robot are described by the Euler-Lagrange equations

$$\textbf{85}\qquad \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\mathbf{q}}}\right]-\frac{\partial L}{\partial \mathbf{q}}=\mathbf{u},$$

where L is the Lagrangian of the system. Since it is assumed that there is no gravity, the potential energy of the robot is a constant, which can be taken to be zero. Hence the Lagrangian equals the kinetic energy K. As is customary, assume that

$$K=\frac{1}{2}\dot{\mathbf{q}}'\mathbf{D}(\mathbf{q})\dot{\mathbf{q}},$$

where the matrix $\mathbf{D}(\mathbf{q})$ is called the **inertia matrix.** This matrix is configuration dependent

but always positive definite. It is reasonable to assume that there exist positive constants α and β such that

$$0 < \alpha \leq \lambda_{\min}[\mathbf{D}(\mathbf{q})] \leq \lambda_{\max}[\mathbf{D}(\mathbf{q})] \leq \beta, \ \forall \mathbf{q}.$$

Substituting for $L = K$ in (85) gives the dynamical equations in the standard form

$$\sum_{j=1}^{n} d_{ij}(\mathbf{q})\,\ddot{q}_j + \sum_{j=1}^{n}\sum_{k=1}^{n} c_{ijk}(\mathbf{q})\,\dot{q}_j\dot{q}_k = u_i,\ i = 1, \cdots, n,$$

where

$$c_{ijk} = \frac{1}{2}\left[\frac{\partial d_{ik}}{\partial q_j} + \frac{\partial d_{ij}}{\partial q_k} - \frac{\partial d_{jk}}{\partial q_i}\right]$$

are called the *Christoffel symbols.* [For further details see Spong and Vidyasagar (1989), Sec. 6.3.] These equations can be written compactly as

$$\mathbf{D}(\mathbf{q})\,\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\,\dot{\mathbf{q}} = \mathbf{u},$$

where $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})$ is an $n \times n$ matrix whose ij-th element is

$$c_{ij}(\mathbf{q}, \dot{\mathbf{q}}) = \sum_{k=1}^{n} c_{ijk}(\mathbf{q}, \dot{\mathbf{q}})\,\dot{q}_k.$$

Of course, by introducing the state variables $\mathbf{x} = \mathbf{q}$, $\mathbf{y} = \dot{\mathbf{q}}$, these equations can be put in the familiar form (5.1.1), namely

$$\dot{\mathbf{x}} = \mathbf{y},\ \dot{\mathbf{y}} = [\mathbf{D}(\mathbf{x})]^{-1}\,[\mathbf{u} - \mathbf{C}(\mathbf{x}, \mathbf{y})\mathbf{y}].$$

Suppose one is given a vector $\mathbf{q}_d$ representing the desired value of the generalized coordinate vector $\mathbf{q}$. If $\mathbf{q}$ is the vector of joint angles of the robot, then $\mathbf{q}_d$ would be the vector of desired joint angles. To make $\mathbf{q}(t)$ approach the desired vector $\mathbf{q}_d$, let us try the control law

$$\mathbf{u} = -\mathbf{K}_p(\mathbf{q} - \mathbf{q}_d) - \mathbf{K}_d\dot{\mathbf{q}} = -\mathbf{K}_p(\mathbf{x} - \mathbf{q}_d) - \mathbf{K}_d\mathbf{y},$$

where $\mathbf{K}_p$ and $\mathbf{K}_d$ are arbitrary positive definite matrices. This control law is known as a PD (proportional plus derivative) control law. With this control law, then the system equations become

$$\mathbf{86} \quad \dot{\mathbf{x}} = \mathbf{y},\ \dot{\mathbf{y}} = -[\mathbf{D}(\mathbf{x})]^{-1}\,[\mathbf{K}_p(\mathbf{x} - \mathbf{q}_d) + \mathbf{K}_d\mathbf{y} + \mathbf{C}(\mathbf{x}, \mathbf{y})\,\mathbf{y}].$$

These equations can be made to look more familiar by reverting briefly to the original variables $\mathbf{q}$ and $\dot{\mathbf{q}}$, and writing them as

$$\mathbf{D}(\mathbf{q})\,\ddot{\mathbf{q}} + [\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{K}_d]\,\dot{\mathbf{q}} + \mathbf{K}_p\mathbf{q} = \mathbf{0},$$

which look like the dynamical equations of a mass-spring-dashpot system.

Now Theorem (79) is used to show that the equilibrium $\mathbf{x} = \mathbf{q}_d$, $\mathbf{y} = \mathbf{0}$ of the system (86) is globally uniformly asymptotically stable . For this purpose, select the Lyapunov function candidate

$$V = \frac{1}{2}[\dot{\mathbf{q}}'\,\mathbf{D}(\mathbf{q})\,\dot{\mathbf{q}} + (\mathbf{q} - \mathbf{q}_d)'\,\mathbf{K}_p\,(\mathbf{q} - \mathbf{q}_d)]$$

$$= \frac{1}{2}[\mathbf{y}'\,\mathbf{D}(\mathbf{x})\,\mathbf{y} + (\mathbf{x} - \mathbf{q}_d)'\,\mathbf{K}_p\,(\mathbf{x} - \mathbf{q}_d)].$$

The first term in V is the kinetic energy while the second term is the potential energy due to the proportional feedback, which acts like a spring. Clearly V is positive definite and radially unbounded. Now let us differentiate V along the system trajectories. For this purpose, note that

$$\mathbf{87} \qquad \frac{d}{dt}[d_{ij}(\mathbf{x})] = \sum_{k=1}^{n} \frac{\partial d_{ij}(\mathbf{x})}{\partial x_k}\,\dot{x}_k = \sum_{k=1}^{n} \frac{\partial d_{ij}(\mathbf{x})}{\partial x_k}\,y_k.$$

Let us define an $n \times n$ matrix $\dot{\mathbf{D}}(\mathbf{x}, \mathbf{y})$ whose ij-th element is the right side of (87). Then, along the trajectories of (86), we have

$$\dot{V} = \mathbf{y}'\,\mathbf{D}(\mathbf{x})\,\dot{\mathbf{y}} + \frac{1}{2}\mathbf{y}'\,\dot{\mathbf{D}}(\mathbf{x}, \mathbf{y})\,\mathbf{y} + \dot{\mathbf{x}}'\,\mathbf{K}_p\,(\mathbf{x} - \mathbf{q}_d)$$

$$= -\mathbf{y}'\,[\mathbf{K}_p\,(\mathbf{x} - \mathbf{q}_d) + \mathbf{K}_d\mathbf{y} + \mathbf{C}(\mathbf{x}, \mathbf{y})\,\mathbf{y}] + \frac{1}{2}\mathbf{y}'\,\dot{\mathbf{D}}(\mathbf{x}, \mathbf{y})\,\mathbf{y} + \mathbf{y}'\mathbf{K}_p\,(\mathbf{x} - \mathbf{q}_d)$$

$$= -\mathbf{y}'\mathbf{K}_d\mathbf{y} + \frac{1}{2}\mathbf{y}'\,[\dot{\mathbf{D}}(\mathbf{x}, \mathbf{y}) - 2\mathbf{C}(\mathbf{x}, \mathbf{y})]\,\mathbf{y}.$$

Next it is shown that the matrix $\dot{\mathbf{D}} - 2\mathbf{C}$ is skew-symmetric, which implies that the last term on the right side is zero. Define $\mathbf{M} := \dot{\mathbf{D}} - 2\mathbf{C}$, and note that

$$2c_{ij} = \sum_{k=1}^{n} 2c_{ijk}\,y_k = \sum_{k=1}^{n}\left[\frac{\partial d_{ik}}{\partial x_j} + \frac{\partial d_{ij}}{\partial x_k} - \frac{\partial d_{jk}}{\partial x_i}\right] y_k.$$

Hence, from (87),

$$m_{ij} = \sum_{k=1}^{n}\left[\frac{\partial d_{ik}}{\partial x_j} - \frac{\partial d_{jk}}{\partial x_i}\right] y_k.$$

Interchanging i and j gives

$$m_{ji} = \sum_{k=1}^{n} \left[\frac{\partial d_{jk}}{\partial x_i} - \frac{\partial d_{ik}}{\partial x_j} \right] y_k = -m_{ij}.$$

Hence $\mathbf{M}$ is skew-symmetric and $\mathbf{y}'\mathbf{M}\mathbf{y} \equiv 0$, so that

$$\dot{V} = -\mathbf{y}'\mathbf{K}_d\mathbf{y}.$$

To complete the example, note that $\dot{V} \leq 0 \; \forall (\mathbf{x}, \mathbf{y}) \in \mathbf{R}^n \times \mathbf{R}^n$. Moreover, the set R of (81) is given by

$$R = \{(\mathbf{x}, \mathbf{y}): \mathbf{y} = \mathbf{0}\} = \mathbf{R}^n \times \{\mathbf{0}\}.$$

Suppose $[\mathbf{x}(t), \mathbf{y}(t)]$ is a trajectory that lies entirely in R. Then

$$\dot{\mathbf{y}} \equiv \mathbf{0} \Rightarrow \mathbf{K}_p(\mathbf{x} - \mathbf{q}_d) \equiv \mathbf{0} \Rightarrow \mathbf{x}(t) = \mathbf{q}_d, \; \forall t \geq 0.$$

Hence R contains no trajectories of the system other than the equilibrium $(\mathbf{q}_d, \mathbf{0})$. It now follows from Theorem (79) that this equilibrium is globally asymptotically stable.

5.3.3 Theorems on Instability

In the two preceding subsections, we presented sufficient conditions for stability and for asymptotic stability. This subsection contains several sufficient conditions for instability.

88 Theorem *The equilibrium* $\mathbf{0}$ *of (5.1.1) is unstable if there exist a* C^1 *decrescent function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *and a time* $t_0 \geq 0$ *such that (i)* $\dot{V}$ *is an lpdf, (ii)* $V(t, \mathbf{0}) = 0 \; \forall t \geq t_0$, *and (iii) there exist points* $\mathbf{x}_0 \neq \mathbf{0}$ *arbitrarily close to* $\mathbf{0}$ *such that* $V(t_0, \mathbf{x}_0) \geq 0$.

Proof To demonstrate that $\mathbf{0}$ is an unstable equilibrium, it must be shown that, for some $\varepsilon > 0$, no $\delta > 0$ exists such that (5.1.10) holds. Since $\dot{V}$ is an lpdf and V is decrescent, there exist a constant $r > 0$, and functions β, γ of class K such that

89 $$V(t, \mathbf{x}) \leq \beta(\|\mathbf{x}\|), \; \dot{V}(t, \mathbf{x}) \geq \gamma(\|\mathbf{x}\|), \; \forall t \geq 0, \; \forall \mathbf{x} \in B_r.$$

It is now shown that, if we let $\varepsilon = r$, then no matter how small we choose $\delta > 0$, there always exists a corresponding $\mathbf{x}_0$ in B_δ such that $\|\mathbf{s}(t, t_0, \mathbf{x}_0)\|$ eventually equals or exceeds ε. Given any $\delta > 0$, pick an $\mathbf{x}_0 \neq \mathbf{0}$ in B_δ such that $V(t, \mathbf{x}_0) \geq 0$; such an $\mathbf{x}_0$ exists by condition (iii). In the interests of brevity, let $\mathbf{x}(t)$ denote $\mathbf{s}(t, t_0, \mathbf{x}_0)$. Then initially $\dot{V}(t_0, \mathbf{x}_0) > 0$, so there exists a $t_1 \geq t_0$ such that $V[t_1, \mathbf{x}(t_1)] =: c > 0$. To show that eventually $\|\mathbf{x}(t)\| \geq \varepsilon$, suppose by way of contradiction that $\mathbf{x}(t) \in B_r \; \forall t \geq t_1$. Then (89) implies that $V[t, \mathbf{x}(t)] \leq \beta(\varepsilon) \; \forall t \geq t_1$. Also, since $\dot{V}(t, \mathbf{x}) \geq 0 \; \forall t, \; \forall \mathbf{x} \in B_r$, it follows that $V[t, \mathbf{x}(t)] \geq V[t_1, \mathbf{x}(t_1)] = c \; \forall t \geq t_1$. This in turn implies, from (89), that $\|\mathbf{x}(t)\| \geq \beta^{-1}(c) \; \forall t \geq t_1$, and that $\dot{V}[t, \mathbf{x}(t)] \geq \gamma[\beta^{-1}(c)] =: d > 0$. Now combining all of these inequalities shows that

90 $$\beta(\varepsilon) \geq V[t, \mathbf{x}(t)] = V[t_1, \mathbf{x}(t_1)] + \int_{t_1}^{t} \dot{V}[t, \mathbf{x}(t)]\, dt \geq c + (t - t_1)d, \ \forall t \geq t_1.$$

However, the inequality (90) is absurd, since the right side is an unbounded function of t while the left side is a fixed constant. This contradiction shows that the assumption is false, i.e. it is *not* true that $\|\mathbf{x}(t)\| < \varepsilon \ \forall t \geq t_1$. In other words, there is a time $t \geq t_1$ at which $\|\mathbf{x}(t)\| \geq \varepsilon$. This shows that the equilibrium $\mathbf{0}$ is unstable. ■

Note that, in contrast with previous theorems, the Lyapunov function V in the present theorem can assume both positive as well as negative values. Also, the inequality (89) requiring that V be a decrescent function places no restrictions on the behavior of $V(t, \mathbf{x})$ when it assumes negative values.

91 Example Consider the system of equations

$$\dot{x}_1 = x_1 - x_2 + x_1 x_2, \dot{x}_2 = -x_2 - x_2^2,$$

and choose the Lyapunov function candidate

$$V(x_1, x_2) = (2x_1 - x_2)^2 - x_2^2.$$

Even though V assumes both positive and negative values, it has the requisite property that it assumes nonnegative values arbitrarily close to the origin. Hence it is a suitable Lyapunov function candidate for applying Theorem (88). Differentiating V gives

$$\dot{V}(x_1, x_2) = 2(2x_1 - x_2)(2\dot{x}_1 - \dot{x}_2) - 2x_2\dot{x}_2 = [(2x_1 - x_2)^2 + x_2^2](1 + x_2).$$

Thus $\dot{V}$ is an lpdf over the ball B_{1-d} for each $d \in (0, 1)$, and all conditions of Theorem (88) are satisfied. It follows that $\mathbf{0}$ is an unstable equilibrium. ■

Remarks Some authors prove a less efficient version of Theorem (88) by showing that $\mathbf{0}$ is an unstable equilibrium if one can find a C^1 function V such that both V and $\dot{V}$ are lpdf's. Actually, it can be shown that if one can find such a function V, then the origin is a **completely unstable** equilibrium; in other words, there exist an $\varepsilon > 0$ and an $r > 0$ such that *every* trajectory starting in B_r other than the trivial trajectory eventually leaves the ball B_ε. While such instability occurs sometimes (e.g., the Van der Pol oscillator), this particular instability theorem is much less useful than Theorem (88).

Alternate sufficient conditions for instability are given by Theorem (92) below and Theorem (99) following.

92 Theorem *The equilibrium* $\mathbf{0}$ *of (5.1.1) is unstable if there exist a* C^1 *function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *and a constant* $r > 0$ *such that (i)* V *is decrescent, (ii)* $V(0, \mathbf{0}) = 0$ *and* $V(0, \cdot)$ *assumes positive values arbitrarily close to the origin, (iii) there exist a positive constant* λ *and a function* $W: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *such that*

93 $\dot{V}(t, \mathbf{x}) = \lambda V(t, \mathbf{x}) + W(t, \mathbf{x})$, and

94 $W(t, \mathbf{x}) \geq 0, \forall t \geq 0, \forall \mathbf{x} \in B_r$.

Proof It is shown that if we choose $\varepsilon = r$, then (5.1.10) cannot be satisfied for any choice of $\delta > 0$. Given $\delta > 0$, choose $\mathbf{x}_0 \neq 0$ in B_δ such that $V(0, \mathbf{x}_0) > 0$, and let $\mathbf{x}(t)$ denote the resulting solution trajectory $\mathbf{s}(t, 0, \mathbf{x}_0)$. Then, whenever $\mathbf{x}(t) \in B_r$, we have

95 $$\frac{d}{dt} V[t, \mathbf{x}(t)] = \lambda V[t, \mathbf{x}(t)] + W[t, \mathbf{x}(t)] \geq \lambda V[t, \mathbf{x}(t)],$$

and therefore

96 $$\frac{d}{dt} \{\exp(-\lambda t) V[t, \mathbf{x}(t)]\} \geq 0.$$

Hence

97 $V[t, \mathbf{x}(t)] \geq V(0, \mathbf{x}_0) \exp(\lambda t)$.

Since the function on the right side is unbounded, $\mathbf{x}(t)$ must eventually leave B_r. Therefore $\mathbf{0}$ is an unstable equilibrium. ■

98 **Example** Consider the system of equations

$$\dot{x}_1 = x_1 + 2x_2 + x_1 x_2^2, \dot{x}_2 = 2x_1 + x_2 - x_1^2 x_2.$$

Let

$$V(x_1, x_2) = x_1^2 - x_2^2.$$

Then V is a suitable Lyapunov function candidate for applying Theorem (92). Now

$$\dot{V}(t, \mathbf{x}) = 2x_1 \dot{x}_1 - 2x_2 \dot{x}_2 = 2x_1^2 - 2x_2^2 + 4x_1^2 x_2^2 = 2V(x_1, x_2) + 4x_1^2 x_2^2.$$

Since $W(x_1, x_2) := 4x_1^2 x_2^2 \geq 0 \, \forall \mathbf{x}$, all conditions of Theorem (92) are satisfied, and $\mathbf{0}$ is an unstable equilibrium. ■

In Theorems (88) and (92), the function V is required to satisfy certain conditions at all points belonging to some neighborhood of the origin. In contrast, the various conditions in Theorem (99) below are only required to hold in a region for which the origin is a *boundary* point. This theorem is generally known as Chetaev's theorem.

99 **Theorem (Chetaev)** *The equilibrium* $\mathbf{0}$ *is unstable if there exist a* C^1 *function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$, *a ball* B_r, *an open set* $\Omega \subseteq B_r$ *and a function* γ *of class K such that*

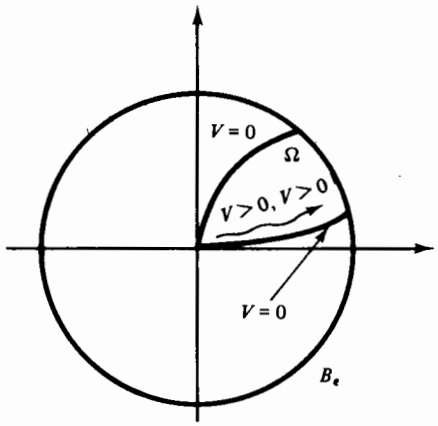


Fig. 5.14

100 $0 < V(t, \mathbf{x}), \forall t \geq 0, \forall \mathbf{x} \in \Omega,$

101 $\sup_{t \geq 0} \sup_{\mathbf{x} \in \Omega} V(t, \mathbf{x}) < \infty,$

102 $0 \in \partial\Omega$ (the boundary of Ω),

103 $V(t, \mathbf{x}) = 0, \forall t \geq 0, \forall \mathbf{x} \in \partial\Omega \cap B_r,$

104 $\dot{V}(t, \mathbf{x}) \geq \gamma(\|\mathbf{x}\|), \forall t \geq 0, \forall \mathbf{x} \in \Omega.$

Proof The situation can be depicted as shown in Figure 5.14. The assumptions on V and $\dot{V}$ imply that, along any nontrivial trajectory starting inside Ω, $V[t, \mathbf{x}(t)]$ increases indefinitely so long as the trajectory stays inside Ω. Since $V = 0$ on $\partial\Omega \cap B_r$, the trajectory cannot escape Ω by moving across $\partial\Omega \cap B_r$. Hence the trajectory must eventually reach the boundary of B_r itself, irrespective of its starting point. This shows that $\mathbf{0}$ is an unstable equilibrium. ■

105 Example This is a continuation of Example (19), regarding the spinning of a rigid body. The motion is described by

$$\dot{x} = ayz, \dot{y} = -bxz, \dot{z} = cxy.$$

Consider an equilibrium of the form $(0, y_0, 0)$, where $y_0 \neq 0$; say $y_0 > 0$ to be specific. First let us translate the coordinates so that the equilibrium under study becomes the origin. This is achieved by rewriting the system equations as

$$\dot{x} = a(y - y_0) + ay_0 z, \dot{y} - \dot{y}_0 = -bxz, \dot{z} = cx(y - y_0) + cxy_0.$$

If $y - y_0$ is denoted by y_s, then the equations become

$$\dot{x} = ay_s z + ay_0 z,\ \dot{y}_s = -bxz,\ \dot{z} = cxy_s + cxy_0.$$

Now apply Theorem (99) with

$$V(x, y, z) = xz,$$

$$B_r = \{(x, y_s, z): x^2 + y_s^2 + z^2 < y_0^2\},$$

$$\Omega = \{(x, y_s, z) \in B_{r/2}: x > 0 \text{ and } z > 0\}.$$

Then Ω is open, and

$$\partial\Omega \cap B_r = \{(x, y_s, z) \in \bar{B}_{r/2}: x = 0 \text{ or } z = 0\},$$

where $\bar{B}_{r/2}$ denotes the closed ball of radius $r/2$. Hence conditions (100) to (103) are satisfied. Finally,

$$\dot{V} = x\dot{z} + \dot{x}z = 2(y_s + y_0)(cx^2 + az^2).$$

If $(x, y_s, z) \in \Omega$, then $y_s + y_0 > 0$, so that (104) is also satisfied. Hence, by Theorem (99), the origin (in the new coordinate system) is an unstable equilibrium.

5.3.4 Concluding Remarks

In this section, several theorems in Lyapunov stability theory have been presented. The favorable aspects of these theorems are:

1. They enable one to draw conclusions about the stability status of an equilibrium *without solving the system equations.*
2. Especially in the theorems on stability and asymptotic stability, the Lyapunov function V has an intuitive appeal as the total energy of the system.

The unfavorable aspects of these theorems are:

3. They represent only sufficient conditions for the various forms of stability. Thus, if a particular Lyapunov function candidate V fails to satisfy the hypotheses on $\dot{V}$, then no conclusions can be drawn, and one has to begin anew with another Lyapunov function candidate.
4. In a general system of nonlinear equations, which do not have the structure of Hamiltonian equations of motion or some other such structure, there is no *systematic* procedure for generating Lyapunov function candidates.

Thus the reader is justified in asking what the role of Lyapunov theory is today. Two remarks may be made in response.

(1) Originally Lyapunov stability theory was advanced as a means of *testing* the stability status of a given system. Nowadays, however, it is increasingly being used to *guarantee* stability. For example, in adaptive control or PD stabilization of robots [see Example (84)], one *first* chooses a Lyapunov function candidate, and then chooses the adaptation law or the control law to *ensure* that the hypotheses of a particular stability theorem are satisfied. In this way, the problem of searching for a Lyapunov function is alleviated. (2) Though the various theorems given here are only sufficient conditions, it is possible to prove so-called *converse theorems*, which state that if the equilibrium has a particular property, then *there exists* a suitable Lyapunov function that would enable us to deduce this property. Usually this Lyapunov function is specified in terms of the solution trajectories of the system, and can be used in perturbational analysis. Roughly speaking, the line of reasoning goes like this: Begin with a system which is easy to analyze (such as a linear system; cf. the next section). Construct a Lyapunov function for the same. Now see under what conditions the same Lyapunov function candidate continues to work for a modified system. We shall see several examples of such an approach in this chapter and the next.

Problem 5.11 For the system of Example (19), suppose it is desired to analyze the stability of an equilibrium of the form $(x_0, 0, 0)$ where $x_0 \neq 0$. Set up a new set of coordinates such that the equilibrium under study is the origin of the new set. Define a suitable Lyapunov function such that the stability of the equilibrium can be established by applying Theorem (14). Repeat for an equilibrium of the form $(0, 0, z_0)$ where $z_0 \neq 0$.

Problem 5.12 Analyze the circuit of Example (48) (Figure 5.7) when the capacitances are also nonlinear. Let q_i denote the charge across the i-th capacitor, and suppose q_i is a (possibly nonlinear) function of the voltage x_i. Assume that $q_i(x_i) = 0$, and define $C_i(x_i) = \partial q_i(x_i)/\partial x_i$. Suppose there exists constants α_i and β_i such that

$$0 < \alpha_i \leq C_i(x_i) \leq \beta_i, \text{ for } i = 1, \cdots, n.$$

Using the total energy stored in the capacitors as a Lyapunov function candidate, analyze the stability of the equilibrium $\mathbf{x} = \mathbf{0}$.

Problem 5.13 Suppose a particle of mass m is moving in a smooth potential field. To simplify the problem, suppose the motion is one-dimensional. Let x denote the position coordinate of the particle, and let $\phi(x)$ denote the potential energy at x. If the only force acting on the particle is due to the potential, then the motion of the particle is described by

$$m\ddot{x} = -\phi'(x) =: f(x),$$

where the prime denotes differentiation with respect to x. Show that every local minimum of the function ϕ is a stable equilibrium.

Problem 5.14 Consider the autonomous differential equation

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

and suppose $\mathbf{f}$ is a C^1 function such that $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. Then there exists a C^1 matrix-valued

function **A** such that [cf. Lemma (2.5.17)]

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\,\mathbf{x},\ \forall \mathbf{x} \in \mathbf{R}^n.$$

(a) Show that if the matrix $\mathbf{A}'(\mathbf{0}) + \mathbf{A}(\mathbf{0})$ is negative definite, then the origin is an exponentially stable equilibrium. More generally, show that if there exists a positive definite matrix $\mathbf{P}$ such that $\mathbf{A}'(\mathbf{0})\mathbf{P} + \mathbf{P}\mathbf{A}(\mathbf{0})$ is negative definite, then the origin is an exponentially stable equilibrium. (Hint: Consider the Lyapunov function candidate $V(\mathbf{x}) = \|\mathbf{x}\|^2$.)

(b) Extend the results in (a) to global stability.

Problem 5.15 Consider the differential equation

$$\ddot{y}(t) + f[y(t)]\,\dot{y}(t) + g[y(t)] = 0.$$

Transform this equation into state variable form by choosing $\dot{x}_1 = y, \dot{x}_2 = \dot{y}$.

(a) Suppose the functions f and g are continuous and satisfy the following conditions for some positive number δ:

$$\sigma g(\sigma) > 0,\ \forall \sigma \in (-\delta, \delta),$$

$$f(\sigma) \geq 0,\ \forall \sigma \in (-\delta, \delta).$$

Show that the equilibrium **0** is stable.

(b) Suppose that the condition on f is strengthened to

$$f(\sigma) > 0,\ \forall \sigma \in (-\delta, \delta).$$

Show that the equilibrium **0** is asymptotically stable.

(c) Show that if, in addition to the conditions in (b), both f and g are continuously differentiable, then the equilibrium **0** is exponentially stable.

(d) Find suitable conditions on f and g to ensure global asymptotic stability and global exponential stability.

Problem 5.16 Consider the system

$$\dot{x}_1 = x_1 + 2x_2^2,\ \dot{x}_2 = 2x_1x_2 + x_2^2.$$

Using the Lyapunov function candidate

$$V(\mathbf{x}) = x_1^2 - x_2^2,$$

show that **0** is an unstable equilibrium.

Problem 5.17 Consider the system

$$\dot{x}_1 = x_1^2 - x_1 x_2, \dot{x}_2 = -x_1^2 - 2x_1 x_2.$$

Using the Lyapunov function candidate

$$V(\mathbf{x}) = x_1(x_1 - x_2),$$

and Theorem (92), show that $\mathbf{0}$ is an unstable equilibrium.

5.4 STABILITY OF LINEAR SYSTEMS

In this section we study the Lyapunov stability of systems described by linear vector differential equations. The results presented here not only enable us to obtain necessary and sufficient conditions for the stability of linear systems, but also pave the way to deriving Lyapunov's linearization method, which is presented in the next section.

5.4.1 Stability and the State Transition Matrix

Consider a system described by the linear vector differential equation

1 $$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\,\mathbf{x}(t),\ t \geq 0.$$

The system (1) is autonomous if $\mathbf{A}(\cdot)$ is constant as a function of time; otherwise it is nonautonomous. It is clear that $\mathbf{0}$ is always an equilibrium of the system (1). Further, $\mathbf{0}$ is an isolated equilibrium if $\mathbf{A}(t)$ is nonsingular for some $t \geq 0$. The general solution of (1) is given by

2 $$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\,\mathbf{x}(t_0),$$

where $\mathbf{\Phi}(\cdot, \cdot)$ is the state transition matrix associated with $\mathbf{A}(\cdot)$ and is the unique solution of the equation

3 $$\frac{d}{dt}\mathbf{\Phi}(t, t_0) = \mathbf{A}(t)\,\mathbf{\Phi}(t, t_0),\ \forall t \geq t_0 \geq 0,$$

4 $$\mathbf{\Phi}(t_0, t_0) = I,\ \forall t_0 \geq 0.$$

With the aid of this explicit characterization of the solutions of (1), it is possible to derive some useful conditions for the stability of the equilibrium $\mathbf{0}$. Since these conditions involve the state transition matrix $\mathbf{\Phi}$, they are not of much computational value, because in general it is impossible to derive an analytical expression for $\mathbf{\Phi}$. Nevertheless, they are of conceptual value, enabling one to understand the mechanisms of stability and instability in linear systems.

5 Theorem *The equilibrium* $\mathbf{0}$ *is stable if and only if for each* $t_0 \geq 0$ *it is true that*

$$\mathbf{6} \qquad \sup_{t \geq t_0} \|\Phi(t, t_0)\|_i =: m(t_0) < \infty,$$

where $\|\cdot\|_i$ *denotes the induced norm of a matrix.*

Proof "If" Suppose (6) is true, and let $\varepsilon > 0$, $t_0 \geq 0$ be specified. If we define $\delta(\varepsilon, t_0)$ as $\varepsilon/m(t_0)$, then

$$\mathbf{7} \qquad \|\mathbf{x}(t_0)\| < \delta \Rightarrow \|\mathbf{x}(t)\| = \|\Phi(t, t_0)\mathbf{x}(t_0)\| \leq \|\Phi(t, t_0)\|_i \ \|\mathbf{x}(t_0)\|$$

$$< m(t_0)\delta = \varepsilon,$$

so that (5.1.10) is satisfied. This shows that the equilibrium $\mathbf{0}$ is stable.

"Only if" Suppose (6) is false, so that $\|\Phi(t, t_0)\|_i$ is an unbounded function of t for some $t_0 \geq 0$. To show that $\mathbf{0}$ is an unstable equilibrium, let $\varepsilon > 0$ be any positive number, and let δ be an arbitrary positive number. It is shown that one can choose an $\mathbf{x}(t_0)$ in the ball B_δ such that the resulting solution $\mathbf{x}(t)$ satisfies $\|\mathbf{x}(t)\| \geq \varepsilon$ for some $t \geq t_0$. Select a δ_1 in the open interval $(0, \delta)$. Since $\|\Phi(t, t_0)\|_i$ is unbounded as a function of t, there exists a $t \geq t_0$ such that

$$\mathbf{8} \qquad \|\Phi(t, t_0)\|_i > \frac{\varepsilon}{\delta_1}.$$

Next, select a vector $\mathbf{v}$ of norm one such that

$$\mathbf{9} \qquad \|\Phi(t, t_0)\mathbf{v}\| = \|\Phi(t, t_0)\|_i.$$

This is possible in view of the definition of the induced matrix norm. Finally, let $\mathbf{x}(t_0) = \delta_1 \mathbf{v}$. Then $\mathbf{x} \in B_\delta$. Moreover,

$$\mathbf{10} \qquad \|\mathbf{x}(t)\| = \|\Phi(t, t_0)\mathbf{x}(t_0)\| = \|\delta_1 \Phi(t, t_0)\mathbf{v}\| = \delta_1 \|\Phi(t, t_0)\|_i > \varepsilon.$$

Hence the equilibrium $\mathbf{0}$ is unstable. ■

Remark: Note that, in the case of linear systems, the instability of the equilibrium $\mathbf{0}$ does indeed imply that some solution trajectories actually "blow up." This is in contrast to the case of nonlinear systems, where the instability of $\mathbf{0}$ can be accompanied by the boundedness of all solutions, as in the Van der Pol oscillator [see Example (5.1.25)].

Necessary and sufficient conditions for uniform stability are given next.

11 Theorem *The equilibrium* $\mathbf{0}$ *is uniformly stable if and only if*

12 $m_0 := \sup_{t_0 \geq 0} m(t_0) = \sup_{t_0 \geq 0} \sup_{t \geq t_0} \|\Phi(t, t_0)\|_i < \infty.$

Proof "If" Suppose m_0 is finite; then, for any $\varepsilon > 0$ and any $t_0 \geq 0$, (5.1.11) is satisfied with $\delta = \varepsilon/m_0$.

"Only if" Suppose $m(t_0)$ is unbounded as a function of t_0. Then at least one component of $\Phi(\cdot, \cdot)$, say the ij-th component, has the property that

13 $\sup_{t \geq t_0} |\phi_{ij}(t, t_0)|$ is unbounded as a function of t_0.

Let $\mathbf{x}_0 = \mathbf{e}_j$, the elementary vector with a 1 in the j-th row and zeros elsewhere. Then (13) implies that the quantity $\|\mathbf{x}(t)\|/\|\mathbf{x}_0\| = \|\Phi(t, t_0)\mathbf{x}_0\|/\|\mathbf{x}_0\|$ cannot be bounded independently of t_0. Hence $\mathbf{0}$ is not a uniformly stable equilibrium. ■

The next theorem characterizes uniform asymptotic stability.

14 Theorem *The equilibrium* $\mathbf{0}$ *is (globally) uniformly asymptotically stable if and only if*

15 $\sup_{t_0 \geq 0} \sup_{t \geq t_0} \|\Phi(t, t_0)\|_i < \infty,$

16 $\|\Phi(t_0 + t, t_0)\|_i \to 0$ as $t \to \infty$, uniformly in t_0.

Remark: The condition (16) can be expressed equivalently as follows: For each $\varepsilon > 0$, there exists a $T = T(\varepsilon)$ such that

17 $\|\Phi(t_0 + t, t_0)\|_i < \varepsilon, \forall t \geq T, \forall t_0 \geq 0.$

Proof "If" By Theorem (11), if (15) holds then the equilibrium $\mathbf{0}$ is uniformly stable. Similarly, if (16) holds, then the ratio $\|\Phi(t, t_0)\mathbf{x}(t_0)\|/\|\mathbf{x}(t_0)\|$ approaches zero uniformly in t_0, so that $\mathbf{0}$ is uniformly attractive. Hence, by definition, $\mathbf{0}$ is uniformly asymptotically stable.

"Only if" This part of the proof is left as an exercise (see Problem 5.18) ■

Theorem (18) below shows that, for linear systems, uniform asymptotic stability is equivalent to exponential stability.

18 Theorem *The equilibrium* $\mathbf{0}$ *is uniformly asymptotically stable if and only if there exist constants* $m, \lambda > 0$ *such that*

19 $\|\Phi(t, t_0)\|_i \leq m \exp[-\lambda(t - t_0)], \forall t \geq t_0 \geq 0.$

Proof "If" Suppose (19) is satisfied. Then clearly (15) and (16) are also true, whence $\mathbf{0}$ is uniformly asymptotically stable by Theorem (14).

"Only if" Suppose (15) and (16) are true. Then there exist finite constants μ and T such that

20 $\|\Phi(t, t_0)\|_i \le \mu, \forall t \ge t_0 \ge 0,$

21 $\|\Phi(t_0 + t, t_0)\|_i \le 1/2, \forall t \ge T, \forall t_0 \ge 0.$

In particular, (21) implies that

22 $\|\Phi(t_0 + T, t_0)\|_i \le 1/2, \forall t_0 \ge 0.$

Now, given any t_0 and any $t \ge t_0$, pick an integer k such that $t_0 + kT \le t < t_0 + (k+1)T$. Then

23 $\Phi(t, t_0) = \Phi(t, t_0 + kT)\,\Phi(t_0 + kT, t_0 + kT - T) \cdots \Phi(t_0 + T, t_0).$

Hence

24 $$\|\Phi(t, t_0)\|_i \le \|\Phi(t, t_0 + kT)\|_i \cdot \prod_{j=1}^{k} \|\Phi(t_0 + jT, t_0 + jT - T)\|_i,$$

where the empty product is taken as one. Now repeated application of (20) and (22) gives

25 $\|\Phi(t, t_0)\|_i \le \mu\, 2^{-k} \le (2\mu)\, 2^{-(t - t_0)/T}.$

Hence (19) is satisfied if we define

26 $m = 2\mu,$

27 $\lambda = \dfrac{\log 2}{T}.$

This completes the proof. ■

In conclusion, this subsection contains several results that relate the stability properties of a linear system to its state transition matrix. Since these results require an explicit expression for the state transition matrix, they are not of much use for testing purposes. Nevertheless, they do provide some insight. For example, Theorem (18), which shows that uniform asymptotic stability is equivalent to exponential stability, is not very obvious on the surface.

5.4.2 Autonomous Systems

Throughout this subsection, attention is restricted to linear *autonomous* systems of the form

28 $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t).$

In this special case, Lyapunov theory is very complete, as we shall see.

29 Theorem *The equilibrium* $\mathbf{0}$ *of (28) is (globally) exponentially stable if and only if all eigenvalues of* $\mathbf{A}$ *have negative real parts. The equilibrium* $\mathbf{0}$ *of (28) is stable if and only if all eigenvalues of* $\mathbf{A}$ *have nonpositive real parts, and in addition, every eigenvalue of* $\mathbf{A}$ *having a zero real part is a simple zero of the minimal polynomial of* $\mathbf{A}$.

Proof The state transition matrix $\Phi(t, t_0)$ of the system (28) is given by

$$\Phi(t, t_0) = \exp[\mathbf{A}(t - t_0)], \tag{30}$$

where $\exp(\cdot)$ is the matrix exponential. Furthermore, $\exp(\mathbf{A}t)$ can be expressed as

$$\exp(\mathbf{A}t) = \sum_{i=1}^{r} \sum_{j=1}^{m_i} p_{ij}(\mathbf{A})\, t^{j-1} \exp(\lambda_i t), \tag{31}$$

where r is the number of distinct eigenvalues of $\mathbf{A}$; $\lambda_1, \cdots, \lambda_r$ are the distinct eigenvalues; m_i is the multiplicity of the eigenvalue λ_i; and p_{ij} are interpolating polynomials. The stated conditions for stability and for asymptotic stability now follow readily from Theorems (5) and (14) respectively. ■

Thus, in the case of linear time-invariant systems of the form (28), the stability status of the equilibrium $\mathbf{0}$ can be ascertained by studying the eigenvalues of $\mathbf{A}$. However, it is possible to formulate an entirely different approach to the problem, based on the use of quadratic Lyapunov functions. This theory is of interest in itself, and is also useful in studying nonlinear systems using linearization methods (see Section 5.5).

Given the system (28), the idea is to choose a Lyapunov function candidate of the form

$$V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x}, \tag{32}$$

where $\mathbf{P}$ is a real symmetric matrix. Then $\dot{V}$ is given by

$$\dot{V}(\mathbf{x}) = \dot{\mathbf{x}}'\mathbf{P}\mathbf{x} + \mathbf{x}'\mathbf{P}\dot{\mathbf{x}} = -\mathbf{x}'\mathbf{Q}\mathbf{x}, \tag{33}$$

where

$$\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}. \tag{34}$$

Equation (34) is commonly known as the **Lyapunov Matrix Equation**. By means of this equation, it is possible to study the stability properties of the equilibrium $\mathbf{0}$ of the system (28). For example, if a pair of matrices $(\mathbf{P}, \mathbf{Q})$ satisfying (34) can be found such that both $\mathbf{P}$ and $\mathbf{Q}$ are positive definite, then both V and $-\dot{V}$ are positive definite functions, and V is radially unbounded. Hence, by Theorem (5.3.45), the equilibrium $\mathbf{0}$ is globally exponentially stable. On the other hand, if a pair $(\mathbf{P}, \mathbf{Q})$ can be found such that $\mathbf{Q}$ is positive definite and $\mathbf{P}$ has at least one nonpositive eigenvalue, then $-\dot{V}$ is positive definite, and V assumes nonpositive values arbitrarily close to the origin. Hence, by Theorem (5.3.88), the origin is an unstable equilibrium.

This, then, is the rationale behind studying Equation (34). There are two possible ways in which (34) can be tackled: (1) Given a particular matrix $\mathbf{A}$, one can pick a particular matrix $\mathbf{P}$ and study the properties of the matrix $\mathbf{Q}$ resulting from (34). (2) Given $\mathbf{A}$, one can pick $\mathbf{Q}$ and study the matrix $\mathbf{P}$ resulting from (34). The latter approach is adopted here, for two reasons — one pragmatic and the other philosophical. The pragmatic reason is that the second approach is the one for which the theory is better developed. On a more philosophical level, one can reason as follows: Given a matrix $\mathbf{A}$, we presumably do not know ahead of time whether or not $\mathbf{0}$ is a stable equilibrium. If we pick $\mathbf{P}$ and study the resulting $\mathbf{Q}$, we would be obliged (because of the available stability theorems) to make an *a priori* guess as to the stability status of $\mathbf{0}$. If we believe that $\mathbf{0}$ is asymptotically stable, then we should pick $\mathbf{P}$ to be positive definite, whereas if we believe that $\mathbf{0}$ is unstable, we should pick $\mathbf{P}$ to be indefinite or even negative definite. On the other hand, if we were to pick $\mathbf{Q}$, there is no need to make such an *a priori* guess as to the stability status of $\mathbf{0}$. The matrix $\mathbf{Q}$ should be always chosen to be positive definite. If the resulting matrix $\mathbf{P}$ is positive definite, then $\mathbf{0}$ is exponentially stable, by Theorem (5.3.45). If, on the other hand, $\mathbf{P}$ turns out to have at least one nonpositive eigenvalue, then $\mathbf{0}$ is an unstable equilibrium, by Theorem (5.3.88).

One difficulty with selecting $\mathbf{Q}$ and trying to find the corresponding $\mathbf{P}$ is that, depending on the matrix $\mathbf{A}$, (34) may not have a unique solution for $\mathbf{P}$. The next result gives necessary and sufficient conditions under which (34) has a unique solution corresponding to each $\mathbf{Q}$.

35 Lemma *Let $\mathbf{A} \in \mathbf{R}^n$, and let $\{\lambda_1, \cdots, \lambda_n\}$ denote the (not necessarily distinct) eigenvalues of $\mathbf{A}$. Then (34) has a unique solution for $\mathbf{P}$ corresponding to each $\mathbf{Q} \in \mathbf{R}^{n \times n}$ if and only if*

$$\textbf{36} \quad \lambda_i + \lambda_j \neq 0, \forall i, j.$$

The proof of this lemma is not difficult, but requires some concepts from linear algebra not heretofore covered. The interested reader is referred to Chen (1984), Appendix F, or to any other standard book on matrix theory which discusses the "Sylvester Equation," which is a generalization of the Lyapunov matrix equation.

On the basis of Lemma (35), one can state the following corollary.

37 Corollary *If for some choice of $\mathbf{Q} \in \mathbf{R}^{n \times n}$ Equation (34) does not have a unique solution for $\mathbf{P}$, then the origin is not an asymptotically stable equilibrium.*

Proof If all eigenvalues of $\mathbf{A}$ have negative real parts, then (36) is satisfied. ■

The following lemma provides an alternate characterization of the solutions of (34). Note that a matrix $\mathbf{A}$ is called **Hurwitz** if all of its eigenvalues have negative real parts. The terminology arises from the fact that, if all eigenvalues of $\mathbf{A}$ have negative real parts, then the characteristic polynomial of $\mathbf{A}$ is a Hurwitz polynomial.

38 Lemma *Let* $\mathbf{A}$ *be a Hurwitz matrix. Then, for each* $\mathbf{Q} \in \mathbf{R}^{n \times n}$, *the corresponding unique solution of (34) is given by*

$$\mathbf{39} \quad \mathbf{P} = \int_0^\infty e^{\mathbf{A}'t} \mathbf{Q} e^{\mathbf{A}t}\, dt.$$

Proof If $\mathbf{A}$ is Hurwitz, then the condition (36) is satisfied, and (34) has a unique solution for $\mathbf{P}$ corresponding to each $\mathbf{Q} \in \mathbf{R}^{n \times n}$. Moreover, if $\mathbf{A}$ is Hurwitz, then the integral on the right side of (39) is well-defined. Let $\mathbf{M}$ denote this integral. It is now shown that

$$\mathbf{40} \quad \mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} = -\mathbf{Q}.$$

By the uniqueness of solutions to (34), it then follows that $\mathbf{P} = \mathbf{M}$.

To prove (40), observe that

$$\mathbf{41} \quad \mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} = \int_0^\infty [\mathbf{A}' e^{\mathbf{A}'t} \mathbf{Q} e^{\mathbf{A}t} + e^{\mathbf{A}'t} \mathbf{Q} e^{\mathbf{A}t} \mathbf{A}]\, dt$$

$$= \int_0^\infty d[e^{\mathbf{A}'t} \mathbf{Q} e^{\mathbf{A}t}] = \left[e^{\mathbf{A}'t} \mathbf{Q} e^{\mathbf{A}t} \right]_0^\infty$$

$$= -\mathbf{Q}.$$

This completes the proof. ■

Note that the above lemma also provides a convenient way to compute infinite integrals of the form (39).

We can now state one of the main results for the Lyapunov matrix equation.

42 Theorem *Given a matrix* $\mathbf{A} \in \mathbf{R}^{n \times n}$, *the following three statements are equivalent:*

(1) $\mathbf{A}$ *is a Hurwitz matrix.*

(2) *There exists* **some** *positive definite matrix* $\mathbf{Q} \in \mathbf{R}^{n \times n}$ *such that (34) has a corresponding unique solution for* $\mathbf{P}$, *and this* $\mathbf{P}$ *is positive definite.*

(3) *For* **every** *positive definite matrix* $\mathbf{Q} \in \mathbf{R}^{n \times n}$, *(34) has a unique solution for* $\mathbf{P}$, *and this solution is positive definite.*

Proof "(3) $\Rightarrow$ (2)" Obvious.

"(2) $\Rightarrow$ (1)" Suppose (2) is true for some particular matrix $\mathbf{Q}$. Then we can apply Theorem (5.3.25) with the Lyapunov function candidate $V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x}$. Then $\dot{V}(\mathbf{x}) = -\mathbf{x}'\mathbf{Q}\mathbf{x}$, and one can conclude that $\mathbf{0}$ is asymptotically stable equilibrium. By Theorem (29), this implies that $\mathbf{A}$ is Hurwitz.

"(1) $\Rightarrow$ (3)" Suppose $\mathbf{A}$ is Hurwitz and let $\mathbf{Q} \in \mathbf{R}^{n\times n}$ be positive definite but otherwise arbitrary. By Lemma (38), Equation (34) has a corresponding unique solution $\mathbf{P}$ given by (39). It only remains to show that $\mathbf{P}$ is positive definite. For this purpose, factor $\mathbf{Q}$ in the form $\mathbf{M}'\mathbf{M}$ where $\mathbf{M}$ is nonsingular. Now it is claimed that $\mathbf{P}$ is positive definite because

43 $\mathbf{x}'\mathbf{P}\mathbf{x} > 0, \forall \mathbf{x} \neq \mathbf{0}.$

With $\mathbf{Q} = \mathbf{M}'\mathbf{M}$, $\mathbf{P}$ becomes

44 $$\mathbf{P} = \int_0^\infty e^{\mathbf{A}'t}\mathbf{M}'\mathbf{M}e^{\mathbf{A}t}\,dt.$$

Thus, for any $\mathbf{x} \in \mathbf{R}^n$,

45 $$\mathbf{x}'\mathbf{P}\mathbf{x} = \int_0^\infty \mathbf{x}'e^{\mathbf{A}'t}\mathbf{M}'\mathbf{M}e^{\mathbf{A}t}\mathbf{x}\,dt = \int_0^\infty \|\mathbf{M}e^{\mathbf{A}t}\mathbf{x}\|_2^2 \geq 0,$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Next, if $\mathbf{x}'\mathbf{P}\mathbf{x} = 0$, then

46 $\mathbf{M}e^{\mathbf{A}t}\mathbf{x} \equiv 0, \forall t \geq 0.$

Substituting $t = 0$ in (46) gives $\mathbf{M}\mathbf{x} = \mathbf{0}$, which in turn implies that $\mathbf{x} = \mathbf{0}$ since $\mathbf{M}$ is nonsingular. Hence $\mathbf{P}$ is positive definite and (1) implies (3). ■

Remarks

1. Theorem (42) is very important in that it enables one to determine the stability status of the equilibrium $\mathbf{0}$ unambiguously, in the following manner: Given $\mathbf{A} \in \mathbf{R}^{n\times n}$, pick $\mathbf{Q} \in \mathbf{R}^{n\times n}$ to be *any* positive definite matrix. (A logical choice is the identity matrix or some other diagonal matrix.) Attempt to solve (34) for $\mathbf{P}$. If (34) has no solution or has a nonunique solution, then $\mathbf{0}$ is not asymptotically stable. If $\mathbf{P}$ is unique but not positive definite, then once again $\mathbf{0}$ is not asymptotically stable. On the other hand, if $\mathbf{P}$ is uniquely determined and positive definite, then $\mathbf{0}$ is an asymptotically stable equilibrium.

2. Theorem (42) states that if $\mathbf{A}$ is a Hurwitz matrix, then whenever $\mathbf{Q}$ is positive definite, the corresponding $\mathbf{P}$ given by (34) is also positive definite. It **does not** say that, whenever $\mathbf{P}$ is positive definite, the corresponding $\mathbf{Q}$ is positive definite. This statement is false in general (see Problem 5.19).

Theorem (42) shows that, if $\mathbf{A}$ is Hurwitz and $\mathbf{Q}$ is positive definite, then the solution $\mathbf{P}$ of (34) is positive definite. The next result shows that, under certain conditions, $\mathbf{P}$ is positive definite even when $\mathbf{Q}$ is only positive semidefinite.

47 Lemma *Suppose* $\mathbf{A} \in \mathbf{R}^{n\times n}$ *and satisfies (36). Suppose* $\mathbf{C} \in \mathbf{R}^{m\times n}$, *and that*

$$\textbf{48}\quad \operatorname{rank}\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{CA}^{n-1} \end{bmatrix} = n.$$

Under these conditions, the equation

$$\textbf{49}\quad \mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{C}'\mathbf{C}$$

has a unique solution for $\mathbf{P}$*; moreover,* $\mathbf{P}$ *is positive definite.*

Proof The uniqueness of $\mathbf{P}$ follows from Lemma (35). To show that $\mathbf{P}$ is positive definite, observe that (49) is of the form (34) with $\mathbf{Q} = -\mathbf{C}'\mathbf{C}$. Now suppose $\mathbf{x}'\mathbf{P}\mathbf{x} = \mathbf{0}$. Then one can repeat the reasoning of (44) to (46) with $\mathbf{M}$ replaced by $\mathbf{C}$. This shows that

$$\textbf{50}\quad \mathbf{x}'\mathbf{P}\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{C}e^{\mathbf{A}t}\mathbf{x} \equiv \mathbf{0},\ \forall t \geq 0.$$

Let $\mathbf{f}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}$. Then $\mathbf{f}(\cdot)$ as well as all of its derivatives are identically zero. In particular,

$$\textbf{51}\quad \mathbf{0} = \begin{bmatrix} \mathbf{f}(0) \\ \dot{\mathbf{f}}(0) \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{d^{n-1}}{dt^{n-1}}\mathbf{f}(0) \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{CA}^{n-1} \end{bmatrix} \mathbf{x}.$$

But from (48), this shows that $\mathbf{x} = \mathbf{0}$. Hence $\mathbf{P}$ is positive definite. ■

Theorem (42) shows that if the equilibrium $\mathbf{0}$ of the system (28) is exponentially stable, then this fact can be ascertained by choosing a quadratic Lyapunov function and applying Theorem (5.3.45). The following result, stated without proof, allows one to prove that, if the equilibrium $\mathbf{0}$ is unstable because some eigenvalue of $\mathbf{A}$ has a positive real part,[1] then this fact can also be ascertained by choosing a quadratic Lyapunov function and applying Theorem (5.3.88).

---

[1] Note that the equilibrium $\mathbf{0}$ can be unstable in another way as well, namely that the minimal polynomial of $\mathbf{A}$ has a multiple zero on the imaginary axis.

52 Lemma *Consider (34), and suppose the condition (36) is satisfied, so that (34) has a unique solution for* $\mathbf{P}$ *corresponding to each* $\mathbf{Q} \in \mathbf{R}^{n \times n}$. *Under these conditions, if* $\mathbf{Q}$ *is positive definite, then* $\mathbf{P}$ *has as many negative eigenvalues as there are eigenvalues of* $\mathbf{A}$ *with positive real part.*

A proof of this lemma can be found in Taussky (1961). Note that, since (36) is assumed to hold, it follows that $\mathbf{A}$ does not have any eigenvalues with zero real part; consequently, all eigenvalues of $\mathbf{A}$ have either a positive real part or a negative real part.

To see how Lemma (52) can be applied, suppose $\mathbf{A}$ satisfies the hypotheses of this lemma, and choose the Lyapunov function candidate (32). Then $\dot{V}$ is given by (33). Now, if $\mathbf{A}$ is Hurwitz, then V is positive definite and $\dot{V}$ is negative definite, and the exponential stability of $\mathbf{0}$ follows by Theorem (5.3.45). On the other hand, if $\mathbf{A}$ has one or more eigenvalues with positive real part, then $\dot{V}$ is negative definite and V assumes negative values arbitrarily close to the origin; thus the instability of the origin can be deduced using Theorem (5.3.88).

5.4.3 Nonautonomous Systems

In the case of linear time-varying systems described by (1), the stability status of the equilibrium $\mathbf{0}$ can be ascertained, in principle at least, by studying the state transition matrix. This is detailed in Section 5.4.1. The purpose of the present subsection is three-fold: (1) to prove the existence of quadratic Lyapunov functions for uniformly asymptotically stable linear systems; (2) to present some simple sufficient conditions for stability, asymptotic stability, and instability, based on the matrix measure; (3) to present necessary and sufficient conditions for the stability of periodic systems.

The Existence of Quadratic Lyapunov Functions

Theorem (42) shows that if the equilibrium $\mathbf{0}$ of the system (28) is exponentially stable, then a quadratic Lyapunov function exists for this system. A similar result is now proved for nonautonomous systems, under the assumption that $\mathbf{0}$ is exponentially stable [or equivalently, uniformly asymptotically stable; see Theorem (18)]. The relevant result is Theorem (64) below. This theorem is based on two preliminary lemmas.

53 Lemma *Suppose* $\mathbf{Q}: \mathbf{R}_+ \to \mathbf{R}^{n \times n}$ *is continuous and bounded, and that the equilibrium* $\mathbf{0}$ *of (1) is uniformly asymptotically stable. Then, for each* $t \geq 0$, *the matrix*

$$\mathbf{P}(t) = \int_t^\infty \Phi'(\tau, t)\, \mathbf{Q}(\tau)\, \Phi(\tau, t)\, d\tau \tag{54}$$

is well-defined; moreover, $\mathbf{P}(t)$ *is bounded as a function of* t.

Proof The hypothesis of uniform asymptotic stability implies that $\mathbf{0}$ is in fact exponentially stable, by Theorem (18). Thus there exist constants m, $\lambda > 0$ such that

55 $\|\Phi(\tau, t)\|_i \leq m \exp[-\lambda(\tau - t)], \forall \tau \geq t \geq 0.$

The bound (55), together with the boundedness of $\mathbf{Q}(\cdot)$, proves the lemma. ■

56 Lemma *Suppose that, in addition to the hypotheses of Lemma (53), the following conditions also hold:*

(1) $\mathbf{Q}(t)$ *is symmetric and positive definite for each* $t \geq 0$*; moreover, there exists a constant* $\alpha > 0$ *such that*

57 $\alpha \mathbf{x}'\mathbf{x} \leq \mathbf{x}'\mathbf{Q}(t)\mathbf{x}, \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$

(2) *The matrix* $\mathbf{A}(\cdot)$ *is bounded; i.e.,*

58 $m_0 := \sup_{t \geq 0} \|\mathbf{A}(t)\|_{i2} < \infty.$

Under these conditions, the matrix $\mathbf{P}(t)$ *defined in (54) is positive definite for each* $t \geq 0$*; moreover, there exists a constant* $\beta > 0$ *such that*

59 $\beta \mathbf{x}'\mathbf{x} \leq \mathbf{x}'\mathbf{P}(t)\mathbf{x}, \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$

Proof Let $\mathbf{x} \in \mathbf{R}^n$, and consider the triple product $\mathbf{x}'\mathbf{P}(t)\mathbf{x}$. Then, from (54),

60 $$\mathbf{x}'\mathbf{P}(t)\mathbf{x} = \int_t^\infty \mathbf{x}'\Phi'(\tau, t)\,\mathbf{Q}(t)\,\Phi(\tau, t)\mathbf{x}\,d\tau = \int_t^\infty \mathbf{s}'(\tau, t, \mathbf{x})\,\mathbf{Q}(t)\,\mathbf{s}(\tau, t, \mathbf{x})\,d\tau,$$

where $\mathbf{s}(\tau, t, \mathbf{x})$ denotes (as before) the solution of (1) evaluated at time τ, corresponding to the initial condition $\mathbf{x}$ at time t. Now (57) and (60) together imply that

61 $$\mathbf{x}'\mathbf{P}(t)\mathbf{x} \geq \alpha \int_t^\infty \|\mathbf{s}(\tau, t, \mathbf{x})\|_2^2\,d\tau.$$

Next, by Theorem (2.5.1), we have

62 $$\|\mathbf{s}(\tau, t, \mathbf{x})\|_2 \geq \|\mathbf{x}\|_2 \exp\left\{-\int_t^\tau \mu_2[-\mathbf{A}(\theta)]\,d\theta\right\}$$

$$\geq \|\mathbf{x}\|_2 \exp\left\{-\int_t^\tau \|\mathbf{A}(\theta)\|_{i2}\,d\theta\right\}$$

$$\geq \|\mathbf{x}\|_2 \exp[-m_0(\tau - t)], \text{ by (58).}$$

Substituting from (62) into (61) gives

$$\mathbf{63} \quad \mathbf{x}'\mathbf{P}(t)\mathbf{x} \geq \alpha \int_t^\infty \mathbf{x}'\mathbf{x} \exp[-2m_0(\tau - t)]\, d\tau = \frac{\alpha \mathbf{x}'\mathbf{x}}{2m_0}.$$

The inequality (59) now follows by taking $\beta = \alpha/2m_0$. ■

64 Theorem *Suppose* $\mathbf{Q}(\cdot)$ *and* $\mathbf{A}(\cdot)$ *satisfy the hypotheses of Lemmas (53) and (56). Then, for each function* $\mathbf{Q}(\cdot)$ *satisfying the hypotheses, the function*

$$\mathbf{65} \quad V(t, \mathbf{x}) = \mathbf{x}'\mathbf{P}(t)\mathbf{x}$$

is a Lyapunov function in the sense of Theorem (5.3.45) for establishing the exponential stability of the equilibrium $\mathbf{0}$.

Proof With $V(t, \mathbf{x})$ defined as above, we have

$$\mathbf{66} \quad \dot{V}(t, \mathbf{x}) = \mathbf{x}' [\dot{\mathbf{P}}(t) + \mathbf{A}'(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}(t)]\, \mathbf{x}.$$

It is easy to verify by differentiating (54) that

$$\mathbf{67} \quad \dot{\mathbf{P}}(t) = -\mathbf{A}'(t)\mathbf{P}(t) - \mathbf{P}(t)\mathbf{A}(t) - \mathbf{Q}(t),$$

Hence

$$\mathbf{68} \quad \dot{V}(t, \mathbf{x}) = -\mathbf{x}'\mathbf{Q}(t)\mathbf{x}.$$

Thus the functions V and $\dot{V}$ satisfy all the conditions of Theorem (5.3.45). ■

Conditions Based on the Matrix Measure

Next we present some conditions for stability and instability based on the matrix measure. The following result proves useful for this purpose.

69 Lemma *With regard to the system (1), the following inequalities hold:*

$$\mathbf{70} \quad \exp\left\{\int_{t_0}^t -\mu[-\mathbf{A}(\tau)]\, d\tau\right\} \leq \|\Phi(t, t_0)\|_i \leq \exp\left\{\int_{t_0}^t \mu[\mathbf{A}(\tau)]\, d\tau\right\},$$

where $\|\cdot\|$ *is any norm on* $\mathbf{R}^n$ *and* $\mu(\cdot)$ *is the corresponding matrix measure on* $\mathbf{R}^{n\times n}$.

The proof is immediate from Theorem (2.5.1).

Many simple sufficient conditions for various forms of stability are ready consequences of Lemma (69). The proofs are straight-forward applications of results in Section 5.4.1 and are left as exercises.

71 Lemma *The equilibrium* $\mathbf{0}$ *of (1) is stable if, for each* t_0, *there exists a finite constant* $m(t_0)$ *such that*

$$\mathbf{72} \quad \int_{t_0}^{t} \mu[\mathbf{A}(\tau)]\, d\tau \leq m(t_0), \ \forall t \geq t_0 \geq 0.$$

The equilibrium $\mathbf{0}$ *is uniformly stable if there exists a finite constant* m_0 *such that*

$$\mathbf{73} \quad \int_{t_0}^{t} \mu[\mathbf{A}(\tau)]\, d\tau \leq m_0, \ \forall t \geq t_0 \geq 0.$$

74 Lemma *The equilibrium* $\mathbf{0}$ *of (1) is asymptotically stable if*

$$\mathbf{75} \quad \int_{t_0}^{t_0+t} \mu[\mathbf{A}(\tau)]\, d\tau \to -\infty \text{ as } t \to \infty, \ \forall t_0 \geq 0.$$

The equilibrium is uniformly asymptotically stable if the convergence in (75) is uniform in t_0, *i.e., if, for every* $m > 0$ *there exists a* T *such that*

$$\mathbf{76} \quad \int_{t_0}^{t_0+t} \mu[\mathbf{A}(\tau)]\, d\tau < -m, \ \forall t \geq T, \ \forall t_0 \geq 0.$$

77 Lemma *The equilibrium* $\mathbf{0}$ *is unstable if there exists a time* $t_0 \geq 0$ *such that*

$$\mathbf{78} \quad \int_{t_0}^{t} \mu[-\mathbf{A}(\tau)]\, d\tau \to -\infty \text{ as } t \to \infty.$$

Remarks

1. As shown in Section 2.2, the measure of a matrix is strongly dependent on the vector norm on $\mathbf{R}^n$ that is used to define it. In Lemmas (71), (74), and (77), the conclusions follow if the indicated conditions are satisfied for *some* norm on $\mathbf{R}^n$. Thus there is a great deal of flexibility in applying these lemmas.

2. The three lemmas provide only sufficient conditions and are by no means necessary. [But in this connection, see Vidyasagar (1978a).] However, they do have the advantage that one does not have to compute the state transition matrix.

3. Lemma (77) is of rather dubious value, since its hypothesis actually assures *complete* instability of the system (1); in other words, *every* nontrivial trajectory of (1) "blows up."

Periodic Systems

This subsection concludes with a discussion of periodic systems.

Suppose that the matrix $\mathbf{A}(t)$ in (1) is periodic. In this case, we know by Theorems (5.1.43) and (5.1.49) that the stability of the equilibrium $\mathbf{0}$ is equivalent to its uniform stability, and that the asymptotic stability of the equilibrium $\mathbf{0}$ is equivalent to its uniform asymptotic stability. Theorem (89) below shows that in the case of *linear* periodic systems, further simplifications are possible.

79 Lemma *Suppose the matrix* $\mathbf{A}(\cdot)$ *in (1) is periodic, and select a constant* $T > 0$ *such that*

$$\mathbf{A}(t+T) = \mathbf{A}(t), \ \forall t \geq 0. \tag{80}$$

Then the corresponding state transition matrix $\mathbf{\Phi}(t, t_0)$ *has the form*

$$\mathbf{\Phi}(t, t_0) = \mathbf{\Psi}(t, t_0) \exp[\mathbf{M}(t - t_0)], \tag{81}$$

where $\mathbf{M}$ *is a constant matrix, and* $\mathbf{\Psi}$ *is periodic in the sense that*

$$\mathbf{\Psi}(t+T, t_0) = \mathbf{\Psi}(t, t_0), \ \forall t \geq 0. \tag{82}$$

Proof Define

$$\mathbf{R} = \mathbf{\Phi}(t_0 + T, t_0), \tag{83}$$

and choose a matrix $\mathbf{M}$ such that

$$\mathbf{R} = \exp(\mathbf{M}T). \tag{84}$$

This is possible since $\mathbf{R}$ is nonsingular. Now it is claimed that the matrix $\mathbf{\Psi}$ defined by

$$\mathbf{\Psi}(t, t_0) = \mathbf{\Phi}(t, t_0) \exp[-\mathbf{M}(t - t_0)] \tag{85}$$

satisfies (82). To see this, proceed as follows:

$$\begin{aligned} \mathbf{\Psi}(t+T, t_0) &= \mathbf{\Phi}(t+T, t_0) \exp[-\mathbf{M}(t+T-t_0)] \\ &= \mathbf{\Phi}(t+T, t_0+T)\, \mathbf{\Phi}(t_0+T, t_0) \exp(-\mathbf{M}T) \exp[-\mathbf{M}(t-t_0)]. \end{aligned} \tag{86}$$

However, by the periodicity of $\mathbf{A}(\cdot)$, we have

87 $\Phi(t+T, t_0+T) = \Phi(t, t_0)$.

Thus (86) simplifies to

88 $\Psi(t+T, t_0) = \Phi(t, t_0)\exp[-\mathbf{M}(t-t_0)] = \Psi(t, t_0)$.

This establishes (82). ■

Once the representation (81) for the state transition matrix is obtained, the results of Section 5.4.1 can be used to obtain necessary and sufficient conditions for stability. For this purpose, note that λ is an eigenvalue of $\mathbf{M}$ if and only if $\exp(\lambda T)$ is an eigenvalue of $\Phi(T, 0)$.

89 Theorem *Consider the system (1), and suppose the matrix* $\mathbf{A}(\cdot)$ *is periodic. Then the equilibrium* $\mathbf{0}$ *of (1) is uniformly asymptotically stable if and only if all eigenvalues of* $\Phi(T, 0)$ *have magnitude less than one. The equilibrium* $\mathbf{0}$ *is uniformly stable if and only if all eigenvalues of* $\Phi(T, 0)$ *have magnitude no larger than one, and all eigenvalues of* $\Phi(T, 0)$ *with a magnitude of one are simple zeros of the minimal polynomial of* $\Phi(T, 0)$.

The proof is elementary and is left as an exercise.

90 Example The purpose of this example is to show that the stability of a nonautonomous system cannot be deduced by studying only the eigenvalues of the matrix $\mathbf{A}(t)$ for each fixed t. Consider the periodic system (1) with

$$\mathbf{A}(t) = \begin{bmatrix} -1 + a\cos^2 t & 1 - a\sin t\cos t \\ -1 - a\sin t\cos t & -1 + a\sin^2 t \end{bmatrix}.$$

Then it can be verified that

$$\Phi(t, 0) = \begin{bmatrix} e^{(a-1)t}\cos t & e^{-t}\sin t \\ -e^{(a-1)t}\sin t & e^{-t}\cos t \end{bmatrix}.$$

Now, the eigenvalues of $\mathbf{A}(t)$ are independent of t and satisfy the characteristic equation

$$\lambda^2 + (2-a)\lambda + (2-a) = 0.$$

Hence, if $a < 2$, then the eigenvalues of $\mathbf{A}(t)$ have negative real parts for each fixed $t \geq 0$, and in fact the eigenvalues are bounded away from the imaginary axis. Nevertheless, if $a > 1$, the system is unstable. To see this, note that the period T equals 2π in this case, and that

$$\Phi(2\pi, 0) = \begin{bmatrix} e^{2(a-1)\pi} & 0 \\ 0 & e^{-2\pi} \end{bmatrix}.$$

The eigenvalues of this matrix are obviously $e^{2(a-1)\pi}$ and $e^{-2\pi}$. If $a > 1$, the first eigenvalue has a magnitude greater than one, and the equilibrium $\mathbf{0}$ is unstable by Theorem (89). ■

In Section 5.8.2 it is shown that, for so-called *slowly varying systems,* it is possible to deduce the stability of the nonautonomous system by studying only the "frozen" systems.

Problem 5.18 Complete the proof of Theorem (14).

Problem 5.19 Construct an example of a Hurwitz matrix $\mathbf{A}$ and a positive definite matrix $\mathbf{P}$ such that $\mathbf{A}'\mathbf{P}+\mathbf{P}\mathbf{A}$ is not negative definite.

Problem 5.20 Consider an RLC network that does not contain any capacitive loops or inductive cutsets. Such a network can be described in state variable form by choosing the capacitor voltages and the inductor currents as the states. Specifically, let $\mathbf{x}_C$ denote the vector of capacitor voltages, and let $\mathbf{x}_L$ denote the vector of inductor cutsets. Then the system equations are of the form

$$\begin{bmatrix}\dot{\mathbf{x}}_C\\ \dot{\mathbf{x}}_L\end{bmatrix} = -\begin{bmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{L}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{R}_{11} & \mathbf{R}_{12}\\ \mathbf{R}_{21} & \mathbf{R}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{x}_C\\ \mathbf{x}_L\end{bmatrix},$$

where $\mathbf{C}$ is the diagonal positive definite matrix of capacitor values; $\mathbf{L}$ is the positive definite matrix of inductor values, including mutual inductances; $\mathbf{R}_{ij}$ are the matrices arising from the resistive subnetwork, with $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ symmetric, and $\mathbf{R}_{21} = -\mathbf{R}_{12}'$ if the resistive network is reciprocal. Using the total energy stored in the capacitors and the inductors as a Lyapunov function candidate, show that the equilibrium $\mathbf{0}$ is stable if both $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are nonnegative definite, and is asymptotically stable if both $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are positive definite.

Problem 5.21 Using the results of Section 5.4.3, determine whether $\mathbf{0}$ is a uniformly stable or uniformly asymptotically stable equilibrium for the system $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\,\mathbf{x}(t)$, for each of the situations below:

(*a*) $$\mathbf{A}(t) = \begin{bmatrix}-2+\sin t^2 & 1\\ \cos t & -1\end{bmatrix}.$$

(*b*) $$\mathbf{A}(t) = \begin{bmatrix}2t-t^2 & 1-t & 3+t\\ t^2 & -t^3 & 0\\ 2 & 5 & 4-t\end{bmatrix}.$$

Problem 5.22 The equilibrium $\mathbf{0}$ of the system (5.1.1) is said to be *bounded* if, for every $\delta > 0$ and every $t_0 \geq 0$, there exists an $\varepsilon = \varepsilon(\delta, t_0)$ such that

$$\|\mathbf{x}_0\| < \delta \Rightarrow \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| < \varepsilon, \forall t \geq t_0.$$

It is said to be *uniformly bounded* if for every $\delta > 0$ there exists an $\varepsilon = \varepsilon(\delta)$ such that

$$\|\mathbf{x}_0\| < \delta, t_0 \geq 0 \Rightarrow \|\mathbf{s}(t, t_0, \mathbf{x}_0)\| < \varepsilon, \forall t \geq t_0.$$

(a) Show that a *linear* system is bounded if and only if it is stable, and that it is uniformly bounded if and only if it is uniformly stable.

(b) Construct examples of nonlinear systems where the equilibrium $\mathbf{0}$ is stable but not bounded, and where the equilibrium $\mathbf{0}$ is bounded but not stable.

Problem 5.23 Generalize Theorem (18) to a class of nonlinear systems. Specifically, suppose there exists a number $d > 0$ such that the solution trajectories of the system (5.1.1) satisfy the bound

$$\| \mathbf{s}(t, t_0, \mathbf{x}_0) \| \leq \mu \| \mathbf{x}_0 \| \sigma(t - t_0), \ \forall \mathbf{x}_0 \in B_d, \ \forall t \geq t_0 \geq 0,$$

where μ is some finite constant and $\sigma(\cdot)$ is a function of class L. Show that the equilibrium $\mathbf{0}$ is exponentially stable.

Problem 5.24 Given a finite collection of $n \times n$ matrices $\mathbf{A}_1, \cdots, \mathbf{A}_k$, define their *convex hull* $\mathbf{S}$ as

$$\mathbf{S} = \{\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{A}_i : \lambda_i \geq 0 \ \forall i, \ \sum_{i=1}^{k} \lambda_i = 1\}.$$

(a) Suppose there exists a positive definite matrix $\mathbf{P}$ such that $\mathbf{A}_i' \mathbf{P} + \mathbf{P} \mathbf{A}_i$ is negative definite for each i between 1 and k. Show that every matrix in the set $\mathbf{S}$ is Hurwitz.

(b) Consider the differential equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t) \mathbf{x}(t), \text{ where } \mathbf{A}(t) \in \mathbf{S} \ \forall t \geq 0.$$

Show that $\mathbf{0}$ is an exponentially stable equilibrium of this system.

5.5 LYAPUNOV'S LINEARIZATION METHOD

In this section, the results of the preceding two sections are combined to obtain one of the most useful results in Lyapunov stability theory, namely the linearization method. The great value of this method lies in the fact that, under certain conditions, it enables one to draw conclusions about a *nonlinear* system by studying the behavior of a *linear* system.

We begin by defining precisely the concept of linearizing a nonlinear system around an equilibrium. Consider first the autonomous system

1 $\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t)]$.

Suppose $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, so that $\mathbf{0}$ is an equilibrium of the system (1), and suppose also that $\mathbf{f}$ is continuously differentiable. Define

$$\mathbf{2} \qquad \mathbf{A} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{0}},$$

i.e., let $\mathbf{A}$ denote the Jacobian matrix of $\mathbf{f}$ evaluated at $\mathbf{x} = \mathbf{0}$. By the definition of the Jacobian, it follows that if we define

$$\mathbf{3} \qquad \mathbf{f}_1(\mathbf{x}) := \mathbf{f}(\mathbf{x}) - \mathbf{A}\mathbf{x},$$

then

$$\mathbf{4} \qquad \lim_{\|\mathbf{x}\| \to 0} \frac{\|\mathbf{f}_1(\mathbf{x})\|}{\|\mathbf{x}\|} = 0,$$

where, to be specific, all norms are taken as Euclidean norms. Alternatively, one can think of

$$\mathbf{5} \qquad \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{f}_1(\mathbf{x})$$

as the Taylor's series expansion of $\mathbf{f}(\cdot)$ around the point $\mathbf{x} = \mathbf{0}$ [recall that $\mathbf{f}(\mathbf{0}) = \mathbf{0}$]. With this notation, the system

$$\mathbf{6} \qquad \dot{\mathbf{z}}(t) = \mathbf{A}\mathbf{z}(t)$$

is referred to as the **linearization** or the **linearized system** of (1) around the equilibrium $\mathbf{0}$.

The development for nonautonomous systems is similar but for an additional technical requirement. Given the nonautonomous system

$$\mathbf{7} \qquad \dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)],$$

suppose that

$$\mathbf{8} \qquad \mathbf{f}(t, \mathbf{0}) = \mathbf{0}, \ \forall t \geq 0,$$

and that $\mathbf{f}$ is a C^1 function. Define

$$\mathbf{9} \qquad \mathbf{A}(t) = \left[\frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{0}},$$

$$\mathbf{10} \qquad \mathbf{f}_1(t, \mathbf{x}) = \mathbf{f}(t, \mathbf{x}) - \mathbf{A}(t)\mathbf{x}.$$

Then, by the definition of the Jacobian, it follows that *for each fixed* $t \geq 0$, it is true that

$$\mathbf{11} \quad \lim_{\|\mathbf{x}\| \to 0} \frac{\|\mathbf{f}_1(t, \mathbf{x})\|}{\|\mathbf{x}\|} = 0.$$

However, it may or may not be true that

$$\mathbf{12} \quad \lim_{\|\mathbf{x}\| \to 0} \sup_{t \geq 0} \frac{\|\mathbf{f}_1(t, \mathbf{x})\|}{\|\mathbf{x}\|} = 0.$$

In other words, the convergence in (11) may or may not be uniform in t. *Provided (12) holds*, the system

$$\mathbf{13} \quad \dot{\mathbf{z}}(t) = \mathbf{A}(t)\,\mathbf{z}(t)$$

is called the **linearization** or **linearized system** of (7) around the equilibrium $\mathbf{0}$.

14 Example Consider the system

$$\dot{x}_1(t) = -x_1(t) + tx_2^2, \dot{x}_2(t) = x_1(t) - x_2(t).$$

In this case, $\mathbf{f}(\cdot)$ is C^1, and

$$\mathbf{A}(t) = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}, \forall t \geq 0.$$

However, the remainder term $\mathbf{f}_1(t, \mathbf{x})$ is given by

$$\mathbf{f}_1(t, \mathbf{x}) = [tx_2^2 \quad 0]'.$$

Hence the uniformity condition (12) does not hold. Accordingly, the system

$$\dot{z}_1(t) = -z_1(t), \dot{z}_2(t) = z_1(t) - z_2(t)$$

is *not* a linearization of the original system. ■

Theorem (15) is the main stability theorem of the linearization method. Since there is nothing to be gained by assuming that the system under study is autonomous, the result is stated for a general nonautonomous system.

15 Theorem *Consider the system (7). Suppose that (8) holds and that $f(\cdot)$ is continuously differentiable. Define $\mathbf{A}(t)$, $\mathbf{f}_1(t, \mathbf{x})$ as in (9), (10), respectively, and assume that (i) (12) holds, and (ii) $\mathbf{A}(\cdot)$ is bounded. Under these conditions, if $\mathbf{0}$ is an exponentially stable equilibrium of the linear system*

$$\mathbf{16} \quad \dot{\mathbf{z}}(t) = \mathbf{A}(t)\,\mathbf{z}(t),$$

then it is also an exponentially stable equilibrium of the system (7).

Proof Since $\mathbf{A}(\cdot)$ is bounded and the equilibrium $\mathbf{0}$ is uniformly asymptotically stable, it follows from Lemma (5.4.56) that the matrix

$$\mathbf{17} \quad \mathbf{P}(t) = \int_t^\infty \Phi'(\tau, t)\,\Phi(\tau, t)\,d\tau$$

is well-defined for all $t \geq 0$; moreover, there exist constants $\alpha, \beta > 0$ such that

$$\mathbf{18} \quad \alpha \mathbf{x}'\mathbf{x} \leq \mathbf{x}'\mathbf{P}(t)\mathbf{x} \leq \beta \mathbf{x}'\mathbf{x}, \ \forall \mathbf{x} \in \mathbf{R}^n, \ \forall t \geq 0.$$

Hence the function

$$\mathbf{19} \quad V(t, \mathbf{x}) = \mathbf{x}'\mathbf{P}(t)\mathbf{x}$$

is a decrescent pdf, and is thus a suitable Lyapunov function candidate for applying Theorem (5.3.45). Calculating $\dot{V}$ for the system (7) gives

$$\mathbf{20} \quad \dot{V}(t, \mathbf{x}) = \mathbf{x}'\,\dot{\mathbf{P}}(t)\,\mathbf{x} + \mathbf{f}'(t, \mathbf{x})\,\mathbf{P}(t)\,\mathbf{x} + \mathbf{x}'\,\mathbf{P}(t)\,\mathbf{f}(t, \mathbf{x})$$

$$= \mathbf{x}'\,[\dot{\mathbf{P}}(t) + \mathbf{A}'(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{A}(t)]\,\mathbf{x} + 2\mathbf{x}'\,\mathbf{P}(t)\,\mathbf{f}_1(t, \mathbf{x}).$$

However, from (17) it can be easily shown that

$$\mathbf{21} \quad \dot{\mathbf{P}}(t) + \mathbf{A}'(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{A}(t) = -I.$$

Hence

$$\mathbf{22} \quad \dot{V}(t, \mathbf{x}) = -\mathbf{x}'\mathbf{x} + 2\mathbf{x}'\,\mathbf{P}(t)\,\mathbf{f}_1(t, \mathbf{x}).$$

In view of (12), one can pick a number $r > 0$ and a $\rho < 0.5$ such that

$$\mathbf{23} \quad \|\mathbf{f}_1(t, \mathbf{x})\| \leq \frac{\rho}{\beta}\,\|\mathbf{x}\|, \ \forall t \geq 0, \ \forall \mathbf{x} \in B_r.$$

Then (23) and (18) together imply that

$$\mathbf{24} \quad |2\mathbf{x}'\,\mathbf{P}(t)\,\mathbf{f}_1(t, \mathbf{x})| \leq \frac{2\rho}{\beta}\,\mathbf{x}'\mathbf{x}, \ \forall t \geq 0, \ \forall \mathbf{x} \in B_r.$$

Therefore

$$\mathbf{25} \quad \dot{V}(t, \mathbf{x}) \leq -(1 - 2\rho)\,\mathbf{x}'\mathbf{x}, \ \forall t \geq 0, \ \forall \mathbf{x} \in B_r.$$

This shows that $-\dot{V}$ is an lpdf. Thus all the hypotheses of Theorem (5.3.45) are satisfied, and we conclude that $\mathbf{0}$ is an exponentially stable equilibrium. ■

26 Corollary *Consider the autonomous system (1). Suppose that* $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, *that* $\mathbf{f}$ *is continuously differentiable, and define* $\mathbf{A}$ *as in (2). Under these conditions,* $\mathbf{0}$ *is an exponentially stable equilibrium of (1) if all eigenvalues of* $\mathbf{A}$ *have negative real parts.*

In the instability counterpart to Theorem (15), it is assumed that the linearized system is autonomous, even if the original nonlinear system is not.

27 Theorem *Consider the system (7). Suppose that (8) holds, that* $\mathbf{f}$ *is* C^1, *and suppose in addition that*

$$\mathbf{28} \quad \mathbf{A}(t) = \left[\frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial \mathbf{x}} \right] \equiv \mathbf{A}_0 \text{ (a constant matrix)}, \ \forall t \geq 0,$$

and that (12) holds. Under these conditions, the equilibrium $\mathbf{0}$ *is unstable if* $\mathbf{A}_0$ *has at least one eigenvalue with a positive real part.*

Proof The proof is given only for the case where $\mathbf{A}_0$ satisfies the condition (5.4.36). The proof of the general case can be obtained from the one given below by using continuity arguments.

Since $\mathbf{A}_0$ is assumed to satisfy (5.4.36) and has at least one eigenvalue with positive real part, it follows from Lemma (5.4.52) that the equation

$$\mathbf{29} \quad \mathbf{A}_0'\mathbf{P} + \mathbf{P}\mathbf{A}_0 = I$$

has a unique solution for $\mathbf{P}$ and that this matrix $\mathbf{P}$ has at least one positive eigenvalue. Now, by arguments entirely analogous to those used in the proof of Theorem (15), it can be shown that if we choose

$$\mathbf{30} \quad V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x},$$

then V assumes positive values arbitrarily close to $\mathbf{0}$, and $\dot{V}$ is an lpdf. Hence, by Theorem (5.3.88), it follows that $\mathbf{0}$ is an unstable equilibrium of the system (7). The details are left as an exercise. ■

Remarks: Theorems (15) and (27) are very useful because they enable one to draw conclusions about the stability status of the equilibrium $\mathbf{0}$ of a given *nonlinear* system by examining a *linear* system. The advantages of these results are self-evident. Some of the limitations of these results are the following: (i) The conclusions based on linearization are purely local in nature; to study *global* asymptotic stability, it is still necessary to resort to Lyapunov's direct method. (ii) In the case where the linearized system is autonomous, if some eigenvalues of $\mathbf{A}$ have zero real parts and the remainder have negative real parts, then linearization techniques are inconclusive, because this case falls outside the scope of both Corollary (26) and Theorem (27). In such a case, the stability of the equilibrium is determined by the higher order terms that are neglected in the linearization process. [In this connection, see Theorem (5.8.1).] (iii) In the case where the linearized system is nonautonomous, if the equilibrium $\mathbf{0}$ is asymptotically stable but not uniformly asymptotically

stable, then linearization is once again inconclusive. It can be shown by means of examples that the assumption of *uniform* asymptotic stability in Theorem (15) is indispensable.

Let us now return to the autonomous system (1). Suppose $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, and define $\mathbf{A}$, $\mathbf{f}_1$ as in (2) and (3) respectively. Finally, suppose $\mathbf{A}$ is a Hurwitz matrix. Then Corollary (26) tells us that $\mathbf{0}$ is an asymptotically stable equilibrium. Now let $\mathbf{Q}$ be an arbitrary positive definite matrix, and solve the corresponding Lyapunov matrix equation

$$\mathbf{31} \quad \mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}.$$

Then $\mathbf{P}$ is also positive definite. Moreover, it can be seen from the proof of Theorem (15) that, for each $\mathbf{Q}$, the corresponding quadratic form

$$\mathbf{32} \quad V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x}$$

is a suitable Lyapunov function for applying Theorem (5.3.45). However, different Lyapunov functions will give rise to different estimates for the domain of attraction of $\mathbf{0}$. The question can thus be asked: What is the "best" choice for $\mathbf{Q}$?

Let us restate the question in another form. Define V by (32). Then, along the trajectories of the system (1), we have

$$\mathbf{33} \quad \dot{V}(\mathbf{x}) = \mathbf{x}'(\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A})\mathbf{x} + 2\mathbf{x}'\mathbf{P}\mathbf{f}_1(\mathbf{x}). = -\mathbf{x}'\mathbf{Q}\mathbf{x} + 2\mathbf{x}'\mathbf{P}\mathbf{f}_1(\mathbf{x}).$$

Now

$$\mathbf{34} \quad -\mathbf{x}'\mathbf{Q}\mathbf{x} \leq -\lambda_{\min}(\mathbf{Q})\|\mathbf{x}\|^2,$$

$$\mathbf{35} \quad |\mathbf{x}'\mathbf{P}\mathbf{f}_1(\mathbf{x})| \leq \lambda_{\max}(\mathbf{P})\|\mathbf{x}\| \cdot \|\mathbf{f}_1(\mathbf{x})\|.$$

Hence

$$\mathbf{36} \quad \dot{V}(\mathbf{x}) \leq -\lambda_{\min}(\mathbf{Q})\|\mathbf{x}\|^2 + 2\lambda_{\max}(\mathbf{P})\|\mathbf{x}\| \cdot \|\mathbf{f}_1(\mathbf{x})\|$$

$$= \|\mathbf{x}\|\,[2\lambda_{\max}(\mathbf{P})\|\mathbf{f}_1(\mathbf{x})\| - \lambda_{\min}(\mathbf{Q})\|\mathbf{x}\|].$$

Now let us choose $r > 0$ such that

$$\mathbf{37} \quad \frac{\|\mathbf{f}_1(\mathbf{x})\|}{\|\mathbf{x}\|} < \frac{\lambda_{\min}(\mathbf{Q})}{2\lambda_{\max}(\mathbf{P})}, \quad \forall \mathbf{x} \in B_r.$$

Then $\dot{V}(\mathbf{x}) < 0$ whenever $\mathbf{x} \in B_r$ and $\mathbf{x} \neq \mathbf{0}$. By Lemma (5.3.40) and the discussion preceding it, every bounded level set of V contained in B_r is also contained in the domain of attraction $D(\mathbf{0})$. Now (37) makes it clear that the larger the ratio $\lambda_{\min}(\mathbf{Q})/\lambda_{\max}(\mathbf{P})$, the larger the possible choice of r. Hence the "best" choice of $\mathbf{Q}$ is one that maximizes the ratio

38 $$\mu(\mathbf{Q}) = \frac{\lambda_{\min}(\mathbf{Q})}{\lambda_{\max}(\mathbf{P})},$$

where $\mathbf{P}$ of course satisfies (31). Clearly $\mu(\cdot)$ is not affected by scaling $\mathbf{Q}$, i.e., $\mu(a\mathbf{Q}) = \mu(\mathbf{Q})\ \forall a > 0$. Hence one can pose the question at hand as follows: Among all positive definite matrices $\mathbf{Q}$ with $\lambda_{\min}(\mathbf{Q}) = 1$, which one results in the smallest value for $\lambda_{\max}(\mathbf{P})$? The answer turns out to be: the identity matrix.

39 **Lemma** *Suppose* $\mathbf{A}$ *is Hurwitz, and let* $\mathbf{M}$ *be the solution of*

40 $$\mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} = -I.$$

Suppose $\mathbf{Q}$ *is positive definite,* $\lambda_{\min}(\mathbf{Q}) = 1$, *and let* $\mathbf{P}$ *satisfy (31). Then*

41 $$\lambda_{\max}(\mathbf{M}) \leq \lambda_{\max}(\mathbf{P}).$$

Proof Since $\lambda_{\min}(\mathbf{Q}) = 1$, the matrix $\mathbf{Q} - I$ has all nonnegative eigenvalues, and is therefore nonnegative definite. Subtracting (40) from (31) gives

42 $$\mathbf{A}'(\mathbf{P} - \mathbf{M}) + (\mathbf{P} - \mathbf{M})\mathbf{A} = -(\mathbf{Q} - I).$$

From Lemma (5.4.38), the solution of (42) is given by

43 $$\mathbf{P} - \mathbf{M} = \int_0^\infty \exp(\mathbf{A}'t)(\mathbf{Q} - I)\exp(\mathbf{A}t)\,dt.$$

Hence $\mathbf{P} - \mathbf{M}$ is nonnegative definite, which implies (41). ■

The preceding discussion shows that, in some sense, the "best" choice for $\mathbf{Q}$ is the identity matrix.

44 **Example** Consider the system

$$\dot{x}_1 = -x_1 + x_2 + x_1x_2,\ \dot{x}_2 = -x_1 + x_2^2.$$

The linearization of this system around $\mathbf{0}$ is

45 $$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Let $\mathbf{A}$ denote the matrix in (45). Then the eigenvalues of $\mathbf{A}$ are $(-1 \pm j\sqrt{3})/2$. Hence, by Corollary (26), $\mathbf{0}$ is an asymptotically stable equilibrium.

To obtain an estimate of the domain of attraction, we first solve the equation

$$\mathbf{A}'\mathbf{P}+\mathbf{P}\mathbf{A}=-I$$

for $\mathbf{P}$. It can be easily verified that

$$\mathbf{P}=\begin{bmatrix} 1 & -1/2 \\ -1/2 & 3/2 \end{bmatrix}.$$

Hence β in (18) can be chosen as the largest singular value of $\mathbf{P}$, which is $(5+\sqrt{5})/4$,or approximately 1.81. To satisfy (23), we must choose $r>0$ in such a way that

$$\mathbf{46}\quad \|\mathbf{f}_1(\mathbf{x})\| < \frac{\rho}{\beta}\|\mathbf{x}\|, \forall \mathbf{x}\in B_r,$$

where $\rho < 0.5$ is some number. Now

$$\frac{\|\mathbf{f}_1(\mathbf{x})\|}{\|\mathbf{x}\|}=\frac{(x_1^2x_2^2+x_2^4)^{1/2}}{(x_1^2+x_2^2)^{1/2}}=|x_2|\leq\|\mathbf{x}\|.$$

Hence, to satisfy (46), we can choose r as close as possible to $1/\beta$, or approximately 0.54. Hence every level set in the ball of radius 0.54 is in the domain of attraction.

47 Example Consider the Van der Pol oscillator described by

$$\dot{x}_1=x_2, \dot{x}_2=-\mu(1-x_1^2)x_2-x_1.$$

The linearization of this system around the equilibrium $\mathbf{0}$ is

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

If $\mu>0$, then both eigenvalues of $\mathbf{A}$ have positive real parts. Hence, by Theorem (27), $\mathbf{0}$ is an unstable equilibrium. In fact, since *all* eigenvalues of $\mathbf{A}$ have positive real parts, one can show, by applying Corollary (26) and reversing the direction of time, that $\mathbf{0}$ is a *completely* unstable equilibrium.

48 Example Consider again the spinning object of Example (5.3.105), and focus attention on an equilibrium of the form $(0, y_0, 0)$ where $y_0\neq 0$. If we define $y_s=y-y_0$, then the equations of motion can be written as

$$\dot{x}=ay_sz+ay_0z, \dot{y}_s=-bxz, \dot{z}=cxy_s+cxy_0.$$

The linearization of this system around the origin (in the new coordinates!) is obtained by neglecting all the higher order terms, and is

$$\begin{bmatrix} \dot{x} \\ \dot{y}_s \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 0 & 0 & a\mathbf{y}_0 \\ 0 & 0 & 0 \\ cy_0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y_s \\ z \end{bmatrix}.$$

The eigenvalues of the **A** matrix above are $0, \pm\sqrt{ac}\,|\,y_0\,|$. If $|\,y_0\,| \neq 0$, then **A** has a positive real eigenvalue. Hence, by Theorem (27), the equilibrium **0**, or $(0, y_0, 0)$ in the original coordinates, is unstable.

49 Application (Feedback Stabilization of Nonlinear Control Systems) Given an autonomous system described by

50 $\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t), \mathbf{u}(t)]$,

where $\mathbf{f}: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$, the objective is to find a feedback control law of the form

51 $\mathbf{u}(t) = \mathbf{g}[\mathbf{x}(t)]$,

in such a way that the equilibrium **0** of the resulting closed-loop system

52 $\dot{\mathbf{x}}(t) = \mathbf{f}\{\mathbf{x}(t), \mathbf{g}[\mathbf{x}(t)]\}$

is asymptotically stable. A solution to this problem is given by the next result, which is a direct consequence of Corollary (26).

53 Theorem *Consider the autonomous system (50) where* $\mathbf{f}: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$. *Suppose* **f** *is* C^1 *and that* $\mathbf{f}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$, *and define*

54 $$\mathbf{A} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{0}, \mathbf{u}=\mathbf{0}}, \quad \mathbf{B} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right]_{\mathbf{x}=\mathbf{0}, \mathbf{u}=\mathbf{0}}.$$

Suppose there exists a matrix $\mathbf{K} \in \mathbf{R}^{m \times n}$ *such that all eigenvalues of* $\mathbf{A} - \mathbf{BK}$ *have negative real parts. Under these conditions, if we apply the control law*

55 $\mathbf{u}(t) = -\mathbf{K}\mathbf{x}(t)$

in (50), then **0** *is an asymptotically stable equilibrium of the resulting system*

56 $\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t), -\mathbf{K}\mathbf{x}(t)]$.

Proof Define the function $\mathbf{h}: \mathbf{R}^n \to \mathbf{R}^n$ by

57 $\mathbf{h}(\mathbf{x}) = \mathbf{f}(\mathbf{x}, -\mathbf{K}\mathbf{x})$.

Then the closed-loop system (56) can be represented as

58 $\dot{\mathbf{x}}(t) = \mathbf{h}[\mathbf{x}(t)]$.

Next, observe that

59 $$\left[\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{0}} = \mathbf{A} - \mathbf{BK},$$

which is Hurwitz by assumption. The desired conclusion now follows from Corollary (26).

60 Application (Singularly Perturbed Systems) In Section 4.3 some results are derived concerning the stability of singularly perturbed *linear* systems. In this application, Lyapunov's indirect method is combined with these earlier results of Section 4.3 to derive some results for singularly perturbed *nonlinear* systems. The proof of Theorem (61) is omitted, because it follows readily from earlier results.

61 Theorem *Consider the system*

62 $$\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t), \mathbf{y}(t)],$$

$$\varepsilon\dot{\mathbf{y}}(t) = \mathbf{g}[\mathbf{x}(t), \mathbf{y}(t)],$$

where $\mathbf{f}: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$, $\mathbf{g}: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^m$ *are continuously differentiable and satisfy*

63 $\mathbf{f}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$, $\mathbf{g}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$.

Define

64 $$\mathbf{A}_{11} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{(\mathbf{0},\mathbf{0})}, \mathbf{A}_{12} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{y}}\right]_{(\mathbf{0},\mathbf{0})}, \mathbf{A}_{21} = \left[\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right]_{(\mathbf{0},\mathbf{0})}, \mathbf{A}_{22} = \left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right]_{(\mathbf{0},\mathbf{0})},$$

and suppose $\mathbf{A}_{22}$ *is nonsingular. Under these conditions, there exists a* C^1 *function* $\mathbf{h}: \mathbf{R}^n \to \mathbf{R}^m$ *such that, in some neighborhood of* $(\mathbf{0}, \mathbf{0})$,

65 $\mathbf{v} = \mathbf{h}(\mathbf{x})$

is the unique solution of

66 $\mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$.

Moreover,

1. *If both* $\mathbf{A}_{22}$ *and* $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ *are Hurwitz, then there is an* ε_0 *such that* $(\mathbf{0}, \mathbf{0})$ *is an asymptotically stable equilibrium of the system (62) whenever* $0 < \varepsilon < \varepsilon_0$.

2. *If at least one eigenvalue of either* $\mathbf{A}_{22}$ *or* $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ *has a positive real part, then there is an* ε_0 *such that* $(\mathbf{0}, \mathbf{0})$ *is an unstable equilibrium of the system (62) whenever* $0 < \varepsilon < \varepsilon_0$.

Problem 5.25 Check the stability status of all the systems in Problems 5.11 to 5.16 using the linearization method.

Problem 5.26 Consider a feedback system described by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t),\ u(t) = -\phi[\mathbf{c}'\mathbf{x}(t)],$$

where ϕ: $\mathbf{R} \to \mathbf{R}$ is continuously differentiable. Such systems are studied further in Section 5.6. Suppose that the equilibrium $\mathbf{0}$ of this system is asymptotically stable whenever ϕ is set equal to a *linear* function of the form

$$\phi(\sigma) = k\sigma,\ k \in [0, \mu].$$

(a) Using Corollary (26), show that the equilibrium $\mathbf{0}$ continues to be an asymptotically stable equilibrium if $\phi(\cdot)$ is any C^1 function whose derivative lies in the interval $[0, \mu]$ in some neighborhood of the origin.

(b) Generalize the results of part (a) to time-varying systems, using Theorem (15).

5.6 THE LUR'E PROBLEM

In this section, we study the stability of an important class of control systems, namely feedback systems whose forward path contains a linear time-invariant subsystem and whose feedback path contains a memoryless (possibly time-varying) nonlinearity. The type of system is shown in Figure 5.15. The forward-path subsystem is described by

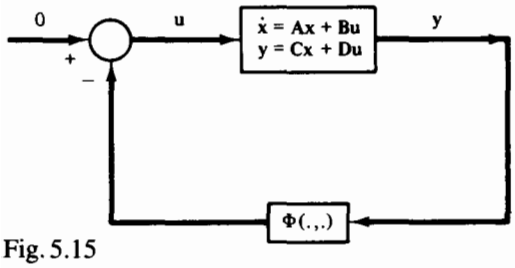


Fig. 5.15

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{1}$$

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

while the feedback subsystem is described by

2 $\mathbf{z}(t) = \Phi[t, \mathbf{y}(t)]$.

Of course, the feedback interconnection is described by

3 $\mathbf{u}(t) = -\mathbf{z}(t)$.

In the above, it is assumed that $\mathbf{x}(t) \in \mathbf{R}^n$, while $\mathbf{u}(t)$, $\mathbf{y}(t)$, $\mathbf{z}(t) \in \mathbf{R}^m$ with $m < n$. Thus it is assumed that both the forward and feedback subsystems are "square" in the sense that they have an equal number of inputs and outputs. It is possible to dispense with this assumption, but at the cost of making the derivations more opaque.

A study of systems of the form (1) – (3) is important for at least two reasons: (i) Many physical systems can be naturally decomposed into a linear part and a nonlinear part. Thus the system description (1) – (3) is widely applicable. (ii) Several comprehensive results are available concerning the stability of such systems (of which only a few are covered in this book).

5.6.1 Problem Statement

In this section, the problem under study is stated precisely, and some preliminary discussion of the problem is given. We begin by reviewing a few elementary concepts from the theory of linear control systems.

4 **Definition** *Suppose n, m, l are given integers, $\mathbf{A} \in \mathbf{R}^{n \times n}$, $\mathbf{B} \in \mathbf{R}^{n \times m}$, $\mathbf{C} \in \mathbf{R}^{l \times n}$, and $\mathbf{D} \in \mathbf{R}^{l \times m}$. Then the pair $(\mathbf{A}, \mathbf{B})$ is said to be* **controllable** *if*

5 $\text{rank}\,[\mathbf{B}\ \mathbf{AB} \cdots \mathbf{A}^{n-1}\mathbf{B}] = n$.

The pair $(\mathbf{C}, \mathbf{A})$ is said to be **observable** *if*

6 $$\text{rank} \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{CA}^{n-1} \end{bmatrix} = n.$$

It is easy to see that the pair $(\mathbf{A}, \mathbf{B})$ is controllable if and only if the pair $(\mathbf{B}', \mathbf{A}')$ is observable.

7 **Definition** *Suppose $\mathbf{H}(\cdot)$ is a proper rational matrix. Then a quadruplet $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ is said to be a* **realization** *of $\mathbf{H}(\cdot)$ if*

8 $\mathbf{H}(s) = \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$.

The realization is said to be **minimal** *if, in addition, the pair $(\mathbf{A}, \mathbf{B})$ is controllable and the*

pair $(\mathbf{C}, \mathbf{A})$ *is observable.*

Now the concept of a sector-bound nonlinearity is introduced.

9 Definition *Suppose* $\Phi: \mathbf{R}_+ \times \mathbf{R}^m \to \mathbf{R}^m$, *and* $a, b \in \mathbf{R}$ *with* $a < b$. *Then* Φ *is said to* **belong to the sector** $[a, b]$ *if (i)* $\Phi(t, \mathbf{0}) = \mathbf{0}\ \forall t \in \mathbf{R}_+$, *and (ii)*

10 $[\Phi(t, \mathbf{y}) - a\mathbf{y}]' [b\mathbf{y} - \Phi(t, \mathbf{y})] \geq 0,\ \forall t \in \mathbf{R}_+,\ \forall \mathbf{y} \in \mathbf{R}^m.$

It is possible to give a graphical interpretation of (10) in the scalar case ($m = 1$). In this case (10) says that, for each fixed $t \in \mathbf{R}_+$, the graph of $\phi(t, y)$ lies between two straight lines of slopes a and b respectively, passing through the origin in $\mathbf{R}^2$. The situation is depicted in Figure 5.16.

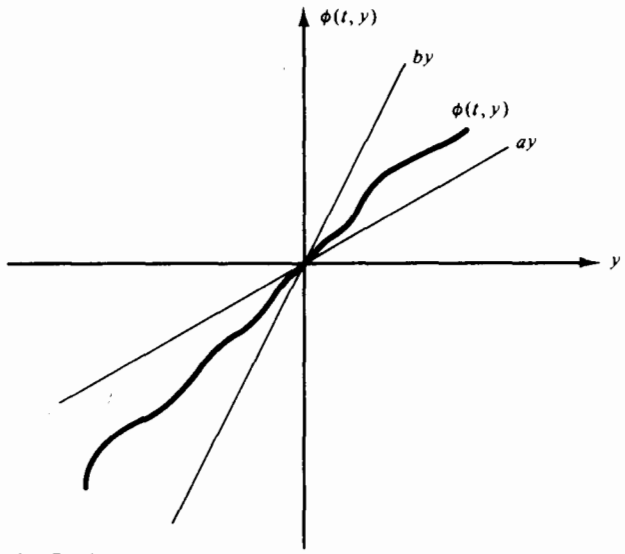


Fig. 5.16

Now the problem under study is stated.

Absolute Stability Problem We are given (i) matrices $\mathbf{A} \in \mathbf{R}^{n \times n}$, $\mathbf{B} \in \mathbf{R}^{n \times m}$, $\mathbf{C} \in \mathbf{R}^{m \times m}$, and $\mathbf{D} \in \mathbf{R}^{m \times m}$, such that the pair $(\mathbf{A}, \mathbf{B})$ is controllable and the pair $(\mathbf{C}, \mathbf{A})$ is observable; and (ii) two real numbers a, b with $a < b$. The problem is to derive conditions involving *only* the transfer matrix $\mathbf{H}(\cdot)$ of (8) and the numbers a, b, such that $\mathbf{x} = \mathbf{0}$ is a *globally* uniformly asymptotically stable equilibrium of the system (1) – (3) for *every* function $\Phi: \mathbf{R}_+ \times \mathbf{R}^m \to \mathbf{R}^m$ belonging to the sector $[a, b]$.

In contrast with the systems studied thus far in this chapter, we are concerned at present not with a *particular* system but an entire *family* of systems, since Φ can be *any* nonlinearity in the sector $[a, b]$. The idea is that no *detailed* information about the nonlinearity is assumed — all that is known is that Φ satisfies (10). For this reason, the problem under

study is referred to as an **absolute stability** problem. It is also known as the Lur'e problem, after the Russian scientist A. I. Lur'e.

The Aizerman and Kalman Conjectures

Suppose it were possible to deduce the stability of a family of *nonlinear time-varying* systems by examining only all the *linear time-invariant* systems within that family. Then the absolute stability problem would be very easy to solve. With this in mind, in 1949 the Russian mathematician M. A. Aizerman made a conjecture regarding the absolute stability problem in the case of strictly proper, single-input, single-output systems (i.e., $\mathbf{D} = \mathbf{0}$ and $m = 1$). In this case, the only *linear time-invariant* maps ϕ satisfying (the scalar version of) (10) are

11 $\phi(t, y) = ky, \ \forall t, y, k \in [a, b]$.

Aizerman's conjecture was that if the system (1) – (3) (with $\mathbf{D} = \mathbf{0}$ and $m = 1$) is globally asymptotically stable for all linear time-invariant maps ϕ of the form (11) as the constant k varies over the interval $[a, b]$, then the same is true for *all* time-invariant nonlinear elements ϕ in the sector $[a, b]$. Unfortunately, while it is a tempting conjecture, it is false in general. [But a modified version of it is true; see Theorem (6.6.126).]

In 1957, R. E. Kalman made another conjecture. Suppose $\phi: \mathbf{R} \to \mathbf{R}$ is a memoryless time-invariant nonlinearity, and is continuously differentiable. Then ϕ is said to **belong to the incremental sector** $[a, b]$ if $\phi(0) = 0$, and in addition,

12 $a \leq \phi'(y) \leq b, \ \forall y \in \mathbf{R}$.

Kalman's conjecture was that if the system (1) – (3) (with $\mathbf{D} = \mathbf{0}$ and $m=1$) is globally asymptotically stable for all ϕ of the form (11), then the same is true for all time-invariant nonlinear elements ϕ belonging to the *incremental* sector $[a, b]$.

It is easy to see that if $\phi: \mathbf{R} \to \mathbf{R}$ belongs to the *incremental* sector $[a, b]$, then it also belongs to the sector $[a, b]$; this is a ready consequence of the mean-value theorem. But the converse is not true in general. Thus the family of nonlinearities ϕ covered by Kalman's conjecture is strictly smaller than that covered by Aizerman's conjecture. So Kalman's conjecture "had a better chance" of being true than Aizerman's conjecture. Moreover, using Lyapunov's linearization method [Corollary (5.5.26)], it can be shown that the following statement *is* true (see Problem 5.26): If the system (1) – (3) is globally asymptotically stable for all ϕ of the form (11), or equivalently, if $\mathbf{A} - \mathbf{B}k\mathbf{C}$ is a Hurwitz matrix for all $k \in [a, b]$, then $\mathbf{x} = \mathbf{0}$ is an asymptotically stable equilibrium of the system (1) – (3) for all time-invariant ϕ belonging to the incremental sector $[a, b]$. Thus the essence of Kalman's conjecture lies in replacing "asymptotically stable" by "*globally* asymptotically stable." Nevertheless, Kalman's conjecture is also false in general.

5.6.2 The Circle Criterion

In this subsection, we present a sufficient condition for absolute stability, known as the circle criterion. The contents of this subsection as well as the next depend in an essential way on the following result of independent interest, known as the Kalman-Yacubovitch lemma.

13 Theorem (Kalman-Yacubovitch) *Consider the system (1), where* $\mathbf{x}(t) \in \mathbf{R}^n$, *and* $\mathbf{y}(t), \mathbf{u}(t) \in \mathbf{R}^m$ *with* $m < n$. *Define* $\mathbf{H}(\cdot)$ *as in (8). Suppose (i) the matrix* $\mathbf{A}$ *is Hurwitz, (ii) the pair* $(\mathbf{A}, \mathbf{B})$ *is controllable, (iii) the pair* $(\mathbf{C}, \mathbf{A})$ *is observable, and (iv)*

$$\textbf{14} \quad \inf_{\omega \in \mathbb{R}} \lambda_{\min}[\mathbf{H}(j\omega) + \mathbf{H}^*(j\omega)] > 0,$$

where * *denotes the conjugate transpose, and* $\lambda_{\min}$ *denotes the smallest eigenvalue of a Hermitian matrix. Under these conditions, there exist a symmetric positive definite matrix* $\mathbf{P} \in \mathbf{R}^{n \times n}$, *matrices* $\mathbf{Q} \in \mathbf{R}^{m \times n}$, $\mathbf{W} \in \mathbf{R}^{m \times m}$, *and an* $\varepsilon > 0$ *such that*

$$\textbf{15} \quad \mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\varepsilon\mathbf{P} - \mathbf{Q}'\mathbf{Q},$$

$$\textbf{16} \quad \mathbf{B}'\mathbf{P} + \mathbf{W}'\mathbf{Q} = \mathbf{C},$$

$$\textbf{17} \quad \mathbf{W}'\mathbf{W} = \mathbf{D} + \mathbf{D}'.$$

The proof of Theorem (13) is given in Appendix B. Note that a transfer matrix $\mathbf{H}(\cdot)$ satisfying (14) is said to be **strictly positive real.**

Now the simplest version of the circle criterion, called the passivity theorem, is presented. It will be seen later that more general versions of the criterion can be derived using a technique known as loop transformation.

18 Theorem (Passivity) *Consider the system (1) – (3), and suppose (i) the matrix* $\mathbf{A}$ *is Hurwitz, (ii) the pair* $(\mathbf{A}, \mathbf{B})$ *is controllable, (iii) the pair* $(\mathbf{C}, \mathbf{A})$ *is observable, (iv)* $\mathbf{H}(\cdot)$ *satisfies (14), and (v)* $\mathbf{\Phi}$ *belongs to the sector* $[0, \infty)$, *i.e.,* $\mathbf{\Phi}(t, \mathbf{0}) = \mathbf{0}\ \forall t \geq 0$, *and*

$$\textbf{19} \quad \mathbf{y}'\mathbf{\Phi}(t, \mathbf{y}) \geq 0,\ \forall t \geq 0,\ \forall \mathbf{y} \in \mathbf{R}^m.$$

Under these conditions, the system (1) – (3) is globally exponentially stable.

Proof Conditions (i) to (iv) of the hypotheses imply that Theorem (13) can be applied. Choose $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{W}$ such that (15) – (17) hold. Define the Lyapunov function candidate

$$\textbf{20} \quad V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x}.$$

Then

$$\textbf{21} \quad \dot{V} = \dot{\mathbf{x}}'\mathbf{P}\mathbf{x} + \mathbf{x}'\mathbf{P}\dot{\mathbf{x}}$$

$$= [\mathbf{Ax} - \mathbf{B}\Phi]' \mathbf{Px} + \mathbf{x}'\mathbf{P}[\mathbf{Ax} - \mathbf{B}\Phi]$$

$$= \mathbf{x}'(\mathbf{A}'\mathbf{P} + \mathbf{PA})\mathbf{x} - \Phi'\mathbf{B}'\mathbf{Px} - \mathbf{x}'\mathbf{PB}\Phi,$$

after substituting for $\mathbf{u}$, and letting Φ denote $\Phi[t, \mathbf{y}(t)]$. Now, from (16) it follows that

22 $\mathbf{B}'\mathbf{P} = \mathbf{C} - \mathbf{W}'\mathbf{Q}.$

Hence

23 $\Phi'\mathbf{B}'\mathbf{Px} = \Phi'\mathbf{Cx} - \Phi'\mathbf{W}'\mathbf{Qx}$

$$= \Phi'(\mathbf{y} - \mathbf{Du}) - \Phi'\mathbf{W}'\mathbf{Qx}$$

$$= \Phi'(\mathbf{y} + \mathbf{D}\Phi) - \Phi'\mathbf{W}'\mathbf{Qx}.$$

Next, substituting from (23) into (21) gives

24 $\dot{V} = \mathbf{x}'(\mathbf{A}'\mathbf{P} + \mathbf{PA})\mathbf{x} - \Phi'(\mathbf{D} + \mathbf{D}')\Phi - \Phi'\mathbf{W}'\mathbf{Qx} - \mathbf{x}'\mathbf{Q}'\mathbf{W}\Phi - \Phi'\mathbf{y} - \mathbf{y}'\Phi.$

Now substitute from (15) and (17) into (24), and observe that $\Phi'\mathbf{y} \geq 0$. This leads to

25 $\dot{V} \leq -\varepsilon\mathbf{x}'\mathbf{Px} - \mathbf{x}'\mathbf{Q}'\mathbf{Qx} - \Phi'\mathbf{W}'\mathbf{W}\Phi - \Phi'\mathbf{W}'\mathbf{Qx} - \mathbf{x}'\mathbf{Q}'\mathbf{W}\Phi$

$$= -\varepsilon\mathbf{x}'\mathbf{Px} - [\mathbf{Qx} + \mathbf{W}\Phi]'[\mathbf{Qx} + \mathbf{W}\Phi]$$

$$\leq -\varepsilon\mathbf{x}'\mathbf{Px}.$$

The global exponential stability of the system now follows from Theorem (5.3.62). ■

Theorem (18) only applies to the case where Φ belongs to the sector $[0, \infty)$. However, using a technique known as "loop transformation," the theorem can be modified to cover the case where Φ belongs to a general sector $[a, b]$. The idea is that, if Φ belongs to the sector $[a, b]$, then $\Phi - aI$ belongs to the sector $[0, b-a]$. Consequently, for each $\delta > 0$, the nonlinearity

26 $\Phi_t = (\Phi - aI)\{I + [1/(b - a + \delta](\Phi - aI)\}^{-1}$

belongs to the sector $[0, \infty)$. See Figure 5.17 for an interpretation of the nonlinearity Φ_t. In the process of modifying the feedback element from Φ to Φ_t, the forward path element gets transformed from $\mathbf{H}(\cdot)$ to

27 $\mathbf{H}_t(s) = \mathbf{H}(s)[I + a\mathbf{H}(s)]^{-1} + [1/(b - a + \delta)]I.$

This can be stated formally as follows:

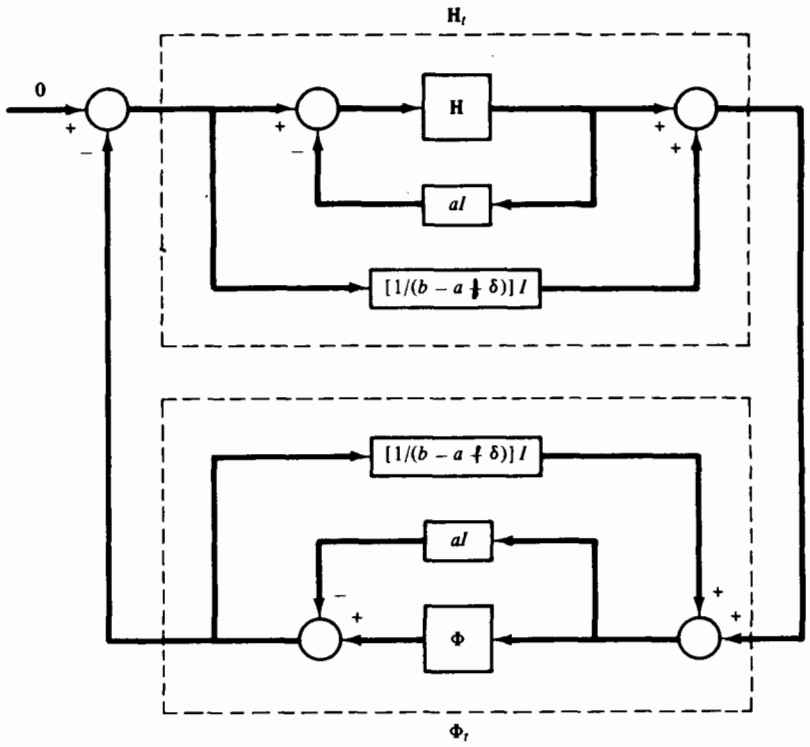


Fig. 5.17

28 Corollary *Consider the system (1) – (3). Suppose (i) the pair* $(\mathbf{A}, \mathbf{B})$ *is controllable, and the pair* $(\mathbf{C}, \mathbf{A})$ *is observable, (ii)* Φ *belongs to the sector* $[a, b]$. *Define*

$$\mathbf{29} \quad \mathbf{H}_a(s) = \mathbf{H}(s)\,[I + a\mathbf{H}(s)]^{-1}.$$

Suppose

$$\mathbf{30} \quad \inf_{\omega \in \mathbb{R}} \lambda_{\min}\{\mathbf{H}_a(j\omega) + \mathbf{H}_a{}^*(j\omega)]/2\} + \frac{1}{b-a} > 0,$$

and all poles of $\mathbf{H}_a(\cdot)$ *have negative real parts. Under these conditions, the system (1) – (3) is exponentially stable.*

Proof The idea is to show that the transformed system satisfies the hypotheses of Theorem (18). Since (30) holds, it follows that

$$\mathbf{31} \quad \inf_{\omega \in \mathbb{R}} \lambda_{\min}[\mathbf{H}_a(j\omega) + \mathbf{H}_a{}^*(j\omega)] + \frac{2}{b-a+\delta} > 0,$$

for sufficiently small $\delta > 0$. Now define $\mathbf{H}_t$ by (27). Then (31) is equivalent to

$$\textbf{32} \quad \inf_{\omega \in \mathbb{R}} \lambda_{\min}[\mathbf{H}_t(j\omega) + \mathbf{H}_t^*(j\omega)] > 0.$$

Thus $\mathbf{H}_t(\cdot)$ satisfies hypothesis (iv) of Theorem (18). As mentioned above, Φ_t satisfies hypothesis (v) of Theorem (18). As for the remaining conditions, a routine calculation shows that a realization for $\mathbf{H}_t(\cdot)$ is given by

$$\textbf{33} \quad \mathbf{A}_t = \mathbf{A} - a\mathbf{B}(I + a\mathbf{D})^{-1}\mathbf{C},$$

$$\mathbf{B}_t = \mathbf{B}(I + a\mathbf{D})^{-1},$$

$$\mathbf{C}_t = (I + a\mathbf{D})^{-1}\mathbf{C},$$

$$\mathbf{D}_t = \mathbf{D}(I + a\mathbf{D})^{-1} + [1/(b - a + \delta)]I.$$

Moreover, it is easy to show that the pair $(\mathbf{A}_t, \mathbf{B}_t)$ is controllable, and that the pair $(\mathbf{C}_t, \mathbf{A}_t)$ is observable, which are respectively hypotheses (ii) and (iii) of Theorem (18). Thus, if all poles of $\mathbf{H}_t(\cdot)$ have negative real parts, then $\mathbf{A}_t$ is Hurwitz, which is the last remaining hypothesis needed to apply Theorem (18). The desired conclusion now follows from Theorem (18). ■

Corollary (28) applies equally well to multi-input, multi-output systems ($m > 1$) and to single-input, single-output systems ($m = 1$). However, in the latter case, it is possible to give an elegant graphical interpretation of the condition (30). This leads to a result commonly known as the circle criterion. To establish the result, it is useful to make the following observation. Suppose $z = x + jy$ is a complex number, and $a, b \in \mathbf{R}$ with $a < b$ and $a \neq 0$ and $b \neq 0$. Then

$$\textbf{34} \quad \text{Re}\,\frac{z}{1 + az} + \frac{1}{b - a} > 0$$

if and only if

$$\textbf{35} \quad \left| z + \frac{b + a}{2ba} \right| > \left| \frac{b - a}{2ba} \right| \text{ if } ba > 0, \; < \left| \frac{b - a}{2ba} \right| \text{ if } ba < 0.$$

This can be established by high school algebra. In fact, both statements are equivalent to

$$\textbf{36} \quad ba(x^2 + y^2) + (b + a)x + 1 > 0.$$

Let $D(a, b)$ denote the closed disk in the complex plane centered at $(b + a)/2ba$ and with radius $(b - a)/2|ba|$. Then the observation is that (34) holds if and only if the complex number z lies *outside* the disk $D(a, b)$ in case $ba > 0$, and lies *in the interior* of the disk $D(a, b)$ in case $ba < 0$.

In Theorem (37) below, reference is made to the (Nyquist) plot of $h(j\omega)$. This is the plot of $h(j\omega)$ as ω increases from $-\infty$ to ∞. If $h(\cdot)$ has a pole on the $j\omega$-axis, the plot is obtained by "indenting" the $j\omega$-axis in such a way that the pole lies to the *left* of the indented $j\omega$-axis; see Figure 5.18.

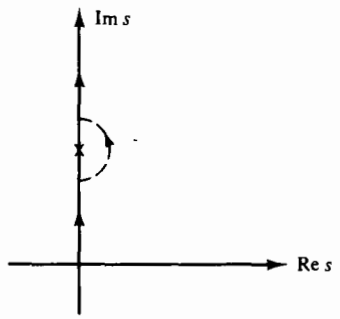


Fig. 5.18

37 Theorem (Circle Criterion) *Consider the system (1) – (3), and suppose*[2] $m = 1$, *(ii) the quadruplet* $(\mathbf{A},\mathbf{b},\mathbf{c},d)$ *is a minimal realization of* $h(\cdot)$, *and (iii)* ϕ *belongs to the sector* $[a, b]$. *Define the disk* $D(a, b)$ *as above. Under these conditions, the system (1) – (3) is globally exponentially stable if one of the following conditions, as appropriate, holds.*

Case (i) $0 < a < b$: The plot of $h(j\omega)$ lies outside and is bounded away from the disk $D(a, b)$; moreover, the plot encircles $D(a, b)$ exactly ν times in the counter-clockwise direction, where ν is the number of eigenvalues of $\mathbf{A}$ with positive real part.

Case (ii) $0 = a < b$: $\mathbf{A}$ is a Hurwitz matrix, and

$$\textbf{38} \qquad \inf_{\omega \in \mathbb{R}} \operatorname{Re} h(j\omega) + \frac{1}{b} > 0.$$

Case (iii) $a < 0 < b$: $\mathbf{A}$ is Hurwitz; the plot of $h(j\omega)$ lies in the interior of the disk $D(a, b)$ and is bounded away from the circumference of $D(a, b)$.

Case (iv): $a < b \leq 0$ Replace $h(\cdot)$ by $-h(\cdot)$, a by $-b$, b by $-a$, and apply (i) or (ii) as appropriate.

Proof *Case (i)*: In this case $ba > 0$. The hypotheses on $h(\cdot)$ imply that (34) holds with z replaced by $h(j\omega)$. Moreover, since $h(j\omega)$ is bounded away from the disk $D(a, b)$, it follows that

---

[2] Since the single-input, single-output case is being considered, matrices are replaced by row or column vectors, or scalars, as appropriate.

39 $$\inf_{\omega \in \mathbb{R}} \operatorname{Re} \frac{h(j\omega)}{1 + a\,h(j\omega)} + \frac{1}{b-a} > 0,$$

which is (30) specialized to the scalar case. Next, the encirclement condition implies that the plot of $h(j\omega)$ encircles the point $-1/a$ exactly ν times in the counter-clockwise direction. Hence by the well-known Nyquist stability criterion [see e.g., Theorem (6.5.35) for a very general version], it follows that all poles of $h_a(\cdot)$ have negative real parts. Since all hypotheses of Corollary (28) hold, the desired conclusion follows.

Case (ii): In this case $h_a = h$, and (38) is the scalar version of (30). Since **A** is Hurwitz, the desired conclusion follows from Corollary (28).

Case (iii): In this case $ba < 0$. Hence the fact that $h(j\omega)$ lies in the *interior* of the disk $D(a, b)$ implies that (34) holds with z replaced by $h(j\omega)$. Moreover, since $h(j\omega)$ is bounded away from the circumference of $D(a, b)$, it follows that (39) holds. Finally, since **A** is Hurwitz, the desired conclusion follows from Corollary (28).

Case (iv): Obvious. ■

An appealing aspect of the circle criterion is its geometric nature, which is reminiscent of the Nyquist criterion. Indeed, if $b - a \to 0$, then the "critical disk" $D(a, b)$ in Case (i) shrinks to the "critical point" $-1/a$ of the Nyquist criterion; in this case the circle criterion reduces to the sufficiency part of the Nyquist criterion. On the other hand, the circle criterion is applicable to time-varying and/or nonlinear systems, whereas the Nyquist criterion is only applicable to linear time-invariant systems.

Another appealing feature of the circle criterion is that it depends only on the transfer function $h(\cdot)$ of the forward path, and not on the particular realization of $h(\cdot)$. This means that if we think of the forward path element as a "black box," then in order to apply the circle criterion it is only necessary to determine the transfer function of this black box, which can be achieved through relatively straight-forward experiments; it is *not* necessary to construct a realization of $h(\cdot)$.

40 Example As an illustration of the circle criterion, suppose the transfer function of the forward-path subsystem in Figure 5.15 is

$$h(s) = \frac{(s+25)^2}{(s+1)(s+2)(s+3)(s+200)}.$$

The plot of $h(j\omega)$ is shown in Figure 5.19, with a portion of it shown in enlargement in Figure 5.20.

Suppose first that the feedback nonlinear element ϕ belongs to the sector $[-5/3, 5]$. The corresponding disk $D(a, b)$ passes through the points -0.2 and 0.6, as shown in Figure 5.19. Moreover, the plot of $h(j\omega)$ lies inside $D(a, b)$. Hence, by Case (iii) of the circle cri-

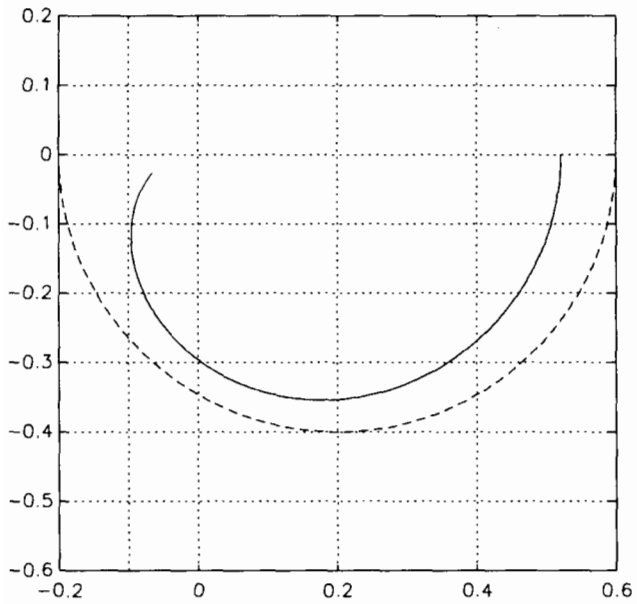


Fig. 5.19

terion, we conclude that the feedback system is globally exponentially stable for all ϕ belonging to the sector $[-5/3, 5]$.

Now suppose ϕ belongs to the sector $[0, 10]$. In this case, (38) is satisfied with $b = 10$. Hence, by Case (ii) of the circle criterion, we can conclude that the feedback system is globally exponentially stable for all ϕ belonging to the sector $[0, 10]$.

At this stage, one might be tempted to combine the above two conclusions, and state that the feedback system is globally exponentially stable for all ϕ belonging to the sector $[-5/3, 10]$, on the basis that

$$[-5/3, 5] \cup [0, 10] = [-5/3, 10].$$

But the statement does not follow. Let $N(a, b)$ denote the set of nonlinear elements belonging to the sector (a, b). Then one can see that

$$N(-5/3, 5) \cup N(0, 10) \neq N(-5/3, 10).$$

As a final application, suppose ϕ belongs to the sector $[4000, 7000]$. The corresponding disk $D(a, b)$ is shown in Figure 5.20. Now it follows from Case (i) of the circle criterion that the feedback system is globally exponentially stable for all ϕ belonging to the sector $[4000, 7000]$.

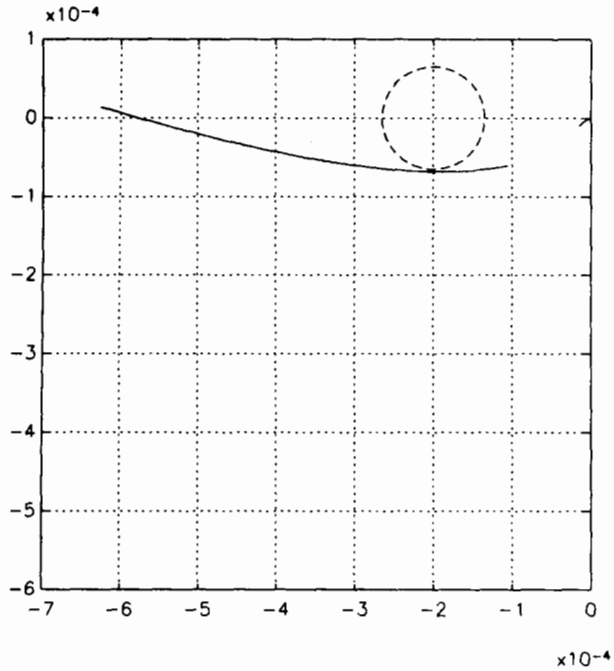


Fig. 5.20

5.6.3 The Popov Criterion

In this section, we derive another criterion for absolute stability, known as the Popov criterion, after the Roumanian scientist V. M. Popov. Unlike the circle criterion, the Popov criterion is applicable only to *autonomous* systems.

The class of systems studied by Popov is described by

41 $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u,$

$\dot{\xi} = u,$

$y = \mathbf{c}\mathbf{x} + d\xi,$

42 $u = -\phi(y),$

where $\mathbf{x} \in \mathbf{R}^n$, ξ, u, y are scalars; and $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}$, d have commensurate dimensions. The nonlinear element ϕ: $\mathbf{R} \to \mathbf{R}$ is a *time-invariant* nonlinearity belonging to the *open* sector $(0, \infty)$. This means that

43 $\phi(0) = 0,\ y\,\phi(y) > 0,\ \forall y \neq 0.$

Notice the differences between the system descriptions (1) – (3) and (41) – (42). In the latter system, there is a pole at the origin, and there is no throughput from the input to the output. Moreover, the nonlinearity belongs to an *open* sector, and not a *closed* sector as in the former system. The system description (41) can be rewritten as

$$\mathbf{44} \quad \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\xi} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \xi \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ 1 \end{bmatrix} u,$$

$$y = [\mathbf{c} \quad d] \begin{bmatrix} \mathbf{x} \\ \xi \end{bmatrix}.$$

Hence its transfer function is

$$\mathbf{45} \quad h(s) = \frac{d}{s} + \mathbf{c}(sI - \mathbf{A})^{-1}\mathbf{b}.$$

46 Theorem (Popov Criterion) *Consider the system (41) – (42) and suppose (i) the matrix* $\mathbf{A}$ *is Hurwitz, (ii) the pair* $(\mathbf{A}, \mathbf{b})$ *is controllable, (iii) the pair* $(\mathbf{c}, \mathbf{A})$ *is observable, (iv)* $d > 0$*, and (v) the nonlinear element* ϕ *belongs to the sector* $(0, \infty)$*. Under these conditions, the system (41) – (42) is globally asymptotically stable if there exists a number* $r > 0$ *such that*

$$\mathbf{47} \quad \inf_{\omega \in \mathbb{R}} \operatorname{Re}[(1 + j\omega r) h(j\omega)] > 0.$$

Remarks The quantity $1 + j\omega r$ is called a "multiplier." Note that if $r = 0$, then (47) reduces to (38) with $b = \infty$. In this sense, and only in this sense, the Popov criterion is a generalization of the circle criterion. One should not carry the comparison too far, since the two criteria apply to distinct classes of systems. Case (ii) of the circle criterion applies to open-loop stable systems, while the Popov criterion applies to systems whose forward-path transfer function has a simple pole at $s = 0$ but are otherwise stable. Moreover, the circle criterion guarantees global *exponential* stability, whereas the Popov criterion only guarantees global *asymptotic* stability.

Proof Note that

$$\mathbf{48} \quad s(sI - \mathbf{A})^{-1} = (sI - \mathbf{A} + \mathbf{A})(sI - \mathbf{A})^{-1} = I + \mathbf{A}(sI - \mathbf{A})^{-1}.$$

Hence

$$\mathbf{49} \quad (1 + rs) h(s) = (1 + rs)\left[\frac{d}{s} + \mathbf{c}(sI - \mathbf{A})^{-1}\mathbf{b}\right]$$

$$= \frac{d}{s} + rd + \mathbf{c}(sI - \mathbf{A})^{-1}\mathbf{b} + r\mathbf{c}\mathbf{b} + r\mathbf{c}\mathbf{A}(sI - \mathbf{A})^{-1}\mathbf{b}.$$

If $s = j\omega$, then the term $d/j\omega$ is purely imaginary, so that

50 $\mathrm{Re}\,[(1+j\omega r)\,h(j\omega)] = \mathrm{Re}\,[r\,(d+\mathbf{cb}) + \mathbf{c}\,(I+r\mathbf{A})\,(j\omega I - \mathbf{A})^{-1}\,\mathbf{b}]$.

Define the transfer function

51 $g(s) = r\,(d+\mathbf{cb}) + \mathbf{c}\,(I+r\mathbf{A})\,(sI-\mathbf{A})^{-1}\,\mathbf{b}$,

and observe that the quadruple $\{\mathbf{A}, \mathbf{b}, \mathbf{c}\,(I+r\mathbf{A}), r\,(d+\mathbf{cb})\}$ is a minimal realization of g. Moreover, in view of (50), (47) is equivalent to

52 $\inf_{\omega\in\mathbb{R}} \mathrm{Re}\,g(j\omega) > 0$.

Hence, by Theorem (13), there exist a symmetric positive definite matrix $\mathbf{P}\in\mathbb{R}^{n\times n}$, a (row) vector $\mathbf{q}\in\mathbb{R}^{1\times n}$, $w\in\mathbb{R}$, and $\varepsilon > 0$, such that

53 $\mathbf{A}'\mathbf{P} + \mathbf{PA} = -\varepsilon\mathbf{P} - \mathbf{q}'\mathbf{q}$,

54 $\mathbf{b}'\mathbf{P} + w\mathbf{q} = \mathbf{c}\,(I+r\mathbf{A})$,

55 $w^2 = r\,(d+\mathbf{cb})$.

To establish the global asymptotic stability of the system (41) – (42), choose the Lyapunov function candidate

56 $V(\mathbf{x}, \xi) = \mathbf{x}'\mathbf{P}\mathbf{x} + d\xi^2 + 2r\,\psi(y)$,

where

57 $\psi(y) = \int_0^y \phi(\sigma)\,d\sigma$.

Since ϕ belongs to the sector $(0, \infty)$, it follows that $\psi(y) \geq 0\ \forall y$. Hence V is positive definite and radially unbounded, and is thus a suitable Lyapunov function candidate for applying Theorem (5.3.56). Now

58 $\dot{V} = \dot{\mathbf{x}}'\mathbf{P}\mathbf{x} + \mathbf{x}'\mathbf{P}\dot{\mathbf{x}} + 2d\xi\dot{\xi} + 2r\,\phi(y)\,\dot{y}$

$= (\mathbf{Ax} - \mathbf{b}\phi)'\,\mathbf{P}\mathbf{x} + \mathbf{x}'\mathbf{P}\,(\mathbf{Ax} - \mathbf{b}\phi) - 2d\xi\phi + 2r\phi\,[\mathbf{c}(\mathbf{Ax} - \mathbf{b}\phi) - d\phi]$.

Now note that $d\xi = y - \mathbf{cx}$. Substituting this relationship in (58) and rearranging gives

59 $\dot{V} = \mathbf{x}'\,(\mathbf{A}'\mathbf{P} + \mathbf{PA})\,\mathbf{x} - 2\phi\mathbf{b}'\mathbf{P}\mathbf{x} + 2\phi\mathbf{c}\,(I+r\mathbf{A})\,\mathbf{x} - 2r\,(d+\mathbf{cb})\,\phi^2 - 2y\phi$.

Now substitute from (53) – (55) into (59). This gives, by familiar arguments,

60 $\dot{V} = -\varepsilon \mathbf{x}'\mathbf{P}\mathbf{x} - (\mathbf{q}\mathbf{x} - w\phi)^2 - r(d + \mathbf{c}\mathbf{b})\phi^2 - 2y\phi.$

Next, since $g(j\omega) \to r(d + \mathbf{cb})$ as $\omega \to \infty$, [cf. (51)], (52) implies that $r(d + \mathbf{cb}) > 0$. Hence from (60), it follows that

61 $\dot{V} \leq -\varepsilon \mathbf{x}'\mathbf{P}\mathbf{x} - 2y\phi \leq 0, \ \forall \mathbf{x}, \xi.$

It is now shown that

62 $\dot{V}(\mathbf{x}, \xi) < 0$ if $(\mathbf{x}, \xi) \neq (\mathbf{0}, 0)$.

If $\mathbf{x} \neq \mathbf{0}$, then $\dot{V} < 0$ since $\mathbf{P} > 0$. If $\mathbf{x} = \mathbf{0}$ but $\xi \neq 0$, then $\dot{y} = d\xi \neq 0$, and $y\phi > 0$ since ϕ belongs to the open sector $(0, \infty)$ [cf. (43)]. Hence once again $\dot{V} < 0$. Now the global asymptotic stability of the system follows from Theorem (5.3.56). ■

63 Corollary *Consider the system (41) – (42). Let all hypotheses be as in Theorem (46), except that ϕ belongs to the sector $(0, k)$ where $k > 0$ is some finite number. Under these conditions, the system (41) – (42) is globally asymptotically stable if there exists a number $r > 0$ such that*

64 $\inf_{\omega \in \mathbb{R}} \operatorname{Re}[(1 + j\omega r)h(j\omega)] + \frac{1}{k} > 0.$

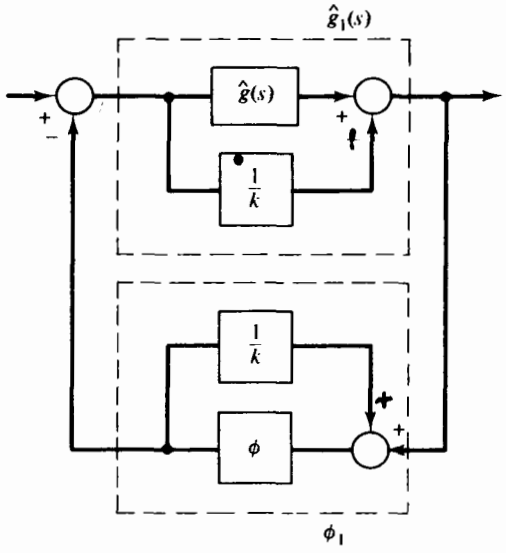


Fig. 5.21

Proof Perform a loop transformation as shown in Figure 5.21, by placing a negative feedback of $-1/k$ around ϕ and a positive feedforward of $1/k$ around h. This leads to

65 $\phi_t = \phi\,[I - (1/k)\phi]^{-1}$,

which belongs to the sector $(0, \infty)$, and

66 $h_t(s) = h(s) + \dfrac{1}{k}$.

Now suppose (64) holds. Then, since

67 $\mathrm{Re}\,[(1 + j\omega r)\,h_t(j\omega)] = \mathrm{Re}\,[(1 + j\omega r)\,h(j\omega)] + \dfrac{1}{k}$,

it follows that

68 $\inf\limits_{\omega \in \mathbb{R}} \mathrm{Re}\,[(1 + j\omega r)\,h_t(j\omega)] > 0.$

The global asymptotic stability of the system now follows from Theorem (46). ■

Like the circle criterion, the Popov criterion can also be given a graphical interpretation. Suppose we plot $\mathrm{Re}\,h(j\omega)$ vs. $\omega\,\mathrm{Im}\,h(j\omega)$ as ω varies from 0 to ∞. Note that, since both $\mathrm{Re}\,h(j\omega)$ and $\omega\,\mathrm{Im}\,h(j\omega)$ are even functions of ω, it is not necessary to plot negative values of ω. The resulting plot is known as the **Popov plot**, in contrast with the Nyquist plot, which is a plot of $\mathrm{Re}\,h(j\omega)$ vs. $\mathrm{Im}\,h(j\omega)$. The inequality (64) states that there exists a nonnegative number r such that the Popov plot of h lies to the right of a straight line of slope $1/r$ passing through the point $-1/k$. If $r = 0$, the straight line is vertical.

69 **Example** Consider a system of the form (41) – (42), with

$$h(s) = \frac{1}{s(s+1)^2}.$$

The Popov plot of h is shown in Figure 5.22. It is clear from this figure that, if $k < 2$, then it is always possible to draw a straight line through the point $-1/k$ such that the plot lies to the right of the straight line. Hence, by Corollary (63), we conclude that the feedback system is globally asymptotically stable for all *time-invariant* nonlinearities in the sector $(0, 2)$.

Problem 5.27 Consider a feedback system of the form (1) – (3) with

$$h(s) = \frac{1}{(s+1)(s+2)(s+3)}.$$

Using the circle criterion, determine several intervals $[a, b]$ such that the feedback system is globally exponentially stable for all ϕ belonging to the sector $[a, b]$.

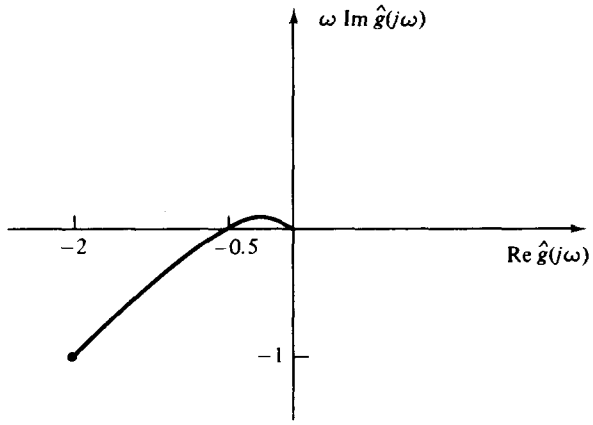


Fig. 5.22

Problem 5.28 Consider a feedback system of the form (1) – (3) with

$$h(s) = \frac{1}{(s-1)(s+3)^2}.$$

Using the circle criterion, determine numbers $0 < a < b$ such that the feedback system is globally exponentially stable for all nonlinearities in the sector $[a, b]$.

Problem 5.29 Consider a system of the form (41) – (42) with

$$h(s) = \frac{1}{s(s+1)}.$$

Using the Popov criterion, show that the feedback system is globally asymptotically stable for all time-invariant nonlinearities ϕ belonging to the sector $(0, k)$ where k is any *finite* number.

5.7 CONVERSE THEOREMS

In this section, the so-called *converse theorems* of Lyapunov stability theory are stated and proved. In the next section these theorems are applied to four problems in control theory, and it is shown that converse theorems lead to elegant solutions to each of these problems.

Though there are several converse theorems [see e.g., Hahn (1967), Sections 48 to 51], only three are stated here, namely those for uniform asymptotic stability, exponential stability, and global exponential stability. In essence, these theorems state that the conditions of Theorems (5.3.25), (5.3.45) and (5.3.62) are necessary as well as sufficient. If the equilibrium **0** has the stability property mentioned in each of these theorems, then there exists a Lyapunov function V satisfying the conditions of the theorem. Since the function V is

constructed in terms of the solution trajectories of the system, the converse theorems cannot really be used to construct an explicit formula for the Lyapunov function, except in special cases (e.g., linear systems; see Section 5.4). However, they can be used in the same way as the Lyapunov functions for stable linear systems are used in Section 5.5: Knowing something about the stability status of System A allows us to conclude something about the stability status of a related System B.

We begin by presenting two preliminary results.

1 Lemma *(Gronwall 1919) Suppose* $a: \mathbf{R}_+ \to \mathbf{R}_+$ *is a continuous function, and* $b, c \geq 0$ *are given constants. Under these conditions, if*

$$2 \qquad a(t) \leq b + \int_0^t c\, a(\tau)\, d\tau, \ \forall t \geq 0,$$

then

$$3 \qquad a(t) \leq b \exp(ct), \ \forall t \geq 0.$$

Remarks The point of the lemma is to convert the implicit bound in (2) to the explicit bound in (3). If b and c are not constants but are themselves functions of t, then the bound (3) needs to be replaced by a more complicated expression. This is known as *Bellman's inequality*; see Bellman (1953).

Proof Define

$$4 \qquad d(t) = b + \int_0^t c\, a(\tau)\, d\tau.$$

Then (2) states that

$$5 \qquad a(t) \leq d(t), \ \forall t \geq 0.$$

Further, from (4) and (5),

$$6 \qquad \dot{d}(t) = c\, a(t) \leq c\, d(t), \ \forall t \geq 0.$$

Now using the integrating factor $\exp(-ct)$ in (6), one can show that (6) implies

$$7 \qquad d(t) \leq d(0) \exp(ct) = b \exp(ct).$$

The conclusion (3) follows from (7) and (8). ■

8 Lemma *(Massera 1949) Suppose* $\sigma(\cdot)$ *is a given function of class L, and* $\lambda > 0$ *is a given constant. Then there exists a* C^∞ *function* $\gamma: \mathbf{R}_+ \to \mathbf{R}_+$ *such that (i)* γ *and* γ' *are both functions of class K, and (ii)*

$$\mathbf{9} \quad \int_0^\infty \gamma[\sigma(\tau)]\,d\tau < \infty, \int_0^\infty \gamma'[\sigma(\tau)]\exp(\lambda\tau)\,d\tau < \infty.$$

Remark Note that, throughout this section, the prime is used to denote the derivative of a function. This is in contrast with the usage in the remainder of the book, where the prime denotes the transpose of a matrix. Since all primed quantities in this section are scalar-valued functions, no confusion should result from this altered convention.

The proof of this lemma is long and technical, and the reader will miss very little by accepting it on faith and moving ahead to Theorem (24).

Proof Observe first that it is enough to prove the lemma under the additional assumption that $\sigma(0) = 1$. To see this, suppose the lemma is true for all functions $\sigma(\cdot)$ of class L with the additional property that $\sigma(0) = 1$, and let $\theta(\cdot)$ be an arbitrary function of class L. Now define $\sigma(t) = \theta(t)/\theta(0)$, and note that $\sigma(0) = 1$. Select a function $\gamma(\cdot)$ such that the conditions of the lemma are true, and define $\phi: \mathbf{R}_+ \to \mathbf{R}_+$ by $\phi(r) = \gamma[r/\theta(0)]$. Then clearly $\phi[\theta(\tau)] = \gamma[\sigma(\tau)]$, and

$$\mathbf{10} \quad \int_0^\infty \phi[\theta(\tau)]\,d\tau < \infty, \int_0^\infty \phi'[\theta(\tau)]\exp(\lambda\tau)\,d\tau < \infty.$$

So suppose $\sigma(0) = 1$. Since $\sigma(\cdot)$ is strictly decreasing and $\sigma(t) \to 0$ as $t \to \infty$, for each integer $n \geq 0$ there exists a unique time t_n such that $\sigma(t_n) = 1/(n+1)$. [Of course, $t_0 = 0$ since $\sigma(0) = 1$.] Now define a continuous function $\bar{\eta}: (0, \infty) \to (0, \infty)$ as follows: (i) $\bar{\eta}(t_n) = 1/n$ for all $n \geq 1$. (ii) In the interval (t_n, t_{n+1}), $\bar{\eta}(t)$ is an affine (i.e., linear plus a constant) function of t. (iii) In the interval $(0, t_1)$, $\bar{\eta}(t) = 1/t^p$, where $0 < p < 1$. A pictorial representation of $\bar{\eta}(\cdot)$ is shown in Figure 5.23. Now $\bar{\eta}$ has the following properties: (i) $\bar{\eta}$ is strictly decreasing, and is thus a one-to-one map of $(0, \infty)$ onto itself; (ii) $\bar{\eta}(t) > \sigma(t)\ \forall t > 0$; (iii) for each number T, we have

$$\mathbf{11} \quad \int_0^T \bar{\eta}(t)\,dt < \infty.$$

This is because, as $t \to \infty$, the function $\bar{\eta}(t)$ "blows up" quite slowly since $p < 1$. Now $\bar{\eta}$ is continuously differentiable up to all orders except at a countable number of values of t. By rounding out the corners, one can replace $\bar{\eta}$ by another function η which is C^∞ and which continues to have the same three properties. Now define $\mu, \gamma: \mathbf{R}_+ \to \mathbf{R}_+$ by

$$\mathbf{12} \quad \mu(r) = \exp[-(\lambda+1)\eta^{-1}(r)] \text{ if } r > 0, \mu(0) = 0,$$

$$\mathbf{13} \quad \gamma(r) = \int_0^r \mu(s)\,ds.$$

It is claimed that this $\gamma(\cdot)$ is the desired function. To see this, note first that $\eta^{-1}: (0, \infty) \to (0, \infty)$ is well-defined, that $\eta^{-1}(r) \to 0$ as $r \to \infty$, and $\eta^{-1}(r) \to \infty$ as $r \to 0$.

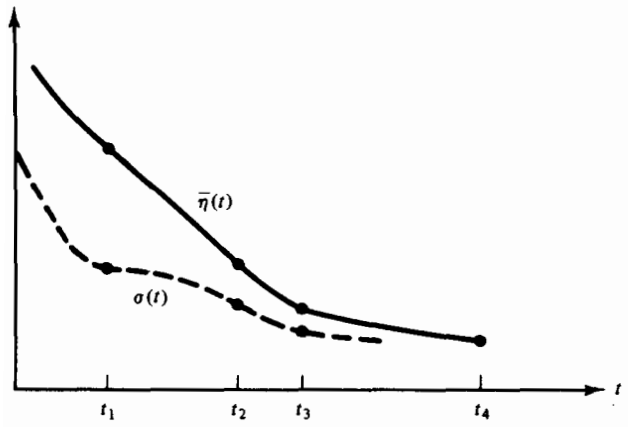


Fig. 5.23

Hence μ is continuous and is in class K, $\mu(0)=0$, and $\mu(r)\to 1$ as $r\to\infty$. Since $\gamma'(r)=\mu(r)$, the second integral in (9) becomes

$$\textbf{14}\quad \int_0^\infty \mu[\sigma(\tau)]\exp(\lambda\tau)\,d\tau =: I_2.$$

To show that I_2 is finite, note that

$$\textbf{15}\quad \mu[\sigma(\tau)] \le \mu[\eta(\tau)], \text{ since } \mu \text{ belongs to class K,}$$

$$= \exp[-(\lambda+1)\eta^{-1}(\eta(\tau))] = \exp[-(\lambda+1)\tau].$$

Now it is clear that I_2 is finite.

Showing that the first integral in (9) is finite requires a bit of work. Since $\sigma(\tau)\le\eta(\tau)\ \forall\tau$, it is enough to show that

$$\textbf{16}\quad I := \int_0^\infty \gamma[\eta(\tau)]\,d\tau < \infty.$$

Now

$$\textbf{17}\quad \gamma[\eta(\tau)] = \int_0^{\eta(\tau)} \mu(s)\,ds = \int_0^{\eta(\tau)} \exp[-(\lambda+1)\eta^{-1}(s)]\,ds.$$

Make the change of variables

18 $s = \eta(v), ds = \eta'(v)\,dv.$

Then $s = \eta(\tau)$ if and only if $v = \tau$, and $s = 0$ if and only if $v = \infty$. Thus

$$\textbf{19} \quad \gamma[\eta(v)] = -\int_{\tau}^{\infty} \exp[-(\lambda+1)v]\,\eta'(v)\,dv,$$

$$\textbf{20} \quad I = -\int_{0}^{\infty}\int_{\tau}^{\infty} \exp[-(\lambda+1)v]\,\eta'(v)\,dv\,d\tau,$$

$$= -\int_{0}^{\infty}\int_{0}^{v} \exp[-(\lambda+1)v]\,\eta'(v)\,d\tau\,dv,$$

upon interchanging the order of integration. Since the integrand in (20) is independent of τ,

$$\textbf{21} \quad I = -\int_{0}^{\infty} v \exp[-(\lambda+1)v]\,\eta'(v)\,dv$$

$$= -\int_{0}^{t_1} v \exp[-(\lambda+1)v]\,\eta'(v)\,dv - \int_{t_1}^{\infty} v \exp[-(\lambda+1)v]\,\eta'(v)\,dv,$$

where t_1 is the first element of the sequence $\{t_n\}$ defined after (10). Over the interval $(0, t_1)$, the quantity $\exp[-(\lambda+1)v]$ is bounded, and

$$\textbf{22} \quad -v\,\eta'(v) = v\,\frac{p}{v^{p+1}} = \frac{p}{v^p}.$$

Since $p < 1$, the integral from 0 to t_1 is finite. Over the interval (t_1, ∞), the quantity $\eta'(v)$ is bounded, and the function $v \exp[-(\lambda+1)v]$ is absolutely integrable. Hence I is finite. ■

Before presenting the converse theorems, a little bit of notation is introduced. If f is a function of several arguments, then D_i denotes the partial derivative of f with respect to the i-th argument. Thus, for example, if $f = f(x, y, z)$, then

$$\textbf{23} \quad D_1 f = \frac{\partial f}{\partial x}, D_2 f = \frac{\partial f}{\partial y}, D_3 f = \frac{\partial f}{\partial z}.$$

At last we come to the converse theorems themselves. The first theorem is a converse of Theorem (5.3.25) on uniform asymptotic stability.

24 Theorem *Consider the system*

$$\textbf{25} \quad \dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], \forall t \geq 0,$$

and suppose that $f(t, \mathbf{0}) = \mathbf{0}\ \forall t \geq 0$ *and that* f *is* C^k *for some integer* $k \geq 1$. *Suppose that there exist a constant* $r > 0$, *a function* ϕ *of class K, and a function* σ *of class L, such that the*

solution trajectories of (25) satisfy

26 $\|\mathbf{s}(t, t_0, \mathbf{x}_0)\| \leq \phi(\|\mathbf{x}_0\|)\sigma(t - t_0), \forall t \geq t_0 \geq 0, \forall \mathbf{x}_0 \in B_r.$

Finally, suppose in addition that, for some finite λ,

27 $\|D_2\mathbf{f}(t, \mathbf{x})\| \leq \lambda, \forall t \geq 0, \forall \mathbf{x} \in B_{\phi(r)\sigma(0)}.$

Under these conditions, there exist a C^k *function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *and* C^∞ *functions* α, β, γ *of class K such that*

28 $\alpha(\|\mathbf{x}\|) \leq V(t, \mathbf{x}) \leq \beta(\|\mathbf{x}\|), \forall t \geq 0, \forall \mathbf{x} \in B_r,$

29 $\dot{V}(t, \mathbf{x}) \leq -\gamma(\|\mathbf{x}\|), \forall t \geq 0, \forall \mathbf{x} \in B_r,$

30 $\sup_{\mathbf{x} \in B_r} \|D_2 V(t, \mathbf{x})\| < \infty.$

Remarks

1. Condition (26) is equivalent to the requirement that $\mathbf{0}$ be a uniformly asymptotically stable equilibrium; cf. Theorem (5.1.61).
2. If $\mathbf{f}$ does not depend explicitly on t, then (27) is automatically satisfied since $\mathbf{f}$ is C^1 and the closure of B_r is compact. Hence this condition only comes into the picture for nonautonomous systems.
3. Conditions (28) and (29) are the same as (5.3.27) and (5.3.28), respectively. Condition (30) is an added bonus, so to speak.
4. The Lyapunov function V is differentiable as many times as is the function $\mathbf{f}$. Thus, if $\mathbf{f}$ is C^∞, then so is V.

Proof For convenience, let $\|\cdot\|$ denote both the Euclidean norm on $\mathbf{R}^n$ as well as the corresponding induced matrix norm on $\mathbf{R}^{n \times n}$.

The solution trajectories of (25) satisfy

31 $$\mathbf{s}(t, t_0, \mathbf{x}_0) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}[\tau, \mathbf{s}(\tau, t_0, \mathbf{x}_0)]\, d\tau, \forall t \geq t_0.$$

Differentiating with respect to $\mathbf{x}_0$ gives

32 $$D_3\mathbf{s}(t, t_0, \mathbf{x}_0) = I + \int_{t_0}^{t} D_2\mathbf{f}[\tau, \mathbf{s}(\tau, t_0, \mathbf{x}_0)]\, D_3\mathbf{s}(\tau, t_0, \mathbf{x}_0)\, d\tau,$$

and as a consequence,

$$\textbf{33}\quad \|D_3\mathbf{s}(t, t_0, \mathbf{x}_0)\| \le 1 + \int_{t_0}^{t} \lambda \|D_3\mathbf{s}(\tau, t_0, \mathbf{x}_0)\|\, d\tau,$$

where we have used (27). Thus, by Gronwall's inequality [Lemma (1)],

$$\textbf{34}\quad \|D_3\mathbf{s}(t, t_0, \mathbf{x}_0)\| \le \exp[\lambda(t-t_0)],\ \forall t \ge t_0 \ge 0.$$

Similarly, differentiating (31) with respect to t_0 and using Leibniz' rule gives

$$\textbf{35}\quad D_2\mathbf{s}(t, t_0, \mathbf{x}_0) = -\mathbf{f}(t_0, \mathbf{x}_0) + \int_{t_0}^{t} D_2\mathbf{f}[\tau, \mathbf{s}(\tau, t_0, \mathbf{x}_0)]\, D_2\mathbf{s}(\tau, t_0, \mathbf{x}_0)\, d\tau.$$

Note that, by (27),

$$\textbf{36}\quad \|\mathbf{f}(t_0, \mathbf{x}_0)\| \le \lambda \|\mathbf{x}_0\| \le \lambda r.$$

Hence, from (35),

$$\textbf{37}\quad \|D_2\mathbf{s}(t, t_0, \mathbf{x}_0)\| \le \lambda r + \int_{t_0}^{t} \lambda \|D_2\mathbf{s}(\tau, t_0, \mathbf{x}_0)\|\, d\tau.$$

Applying Gronwall's inequality again results in

$$\textbf{38}\quad \|D_2\mathbf{s}(t, t_0, \mathbf{x}_0)\| \le \lambda r \exp[\lambda(t-t_0)],\ \forall t \ge t_0 \ge 0.$$

Next, select a C^∞ function γ such that both γ and γ' are of class K, and

$$\textbf{39}\quad \int_0^\infty \gamma[\sigma_1(\tau)]\, d\tau < \infty,\ \int_0^\infty \gamma'[\sigma_1(\tau)] \exp(\lambda\tau)\, d\tau < \infty,$$

where

$$\textbf{40}\quad \sigma_1(\tau) = \phi(r)\, \sigma(\tau).$$

Such a function $\gamma(\cdot)$ exists by Massera's lemma. Now define

$$\textbf{41}\quad V(t, \mathbf{x}) = \int_t^\infty \gamma[\|\mathbf{s}(\tau, t, \mathbf{x})\|]\, d\tau.$$

It is claimed that this function V satisfies (28) to (30). First, whenever $\mathbf{x} \in B_r$, (26) implies that

$$\textbf{42}\quad \|\mathbf{s}(\tau, t, \mathbf{x})\| \le \phi(r)\, \sigma(\tau - t) = \sigma_1(\tau - t),$$

so that $V(t, \mathbf{x})$ is well-defined. Next, it is shown that V is C^k. From (31) and the fact that $\mathbf{f}$ is C^k, it follows that the function[3] $\|\mathbf{s}(\tau, t, \mathbf{x})\|^2 = \mathbf{s}^T\mathbf{s}$ is C^k, so V is C^k if the integrand in (41)

[3] In this proof we use $(\cdot)^T$ to denote the transpose, since the prime is used to denote the derivative.

can be expressed as a C^k function of $\|\mathbf{s}\|^2$ and the integral converges uniformly. For this purpose, define

43 $\delta(r) = \gamma(\sqrt{r})$,

44 $\gamma(r) = \delta(r^2)$.

Then δ is C^∞ whenever $r \neq 0$. Now

45 $\delta'(r) = \dfrac{1}{2\sqrt{r}}\gamma'(\sqrt{r})$.

If we use the construction in the proof of Lemma (8), then

46 $\gamma'(r) = \mu(r) = \exp[-(\lambda+1)\eta^{-1}(r)]$.

Since $\eta^{-1}(r) \to \infty$ as $r \to 0$, we see that δ is C^∞ at $r = 0$ also. Hence the integrand in (41) is a C^k function. To show that V is itself a C^k function, the uniform convergence of the integral must be demonstrated. For this purpose, note that

47 $D_1 V(t, \mathbf{x}) = -\gamma(\|\mathbf{x}\|) + \int_t^\infty D_2\gamma[\|\mathbf{s}(\tau, t, \mathbf{x})\|]\, d\tau$.

But, since $\|\mathbf{s}\| = (\mathbf{s}^T\mathbf{s})^{1/2}$,

48 $D_2\gamma[\|\mathbf{s}(\tau, t, \mathbf{x})\|] = \dfrac{\gamma'[\|\mathbf{s}(\tau, t, \mathbf{x})\|]\,[\mathbf{s}(\tau, t, \mathbf{x})]^T D_2\mathbf{s}(\tau, t, \mathbf{x})}{\|\mathbf{s}(\tau, t, \mathbf{x})\|}$,

49 $\|D_2\gamma[\|\mathbf{s}(\tau, t, \mathbf{x})\|]\| \leq \gamma'[\|\mathbf{s}(\tau, t, \mathbf{x})\|]\, \|D_2\mathbf{s}(\tau, t, \mathbf{x})\|$.

Substituting this inequality in (47) gives

50 $|D_1 V(t, \mathbf{x})| \leq \gamma(\|\mathbf{x}\|) + \int_t^\infty \gamma'[\|\mathbf{s}(\tau, t, \mathbf{x})\|]\, \|D_2\mathbf{s}(\tau, t, \mathbf{x})\|\, d\tau$.

But, from (26),

51 $\|\mathbf{s}(\tau, t, \mathbf{x})\| \leq \phi(r)\,\sigma(\tau - t) = \sigma_1(\tau - t)$.

Using (51) and (38) in (50) gives

52 $|D_1 V(t, \mathbf{x})| \leq \gamma(\|\mathbf{x}\|) + \int_t^\infty \gamma'[\sigma_1(\tau - t)]\,\lambda r \exp[\lambda(\tau - t)]\, d\tau < \infty$.

Similarly,

$$\textbf{53}\quad \|D_2V(t,\mathbf{x})\| \le \int_t^\infty \gamma'[\|\mathbf{s}(\tau,t,\mathbf{x})\|]\,\|D_3\mathbf{s}(\tau,t,\mathbf{x})\|\,d\tau$$

$$\le \int_t^\infty \gamma'[\sigma_1(\tau-t)]\exp[\lambda(\tau-t)]\,d\tau < \infty,$$

where we have used (34). Thus V is C^k, and in fact V satisfies (30).

To prove that V is decrescent, observe that, for each $\mathbf{x}\in B_r$,

$$\textbf{54}\quad V(t,\mathbf{x}) \le \int_t^\infty \gamma[\sigma_1(\tau-t)]\,d\tau < \infty.$$

Hence the function $\beta\colon \mathbf{R}_+ \to \mathbf{R}_+$ defined by

$$\textbf{55}\quad \beta(d) = \sup_{\|\mathbf{x}\|\le d}\ \sup_{t\ge 0} V(t,\mathbf{x})$$

is well-defined for $d < r$ and can be extended to all of $\mathbf{R}_+$. If β is not C^∞, it can be bounded above by a C^∞ function.

To prove that V is an lpdf, note that (27) implies that

$$\textbf{56}\quad \|\mathbf{f}(t,\mathbf{x})\| \le \lambda\|\mathbf{x}\| \le \lambda r,\ \forall \mathbf{x}\in B_r.$$

Hence, from (31),

$$\textbf{57}\quad \|\mathbf{s}(\tau,t,\mathbf{x})\| \ge \|\mathbf{x}\| - \lambda r(\tau-t).$$

In particular,

$$\textbf{58}\quad \|\mathbf{s}(\tau,t,\mathbf{x})\| \ge \|\mathbf{x}\|/2 \text{ for } \tau\in[t,\ t+\|\mathbf{x}\|/2\lambda r].$$

Hence, from (41),

$$\textbf{59}\quad V(t,\mathbf{x}) \ge \frac{\|\mathbf{x}\|}{2\lambda r}\gamma(\|\mathbf{x}\|/2) =: \alpha(\|\mathbf{x}\|).$$

Clearly $\alpha(d) = d\gamma(d/2)/2\lambda r$ is C^∞.

Finally, to prove that $\dot{V}$ is locally negative definite, observe from (41) that

$$\textbf{60}\quad V[\tau,\mathbf{s}(\tau,t,\mathbf{x})] = \int_\tau^\infty \gamma\{\|\mathbf{s}[\theta,\tau,\mathbf{s}(\tau,t,\mathbf{x})]\|\}\,d\theta = \int_\tau^\infty \gamma[\|\mathbf{s}(\theta,t,\mathbf{x})\|]\,d\theta.$$

Hence

61 $$\frac{d}{d\tau} V[\tau, \mathbf{s}(\tau, t, \mathbf{x})] = -\gamma[\|\mathbf{s}(\tau, t, \mathbf{x})\|].$$

In other words,

62 $$\dot{V}(t, \mathbf{x}) = -\gamma(\|\mathbf{x}\|),$$

and γ is C^∞. This completes the proof. ■

The next result provides the converse of Theorem (5.3.45) for exponential stability.

63 Theorem *Consider the system (25), and suppose* $\mathbf{f}$ *is* C^k, *and that* $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}\ \forall t \geq 0$. *Suppose there exist constants* $\mu, \delta, r > 0$ *such that*

64 $$\|\mathbf{s}(\tau, t, \mathbf{x})\| \leq \mu \|\mathbf{x}\| \exp[-\delta(\tau - t)],\ \forall \tau \geq t \geq 0,\ \forall \mathbf{x} \in B_r.$$

Finally, suppose that, for some finite constant λ,

65 $$\|D_2\mathbf{f}(t, \mathbf{x})\| \leq \lambda,\ \forall t \geq 0,\ \forall \mathbf{x} \in B_{\mu r}.$$

Under these conditions, there exist a C^k *function* $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *and constants* $a, b, c, m > 0, p > 1$, *such that*

66 $$a\|\mathbf{x}\|^p \leq V(t, \mathbf{x}) \leq b\|\mathbf{x}\|^p,\ \dot{V}(t, \mathbf{x}) \leq -c\|\mathbf{x}\|^p,\ \forall t \geq 0,\ \forall \mathbf{x} \in B_r,$$

67 $$\|D_2 V(t, \mathbf{x})\| \leq m\|\mathbf{x}\|^{p-1},\ \forall t \geq 0,\ \forall \mathbf{x} \in B_r.$$

Proof Choose the constant $p > 1$ such that

68 $$(p - 1)\delta > \lambda,$$

and let

69 $$\gamma(r) = r^p,\ \sigma_1(t) = \mu r \exp(-\delta t).$$

Then, as can be verified readily, the condition (39) is satisfied. Hence, by Theorem (24), the function

70 $$V(t, \mathbf{x}) = \int_t^\infty \gamma[\|\mathbf{s}(\tau, t, \mathbf{x})\|]\, d\tau = \int_t^\infty \|\mathbf{s}(\tau, t, \mathbf{x})\|^p\, d\tau$$

is a Lyapunov function. Showing that V satisfies (66) is quite straight-forward. First,

71 $$V(t, \mathbf{x}) \leq \int_t^\infty \mu^p \|\mathbf{x}\|^p \exp[-p\delta(\tau - t)]\, d\tau = \frac{\mu^p}{p\delta} \|\mathbf{x}\|^p.$$

Next, from (64) and (65),

72 $\|\mathbf{f}[\tau, \mathbf{s}(\tau, t, \mathbf{x})]\| \leq \lambda \|\mathbf{s}(\tau, t, \mathbf{x})\| \leq \lambda\mu \|\mathbf{x}\|, \forall \tau \geq t \geq 0.$

Hence, from (31),

73 $\|\mathbf{s}(\tau, t, \mathbf{x})\| \geq \dfrac{\|\mathbf{x}\|}{2}, \ \forall \tau \in [t, t + 1/2\lambda\mu].$

Therefore (70) implies that

74 $V(t, \mathbf{x}) \geq \displaystyle\int_t^{t+1/2\lambda\mu} \frac{\|\mathbf{x}\|^p}{2^p} d\tau = \frac{1}{2^{p+1}\lambda\mu} \|\mathbf{x}\|^p.$

As shown in (62),

75 $\dot{V}(t, \mathbf{x}) = -\gamma(\|\mathbf{x}\|) = -\|\mathbf{x}\|^p.$

Finally, using (53), and noting that $\gamma'(r) = pr^{p-1}$, gives

76 $\|D_2 V(t, \mathbf{x})\| \leq \displaystyle\int_0^{\infty} p\mu^{p-1} \|\mathbf{x}\|^{p-1} \exp[-(p-1)\delta\tau] \exp(\lambda\tau)\, d\tau$

$$= \text{constant} \cdot \|\mathbf{x}\|^{p-1}.$$

This completes the proof. ■

The next corollary shows that, for an exponentially stable equilibrium, it is possible to construct a "quadratic type" Lyapunov function.

77 Corollary *Suppose all hypotheses of Theorem (63) are satisfied. Then there exist a* C^k *function* $W: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ *and constants* $\alpha, \beta, \gamma, \mu > 0$ *such that*

78 $\alpha\|\mathbf{x}\|^2 \leq W(t, \mathbf{x}) \leq \beta\|\mathbf{x}\|^2, \ \dot{W}(t, \mathbf{x}) \leq -\gamma\|\mathbf{x}\|^2, \ \forall t \geq 0, \ \forall \mathbf{x} \in B_r,$

79 $\|D_2 W(t, \mathbf{x})\| \leq \mu\|\mathbf{x}\|, \ \forall t \geq 0, \ \forall \mathbf{x} \in B_r.$

Proof Construct V as in Theorem (63), and let

80 $W(t, \mathbf{x}) = [V(t, \mathbf{x})]^{2/p}.$

The details are left as an exercise. ■

Theorem (63) and Corollary (77) can be extended in a totally straight-forward manner to *global* exponential stability.

81 Theorem *Consider the system (25). Suppose* $\mathbf{f}$ *is* C^k, *and that* $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}\ \forall t \geq 0$. *Suppose there exist constants* $\mu, \delta, \lambda > 0$ *such that*

82 $\|\mathbf{s}(\tau, t, \mathbf{x})\| \leq \mu \|\mathbf{x}\| \exp[-\delta(\tau - t)],\ \forall \tau \geq t \geq 0,\ \forall \mathbf{x} \in \mathbf{R}^n,$

83 $\|D_2\mathbf{f}(t, \mathbf{x})\| \leq \lambda,\ \forall t \geq 0,\ \forall \mathbf{x} \in \mathbf{R}^n.$

Under these conditions, there exist a C^k *function* W *and constants* $\alpha, \beta, \gamma, \mu > 0$ *such that*

84 $\alpha \|\mathbf{x}\|^2 \leq W(t, \mathbf{x}) \leq \beta \|\mathbf{x}\|^2,\ \dot{W}(t, \mathbf{x}) \leq -\gamma \|\mathbf{x}\|^2,\ \forall t \geq 0,\ \forall \mathbf{x} \in \mathbf{R}^n,$

85 $\|D_2 W(t, \mathbf{x})\| \leq \mu \|\mathbf{x}\|,\ \forall t \geq 0,\ \forall \mathbf{x} \in \mathbf{R}^n.$

Proof Simply follow the proof of Theorem (63) and replace B_r and $B_{\mu r}$ by $\mathbf{R}^n$ throughout. In this way, it follows that the function V defined in (70) satisfies (66) and (67) for *all* $\mathbf{x} \in \mathbf{R}^n$. Now define W by (80). ■

Note that the condition (83) requires the function $\mathbf{f}$ to be *globally* Lipschitz continuous; this is a much more restrictive condition than (65), especially for autonomous systems.

If we attempt to extend Theorem (24) to global uniform asymptotic stability, then we run into the difficulty that the function V defined in (41) may not be decrescent. The problem arises in (58), where the upper bound on τ approaches t as $r \to \infty$. In the case of exponential stability, the difficulty does not arise, because the function $\phi(\|\mathbf{x}\|)$ is linear in $\|\mathbf{x}\|$; see (72) to (74). Hence Theorem (24) can be extended to cover global stability *provided* the function $\phi: \mathbf{R}_+ \to \mathbf{R}_+$ defined by

86 $$\phi(r) = \sup_{\mathbf{x} \in B_r} \sup_{t \geq 0} \sup_{\tau \geq t} \|\mathbf{s}(\tau, t, \mathbf{x})\|$$

can be bounded by a linear function of r.

5.8 APPLICATIONS OF CONVERSE THEOREMS

In this section, the converse theorems derived in the preceding section are used to solve four problems in control theory.

5.8.1 Exponential Stability of Nonlinear Systems

1 Theorem *Consider the system*

2 $\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t)],$

where $\mathbf{f}$ *is* C^2, *and* $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. *Define*

$$\mathbf{3} \qquad \mathbf{A} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{0}}$$

Then $\mathbf{0}$ *is an exponentially stable equilibrium of (2) if and only if the linearized system*

$$\mathbf{4} \qquad \dot{\mathbf{z}}(t) = \mathbf{A}\mathbf{z}(t)$$

is (globally) exponentially stable.

Theorem (1) resolves one of the issues left hanging in Corollary (5.5.26). If $\mathbf{A}$ is defined as above and if all eigenvalues of $\mathbf{A}$ have negative real parts, then $\mathbf{0}$ is an exponentially stable equilibrium. If some eigenvalues of $\mathbf{A}$ have a positive real part, then $\mathbf{0}$ is an unstable equilibrium. But what if all eigenvalues of $\mathbf{A}$ have nonpositive real parts, but some have a zero real part? In such a case, depending on the nature of the neglected higher order terms, it is possible for the origin to be asymptotically stable. But Theorem (1) shows that, even if the origin is asymptotically stable, it cannot be *exponentially* stable.

Proof "If" This is just Corollary (5.5.26).

"Only if" Suppose $\mathbf{0}$ is an exponentially stable equilibrium of (25). Then, by Corollary (5.7.77), there exists a C^2 function $V: \mathbf{R}^n \to \mathbf{R}$ and constants $\alpha, \beta, \gamma, \mu$, and $r > 0$ such that

$$\mathbf{5} \qquad \alpha \|\mathbf{x}\|^2 \leq V(t, \mathbf{x}) \leq \beta \|\mathbf{x}\|^2, \dot{V}(t, \mathbf{x}) \leq -\gamma \|\mathbf{x}\|^2, \forall \mathbf{x} \in B_r,$$

$$\mathbf{6} \qquad \|\nabla V(t, \mathbf{x})\| \leq \mu \|\mathbf{x}\|, \forall \mathbf{x} \in B_r.$$

[Compare (5.7.78) and (5.7.79).] Expand V and $\dot{V}$ in a Taylor series around $\mathbf{x} = \mathbf{0}$. Now $V(\mathbf{0}) = \mathbf{0}$ and $\dot{V}(\mathbf{0}) = \mathbf{0}$. Also, since both V and $\dot{V}$ are sign-definite, there cannot be a linear term in the Taylor series expansion. In other words, V and $\dot{V}$ are of the form

$$\mathbf{7} \qquad V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x} + V_1(\mathbf{x}),$$

$$\mathbf{8} \qquad \dot{V}(\mathbf{x}) = -\mathbf{x}'\mathbf{Q}\mathbf{x} + W_1(\mathbf{x}),$$

where both $\mathbf{P}$ and $\mathbf{Q}$ are symmetric and positive definite. Expand $\mathbf{f}$ in the form

$$\mathbf{9} \qquad \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{f}_1(\mathbf{x}),$$

where

$$\mathbf{10} \qquad \lim_{\|\mathbf{x}\| \to 0} \frac{\|\mathbf{f}_1(\mathbf{x})\|}{\|\mathbf{x}\|} = 0.$$

Now using (10) and (7) gives

11 $\dot{V}(\mathbf{x}) = \nabla V(\mathbf{x})\,\mathbf{f}(\mathbf{x}) = \mathbf{x}'\,[\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A}]\,\mathbf{x} + \mathbf{h}(\mathbf{x})$,

where $\mathbf{h}(\mathbf{x})$ denotes a term which decays more rapidly than a quadratic. Comparing (8) and (11) shows that

12 $\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$.

Finally, (12) and (5) show that $S(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x}$ is a suitable Lyapunov function for applying Theorem (5.3.45) to establish the global exponential stability of the linearized system (4). ■

5.8.2 Slowly Varying Systems

Consider a general nonautonomous system described by

13 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)], \forall t \geq 0.$

If $r \in \mathbf{R}_+$ is any fixed number, then we can think of the *autonomous* system

14 $\dot{\mathbf{x}}(t) = \mathbf{f}[r, \mathbf{x}(t)], \forall t \geq 0,$

as a particular case of the system (13), with its time dependence "frozen" at time r. Example (5.4.90) shows that, even if each of the frozen systems is exponentially stable, the overall system can be unstable. But it is now shown that if each frozen system is exponentially stable *and the system is slowly varying*, then (13) is indeed exponentially stable. For this purpose, a little notation is needed. As usual, let $\mathbf{s}(\tau, t, \mathbf{x})$ denote the solution of (13) starting at time t and state $\mathbf{x}$, and evaluated at time τ. Let $\mathbf{s}_r(\tau, t, \mathbf{x})$ denote the solution of the *frozen* system (14), starting at time t and state $\mathbf{x}$, and evaluated at time τ.

15 **Theorem** *Consider the system (13). Suppose (i)* $\mathbf{f}$ *is* C^1, *and (ii)*

16 $\sup_{\mathbf{x} \in \mathbf{R}^n} \sup_{t \geq 0} \| D_2 \mathbf{f}(t, \mathbf{x}) \| =: \lambda < \infty,$

(iii) there exist constants $\mu, \delta > 0$ *such that*

17 $\| \mathbf{s}_r(\tau, t, \mathbf{x}_0) \| \leq \mu \| \mathbf{x}_0 \| \exp[-\delta(\tau - t)], \forall \tau \geq t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n, \forall r \in \mathbf{R}_+.$

Finally, suppose there is a constant $\varepsilon > 0$ *such that*

18 $\| D_1 \mathbf{f}(t, \mathbf{x}) \| \leq \varepsilon \| \mathbf{x} \|, \forall t \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$

Then the nonautonomous system (13) is globally exponentially stable provided

19 $\varepsilon < \dfrac{\delta[(p-1)\delta - \lambda]}{p\mu^p},$

where $p > 1$ *is any number such that* $(p-1)\delta > \lambda$.

Remarks To put the conditions of the theorem in better perspective, consider the linear time-varying system

20 $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\,\mathbf{x}(t),$

and suppose $\mathbf{A}(t)$ is a Hurwitz matrix for each fixed t. Then λ is the maximum of $\|\mathbf{A}(t)\|$ with respect to t; $-\delta$ is the largest (i.e., the least negative) of the real parts of the eigenvalues of $\mathbf{A}(t)$, as t varies; and μ is the maximum of the condition number of $\mathbf{A}(t)$ as t varies.

Proof We begin by estimating the rate of variation of the function $\mathbf{s}_r(\tau, 0, \mathbf{x})$ with respect to r. From (14), it follows that

21 $$\mathbf{s}_r(\tau, 0, \mathbf{x}) = \mathbf{x} + \int_0^\tau \mathbf{f}[r, \mathbf{s}_r(\sigma, 0, \mathbf{x})]\, d\sigma.$$

Differentiating with respect to r gives

22 $$\frac{\partial}{\partial r}\mathbf{s}_r(\tau, 0, \mathbf{x}) = \int_0^\tau D_1\mathbf{f}[r, \mathbf{s}_r(\sigma, 0, \mathbf{x})]\, d\sigma + \int_0^\tau D_2\mathbf{f}[r, \mathbf{s}_r(\sigma, 0, \mathbf{x})]\, \frac{\partial}{\partial r}\mathbf{s}_r(\sigma, 0, \mathbf{x})\, d\sigma.$$

For conciseness, define

23 $g(\tau) = \|\partial \mathbf{s}_r(\tau, 0, \mathbf{x})/\partial r\|,$

and note from (18) that

24 $\|D_1\mathbf{f}[r, \mathbf{s}_r(\sigma, 0, \mathbf{x})]\| \le \varepsilon\|\mathbf{s}_r(\sigma, 0, \mathbf{x})\| \le \varepsilon\mu\|\mathbf{x}\| \exp(-\delta\sigma).$

Using (24) and (16) in (22) gives

25 $$g(\tau) \le \int_0^\tau \varepsilon\mu\|\mathbf{x}\| \exp(-\delta\sigma)\, d\sigma + \int_0^\tau \lambda g(\sigma)\, d\sigma$$

$$\le \frac{\varepsilon\mu\|\mathbf{x}\|}{\delta} + \int_0^\tau \lambda g(\sigma)\, d\sigma.$$

Applying Gronwall's lemma to (25) gives

26 $$\|\partial \mathbf{s}_r(\tau, 0, \mathbf{x})/\partial r\| = g(\tau) \le \frac{\varepsilon\mu\|\mathbf{x}\|}{\delta} \exp(\lambda\tau),\ \forall \tau \ge 0.$$

Next, for each $r \in \mathbf{R}_+$, define a Lyapunov function $V_r: \mathbf{R}^n \to \mathbf{R}$ for the system (14) as in Theorem (5.7.63). Select $p > 1 + \lambda/\delta$ [i.e., $(p-1)\delta > \lambda$], and define

27 $$V_r(\mathbf{x}) = \int_0^\infty \| \mathbf{s}_r(\tau, 0, \mathbf{x}) \|^p \, d\tau.$$

This is the same function as in (5.7.70), since the system (14) is autonomous. At this stage, replace r by t, and define $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ by

28 $$V(t, \mathbf{x}) = \int_0^\infty \| \mathbf{s}_t(\tau, 0, \mathbf{x}) \|^p \, d\tau.$$

Then, as shown in the proof of Theorem (5.7.63), it follows in analogy with (5.7.71), (5.7.74), and (5.7.75) that

29 $$\frac{1}{2^{p+1}\lambda\mu} \| \mathbf{x} \|^p \leq V(t, \mathbf{x}) \leq \frac{\mu^p}{p\delta} \| \mathbf{x} \|^p,$$

30 $$D_2 V(t, \mathbf{x}) \mathbf{f}(t, \mathbf{x}) = -\| \mathbf{x} \|^p.$$

Let us compute the derivative $\dot{V}(t, \mathbf{x})$ along the trajectories of (13). By definition,

31 $$\dot{V}(t, \mathbf{x}) = D_1 V(t, \mathbf{x}) + D_2 V(t, \mathbf{x}) \mathbf{f}(t, \mathbf{x}) = D_1 V(t, \mathbf{x}) - \| \mathbf{x} \|^p.$$

It only remains to estimate $D_1 V(t, \mathbf{x})$. Let $\gamma := p/2$. Then, from (28),

32 $$D_1 V(t, \mathbf{x}) = \int_0^\infty \frac{\partial}{\partial t} [\mathbf{s}_t'(\tau, 0, \mathbf{x}) \mathbf{s}_t(\tau, 0, \mathbf{x})]^\gamma \, d\tau$$

$$= \int_0^\infty 2\gamma \, [\mathbf{s}_t'(\tau, 0, \mathbf{x}) \mathbf{s}_t(\tau, 0, \mathbf{x})]^{\gamma - 1} \mathbf{s}_t'(\tau, 0, \mathbf{x}) \frac{\partial}{\partial t} \mathbf{s}_t(\tau, 0, \mathbf{x}) \, d\tau,$$

33 $$| D_1 V(t, \mathbf{x}) | \leq \int_0^\infty 2\gamma \, \| \mathbf{s}_t(\tau, 0, \mathbf{x}) \|^{2\gamma - 1} \, \| \frac{\partial}{\partial t} \mathbf{s}_t(\tau, 0, \mathbf{x}) \| \, d\tau.$$

Now use the bounds (17) for $\| \mathbf{s}_t(\tau, 0, \mathbf{x}) \|$ and (26) for $\| \partial \mathbf{s}_t(\tau, 0, \mathbf{x})/\partial t \|$, and note that $2\gamma = p$. This gives

34 $$| D_1 V(t, \mathbf{x}) | \leq \int_0^\infty p\mu^{p-1} \| \mathbf{x} \|^{p-1} \frac{\varepsilon\mu \| \mathbf{x} \|}{\delta} \exp[-(p-1)\delta\tau + \lambda\tau] \, d\tau$$

$$= \frac{p\varepsilon\mu^p}{\delta[(p-1)\delta - \lambda]} \| \mathbf{x} \|^p.$$

Let m denote the constant multiplying $\| \mathbf{x} \|^p$ on the right side, and note that $m < 1$ by (19). Finally, from (31),

35 $\dot{V}(t, \mathbf{x}) \leq -(1-m)\|\mathbf{x}\|^p$.

Now (29) and (35) show that V is a suitable Lyapunov function for applying Theorem (5.3.62) to conclude global exponential stability. ■

5.8.3 Observer-Controller Stabilization

In this subsection, it is shown that a well-known strategy for stabilizing linear time-invariant systems also works for nonlinear systems.

As a motivation for studying the problem, consider a *linear* time-invariant system described by

36 $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$, $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)$.

Suppose the system is stabilizable and detectable [see Chen (1986) or Kailath (1980) for definitions of these terms]. By the assumption of stabilizability, there exists a matrix $\mathbf{K}$ such that $\mathbf{A} - \mathbf{B}\mathbf{K}$ is Hurwitz. Hence, if we could apply the feedback control

37 $\mathbf{u}(t) = -\mathbf{K}\mathbf{x}(t)$,

then the resulting closed-loop system would be stable. However, $\mathbf{x}(t)$ cannot be measured directly, and only $\mathbf{y}(t)$ is available for control purposes. To overcome this difficulty, one can set up a *detector*, which is a system of the form

38 $\dot{\mathbf{z}}(t) = \mathbf{A}\mathbf{z}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{F}[\mathbf{y}(t) - \mathbf{C}\mathbf{z}(t)]$,

where $\mathbf{F}$ is called the *filter gain*. By the assumption of detectability, there exists a matrix $\mathbf{F}$ such that $\mathbf{A} - \mathbf{F}\mathbf{C}$ is Hurwitz. For such a choice of $\mathbf{F}$, it follows that $\mathbf{z}(t) - \mathbf{x}(t) \to \mathbf{0}$ as $t \to \infty$; in other words, asymptotically $\mathbf{z}(t)$ becomes an accurate estimate of $\mathbf{x}(t)$. With this in mind, suppose one implements the control law

39 $\mathbf{u}(t) = -\mathbf{K}\mathbf{z}(t)$.

Then the closed-loop system is described by

40 $$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & -\mathbf{B}\mathbf{K} \\ \mathbf{F}\mathbf{C} & \mathbf{A} - \mathbf{B}\mathbf{K} - \mathbf{F}\mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}.$$

Let $\mathbf{M}$ denote the square matrix in (40), and define

41 $$\mathbf{T} = \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}.$$

Then

$$\mathbf{42}\quad \mathbf{T}^{-1}\mathbf{M}\mathbf{T}=\begin{bmatrix}\mathbf{A}-\mathbf{BK} & -\mathbf{BK}\\ 0 & \mathbf{A}-\mathbf{FC}\end{bmatrix}.$$

This shows that the spectrum of $\mathbf{M}$ (i.e., the set of eigenvalues of $\mathbf{M}$) is just the union of the spectra of $\mathbf{A}-\mathbf{BK}$ and $\mathbf{A}-\mathbf{FC}$. Since both matrices are Hurwitz, it follows that $\mathbf{M}$ is also Hurwitz. The conclusion is that the stabilizing control law (39) continues to do the job even if the true state $\mathbf{x}(t)$ is replaced by the estimated state $\mathbf{z}(t)$. For this reason, the strategy is known as *observer-controller stabilization.* This is sometimes called the (deterministic) separation theorem.

The preceding proof is very much a "linear time-invariant" proof, being based on eigenvalue arguments. Thus it is perhaps surprising that a similar result also holds for nonlinear nonautonomous systems. Suppose the system to be stabilized is described by

$$\mathbf{43}\quad \dot{\mathbf{x}}(t)=\mathbf{f}[t,\ \mathbf{x}(t),\ \mathbf{u}(t)],\ \mathbf{y}(t)=\mathbf{g}[t,\ \mathbf{x}(t)],$$

where $\mathbf{x}(t)\in\mathbf{R}^n$, $\mathbf{u}(t)\in\mathbf{R}^m$, $\mathbf{y}(t)\in\mathbf{R}^l$, $\mathbf{f}\colon \mathbf{R}_+\times\mathbf{R}^n\times\mathbf{R}^m\to\mathbf{R}^n$, and $\mathbf{g}\colon \mathbf{R}_+\times\mathbf{R}^n\to\mathbf{R}^l$. It is assumed that $\mathbf{f}$ is C^1 and that $\mathbf{f}(t,\ \mathbf{0},\ \mathbf{0})=\mathbf{0}$, $\forall t\geq 0$, Now suppose $\mathbf{h}\colon \mathbf{R}_+\times\mathbf{R}^n\to\mathbf{R}^m$ is continuous, and define the control law

$$\mathbf{44}\quad \mathbf{u}(t)=\mathbf{h}[t,\ \mathbf{x}(t)]$$

This control law is said to **stabilize** the system (43) if $\mathbf{h}(t,\ \mathbf{0})=\mathbf{0}\ \forall t\geq 0$, and $\mathbf{0}$ is a uniformly asymptotically stable equilibrium of the closed-loop system

$$\mathbf{45}\quad \dot{\mathbf{x}}(t)=\mathbf{f}\{t,\ \mathbf{x}(t),\ \mathbf{h}[t,\ \mathbf{x}(t)]\}.$$

Now a nonlinear analog of detectability is defined. The system (43) is said to be **weakly detectable** if one can find a function $\mathbf{r}\colon \mathbf{R}_+\times\mathbf{R}^n\times\mathbf{R}^m\times\mathbf{R}^l\to\mathbf{R}^n$ such that

$$\mathbf{46}\quad \dot{\mathbf{z}}(t)=\mathbf{r}[t,\ \mathbf{z}(t),\ \mathbf{u}(t),\ \mathbf{y}(t)]=\mathbf{r}\{t,\ \mathbf{z}(t),\ \mathbf{u}(t),\ \mathbf{g}[t,\ \mathbf{x}(t)]\}$$

acts as a "weak detector" for the system (43). This means that (i) $\mathbf{r}(t,\ \mathbf{0},\ \mathbf{0},\ \mathbf{0})=\mathbf{0}$, and (ii) there exist a C^1 function $W\colon \mathbf{R}_+\times\mathbf{R}^n\times\mathbf{R}^n\to\mathbf{R}$, class K functions α, β, γ and a number $\rho>0$ such that

$$\mathbf{47}\quad \alpha(\|\mathbf{x}-\mathbf{z}\|)\leq W(t,\ \mathbf{x},\ \mathbf{z})\leq\beta(\|\mathbf{x}-\mathbf{z}\|),\ \forall t\geq 0,\ \forall(\mathbf{x},\ \mathbf{z})\in B_\rho\times B_\rho,$$

$$\mathbf{48}\quad D_1W(t,\ \mathbf{x},\ \mathbf{z})+D_2W(t,\ \mathbf{x},\ \mathbf{z})\,\mathbf{f}(t,\ \mathbf{x},\ \mathbf{u})+D_3W(t,\ \mathbf{x},\ \mathbf{z})\,\mathbf{r}[t,\ \mathbf{x},\ \mathbf{u},\ \mathbf{g}(t,\ \mathbf{x})]$$

$$\leq-\gamma(\|\mathbf{x}-\mathbf{z}\|),\ \forall t\geq 0,\ \forall(\mathbf{x},\ \mathbf{z},\ \mathbf{u})\in B_\rho\times B_\rho\times B_\rho.$$

These assumptions mean that if initially $\|\mathbf{x}(t_0)\|<\rho$, $\|\mathbf{z}(t_0)\|<\rho$, and if $\|\mathbf{u}(t)\|<\rho\ \forall t\geq t_0$, *and if* the resulting trajectory $[\mathbf{x}(t),\ \mathbf{z}(t)]$ does not leave $B_\rho\times B_\rho$, then actually $\mathbf{x}(t)-\mathbf{z}(t)\to\mathbf{0}$ as $t\to\infty$. As the name implies, weak detectability is a weaker property than detectability, though for linear systems the two concepts are equivalent. This is

discussed further in Vidyasagar (1980b).

Now it is possible to state the main result:

49 Theorem *Suppose the system (43) is weakly detectable, and that with the control law (44), the equilibrium* $\mathbf{x}=\mathbf{0}$ *of the system (45) is uniformly asymptotically stable. Suppose there exist finite constants* r, λ, μ *such that*

$$\mathbf{50}\quad \sup_{t\geq 0}\sup_{(\mathbf{x},\mathbf{u})\in B_r\times B_r}\max\{\|D_2\mathbf{f}(t,\mathbf{x},\mathbf{u})\|,\ \|D_3\mathbf{f}(t,\mathbf{x},\mathbf{u})\|\}\leq\lambda,$$

$$\mathbf{51}\quad \sup_{t\geq 0}\sup_{\mathbf{x}\in B_r}\|D_2\mathbf{h}(t,\mathbf{x})\|\leq\mu.$$

Then the origin in $\mathbf{R}^n\times\mathbf{R}^n$ *is a uniformly asymptotically stable equilibrium of the system*

$$\mathbf{52}\quad \dot{\mathbf{x}}(t)=\mathbf{r}\{t,\mathbf{x}(t),\mathbf{h}[t,\mathbf{z}(t)]\},$$

$$\mathbf{53}\quad \dot{\mathbf{z}}(t)=\mathbf{r}\{t,\mathbf{x}(t),\mathbf{h}[t,\mathbf{z}(t)],\mathbf{g}[t,\mathbf{x}(t)]\}.$$

Proof It is first shown that $(\mathbf{0},\mathbf{0})\in\mathbf{R}^n\times\mathbf{R}^n$ is uniformly stable; then it is shown that it is uniformly attractive.

To prove that the origin is uniformly stable, suppose an $\varepsilon>0$ is specified; the objective is to construct a $\delta>0$ such that

$$\mathbf{54}\quad \|\mathbf{x}_0\|<\delta,\ \|\mathbf{z}_0\|<\delta\Rightarrow\|\mathbf{x}(t)\|<\varepsilon,\ \|\mathbf{z}(t)\|<\varepsilon,\ \forall t\geq t_0,$$

where $\mathbf{x}(t)=\mathbf{x}(t,t_0,\mathbf{x}_0,\mathbf{z}_0)$, $\mathbf{z}(t)=\mathbf{z}(t,t_0,\mathbf{x}_0,\mathbf{z}_0)$ denote the solution of (52) – (53) corresponding to the initial condition

$$\mathbf{55}\quad \mathbf{x}(t_0)=\mathbf{x}_0,\ \mathbf{z}(t_0)=\mathbf{z}_0.$$

By assumption, $\mathbf{x}=\mathbf{0}$ is a uniformly asymptotically stable equilibrium of the system (45). Now (50) and (51) ensure that the function $(t,\mathbf{x})\mapsto\mathbf{f}[t,\mathbf{x},\mathbf{h}(t,\mathbf{x})]$ satisfies all the hypotheses of Theorem (5.7.24). Hence there exists a C^1 function $V:\mathbf{R}_+\times\mathbf{R}^n\to\mathbf{R}_+$, C^∞ functions θ,ϕ,ψ of class K, and a finite constant $L>0$, such that

$$\mathbf{56}\quad \theta(\|\mathbf{x}\|)\leq V(t,\mathbf{x})\leq\phi(\|\mathbf{x}\|),\ \forall t\geq 0,\ \forall\mathbf{x}\in B_r,$$

$$\mathbf{57}\quad D_1V(t,\mathbf{x})+D_2V(t,\mathbf{x})\mathbf{f}[t,\mathbf{x},\mathbf{h}(t,\mathbf{x})]\leq-\psi(\|\mathbf{x}\|),\ \forall t\geq 0,\ \forall\mathbf{x}\in B_r,$$

$$\mathbf{58}\quad \|D_2V(t,\mathbf{x})\|\leq L,\ \forall t\geq 0,\ \forall\mathbf{x}\in B_r.$$

Note that the same B_r appears in (51) as well as (56) – (58). This cuts down on the proliferation of symbols, without any loss of generality. By the same token, it can be assumed that $\rho=r$, where ρ appears in (47) – (48), and that $\varepsilon\leq r$. The construction of the quantity δ is achieved in several stages. First, select

59 $\delta_1 = \min\{\varepsilon, \rho, \rho/\mu\}$,

where μ appears in (51). Next, select ε_1 such that $\phi(\varepsilon_1) \leq \theta(\delta_1/2)$, and select δ_2 such that

60 $\delta_2 = \min\{\varepsilon_1, r, \psi(\varepsilon_1)/L\lambda\mu\}$,

where L, λ, μ are defined in (58), (50), and (51) respectively. Next, choose $\delta_3 > 0$ such that $\beta(\delta_3) < \alpha(\delta_2)$, and define

61 $\delta = \min\{\rho, \delta_3/2\}$.

To show that the above choice of δ satisfies (54), it is first shown that

62 $\|\mathbf{x}(t) - \mathbf{z}(t)\| < \delta_2, \ \forall t \geq t_0$,

63 $\dfrac{d}{dt} W[t, \mathbf{x}(t), \mathbf{z}(t)] \leq 0, \ \forall t \geq t_0$,

64 $\dfrac{d}{dt} V[t, \mathbf{x}(t)] < 0$, whenever $\varepsilon_1 \leq \|\mathbf{x}(t)\| < \varepsilon/2, \ \forall t \geq t_0$.

To follow the proof, it is helpful to observe that

65 $\delta \leq \dfrac{\delta_3}{2}, \ \delta_3 < \delta_2 < \varepsilon_1 < \dfrac{\delta_1}{2} \leq \dfrac{\varepsilon}{2}$.

In order to prove (62) – (64), observe first that all three statements are true at $t = t_0$. Since $\|\mathbf{x}(t_0)\| = \|\mathbf{x}_0\| < \delta$ and $\|\mathbf{z}(t_0)\| = \|\mathbf{z}_0\| < \delta$, it follows that $\|\mathbf{x}_0 - \mathbf{z}_0\| < 2\delta \leq \delta_3 < \delta_2$, which is (62) at $t = t_0$. Next, $\|\mathbf{h}(t_0, \mathbf{x}_0)\| \leq \mu \|\mathbf{x}_0\| < \mu\delta < \mu\delta_1 < \rho$. Hence (48) is applicable [with $\mathbf{u} = \mathbf{h}(t_0, \mathbf{x}_0)$], and (63) holds at $t = t_0$. Finally, (64) holds vacuously since $\|\mathbf{x}_0\| < \varepsilon_1$. Now suppose (62) – (64) hold for all $t \in [t_0, T]$; it is shown that they also hold for all $t \in [T, T + \tau]$ for some sufficiently small positive τ. By assumption (62) – (64) hold for all $t \in [t_0, T]$. Hence, in particular,

66 $\|\mathbf{x}(t) - \mathbf{z}(t)\| < \delta_2, \ \forall t \in [t_0, T]$,

67 $\dfrac{d}{dt} V[t, \mathbf{x}(t)] < 0$, whenever $\varepsilon_1 \leq \|\mathbf{x}(t) \leq \varepsilon, \ \forall t \in [t_0, T]$.

Hence

68 $\theta[\|\mathbf{x}(t)\|] \leq V[t, \mathbf{x}(t)] \leq \max\{V(t_0, \mathbf{x}_0), \phi(\varepsilon_1)\} \leq \phi(\varepsilon_1) < \theta(\delta_1/2)$,

and therefore

69 $\|\mathbf{x}(t)\| < \delta_1/2, \ \forall t \in [t_0, T]$.

Since the solution trajectories of the system (52) – (53) are continuous, it follows that for

sufficiently small positive τ, the analogs of (66) and (69) also hold for $t \in [T, T+\tau]$. This establishes the "extensibility" of (62). To do the same for (63), observe that, for $t \in [T, T+\tau]$, we have $\|\mathbf{x}(t)\| < \delta_1/2$ and $\|\mathbf{x}(t) - \mathbf{z}(t)\| < \delta_2 < \delta_1/2$, from which it follows that $\|\mathbf{z}(t)\| < \delta_1$. Hence (50) and (59) imply that $\|\mathbf{h}[t, \mathbf{z}(t)]\| < \rho$. Thus, by applying (48) with $\mathbf{u} = \mathbf{h}[t, \mathbf{z}(t)]$, we get

$$\mathbf{70} \quad \frac{d}{dt} W[t, \mathbf{x}(t), \mathbf{z}(t)] \leq 0, \forall t \in [T, T+\tau],$$

establishing the extensibility of (63). Finally, to extend (64), observe that if $\varepsilon_1 \leq \|\mathbf{x}(t)\| < \varepsilon/2$, we have

$$\mathbf{71} \quad \frac{d}{dt} V[t, \mathbf{x}(t) = D_1 V[t, \mathbf{x}(t)] + D_2 V[t, \mathbf{x}(t)] \mathbf{f}\{t, \mathbf{x}(t), \mathbf{h}[t, \mathbf{z}(t)]\}$$

$$= D_1 V[t, \mathbf{x}(t)] + D_2 V[t, \mathbf{x}(t)] \mathbf{f}\{t, \mathbf{x}(t), \mathbf{h}[t, \mathbf{x}(t)]\}$$

$$+ D_2 V[t, \mathbf{x}(t)] (\mathbf{f}\{t, \mathbf{x}(t), \mathbf{h}[t, \mathbf{z}(t)]\} - \mathbf{f}\{t, \mathbf{x}(t), \mathbf{h}[t, \mathbf{x}(t)]\})$$

$$\leq -\psi[\|\mathbf{x}(t)\|] + L\lambda \|\mathbf{h}[t, \mathbf{z}(t)] - \mathbf{h}[t, \mathbf{x}(t)]\|, \text{ by (58) and (50)}$$

$$< -\psi(\varepsilon_1) + L\lambda\mu\delta_2 \leq 0 \text{ by (60).}$$

This extends (64). This reasoning shows that there is no "first time" T at which (62) – (64) *fail* to hold; i.e., these equations hold for *all* $t \geq t_0$, as claimed.

With the aid of (62) – (64), it is easy to establish (54), which is uniform stability. Using (62) – (64), one can prove (66) and (69) as before, with the interval $[t_0, T]$ replaced by $[t_0, \infty)$. Finally, since $\|\mathbf{z}(t)\| \leq \|\mathbf{x}(t)\| + \|\mathbf{z}(t) - \mathbf{x}(t)\| < (\delta_1/2 + \delta_2) \leq \delta_1 \leq \varepsilon$, (54) follows, and the origin in $\mathbf{R}^n \times \mathbf{R}^n$ is a uniformly stable equilibrium.

To conclude the proof, it is shown that the origin in $\mathbf{R}^n \times \mathbf{R}^n$ is uniformly attractive. Pick any $\varepsilon > 0$, and construct a corresponding $\delta > 0$ as above. It is shown that $B_\delta \times B_\delta$ is a region of uniform attraction. Suppose $\mathbf{x}_0, \mathbf{z}_0 \in B_\delta$. Then, since $\|\mathbf{z}(t)\| < \delta_1 \; \forall t \geq t_0$, it follows from (50) and (59) that $\|\mathbf{h}[t, \mathbf{z}(t)]\| < \rho$. It has already been established that $\|\mathbf{x}(t) - \mathbf{z}(t)\| < \delta_1 < \rho$, $\forall t \geq t_0$. Hence (48) holds all along the trajectory, from which it follows that $\|\mathbf{x}(t_0 + t) - \mathbf{z}(t_0 + t)\| \to 0$ as $t \to \infty$, uniformly with respect to t_0. Equivalently, there exists a function $\bar{\sigma}$ of class L such that

$$\mathbf{72} \quad \|\mathbf{x}(t) - \mathbf{z}(t)\| \leq \bar{\sigma}(t - t_0), \forall t \geq t_0.$$

Now, by a slight modification of (71), it follows that

73 $\dfrac{d}{dt}V[t, \mathbf{x}(t)] \leq -\psi(\phi^{-1}\{V[t, \mathbf{x}(t)]\}) + L\lambda\mu \|\mathbf{x}(t) - \mathbf{z}(t)\|, \forall t \geq t_0.$

To cut down on the number of parentheses, let η denote $\psi \cdot \phi^{-1}$, and note that η is also a class K function. Next, fix t_0, and define

74 $v(t) = V[t_0 + t, \mathbf{x}(t_0 + t)],$

75 $\sigma(t) = L\lambda\mu\bar{\sigma}(t).$

Then it follows from (72) and (73) that

76 $\dot{v}(t) \leq \bar{\eta}[v(t)] + \sigma(t).$

The proof is complete if it can be shown that $v(t) \to 0$ as $t \to \infty$. Unfortunately, this step requires rather advanced concepts; hence only a sketch of the proof is given here. First, a so-called "comparison equation" is set up, namely

77 $\dot{e}(t) = \bar{\eta}[e(t)] + \sigma(t).$

It can be shown that

78 $e(0) \geq v(0) \Rightarrow e(t) \geq v(t), \forall t \geq 0.$

This is called the **comparison principle**; see e.g., Walter (1970). Then, since $\sigma(t) \to 0$, it can be shown, using a generalization of the invariance arguments of Lemma (5.3.71), that $e(t) \to 0$ as $t \to \infty$; see Levin (1960). Now (78) implies that $v(t) \to 0$. Hence there exists a function $\bar{\eta}$ of class L such that

79 $\|\mathbf{x}(t_0 + t)\| \leq \bar{\eta}(t), \forall t \geq 0.$

Finally, (72) and (79) together show that the origin is uniformly attractive. ■

80 **Example** Consider the system

$$\dot{x}_1 = -x_1 + ux_2, \dot{x}_2 = x_1^2 - x_2^3, y = x_2.$$

The linearization of this system around $\mathbf{x} = \mathbf{0}$, $u = 0$ is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} u, \; y = [0 \quad 1] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Note that the linearized system is not stabilizable, but is detectable. Hence the feasibility of the observer-controller strategy cannot be established by linearization arguments.

Suppose we could apply the control law

$$u = -x_1 + x_2.$$

The resulting closed-loop system is described by

$$\dot{x}_1 = -x_1 - x_1 x_2 + x_2^2, \dot{x}_2 = x_1^2 - x_2^3.$$

To test the stability of this system, choose the Lyapunov function candidate

$$V = x_1^2 + x_2^2.$$

Then

$$\dot{V} = -2(x_1^2 - x_1 x_2^2 + x_2^4) = -2[(x_1 - 0.5x_2^2)^2 + 0.75x_2^4].$$

Thus $\mathbf{x} = \mathbf{0}$ is an asymptotically stable equilibrium (but *not* exponentially stable).

However, the afore-mentioned control law cannot be implemented because x_1 cannot be measured directly. To get around this difficulty, let us set up the system

$$\dot{z}_1 = -z_1 + uy, \dot{z}_2 = z_1^2 - z_2 + y - y^3.$$

To test whether this system is a weak detector, choose the function

$$W(\mathbf{x}, \mathbf{z}) = (z_1 - x_1)^2 + (z_2 - x_2)^2.$$

Then

$$\begin{aligned}\dot{W}(\mathbf{x}, \mathbf{z}) &= 2[(z_1 - x_1)(\dot{z}_1 - \dot{x}_1) + (z_2 - x_2)(\dot{z}_2 - \dot{x}_2)] \\ &= -2[(z_1 - x_1)^2 + (z_2 - x_2)(z_1^2 - x_1^2 + z_2 - x_2)] \\ &= -[z_1 - x_1 \quad z_2 - x_2]\begin{bmatrix} 2 & z_1 + x_1 \\ z_1 + x_1 & 2 \end{bmatrix}\begin{bmatrix} z_1 - x_1 \\ z_2 - x_2 \end{bmatrix}.\end{aligned}$$

The coefficient matrix

$$\mathbf{M} = \begin{bmatrix} 2 & z_1 + x_1 \\ z_1 + x_1 & 2 \end{bmatrix}$$

is positive definite if $\|\mathbf{x}\|$, $\|\mathbf{z}\|$ are sufficiently small. Therefore it follows that (48) is satisfied, and that the system above is a weak detector.

Now, by Theorem (49), it follows that the implementable control law

$$u = -z_1 + z_2$$

stabilizes the system.

Before leaving the example, two comments are in order. First, it is worth noting that the coefficient matrix $\mathbf{M}(\mathbf{x},\mathbf{z})$ is positive definite only when *both* $\|\mathbf{x}\|$ and $\|\mathbf{z}\|$ are small — it is not enough that the quantity $\|\mathbf{x}-\mathbf{z}\|$ be small. Thus the system above is only a weak detector, and not a true detector. Second, since x_2 is available directly, it is somewhat extravagant to set up a detector for it. Instead one might think of setting up a "reduced-order" detector for x_1 alone. The theory to cover this situation is not available at present.

5.8.4 Stability of Hierarchical Systems

In this section, we study the stability of systems of the form

$$\dot{\mathbf{x}}_1(t) = \mathbf{f}_1[t, x_1(t)],$$

$$\dot{\mathbf{x}}_2(t) = \mathbf{f}_2[t, \mathbf{x}_1(t), \mathbf{x}_2(t)],$$

$$\vdots$$

81

$$\dot{\mathbf{x}}_l(t) = \mathbf{f}_l[t, \mathbf{x}_1(t), \cdots, \mathbf{x}_l(t)].$$

Such a system is said to be in **hierarchical** or **triangular form**, since the differential equation governing $\mathbf{x}_i(t)$ depends only on $\mathbf{x}_1(t), \cdots, \mathbf{x}_i(t)$, but not on $\mathbf{x}_j(t)$ for $j > i$. Given an arbitrary differential equation of the form (13), there exist systematic procedures for renumbering and regrouping the variables $x_1, \cdots, x_n$ in such a way that the system equations assume the hierarchical form (81); see Vidyasagar (1980c). The objective is to deduce the stability properties of the system (81) by studying only the simplified systems

82 $$\dot{\mathbf{x}}_i = \mathbf{f}_i[t, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{x}_i(t)],$$

for $i = 1, \cdots, l$. Comparing (82) with the i-th equation in (81), one sees that the l equations in (81) have been *decoupled* by setting $\mathbf{x}_j = 0$ for $j = 1, \cdots, i-1$ in the i-th equation. For this reason, (82) is referred to as the i-th **isolated subsystem.**

Now the main result of this subsection is stated. To streamline the presentation, two notational conventions are employed, namely:

$$\textbf{83}\quad \bar{\mathbf{x}}_i = \begin{bmatrix} \mathbf{x}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_i \end{bmatrix}, \mathbf{x} = \bar{\mathbf{x}}_l = \begin{bmatrix} \mathbf{x}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_l \end{bmatrix}.$$

In other words, $\bar{\mathbf{x}}_i$ is the state vector of the first i equations in (81), whereas $\mathbf{x} = \bar{\mathbf{x}}_l$ is the state vector of the overall system.

84 Theorem *Consider the system (81). Suppose each function* $\mathbf{f}_i$ *is* C^1, *and that the following conditions are satisfied for each* $i \in \{1, \cdots, l\}$:

$$\textbf{85}\quad \mathbf{f}_i(t, \mathbf{0}, \cdots, \mathbf{0}) = \mathbf{0}, \forall t \geq 0,$$

and in addition, there exist constants $\lambda < \infty$ *and* $r > 0$ *such that*

$$\textbf{86}\quad \sup_{t \geq 0} \sup_{\bar{\mathbf{x}}_i \in B_r} \| D_2 \mathbf{f}_i(t, \bar{\mathbf{x}}_i) \| \leq \lambda.$$

Under these conditions, $\mathbf{x} = \mathbf{0}$ *is a uniformly asymptotically stable equilibrium of the system (81) if and only if* $\mathbf{x}_i = \mathbf{0}$ *is a uniformly asymptotically stable equilibrium of the system (82) for each* $i \in \{1, \cdots, l\}$.

Remarks Note that, if the system (81) is autonomous and each function $\mathbf{f}_i$ is autonomous, then (86) is automatically satisfied. Hence the theorem requires only very mild conditions in order to be applicable.

Proof It is helpful to distinguish between the solution trajectories of (81) and of (82) through appropriate notation. Let $\mathbf{x}_i^{(I)}(t)$ denote the solutions of (82), where the superscript (I) is supposed to suggest "isolated." $\mathbf{x}_i(t)$, without the superscript, denotes the corresponding component of the solution of (81). Also, while the concept of uniform asymptotic stability is independent of the particular norm used, it is convenient to take $\| \cdot \|$ to be the norm $\| \cdot \|_\infty$ defined in Example (2.1.9). With this choice of norm, it follows that

$$\textbf{87}\quad \| \mathbf{x} \| = \max_{1 \leq i \leq l} \{ \| \mathbf{x}_1 \|, \cdots, \| \mathbf{x}_l \| \}.$$

"Only if" Suppose $\mathbf{x} = \mathbf{0}$ is a uniformly asymptotically stable equilibrium of (81). Then, by Theorem (5.1.61), there exist functions η of class K and σ of class L, and a constant $\rho > 0$ such that

$$\textbf{88}\quad \| \mathbf{x}(t) \| \leq \eta[\| \mathbf{x}(t_0) \|] \sigma(t - t_0), \forall t \geq t_0, \forall \mathbf{x}(t_0) \in B_\rho.$$

In particular, suppose

89 $\mathbf{x}(t_0) = [\mathbf{0}' \cdots \mathbf{0}' \ \mathbf{x}'_{i0} \ \mathbf{0}' \cdots \mathbf{0}']', \mathbf{x}_{i0} \in B_\rho,$

where $\mathbf{x}'_{i0}$ occurs in the i-th block. Then, in view of (85), it follows that the first $i-1$ blocks of the solution $\mathbf{x}(\cdot)$ equal zero, while the i-th block equals $\mathbf{x}_i^{(I)}(\cdot)$. Now (88) implies that

90 $\| \mathbf{x}_i^{(I)}(t) \| \leq \eta(\| \mathbf{x}_{i0} \|) \sigma(t - t_0), \forall t \geq t_0.$

This shows that $\mathbf{x}_i = \mathbf{0}$ is a uniformly asymptotically stable equilibrium of (82). The argument can be repeated for each i.

"If" The proof is given for the case $l = 2$; the general case follows by induction.

By assumption, there exist functions η_1 of class K, σ_1 of class L, and a constant $r > 0$, such that

91 $\| \mathbf{x}_1^{(I)}(t) \| \leq \eta_1[\| \mathbf{x}_{10} \|] \sigma_1(t - t_0), \forall t \geq t_0, \forall \mathbf{x}_{10} \in B_r,$

where $\mathbf{x}_1^{(I)}(\cdot)$ denotes the solution of (82) with $i = 1$, and with the initial condition $\mathbf{x}_1(t_0) = \mathbf{x}_{10}$. Similarly, the hypothesis on the second isolated subsystem, combined with Theorem (5.7.24), allow one to conclude the existence of a C^1 function V, class K functions α, β, γ, and constants $r > 0$ and $L < \infty$, such that

92 $\alpha(\| \mathbf{x}_2 \|) \leq V(t, \mathbf{x}_2) \leq \beta(\| \mathbf{x}_2 \|), \forall t \geq 0, \forall \mathbf{x}_2 \in B_r,$

93 $\dot{V}^{(I)}(t, \mathbf{x}_2) := D_1 V(t, \mathbf{x}_2) + D_2 V(t, \mathbf{x}_2) \mathbf{f}_2(t, \mathbf{0}, \mathbf{x}_2) \leq -\gamma(\| \mathbf{x}_2 \|), \forall t \geq 0, \forall \mathbf{x}_2 \in B_r,$

94 $\| D_2 V(t, \mathbf{x}_2) \| \leq L, \forall t \geq 0, \forall \mathbf{x}_2 \in B_r.$

Now consider the system

95 $\dot{\mathbf{x}}_1(t) = \mathbf{f}_1[t, \mathbf{x}_1(t)], \dot{\mathbf{x}}_2(t) = f_2[t, \mathbf{x}_1(t), \mathbf{x}_2(t)].$

It is to be shown that $\mathbf{x} = \mathbf{0}$ is a uniformly asymptotically stable equilibrium of this system. The proof is divided into two parts, namely establishing (i) uniform stability, and (ii) uniform attractivity. To establish uniform stability, suppose $\varepsilon > 0$ is given; it is necessary to determine $\delta > 0$ such that

96 $\| \mathbf{x}_{10} \| < \delta, \| \mathbf{x}_{20} \| < \delta \Rightarrow \| \mathbf{x}_1(t + t_0) \| < \varepsilon, \| \mathbf{x}_2(t + t_0) \| < \varepsilon, \forall t \geq t_0,$

where $\mathbf{x}_i(\cdot)$ denotes the solution of (81) corresponding to the initial condition $\mathbf{x}_i(t_0) = \mathbf{x}_{i0}$. It is easy to see that $\mathbf{x}_1(\cdot) = \mathbf{x}_i^{(I)}(\cdot)$, so the real challenge is to analyze the behavior of $\mathbf{x}_2(\cdot)$. For this purpose, it can be assumed without loss of generality that the constants r appearing in (86) and (91) – (94) are all the same. Given $\varepsilon > 0$, first choose $\varepsilon_1 > 0$ such that $\beta(\varepsilon_1) < \alpha(\varepsilon)$, and then choose $\delta_1 > 0$ such that

97 $\delta_1 < \min\{\varepsilon_1, \gamma(\varepsilon_1)/L\lambda\varepsilon\}$.

Finally, choose $\delta > 0$ such that

98 $\delta < \min\{r, \delta_1, \eta_1^{-1}[\varepsilon/\sigma_1(0)]\}$.

Suppose $\|\mathbf{x}_{10}\| < \delta$, $\|\mathbf{x}_{20}\| < \delta$. Since $\mathbf{x}_1(\cdot) = \mathbf{x}_1^{(I)}(\cdot)$, (91) and (98) together imply that $\|\mathbf{x}_1(t)\| < \varepsilon\ \forall t \geq t_0$. To get an estimate for $\|\mathbf{x}_2(\cdot)\|$, it is claimed that

99 $\dfrac{d}{dt}V[t, \mathbf{x}_2(t)] \leq 0$, whenever $\varepsilon_1 \leq \|\mathbf{x}_2(t)\| \leq r$.

To see this, observe that

100
$$\begin{aligned}\frac{d}{dt}V[t, \mathbf{x}_2(t)] &= D_1V[t, \mathbf{x}_2(t)] + D_2V[t, \mathbf{x}_2(t)]\,\mathbf{f}_2[t, \mathbf{x}_1(t), \mathbf{x}_2(t)] \\ &= \dot{V}^{(I)}[t, \mathbf{x}_2(t)] + D_2V[t, \mathbf{x}_2(t)]\,\{\mathbf{f}_2[t, \mathbf{x}_1(t), \mathbf{x}_2(t)] - \mathbf{f}_2[t, \mathbf{0}, \mathbf{x}_2(t)]\} \\ &\leq -\gamma[\|\mathbf{x}_2(t)\|] + L\lambda\|\mathbf{x}_1(t)\| \\ &\leq -\gamma[\|\mathbf{x}_2(t)\|] + L\lambda\varepsilon \leq 0 \text{ if } \|\mathbf{x}_2(t)\| \geq \varepsilon_1.\end{aligned}$$

Hence, by arguments which parallel those in the proof of Theorem (49), it follows that

101 $V[t, \mathbf{x}_2(t)] \leq \beta(\varepsilon_1)$, $\forall t \geq t_0$,

whence it follows that $\|\mathbf{x}_2(t)\| < \varepsilon\ \forall t \geq t_0$. Thus $\mathbf{x} = \mathbf{0}$ is a uniformly stable equilibrium.

To show that $\mathbf{x} = \mathbf{0}$ is uniformly attractive, select an $\varepsilon > 0$, and select $\delta > 0$ in accordance with (98). Suppose $\|\mathbf{x}(t_0)\| < \delta$. Modify (100) to

102 $\dfrac{d}{dt}V[t, \mathbf{x}_2(t)] \leq -\gamma(\beta^{-1}\{V[t, \mathbf{x}_2(t)]\}) + L\lambda\eta_1(\delta)\sigma_1(t - t_0)$.

This inequality is very similar to (73). Mimicking those arguments shows that $V[t + t_0, \mathbf{x}_2(t + t_0)] \to 0$ as $t \to \infty$, uniformly in t_0. The details are left as an exercise. This shows that B_δ is a region of uniform attractivity, and completes the proof. ■

Using appropriate converse theorems, it is possible to establish theorems regarding other forms of stability.

103 Theorem *Consider the system (81). Suppose each function* $\mathbf{f}_i$ *is* C^1, *and satisfies (85) and (86). Under these conditions,* $\mathbf{x} = \mathbf{0}$ *is an exponentially stable equilibrium of the system (81) if and only if* $\mathbf{x}_i = \mathbf{0}$ *is an exponentially stable equilibrium of (82) for each* $i \in \{1, \cdots, l\}$.

Proof "Only if" This part of the proof closely follows the corresponding part of the proof of Theorem (84), and is left as an exercise.

"If" Since exponential stability implies uniform asymptotic stability, Theorem (84) implies that $\mathbf{x} = \mathbf{0}$ is a uniformly asymptotically stable equilibrium. It only remains to show that solution trajectories converge to $\mathbf{x} = \mathbf{0}$ exponentially fast. Now $\mathbf{x}_1(\cdot) = \mathbf{x}_1^{(I)}(\cdot)$. Hence, by the hypothesis of exponential stability, there exist constants a, b, $r > 0$ such that [cf. (5.1.37)]

104 $\|\mathbf{x}_1(t)\| \leq a \|\mathbf{x}_{10}\| \exp[-b(t-t_0)], \forall t \geq t_0, \forall \mathbf{x}_{10} \in B_r.$

Next, since $\mathbf{x}_2 = \mathbf{0}$ is assumed to be an exponentially stable equilibrium of the system (82) with $i = 2$, it follows from Corollary (5.7.77) that there exist a C^1 function V and constants α, $\beta, \gamma, r > 0$ and $L < \infty$ such that

105 $\alpha \|\mathbf{x}\|^2 \leq V(t, \mathbf{x}_2) \leq \beta \|\mathbf{x}\|^2, \forall t \geq 0, \forall \mathbf{x}_2 \in B_r,$

106 $\dot{V}^{(I)}(t, \mathbf{x}_2) \leq -\gamma \|\mathbf{x}\|^2, \forall t \geq 0, \forall \mathbf{x}_2 \in B_r,$

107 $\|D_2 V(t, \mathbf{x}_2)\| \leq L \|\mathbf{x}_2\|, \forall t \geq 0, \forall \mathbf{x}_2 \in B_r.$

Without loss of generality, it is assumed that the same constant r appears in (104) as well as in (105) – (107). Now (102) becomes

108 $$\frac{d}{dt} V[t, \mathbf{x}_2(t)] \leq -(\gamma/\beta) V[t, \mathbf{x}_2(t)] + L\lambda a \|\mathbf{x}_{10}\| \|\mathbf{x}_2(t)\| \exp[-b(t-t_0)],$$

$$\leq -(\gamma/\beta) V[t, \mathbf{x}_2(t)]$$

$$+ (L\lambda a/\sqrt{\alpha}) \|\mathbf{x}_{10}\| \{V[t, \mathbf{x}_2(t)]\}^{1/2} \exp[-b(t-t_0)], \forall t \geq t_0.$$

Define $W(t) = V[t + t_0, \mathbf{x}_2(t + t_0)]^{1/2}$. Then it readily follows from (108) that

109 $2\dot{W}(t) \leq -(\gamma/\beta) W(t) + (L\lambda a/\sqrt{\alpha}) \|\mathbf{x}_{10}\| \exp[-b(t-t_0)], \forall t \geq t_0.$

Note that

110 $W(0) = [V(t_0, \mathbf{x}_{20})]^{1/2} \leq \sqrt{\beta} \|\mathbf{x}_{20}\|.$

From (109) and (110), it follows that there exist constants η, μ such that

111 $W(t) \leq \eta [\|\mathbf{x}_{10}\| + \|\mathbf{x}_{20}\|] \exp(-\mu t), \forall t \geq 0.$

In turn, (111) implies that

112 $$\|\mathbf{x}_2(t)\| \le (1/\sqrt{\alpha})\, W(t-t_0) \le (\eta/\sqrt{\alpha})\,[\|\mathbf{x}_{10}\| + \|\mathbf{x}_{20}\|]\exp[-\mu(t-t_0)],\ \forall t \ge t_0.$$

Now (104) and (112) together establish that the equilibrium $\mathbf{x}=\mathbf{0}$ is exponentially stable. ■

113 Theorem *Consider the system (81). Suppose each function* $\mathbf{f}_i$ *is* C^1, $\mathbf{f}_i(t, \mathbf{0})=\mathbf{0}$, $\forall t \ge 0$, *and*

114 $$\sup_{t \ge 0} \sup_{\bar{\mathbf{x}}} \|D_2\mathbf{f}_i(t, \bar{\mathbf{x}})\| < \infty,$$

for $i \in \{1, \cdots, l\}$. *Under these conditions,* $\mathbf{x}=\mathbf{0}$ *is a globally exponentially stable equilibrium of the system (81) if and only if* $\mathbf{x}_i=\mathbf{0}$ *is a globally exponentially stable equilibrium of the system (82) for each* $i \in \{1, \cdots, l\}$.

Proof "Only if" This part of the proof is left as an exercise.

"If" Let $r=\infty$ in (104) to (107) and proceed as in the proof of Theorem (103). ■

Problem 5.30 Consider the system

$$\dot{\mathbf{x}}(t)=\mathbf{f}[\mathbf{x}(t), \mathbf{u}(t)],\ \forall t \ge 0,$$

where $\mathbf{x}(t) \in \mathbf{R}^n$, $\mathbf{u}(t) \in \mathbf{R}^m$. Suppose $\mathbf{f}$ is C^2, and that $\mathbf{f}(t, \mathbf{0}, \mathbf{0})=\mathbf{0}$, $\forall t \ge 0$. Define

$$\mathbf{A}=\left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{(\mathbf{x},\mathbf{u})=(\mathbf{0},\mathbf{0})}, \mathbf{B}=\left[\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right]_{(\mathbf{x},\mathbf{u})=(\mathbf{0},\mathbf{0})}.$$

Recall that the pair $(\mathbf{A}, \mathbf{B})$ is said to be **stabilizable** if there exists a matrix $\mathbf{K} \in \mathbf{R}^{m \times n}$ such that $\mathbf{A}-\mathbf{BK}$ is a Hurwitz matrix.

(a) Prove the following statement: There exists a C^2 function $\mathbf{r}: \mathbf{R}^n \to \mathbf{R}^m$ such that $\mathbf{x}=\mathbf{0}$ is an exponentially stable equilibrium of the system

$$\dot{\mathbf{x}}(t)=\mathbf{f}\{\mathbf{x}(t), \mathbf{r}[t, \mathbf{x}(t)]\},$$

if and only if the pair $(\mathbf{A}, \mathbf{B})$ is stabilizable. [Hint: Use Theorem (1).]

(b) Construct an example to show that (a) is false if "exponentially stable" is replaced by "asymptotically stable." [Hint: See Example (80).]

Problem 5.31 Using Corollary (5.7.77) and Lemma (5.4.53), state and prove an extension of Theorem (1) to time-varying systems.

Problem 5.32 Using the results of Problem 5.31, extend the results of Problem 5.30 to time-varying systems.

Problem 5.33 Consider a modification of the linear system studied in Example (5.4.90), namely

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

where

$$\mathbf{A}(t) = \begin{bmatrix} -1 + a\cos^2 \lambda t & 1 - a \sin \lambda t \cos \lambda t \\ -1 - a \sin \lambda t \cos \lambda t & -1 + a \sin^2 \lambda t \end{bmatrix}.$$

(a) Find the state transition matrix. (Hint: see the form given in the Example and modify it suitably.)

(b) Using Lemma (5.4.79) and Theorem (5.4.89), find a range of values of the pair (a, λ) for which the above system is asymptotically stable.

(c) Use the result of Theorem (113) regarding slowly varying systems to construct suitable bounds on a and λ which assure that the system is asymptotically stable. Compare with the results obtained in part (b).

Problem 5.34 Prove the "only if" parts of Theorems (103) and (113).

Problem 5.35 Give a proof of Theorem (103) based on Problem 5.31 in the case where each function $\mathbf{f}_i$ is C^2.

5.9 DISCRETE-TIME SYSTEMS

Until now the emphasis in this chapter has been on continuous-time systems described by differential equations. In the present section the focus is on discrete-time systems, described by a recursive relationship of the form

$$\mathbf{1} \qquad \mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k),$$

where $\mathbf{x}_k \in \mathbf{R}^n$, and $\mathbf{f}_k : \mathbf{R}^n \to \mathbf{R}^n$ for all $k \geq 0$. Note that (1) always has exactly one solution corresponding to an initial condition of the form $\mathbf{x}(k_0) = \mathbf{x}_0$; this solution, evaluated at the k-th instant of time ($k \geq k_0$), is denoted by $\mathbf{s}(k, k_0, \mathbf{x}_0)$. If, in addition, it is assumed that $\mathbf{f}_k$ is a continuous function for all k, then $\mathbf{s}(k, k_0, \cdot)$ is also continuous for each pair (k, k_0) with $k \geq k_0$. Thus existence, uniqueness, and continuous dependence of solutions of recursion relations is really not an issue, in contrast with the case of differential equations.

The objective of the present section is to define various concepts of stability for the equilibria of the system (1), and to present various stability theorems. Since the details of the definitions, theorems, and proofs very closely parallel those of their continuous-time counterparts, very few details are given, and all proofs are left as exercises.

A point $\mathbf{x}_0 \in \mathbf{R}^n$ is called an **equilibrium** of the system (1) if

$$\mathbf{2} \qquad \mathbf{f}_k(\mathbf{x}_0) = \mathbf{x}_0, \ \forall k \geq 0,$$

i.e., if $\mathbf{x}_0$ is a *fixed point* of the map $\mathbf{f}_k$ for each $k \geq 0$. Clearly, if (2) holds, then

3 $\mathbf{s}(k, k_0, \mathbf{x}_0) = \mathbf{x}_0, \forall k \geq k_0 \geq 0.$

One can assume, without loss of generality, that the equilibrium of interest is the origin, i.e., that

4 $\mathbf{f}_k(\mathbf{0}) = \mathbf{0}, \forall k \geq 0.$

Suppose $V: Z_+ \times \mathbf{R}^n \to \mathbf{R}$, where Z_+ denotes the set of nonnegative integers. Then, along the solution trajectories of (1), define

5 $V^*_k = V[k, \mathbf{s}(k, k_0, \mathbf{x}_0)].$

The forward difference of the sequence $\{V^*_k\}$ is

6 $\Delta V^*_k = V^*_{k+1} - V^*_k = V[k+1, \mathbf{s}(k+1, k, \mathbf{x}_k)] - V(k, \mathbf{x}_k).$

With this in mind, we define the **forward difference function** $\Delta V: Z_+ \times \mathbf{R}^n \to \mathbf{R}$ as follows:

7 $\Delta V(k, \mathbf{x}) = V[k+1, \mathbf{s}(k+1, k, \mathbf{x})] - V(k, \mathbf{x}).$

Obviously, ΔV depends on both the function V and the system (1). Note that, along the trajectories of (1), we have

8 $V^*_k = V^*_j + \sum_{i=j}^{k-1} \Delta V(i, \mathbf{x}_i).$

8 **Definitions** *The equilibrium* $\mathbf{0}$ *of the system (1) is* **stable** *if, for each* $\varepsilon > 0$ *and each* $k_0 \geq 0$, *there exists a* $\delta = \delta(\varepsilon, k_0)$ *such that*

9 $\|\mathbf{x}_0\| < \delta(\varepsilon, k_0) \Rightarrow \|\mathbf{s}(k, k_0, \mathbf{x}_0)\| < \varepsilon, \forall k \geq k_0.$

The equilibrium $\mathbf{0}$ *is* **uniformly stable** *if, for each* $\varepsilon > 0$ *there exists a* $\delta = \delta(\varepsilon)$ *such that*

10 $k_0 \geq 0, \|\mathbf{x}_0\| < \delta(\varepsilon) \Rightarrow \|\mathbf{s}(k, k_0, \mathbf{x}_0)\| < \varepsilon, \forall k \geq k_0.$

11 **Definitions** *The equilibrium* $\mathbf{0}$ *is* **attractive** *if, for each* $k_0 \geq 0$, *there exists an* η_{k_0} *such that*

12 $\|\mathbf{x}_0\| < \eta_{k_0} \Rightarrow \mathbf{s}(k, k_0, \mathbf{x}_0) \to \mathbf{0}$ as $k \to \infty.$

It is **uniformly attractive** *if there exists an* $\eta > 0$ *such that*

13 $\|\mathbf{x}_0\| < \eta, k_0 \geq 0 \Rightarrow \mathbf{s}(k_0 + k, k_0, \mathbf{x}_0) \to \mathbf{0}$ as $k \to \infty$, uniformly in $k_0, \mathbf{x}_0.$

Equivalently, $\mathbf{0}$ *is uniformly attractive if there exists an* $\eta > 0$, *such that for each* $\varepsilon > 0$ *there exists an* $m = m(\varepsilon)$ *such that*

14 $\|\mathbf{x}_0\| < \eta, k_0 \geq 0 \Rightarrow \|\mathbf{s}(k_0 + k, k_0, \mathbf{x}_0)\| < \varepsilon, \forall k \geq m(\varepsilon).$

The equilibrium $\mathbf{0}$ *is* **asymptotically stable** *if it is stable and attractive; it is* **uniformly asymptotically stable** *if it is uniformly stable and uniformly attractive.*

15 Definition *The equilibrium* $\mathbf{0}$ *of (1) is* **exponentially stable** *if there exist constants* $\eta, a > 0$ *and* $\rho < 1$ *such that*

16 $\|\mathbf{x}_0\| < \eta, k_0 \geq 0 \Rightarrow \|\mathbf{s}(k_0 + k, k_0, \mathbf{x}_0)\| \leq a\|\mathbf{x}_0\|\rho^k, \forall k \geq 0.$

17 Definition *The equilibrium* $\mathbf{0}$ *of (1) is* **globally uniformly asymptotically stable** *if (i) it is uniformly stable, and (ii) for each* $\eta, \varepsilon > 0$, *there exists an* $m = m(\eta, \varepsilon)$ *such that*

18 $\|\mathbf{x}_0\| < \eta, k_0 \geq 0 \Rightarrow \|\mathbf{s}(k_0 + k, k_0, \mathbf{x}_0)\| < \varepsilon, \forall k \geq m.$

It is **globally exponentially stable** *if there exist constants* $a > 0$ *and* $\rho < 1$ *such that*

19 $\|\mathbf{s}(k_0 + k, k_0, \mathbf{x}_0)\| \leq a\|\mathbf{x}_0\|\rho^k, \forall k, k_0 \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$

So much for the definitions of various forms of stability. The stability theorems for discrete-time systems are also reminiscent of their continuous-time counterparts. To state these one needs the concepts of positive definiteness, etc.

20 Definitions *A function* $V: Z_+ \times \mathbf{R}^n \to \mathbf{R}$ *is a* **locally positive definite function (lpdf)** *if (i)* $V(k, \mathbf{0}) = 0\ \forall k \geq 0$, *and (ii) there exists a constant* $r > 0$ *and a function* α *of class K such that*

21 $\alpha(\|\mathbf{x}\|) \leq V(k, \mathbf{x}), \forall k \geq 0, \forall \mathbf{x} \in B_r.$

V is **decrescent** *if there is a function* β *of class K and a constant* $r > 0$ *such that*

22 $V(k, \mathbf{x}) \leq \beta(\|\mathbf{x}\|), \forall k \geq 0, \forall \mathbf{x} \in B_r.$

V is a **positive definite function (pdf)** *if (i)* $V(k, \mathbf{0}) = 0\ \forall k \geq 0$, *and (ii) there is a function* α *of class K such that*

23 $\alpha(\|\mathbf{x}\|) \leq V(k, \mathbf{x}), \forall k \geq 0, \forall \mathbf{x} \in \mathbf{R}^n.$

V is **radially unbounded** *if* $V(k, \mathbf{x}) \to \infty$ *as* $\|\mathbf{x}\| \to \infty$, *uniformly in k. V is* **radially unbounded** *if* $V(k, \mathbf{x}) \to \infty$ *as* $\|\mathbf{x}\| \to \infty$, *uniformly in k.*

24 Theorem (Stability) *The equilibrium* $\mathbf{0}$ *of (1) is stable if there exist a function* $V: Z_+ \times \mathbf{R}^n \to \mathbf{R}$ *and a constant* $r > 0$ *such that (i) V is an lpdf, and (ii)*

25 $\Delta V(k, \mathbf{x}) \leq 0, \forall k \geq 0, \forall \mathbf{x} \in B_r.$

If, in addition, V is decrescent, then $\mathbf{0}$ *is uniformly stable.*

26 Theorem (Asymptotic Stability) *The equilibrium* $\mathbf{0}$ *of (1) is uniformly asymptotically stable if there is a decrescent lpdf* $V: Z_+ \times \mathbf{R}^n \to \mathbf{R}$ *such that* $-\Delta V$ *is an lpdf.*

27 Theorem (Global Asymptotic Stability) *The equilibrium* $\mathbf{0}$ *of (1) is globally uniformly asymptotically stable if there is a function* $V: Z_+ \times \mathbf{R}^n \to \mathbf{R}$ *such that (i) V is a pdf, decrescent, and radially unbounded, and (ii)* $-\Delta V$ *is a pdf.*

28 Theorem (Exponential Stability) *Suppose there exist a function* $V: Z_+ \times \mathbf{R}^n \to \mathbf{R}$, *and constants* $a, b, c, r > 0$ *and* $p > 1$ *such that*

29 $a \|\mathbf{x}\|^p \le V(k, \mathbf{x}) \le b \|\mathbf{x}\|^p, \Delta V(k, \mathbf{x}) \le -c \|\mathbf{x}\|^p, \forall k \ge 0, \forall \mathbf{x} \in B_r.$

Then $\mathbf{0}$ *is an exponentially stable equilibrium. If it is possible to replace* B_r *by* $\mathbf{R}^n$ *in (29), then* $\mathbf{0}$ *is globally exponentially stable.*

Now consider a linear shift-invariant system described by

30 $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k, k \ge 0.$

If we try the obvious Lyapunov function candidate

31 $V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x},$

then

32 $\Delta V(\mathbf{x}) = \mathbf{x}'\mathbf{A}'\mathbf{P}\mathbf{A}\mathbf{x} - \mathbf{x}'\mathbf{P}\mathbf{x} = \mathbf{x}'(\mathbf{A}'\mathbf{P}\mathbf{A} - \mathbf{P})\mathbf{x}.$

Hence the discrete-time Lyapunov matrix equation is

33 $\mathbf{A}'\mathbf{P}\mathbf{A} - \mathbf{P} = -\mathbf{Q}.$

34 Theorem *The equilibrium* $\mathbf{0}$ *of (30) is (globally) asymptotically stable if and only if all eigenvalues of* $\mathbf{A}$ *have magnitude less than one.*

A matrix $\mathbf{A}$ whose eigenvalues all lie inside the open unit disk is called a **contractive** matrix.

35 Theorem *With respect to Equation (33), the following three statements are equivalent:*

1. *All eigenvalues of* $\mathbf{A}$ *have magnitude less than one.*
2. *There exists a positive definite matrix* $\mathbf{Q}$ *such that (33) has a unique solution for* $\mathbf{P}$, *and this* $\mathbf{P}$ *is positive definite.*
3. *For each positive definite matrix* $\mathbf{Q}$, *(33) has a unique solution for* $\mathbf{P}$, *and this solution is positive definite.*

The proof of Theorem (35) is facilitated by the following lemma.

36 Lemma *Suppose* $\mathbf{A}$ *is contractive and* $\mathbf{Q}$ *is a given matrix. Then (33) has a unique solution, given by*

$$\mathbf{37} \quad \mathbf{P} = \sum_{i=0}^{\infty} (\mathbf{A}')^i \mathbf{Q} \mathbf{A}^i.$$

Now consider the notion of linearizing a nonlinear system of the form (1) around the equilibrium $\mathbf{0}$ [assuming of course that (4) is true]. Suppose $\mathbf{f}_k$ is C^1 for each $k \geq 0$, and define

$$\mathbf{38} \quad \mathbf{A}_k = \left[\frac{\partial \mathbf{f}_k}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{0}}, \quad \mathbf{f}_{1k}(\mathbf{x}) = \mathbf{f}_k(\mathbf{x}) - \mathbf{A}_k \mathbf{x}.$$

Suppose it is true that

$$\mathbf{39} \quad \lim_{\|\mathbf{x}\| \to 0} \sup_{k \geq 0} \frac{\|\mathbf{f}_{1k}(\mathbf{x})\|}{\|\mathbf{x}\|} = 0.$$

Then the system

$$\mathbf{40} \quad \mathbf{z}_{k+1} = \mathbf{A}_k \mathbf{z}_k$$

is called the **linearization** of (1) around $\mathbf{0}$.

41 Theorem *If* $\mathbf{0}$ *is a uniformly asymptotically stable equilibrium of the linear system (40), then it is also a uniformly asymptotically stable equilibrium of (1).*

42 Theorem *Consider the autonomous system*

$$\mathbf{43} \quad \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k),$$

where $\mathbf{f}$ *is* C^1 *and* $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. *Define*

$$\mathbf{44} \quad \mathbf{A} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{0}}.$$

If $\mathbf{A}$ *is contractive, then* $\mathbf{0}$ *is an asymptotically stable equilibrium of (43). If* $\mathbf{A}$ *has at least one eigenvalue with magnitude greater than one, then* $\mathbf{0}$ *is an unstable equilibrium of (43).*

Problem 5.36 State and prove the discrete-time analogs of the converse theorems of Section 5.7.

Notes and References

The concept of Lyapunov stability was first formulated by the Russian mathematician A. M. Lyapunov in a famous memoir published in Russian in 1892. This memoir was translated into French in 1907 and is available in the Western world from Princeton University Press; see Lyapunov (1892). Lyapunov stated and proved the basic stability theorem, and also proved the linearization method (Section 5.5), the results concerning the Lyapunov matrix equation for analyzing linear systems (Section 5.4), as well as several other results which are no longer popular. The symbol V for the function which today we call the Lyapunov function was introduced by Lyapunov himself. It is a good guess that he was thinking of potential energy when he used this symbol, since in problems of particles moving in potential fields, the potential energy is a logical Lyapunov function candidate (see Problem 5.13, Section 5.3). Whatever his motivation, the symbol V is now so firmly entrenched that no one would seriously think of using anything else, though some intrepid souls have been known to use v.

Originally Lyapunov did not distinguish between stability and uniform stability, etc. The awareness that the distinction was important came gradually through the efforts of several researchers, the most notable of whom were Malkin and Massera. The asymptotic stability theorems that are based on the idea of invariant sets are due to Barbashin and Krasovskii (1952) in the Soviet Union and LaSalle (1960) in the United States.

The contents of Section 5.6 on the Lur'e problem again represent the collective efforts of many individuals. The Popov criterion was originally proved by Popov (1961) using a method quite different from what is given here. This method is now called the hyperstability method; see Popov (1973). The proof given here, which demonstrates the existence of a Lyapunov function if certain frequency-domain conditions are satisfied, requires the so-called Kalman-Yacubovitch lemma, discovered independently by Kalman (1962) and Yacubovitch (1964). The circle criterion, based on the Kalman-Yacubovitch lemma, was proved by Narendra and Goldwyn (1964). See Narendra and Taylor (1973) for a book-length treatment of this topic.

Malkin and Massera were again instrumental in proving that some of the Lyapunov theorems were "reversible" and proved most of the contents of Section 5.7. The applications of the converse theorems found in Section 5.8 are due to various authors. The analysis of slowly-varying systems is found in the thesis of Barman (1973), with earlier efforts for linear systems due to Desoer (1969) and (1970). The proof that the observer-controller strategy works for nonlinear systems is due to Vidyasagar (1980b). The result on the stability of hierarchical systems is found in Vidyasagar (1980c), and is a generalization of an earlier result due to Michel et al. (1978).

6. INPUT-OUTPUT STABILITY

In this chapter, we present the basic results of input-output stability theory. This theory is much more recent in origin than Lyapunov theory, having been pioneered by Sandberg and Zames in the 1960's [see Sandberg (1964, 1965a, 1965b), Zames (1966a, 1966b)]. While this chapter contains most of the principal results of the subject, the treatment is by no means encyclopaedic. The reader is referred to Desoer and Vidyasagar (1975) for a thorough discussion of the subject.

Before proceeding to the study of input-output *stability*, it is necessary to reconcile the input-output *approach* to system analysis and the state variable methods employed in the preceding chapters. The methods of the preceding four chapters are predicated on the system under study being governed by a set of *differential equations* which describe the time evolution of the system *state variables*. In contrast, the systems encountered in this chapter are assumed to be described by an *input-output mapping* that assigns, to each *input*, a corresponding *output*. In view of this seeming dichotomy, it is important to realize that an input-output representation and a state variable representation are two different ways of looking at the *same system*—the two types of representation are used because they each give a different kind of insight into how the system works. It is now known that, not only does there exist a close relationship between the input-output representation and the state representation of a given system, but that there also exists a very close relationship between the kinds of *stability results* that one can obtain using the two approaches.

At this stage one may well ask: Why not simply use one of the two approaches—why use both? The answer is that while the two approaches are related, they are not equivalent. Since, in analyzing a system, we would like to have as many answers as we can, it is desirable to have both the approaches at our disposal, each yielding its own set of insights and information.

Finally, it should be mentioned that many of the arguments and proofs in input-output theory are conceptually clearer than their Lyapunov stability counterparts. Compare the proofs of the circle criterion and the Popov criterion in the two approaches, for example. Also, analyzing distributed systems (e.g., systems containing delays) in an input-output setting is no more complicated than analyzing lumped systems. In contrast, in the case of Lyapunov stability, analyzing time-delay systems, for example, is substantially more complicated than analyzing ordinary differential equations [see e.g., Hale (1977)]. On the other hand, at a first glance at least, understanding input-output theory would appear to require a greater background in mathematics than understanding Lyapunov theory, since input-output theory makes reference to advanced concepts such as Lebesgue spaces and so on. In many ways, this impression is misleading. As shown subsequently, there are only a few places where the full power of Lebesgue theory is needed, and almost everywhere one can

get by with the more familiar notion of Riemann integration. Part of the objective of this chapter is to make input-output theory as accessible to the student as Lyapunov theory.

6.1 L_p-SPACES AND THEIR EXTENSIONS

In this section, a brief introduction is given to the Lebesgue spaces L_p and their extensions, and to the concepts of truncations and causality. Much of input-output stability theory, including the problem statements and the stability theorems, is couched in these terms, so that a certain degree of familiarity with these concepts is necessary to appreciate input-output theory. On the other hand, most of the input-output results given here do not require any deep results from Lebesgue theory other than the completeness of the L_p spaces, so that the pedestrian treatment given below is sufficient for the present purposes.

6.1.1 L_p-Spaces

The reader is undoubtedly familiar with the idea of Riemann integration, as taught in undergraduate calculus. While this is a fine idea, there are many "reasonable" functions that are not Riemann integrable. For example, suppose $f: \mathbf{R}_+ \to \mathbf{R}_+$ is defined by $f(x) = 1$ if x is rational, and $f(x) = 0$ if x is irrational. Then every interval contains an x such that $f(x) = 0$ and a y such that $f(y) = 1$. Hence f is not Riemann integrable. This shows the need for a more general concept of integration.

There is another argument as well: Consider the set $C[0, 1]$ of continuous real-valued functions defined on the interval $[0, 1]$, together with the norm

$$\mathbf{1} \qquad \|f\|_\infty = \max_{t \in [0,1]} |f(t)|.$$

(The reason for the subscript ∞ becomes apparent later.) Then convergence in the norm $\|\cdot\|_\infty$ is just uniform convergence over $[0, 1]$. Now it is well-known that if a sequence of continuous functions converges uniformly to a function $f(\cdot)$, then $f(\cdot)$ is also continuous; see e.g., Royden (1963). Hence the normed space $\{C[0, 1], \|\cdot\|_\infty\}$ is a Banach space in the sense of Definition (2.1.24): In contrast, suppose we define the norm

$$\mathbf{2} \qquad \|f\|_2 = \left[\int_0^1 |f(t)|^2 \, dt \right]^{1/2}.$$

Then convergence in the norm $\|\cdot\|_2$ is just mean-squared convergence. Now there exist discontinuous functions $f(\cdot)$ which are the limit, in the norm $\|\cdot\|_2$, of a sequence of continuous functions. For example, let

$$\mathbf{3} \qquad f(t) = \begin{cases} 0, & t \in [0, 0.5), \\ 1, & t \in [0.5, 1]. \end{cases}$$

Then the Fourier series expansion of $f(\cdot)$ converges to $f(\cdot)$ in the norm $\|\cdot\|_2$. Thus the

normed space $\{C[0, 1], \|\cdot\|_2\}$ is *not* complete. This leads us to ask how one may enlarge the set $C[0, 1]$ such that the resulting set is in fact complete under the norm $\|\cdot\|_2$. This too is another motivation for studying L_p-spaces. Unfortunately even a modest discussion of the Lebesgue theory of measure and integration would either be very lengthy, or if it is too short, potentially misleading. We mention that a function $f: \mathbf{R}_+ \to \mathbf{R}$ is **measurable** if and only if $f(t)$ is the limit of a sequence of staircase (or piecewise-constant) functions at all t except those belonging to a set of measure zero. A reader who has no prior acquaintance with the Lebesgue theory of measure and integration can simply think of all functions encountered below as piecewise-continuous functions, and of all integrals as Riemann integrals. This would lead to no conceptual difficulties and no loss of insight, except that occasionally some results from Lebesgue theory would have to be accepted on faith.

4 Definition *For each real $p \in [1, \infty)$, the set $L_p[0, \infty) = L_p$ consists of all measurable functions $f(\cdot): \mathbf{R}_+ \to \mathbf{R}$ such that*

$$\int_0^\infty |f(t)|^p \, dt < \infty. \tag{5}$$

The set $L_\infty[0, \infty) = L_\infty$ consists of all measurable functions $f(\cdot): \mathbf{R}_+ \to \mathbf{R}$ that are essentially bounded[1] *on $[0, \infty)$.*

Thus, for $p \in [1, \infty)$, $L_p[0, \infty) = L_p$ denotes the set of all measurable functions whose p-th powers are absolutely integrable over $[0, \infty)$, while $L_\infty[0, \infty) = L_\infty$ denotes the set of essentially bounded measurable functions.

6 Example The function $f(t) = e^{-\alpha t}$, $\alpha > 0$, belongs to L_p for all $p \in [1, \infty]$. The function $g(t) = 1/(t+1)$ belongs to L_p for each $p > 1$ but not to L_1. The function

$$f_p(t) = \left[\frac{1}{t^{1/2}(1 + \log t)} \right]^{2/p}$$

where $p \in [1, \infty)$, belongs to the set L_p, but does *not* belong to L_q for any $q \neq p$. ■

7 Definition *For $p \in [1, \infty)$, the function $\|\cdot\|_p: L_p \to \mathbf{R}_+$ is defined by*

$$\|f(\cdot)\|_p = \left[\int_0^\infty |f(t)|^p \, dt \right]^{1/p}. \tag{8}$$

The function $\|\cdot\|_\infty: L_\infty \to \mathbf{R}_+$ is defined by[2]

---

[1] "Essentially bounded" means "bounded except on a set of measure zero." The reader need not worry unduly about the distinction between "bounded" and "essentially bounded."

[2] The notation "ess. sup." stands for "essential supremum," i.e., supremum except on a set of measure zero.

9 $$\|f(\cdot)\|_\infty = \operatorname*{ess.\,sup}_{t\in[0,\infty)} |f(t)|.$$

Definition (7) introduces the functions $\|\cdot\|_p$ for $p\in[1,\infty]$, which map the set L_p into the half-line $\mathbf{R}_+$. Note that, by virtue of Definition (4), the right sides of (8) and (9) are well-defined and finite for each $f(\cdot)\in L_p$.

10 Fact (Minkowski's Inequality) *Let $p\in[1,\infty]$, and suppose $f(\cdot), g(\cdot)\in L_p$. Then $f(\cdot)+g(\cdot)\in L_p$, and*

11 $$\|f(\cdot)+g(\cdot)\|_p \le \|f(\cdot)\|_p + \|g(\cdot)\|_p.$$

For a detailed proof, see Royden (1963).

12 Fact *For each $p\in[1,\infty]$, the pair $(L_p, \|\cdot\|_p)$ is a normed linear space in the sense of Definition (2.1.8).*

This fact follows readily from Minkowski's inequality.

13 Fact *For each $p\in[1,\infty]$, the normed linear space $(L_p, \|\cdot\|_p)$ is complete and is hence a Banach space. For $p=2$, the norm $\|\cdot\|_2$ corresponds to the inner product*

14 $$<f(\cdot), g(\cdot)> = \int_0^\infty f(t)g(t)\,dt.$$

Thus L_2 is a Hilbert space under the inner product of (14).

Fact (13) brings out one of the main reasons for dealing with L_p-spaces in studying input-output stability. We could instead work in the spaces $C_p[0,\infty)$ of all *continuous* functions $f(\cdot)$ in L_p, and of course the pair $(C_p, \|\cdot\|_p)$ is also a normed space. However, it is not complete except for the special case of $p=\infty$. Now some results in input-output theory, such as those on the well-posedness of feedback systems, require that the problem be set up in a Banach space so that one can apply results such as the contraction mapping principle. In such a case, it is preferrable to work with the L_p-spaces. Note that, for each $p\in[1,\infty)$, the space L_p is precisely the completion of the space C_p; in other words, each function in L_p can be approximated arbitrarily closely by a continuous function, provided $p<\infty$. Each function in L_p, $p<\infty$, can be arbitrarily closely approximated by a piecewise-constant, or staircase function. However, neither statement is true of L_∞.

15 Fact (Hölder's Inequality) *Let $p, q\in[1,\infty]$, and suppose*

16 $$\frac{1}{p}+\frac{1}{q}=1.$$

(Note that we take $p=\infty$ if $q=1$ and vice versa.) Suppose $f(\cdot)\in L_p$ and $g(\cdot)\in L_q$. Then the function $h:\mathbf{R}_+\to\mathbf{R}$ defined by

17 $h(t) := f(t)\, g(t)$

belongs to L_1. *Moreover,*

$$\textbf{18} \quad \int_0^\infty |f(t)\, g(t)|\, dt \le \left[\int_0^\infty |f(t)|^p\, dt\right]^{1/p} \left[\int_0^\infty |g(t)|^q\, dt\right]^{1/q}.$$

The inequality (18) can be expressed more concisely as

$$\textbf{19} \quad \|f(\cdot)\, g(\cdot)\|_1 \le \|f(\cdot)\|_p \|g(\cdot)\|_q.$$

6.1.2 Extended L_p-Spaces

20 Definition *Suppose* $f : \mathbf{R}_+ \to \mathbf{R}$. *Then for each* $T \in \mathbf{R}_+$, *the function* $f_T : \mathbf{R}_+ \to \mathbf{R}$ *is defined by*

$$\textbf{21} \quad f_T(t) = \begin{cases} f(t), & 0 \le t \le T \\ 0, & T < t \end{cases}$$

and is called the **truncation** *of* f *to the interval* $[0, T]$.

22 Definition *The set* L_{pe} *consists of all measurable functions* $f : \mathbf{R}_+ \to \mathbf{R}$ *with the property that* $f_T \in L_p$ *for all finite* T, *and is called the* **extension of** L_p *or the* **extended** L_p**-space.**

Thus the set L_{pe} consists of all measurable functions f which have the property that every truncation of f belongs to L_p, although f itself may or may not belong to L_p. It is easy to see that L_p is a subset of L_{pe}.

23 Example Let $f(\cdot)$ be defined by

$$f(t) = t.$$

Then, for each *finite* value of T, the function f_T belongs to all the spaces L_p, for each $p \in [1, \infty]$. Hence the original function f belongs to L_{pe} for each $p \in [1, \infty]$. However, f itself does not belong to any of the *unextended* spaces L_p. ■

The relationship between extended and unextended spaces is brought out in the next lemma.

24 Lemma *For each* $p \in [1, \infty]$, *the set* L_{pe} *is a linear vector space over the real numbers. For each fixed* p *and* $f \in L_{pe}$, *(i)* $\|f_T\|_p$ *is a nondecreasing function of* T, *and (ii)* $f \in L_p$ *if and only if there exists a finite constant* m *such that*

$$\textbf{25} \quad \|f_T\|_p \le m, \ \forall T \ge 0.$$

In this case,

26 $$\|f\|_p = \lim_{T\to\infty} \|f_T\|_p.$$

The proof is almost obvious and is left as an exercise.

Thus, in summary, the extended space L_{pe} is a linear space that contains the unextended space L_p as a subset. Notice however that L_p is a normed space while L_{pe} is not.

In order to deal with multi-input multi-output systems, we introduce the set L_p^n which consists of all n-tuples $f = [f_1 \cdots f_n]'$ where $f_i \in L_p$ for each i. The norm on L_p^n is defined by

27 $$\|f\|_p := \left[\sum_{i=1}^{n} \|f_i\|_p^2\right]^{1/2}.$$

In other words, the norm of a vector-valued function is the square root of the sum of the norms of the components of the vector. This definition is to some extent arbitrary but has the advantage that L_2^n is a Hilbert space with this definition. The symbol L_{pe}^n is defined analogously. Note that hereafter the symbols $\|f_T\|_p$ and $\|f\|_{Tp}$ are used interchangeably. Also, the same symbol $\|\cdot\|_p$ is used to denote the norm on L_p^n for all integers n.

6.1.3 Causality

This section is concluded with an introduction to the concept of causality. If we think of a mapping A as representing a system and of Af as the output of the system corresponding to the input f, then a causal system is one where the value of the output at time t depends only on the values of the input up to time t. This is made precise next.[3]

28 **Definition** *A mapping $A : L_{pe}^n \to L_{pe}^m$ is said to be* **causal** *if*

29 $$(Af)_T = (Af_T)_T, \ \forall T \geq 0, \ \forall f \in L_{pe}^n.$$

An alternative formulation of causality is provided by the following.

30 **Lemma** *Let $A : L_{pe}^n \to L_{pe}^m$. Then A is causal in the sense of Definition (28) if and only if*

31 $$f, g \in L_{pe}^n, f_T = g_T \Rightarrow (Af)_T = (Ag)_T, \ \forall T \geq 0.$$

Proof "If" Suppose A satisfies (31). Then it must be shown that A satisfies (29). Accordingly, let $f \in L_{pe}^n$ and $T \geq 0$ be arbitrary. Then clearly $f_T = (f_T)_T$, so that by (31) we have

---

[3] Throughout this chapter, for the most part bold-faced symbols are not used, since the various quantities can be either scalars or vectors.

32 $(Af)_T = (Af_T)_T$.

Since the above holds for all $f \in L_{pe}^n$ and all $T \geq 0$, A is causal.

"Only if" Suppose A is causal in the sense of Definition (28). Suppose f, $g \in L_{pe}^n$ and that $f_T = g_T$ for some $T \geq 0$; it must be shown that $(Af)_T = (Ag)_T$. For this purpose, note that, by (29), we have

33 $(Af)_T = (Af_T)_T$, $(Ag)_T = (Ag_T)_T$.

Moreover, since $f_T = g_T$, (33) implies that

34 $(Af)_T = (Ag)_T$,

which was the thing to be proved. ■

Definition (28) and Lemma (30) provide two alternative but entirely equivalent interpretations of causality. Definition (28) states that a mapping A is causal if any truncation of Af to an interval $[0, T]$ depends only on the corresponding truncation f_T. To put it another way, the values of $(Af)(t)$ over $[0, T]$ depend only on the values of $f(t)$ over $[0, T]$. Lemma (30) states that A is causal if, whenever two inputs are equal over an interval $[0, T]$, the corresponding outputs are also equal over the same interval.

Problem 6.1 Determine whether each of the following functions below belongs to any of the unextended spaces L_p and to any of the extended spaces L_{pe}:

(a) $f(t) = t \exp(-t)$

(b) $f(t) = \tan t$.

(c) $f(t) = \exp(t^2)$.

(d) $$f(t) = \begin{cases} 0, & \text{if } t = 0, \\ 1/t^2, & \text{if } 0 < t < 1, \\ 1/t, & \text{if } 1 \leq t. \end{cases}$$

Problem 6.2 Prove Lemma (24).

Problem 6.3 Suppose $H: L_{pe} \to L_{pe}$ has the form

$$(Hf)(t) = \int_0^\infty h(t, \tau) f(\tau)\, d\tau.$$

Show that the operator H is causal if and only if

$h(t, \tau) = 0$ whenever $t < \tau$.

6.2 DEFINITIONS OF INPUT-OUTPUT STABILITY

In this section, the basic definitions of input-output stability are introduced and illustrated.

It is traditional to define input-output stability as a property of relations rather than operators. Suppose X is a set. A **binary relation** on X is a subset of X^2. Suppose R is a binary relation on X. Then we say that $x \in X$ is related to $y \in X$ if the ordered pair $(x, y) \in R$. Suppose $A : X \to X$ is a mapping. Then A defines a binary relation R_A on X, namely

1 $$R_A = \{(x, Ax): x \in X\}.$$

The converse need not be true: Not all binary relations are of the form (1). Suppose R is a binary relation on X. Then, for a particular $x \in X$, there might not exist any $y \in X$ such that $(x, y) \in R$; or else there might exist many (possibly infinitely many) $y \in X$ such that $(x, y) \in R$. Thus relations make it easy to think about and to work with "multi-valued" mappings which might assign more than one value to an argument, and with "partial" mappings whose domain need not equal all of X.

2 **Definition** *Suppose R is a binary relation on L_{pe}. Then R is said to be L_p-**stable** if*

3 $$(x, y) \in R, x \in L_p \Rightarrow y \in L_p.$$

*R is L_p-**stable with finite gain** (**wfg**) if it is L_p-stable, and in addition there exist finite constants γ_p and b_p such that*

4 $$(x, y) \in R, x \in L_p \Rightarrow \|y\|_p \leq \gamma_p \|x\|_p + b_p.$$

*R is L_p-**stable with finite gain and zero bias** (**wb**) if it is L_p-stable, and in addition there exists a finite constant γ_p such that*

5 $$(x, y) \in R, x \in L_p \Rightarrow \|y\|_p \leq \gamma_p \|x\|_p.$$

Since the abbreviation "wfgazb" is unpronounceable, let us agree to use "wb" as a short form for "with finite gain and zero bias." One might interpret "wb" as "without bias."

Remarks

1. Clearly L_p-stability wb implies implies L_p-stability wfg, which in turn implies L_p-stability. Example (7) below shows that these three concepts are indeed independent.
2. Suppose that, for a particular $x \in L_p$, no $y \in L_{pe}$ exists such that $(x, y) \in R$. Then the conditions (3), (4) and (5) are deemed to be satisfied vacuously.

6 Definition *Suppose* $A : L_{pe}^n \to L_{pe}^n$. *Then the map A is said to be* L_p**-stable (wfg, wb)** *if and only if the corresponding binary relation* R_A *on* L_{pe}^n *defined by (1) is* L_p*-stable (wfg, wb).*

7 Example Consider the functions f, g, $h : \mathbf{R} \to \mathbf{R}$ defined by

$$f(r) = r^2,\ g(r) = r+1,\ h(r) = \log(1+r^2).$$

Define an operator $F : L_{\infty e} \to L_{\infty e}$ by

$$(Fx)(t) = f[x(t)],\ \forall t \geq 0,$$

and define G, $H : L_{\infty e} \to L_{\infty e}$ analogously. Then F is L_∞-stable but not L_∞-stable with finite gain. Note that $x \in L_\infty$ implies that $Fx \in L_\infty$, but no constants γ_∞, b_∞ can be found such that (4) holds. This is because the function $f(r)$ cannot be bounded by any straight line of the form $\gamma r + b$; see Figure 6.1. G is L_∞-stable wfg, but not L_∞-stable wb, since $G(0) \neq 0$. H is L_∞-stable wb.

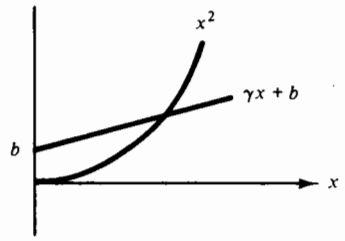


Fig. 6.1

8 Example Consider the mapping A defined by

$$(Af)(t) = \int_0^t \exp[-\alpha(t-\tau)] f(\tau)\, d\tau,\ \forall t \geq 0,$$

where $\alpha > 0$ is a given constant, and suppose it is desired to study the L_∞-stability of this system. First it is established that A maps the extended space $L_{\infty e}$ into itself, so that A is in the class of mappings covered by Definition (6). Accordingly, suppose $f \in L_{\infty e}$. Then $\|f_T\|_\infty$ is finite for all finite T. Hence, for each finite T, there exists a finite constant m_T such that

$$|f(t)| \leq m_T \text{ a.e.},\ \forall t \in [0, T].$$

Here the term "a.e." stands for "almost everywhere," i.e., except on a set of measure zero; hereafter this phrase is omitted, it being implicitly understood. To show that $g := Af \in L_{\infty e}$, suppose T is some finite number. Then

$$|g(t)| \leq \int_0^t \exp[-\alpha(t-\tau)]\,|f(\tau)|\,d\tau$$

$$\leq \int_0^t m_T \exp[-\alpha(t-\tau)]\,d\tau$$

$$\leq \frac{m_T}{\alpha}.$$

Hence $g(\cdot)$ is bounded over $[0, T]$. Since the argument can be repeated for each T, it follows that $g \in L_{\infty e}$, i.e., $A : L_{\infty e} \to L_{\infty e}$.

Next, it is shown that A is L_∞-stable wb in the sense of Definition (6). Suppose $f \in L_\infty$. Then there exists a finite constant m such that

$$|f(t)| \leq m, \forall t \geq 0.$$

Using exactly the same reasoning as before, it can be shown that

$$|Af(t)| \leq \frac{m}{\alpha}, \forall t \geq 0.$$

Hence (5) is satisfied with $\gamma_\infty = 1/\alpha$, which shows that A is L_∞-stable wb. ■

9 Example Consider the system whose input-output relationship is

$$(Af)(t) = \int_0^t \exp(t-\tau) f(\tau)\,d\tau.$$

This mapping A also maps $L_{\infty e}$ into itself. To see this, let $f \in L_{\infty e}$. Then, for each finite T, there exists a finite constant m_T such that

$$|f(t)| \leq m_T, \forall t \in [0, T].$$

Thus, whenever $t \in [0, T]$, we have

$$\textbf{10} \qquad |(Af)(t)| \leq m_T \int_0^t |\exp(t-\tau)|\,d\tau \leq m_T(e^T - 1).$$

Now, for each *finite* T, the right side of (10) is a finite number. Hence $Af \in L_{\infty e}$.

On the other hand, A is *not* L_∞-stable, as can be seen by setting $f(t) \equiv 1, \forall t$. Then $f(\cdot) \in L_\infty$, but

$$(Af)(t) = \int_0^t \exp(t-\tau)\, d\tau = e^t - 1,$$

which clearly does not belong to L_∞ (even though it does belong to $L_{\infty e}$). Hence there is at least one input in L_∞ for which the corresponding output does not belong to L_∞. This shows that A is not L_∞-stable. ■

Remarks Example (9) illustrates the advantages of setting up the input-output stability problem in extended L_p-spaces. If we deal exclusively with L_p-stable systems, then such a system can be represented as a mapping from L_p (the unextended space) into itself, rather than as a mapping from L_{pe} into itself. However, if we are interested in studying unstable systems (for example, the feedback stability of systems containing unstable subsystems), then we must have a way of mathematically describing such a system. This is accomplished by treating such a system as a mapping from L_{pe} into itself.

The next result shows that, for causal operators, stability in its various forms has some useful consequences.

11 Lemma *Suppose $A: L_{pe}^n \to L_{pe}^m$ is causal and L_p-stable wfg, and choose constants γ_p and b_p such that*

12 $\|Ax\|_p \le \gamma_p \|x\|_p + b_p, \ \forall x \in L_p^n.$

Then

13 $\|Ax\|_{Tp} \le \gamma_p \|x\|_{Tp} + b_p, \ \forall T \ge 0, \ \forall x \in L_{pe}^n.$

Proof Given arbitrary $x \in L_{pe}^n$, $T \in \mathbf{R}_+$, note that $x_T \in L_p^n$. Hence $Ax_T \in L_p^m$, and

14 $\|Ax_T\|_p \le \gamma_p \|x_T\|_p + b_p = \gamma_p \|x\|_{Tp} + b_p.$

However, since A is causal, $(Ax_T)_T = (Ax)_T$. Hence

15 $\|Ax\|_{Tp} = \|(Ax)_T\|_p = \|Ax_T\|_{Tp} \le \|Ax_T\|_p.$

Now (13) follows from (14) and (15). ■

The analog of Lemma (11) for L_p-stable operators wb follows simply by setting $b_p = 0$. Note that one can also define a notion of causality for relations, and prove a result similar to Lemma (11); the details are left to the reader.

16 Lemma *Suppose $A: L_{pe}^n \to L_{pe}^m$ is linear. Then A is L_p-stable wfg if and only if it is L_p-stable wb.*

Proof "If" Obvious.

"Only if" Suppose A is L_p-stable wfg, and select constants γ_p, b_p such that (12) holds. Let k be any number. Then, by (12) and linearity,

17 $\|A(kx)\|_p \leq \gamma_p \|kx\|_p + b_p$,

18 $\|Ax\|_p \leq \gamma_p \|x\|_p + (b_p/k)$.

Letting $k \to \infty$ shows that A is L_p-stable wb. ■

Next, the notion of L_p-gain is made precise. Up to now, if a relation R is L_p-stable, then γ_p can be *any* constant such that (4) holds. Now this arbitrariness is removed.

19 Definition *Suppose R is a binary relation on L_{pe}^n. If R is L_p-stable wfg, then the L_p-***gain of** *R is defined as*

20 $\gamma_p(R) := \inf\{\gamma_p : \exists\, b_p \geq 0$ such that (4) holds$\}$.

*If R is L_p-stable, then the L_p-***gain with zero bias of** *R is defined as*

21 $\gamma_p(R) := \inf\{\gamma_p :$ (5) holds$\}$.

Note that both quantities are denoted by $\gamma_p(R)$, the context making clear which is meant. If $A : L_{pe}^n \to L_{pe}^n$, then the quantity $\gamma_p(A)$ is defined to be $\gamma_p(R_A)$.

Thus far we have discussed what might be called "open-loop" stability. But one of the main applications of input-output theory is the stability analysis of feedback systems, of the form shown in Figure 6.2. In this system, it is assumed that u_1, e_1, y_2 are vector-valued functions with n_1 components each, and that u_2, e_2, y_1 are vector-valued functions with n_2 components each. To be specific, suppose $p \in [1, \infty]$ is given, and that

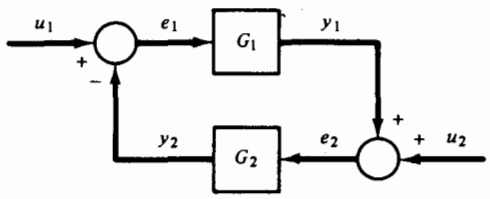


Fig. 6.2

22 $u_1, e_1, y_2 \in L_{pe}^{n_1}$,

23 $u_2, e_2, y_1 \in L_{pe}^{n_2}$.

Then the overall feedback system is described by the equations

24 $e_1 = u_1 - y_2,\ e_2 = u_2 + y_1,\ y_1 = G_1 e_1,\ y_2 = G_2 e_2.$

These equations can be written in a more compact form. Let $n = n_1 + n_2$, and define $u, e, y \in L_{pe}^n$ by

25 $$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},\ e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix},\ y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Define maps $G, F : L_{pe}^n \to L_{pe}^n$ by

26 $$G = \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix},\ F = \begin{bmatrix} 0 & I_{n_1} \\ -I_{n_2} & 0 \end{bmatrix},$$

and note that F is a constant matrix, known as the *interconnection matrix*. Then the equations of the feedback system can be expressed as

27 $e = u - Fy,\ y = Ge.$

On the basis of (27), we can define two binary relations R_{ue} and R_{uy} on L_{pe}^n. To define R_{ue}, eliminate y from (27) to obtain

28 $e = u - FGe$, or $(I + FG)e = u.$

To define R_{uy}, eliminate e from (27) to obtain

29 $y = G(u - Fy).$

Note that we *cannot* in general expand $G(u - Fy)$ as $Gu - GFy$, since G need not be linear. Now the binary relations R_{ue} and R_{uy} are defined as follows:

30 $R_{ue} = \{(u, e) \in L_{pe}^{2n} : e + FGe = u\},$

31 $R_{uy} = \{(u, y) \in L_{pe}^{2n} : y = G(u - Fy)\}.$

32 Definition *The feedback system (27) is* L_p**-stable (u-e)** *if the relation* R_{ue} *is* L_p*-stable; it is* L_p**-stable (u-y)** *if the relation* R_{uy} *is* L_p*-stable. Finally, it is* L_p**-stable (u-e-y)** *if both* R_{ue} *and* R_{uy} *are* L_p*-stable.*

The next lemma shows that we do not need quite so many concepts of stability.

33 Lemma *The following three statements are equivalent:*

(i) R_{ue} *is* L_p*-stable.*

(ii) *R_{uy} is L_p-stable.*

(iii) *Both R_{ue} and R_{uy} are L_p-stable.*

Proof Clearly it is enough to show that (i) and (ii) are equivalent.

(i) $\Rightarrow$ (ii): Suppose (i) is true, i.e., that R_{ue} is L_p-stable. Let $(u, y) \in R_{uy}$ be arbitrary. Then, from (27), the ordered pair

$$\textbf{34} \quad (u, u - Fy) \in R_{ue}.$$

Now suppose $u \in L_p^n$ (not L_{pe}^n). It is desired to show that $y \in L_p^n$. For this purpose, note that $u \in L_p^n$ and the L_p-stability of R_{ue} together imply that

$$\textbf{35} \quad e = u - Fy \in L_p^n.$$

But since F is just a constant nonsingular matrix, (35) implies that

$$\textbf{36} \quad y = F^{-1}(u - e) \in L_p^n.$$

Since the same argument can be repeated for every $u \in L_p^n$, it follows that

$$\textbf{37} \quad (u, y) \in L_{pe}^{2n},\ u \in L_p^n \Rightarrow y \in L_p^n.$$

In other words, R_{uy} is L_p-stable.

(ii) $\Rightarrow$ (i): The proof is quite similar to the above. Suppose R_{uy} is L_p-stable, and suppose $(u, e) \in R_{ue}$. Then $(u, Ge) \in R_{uy}$. If $u \in L_p^n$, then the L_p-stability of R_{uy} implies that $Ge \in L_p^n$. Since F is just a constant matrix, this in turn implies that $FGe \in L_p^n$. Finally $e = u - FGe \in L_p^n$. Hence R_{ue} is L_p-stable. ■

Lemma (33) allows us to say simply "the feedback system is L_p-stable" without specifying whether we mean u-e, u-y, or u-e-y. It is left to the reader to modify Definition (32) to define L_p-stability wfg and wb, and to prove the analogs of Lemma (33).

In Lemma (33), the proof of the implication (ii) $\Rightarrow$ (i) depends only on the fact that F is a constant matrix, whereas the proof of the implication (i) $\Rightarrow$ (ii) depends on the fact that F is a constant *nonsingular* matrix. The nonsingularity of F is a special feature of the feedback configuration, and is not true in general. For instance, consider the system shown in Figure 6.3, known as a multi-controller configuration. In this case,

$$\textbf{38} \quad F = \begin{bmatrix} 0 & I & I \\ -I & 0 & 0 \\ -I & 0 & 0 \end{bmatrix},$$

which is singular. Hence, in an arbitrary interconnected system, u-y stability implies u-e stability, but the converse need not be true.

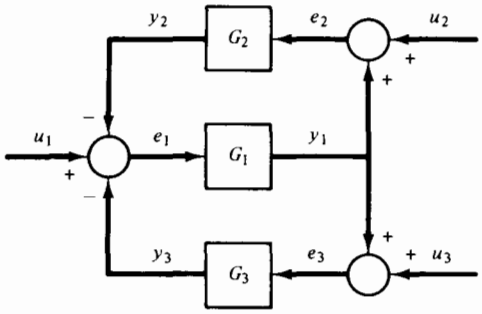


Fig. 6.3

Now that the stability definitions are out of the way, we can at last discuss the reason for introducing relations, and speaking of the stability of relations (rather than operators). The reason simply is to divorce the notion of *stability* from that of *well-posedness*. Stability is defined above. Roughly speaking, well-posedness corresponds to the existence and uniqueness of solutions for e and y for each choice of u, though one may wish to add requirements of causality, continuous dependence, etc. Thus a possible definition of well-posedness is that $(I+FG)^{-1}$ is a well-defined causal map from L^n_{pe} into itself. This would allow one to solve (28) for e and write

39 $e=(I+FG)^{-1}u.$

But well-posedness places no restrictions on stability. Conversely, u-e stability of (27) means that, for each $u\in L^n_p$, *if any* $e\in L^n_{pe}$ satisfy (28), then such e must in fact belong to L^n_p. If, for a particular $u\in L^n_p$, no $e\in L^n_{pe}$ satisfies (28), then this condition is deemed to be satisfied vacuously. In this way, stability and well-posedness become independent concepts. This is desirable since well-posedness can be ensured under quite mild conditions, whereas stability is more difficult to analyze. Roughly speaking, the feedback system is well-posed if either G_1 or G_2 contains some element of smoothing; see Vidyasagar (1980a) or Saeki and Araki (1982).

6.3 RELATIONSHIPS BETWEEN I/O AND LYAPUNOV STABILITY

In Chapter 5 we studied Lyapunov stability, which is defined for unforced ordinary differential equations; there is no input, and the system evolves under the influence of a nonzero initial state. In contrast, in the present chapter, the word "state" is not mentioned at all, and attention is focused on the influence of inputs upon outputs. It is therefore worthwhile to relate the two types of stability, and to underline the point that both approaches tackle different facets of the same underlying issue. This is the objective of the present section.

We begin with a discussion of open-loop stability, and then study feedback systems. Let us begin with a discussion of linear systems of the form

1 $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$,

2 $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)$,

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are all constant matrices. The transfer matrix of this system is

3 $\hat{\mathbf{H}}(s) = \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$.

The system (1-2) is said to be **stabilizable** if there exists a matrix K such that $\mathbf{A} - \mathbf{B}K$ is Hurwitz, and is **detectable** if there exists a matrix F such that $\mathbf{A} - \mathbf{F}\mathbf{C}$ is Hurwitz. As for the transfer matrix $\hat{\mathbf{H}}$, it will be seen in the next section that it represents an L_∞-stable (i.e., BIBO stable) system if and only if all poles of $\hat{\mathbf{H}}$ have negative real parts. Now the following result is well-known [Kailath (1980), Chen (1986)]:

4 **Theorem** *Suppose the system (1-2) is stabilizable and detectable. Under these conditions, the system is L_∞-stable if and only if the associated unforced system*

5 $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$

is (globally) asymptotically stable.

For an extension of Theorem (4) to linear time-varying systems, see Silverman and Anderson (1966) for the continuous-time case and Anderson (1982) for the discrete-time case.

The converse of Theorem (4) is even simpler.

6 **Theorem** *Suppose the system (5) is asymptotically stable. Then the system (1-2) is L_p-stable for each $p \in [1, \infty]$.*

Now let us consider nonlinear systems of the form

7 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t), \mathbf{u}(t)]$, $\mathbf{y}(t) = \mathbf{g}[t, \mathbf{x}(t), \mathbf{u}(t)]$, $\forall t \geq 0$.

Suppose $\mathbf{f}: \mathbf{R}_+ \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ and $\mathbf{g}: \mathbf{R}_+ \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^l$ are unbiased in the sense that

8 $\mathbf{f}(t, \mathbf{0}, \mathbf{0}) = \mathbf{0}$, $\mathbf{g}(t, \mathbf{0}, \mathbf{0}) = \mathbf{0}$, $\forall t \geq 0$.

This ensures that $\mathbf{0}$ is an equilibrium of the unforced system

9 $\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t), \mathbf{0}]$, $\forall t \geq 0$.

To make the theorem statements more compact, some notation and definitions are now introduced. The first one is a "local" version of Definition (6.2.2).

10 **Definition** *The system (7) is* **small signal L_p-stable wb** *if there exist constants $r_p > 0$ and $\gamma_p < \infty$ such that*

11 $\mathbf{x}(0)=\mathbf{0}$, $\|\mathbf{u}(t)\| \le r_p\ \forall t \ge 0$, $\mathbf{u} \in L_p^m \Rightarrow \mathbf{y} \in L_p^l$ and $\|\mathbf{y}\|_p \le \gamma_p \|\mathbf{u}\|_p$.

Note that, even if $\|\mathbf{u}(t)\| \le r_p\ \forall t \ge 0$, (which is equivalent to saying that $\|\mathbf{u}\|_\infty \le r_p$), the norm $\|\mathbf{u}\|_p$ can be arbitrarily large if $p < \infty$. Hence small signal L_p-stability does not necessarily mean that the inputs have to be small in the sense of the L_p-norm; rather, only their instantaneous values need to be small.

The next two concepts involve the state as well, but are restricted to autonomous systems. It is possible to extend these notions to nonautonomous systems, but at the price of a considerable increase in complexity. To define these concepts, let $\mathbf{s}(t, \tau, \mathbf{x}, \mathbf{u})$ denote the solution of (7) evaluated at time t, starting at time τ in the initial state $\mathbf{x}$, and with the input $\mathbf{u}$.

12 Definition *The system (7) is* **reachable** *if there exist a function* β *of class K and a constant* $r > 0$ *such that, for each* $\mathbf{x} \in B_r$, *there exist a time* t^* *and an input* $\mathbf{u}^*$ *such that* $\|\mathbf{u}^*\|_\infty \le \beta(\|\mathbf{x}\|)$ *and* $\mathbf{s}(t^*, 0, \mathbf{0}, \mathbf{u}^*) = \mathbf{x}$. *It is* **globally reachable** *if the preceding statement holds for all* $\mathbf{x} \in \mathbf{R}^n$.

The conditions of the definition mean simply that every state $\mathbf{x}$ in B_r can be reached within a finite time by applying an input whose L_∞-norm is bounded by $\beta(\|\mathbf{x}\|)$.

13 Definition *The system (7) is* **uniformly observable** *if there exists a function* α *of class K such that, with* $\mathbf{u}(t) \equiv \mathbf{0}$, *we have*

14 $\|\mathbf{g}[\cdot, \mathbf{s}(\cdot, 0, \mathbf{x}, \mathbf{0})]\|_2 \ge \alpha(\|\mathbf{x}\|)$.

The inequality (14) can be stated more concisely (but less precisely) as $\|\mathbf{y}\|_2 \ge \alpha(\|\mathbf{x}\|)$. If $\mathbf{y}(\cdot) \notin L_2^l$, then (14) is deemed to be satisfied since, loosely speaking, $\|\mathbf{y}\|_2 = \infty$. For a linear time-invariant system, uniform observability is equivalent to the standard notion of observability. This is because, with zero input, the output depends only on the initial state. If the system is unobservable, then there exists a nonzero initial state which (with zero input) will produce an identically zero output, so that (14) is violated. On the other hand, if the system is observable, then it is not difficult to show that (14) is satisfied.

Now the main theorems relating input-output stability and Lyapunov stability of the system (7) are presented. Theorem (15) is a nonlinear analog of Theorem (6).

15 Theorem *Suppose* $\mathbf{0}$ *is an exponentially stable equilibrium of (9), that* $\mathbf{f}$ *is* C^1, *and that* $\mathbf{f}$, $\mathbf{g}$ *are locally Lipschitz continuous at* $(\mathbf{0}, \mathbf{0})$, *i.e., suppose there exist finite constants* k_f, k_g, r *such that*

16 $\|\mathbf{f}(t, \mathbf{x}, \mathbf{u}) - \mathbf{f}(t, \mathbf{z}, \mathbf{v})\| \le k_f [\|\mathbf{x}-\mathbf{z}\| + \|\mathbf{u}-\mathbf{v}\|]$, $\forall t \ge 0$, $\forall (\mathbf{x}, \mathbf{u}), (\mathbf{z}, \mathbf{v}) \in B_r$,

17 $\|\mathbf{g}(t, \mathbf{x}, \mathbf{u}) - \mathbf{g}(t, \mathbf{z}, \mathbf{v})\| \le k_g [\|\mathbf{x}-\mathbf{z}\| + \|\mathbf{u}-\mathbf{v}\|]$, $\forall t \ge 0$, $\forall (\mathbf{x}, \mathbf{u}), (\mathbf{z}, \mathbf{v}) \in B_r$,

Then the system (7) is small signal L_p*-stable wb for each* $p \in [1, \infty]$. *If* $\mathbf{0}$ *is a globally exponentially stable equilibrium, and (16) and (17) hold with* B_r *replaced by* $\mathbf{R}^{(n+m)}$, *then the system (7) is* L_p*-stable wb for all* $p \in [1, \infty]$.

Proof The condition (16) implies that

18 $\|D_2\mathbf{f}(t, \mathbf{x}, \mathbf{u})\| \le k_f, \forall t \ge 0, \forall (\mathbf{x}, \mathbf{u}) \in B_r,$

where $D_2\mathbf{f}$ denotes the partial derivative of $\mathbf{f}$ with respect to its second argument. Thus all the hypotheses of Corollary (5.7.77) are satisfied, and there exist a C^1 function $V: \mathbf{R}_+ \times \mathbf{R}^n \to \mathbf{R}$ and constants $\alpha, \beta, \gamma, s > 0$ such that

19 $\alpha^2 \|\mathbf{x}\|^2 \le V(t, \mathbf{x}) \le \beta^2 \|\mathbf{x}\|^2, \dot{V}_u(t, \mathbf{x}) \le -\|\mathbf{x}\|^2, \forall t \ge 0, \forall \mathbf{x} \in B_s,$

20 $\|D_2 V(t, \mathbf{x})\| \le \gamma \|\mathbf{x}\|, \forall t \ge 0, \forall \mathbf{x} \in B_s,$

where

21 $\dot{V}_u(t, \mathbf{x}) = D_1 V(t, \mathbf{x}) + D_2 V(t, \mathbf{x})\, \mathbf{f}(t, \mathbf{x}, \mathbf{0})$

denotes the derivative of V along the trajectories of the unforced system (9).

Let $\delta = \min\{r, s\}$, and suppose $\mathbf{u}$ is a fixed input with the property that

22 $$|\mathbf{u}(t)| \le \min\left\{ \frac{\alpha\delta}{\beta\gamma k_f}, r \right\} =: \mu,$$

and suppose $\mathbf{x}(0) = \mathbf{0}$.

Evaluate the derivative of $V[t, \mathbf{x}(t)]$ along the trajectories of the forced system (7), and denote it by $\dot{V}_f$. Then

23
$$\begin{aligned}
\dot{V}_f(t, \mathbf{x}) &= D_1 V(t, \mathbf{x}) + D_2 V(t, \mathbf{x})\, \mathbf{f}(t, \mathbf{x}, \mathbf{u}) \\
&= D_1 V(t, \mathbf{x}) + D_2 V(t, \mathbf{x})\, \mathbf{f}(t, \mathbf{x}, \mathbf{0}) + D_2 V(t, \mathbf{x})\, [\mathbf{f}(t, \mathbf{x}, \mathbf{u}) - \mathbf{f}(t, \mathbf{x}, \mathbf{0})] \\
&= \dot{V}_u(t, \mathbf{x}) + D_2 V(t, \mathbf{x})\, [\mathbf{f}(t, \mathbf{x}, \mathbf{u}) - \mathbf{f}(t, \mathbf{x}, \mathbf{0})] \\
&\le -\|\mathbf{x}\|^2 + \gamma k_f \|\mathbf{x}\| \cdot \|\mathbf{u}\|, \text{ if } (\mathbf{x}, \mathbf{u}) \in B_\delta,
\end{aligned}$$

where in the last step we use (16), (19) and (20). Now, since $\mathbf{x}(0) = \mathbf{0}$, there exists a time $T > 0$ such that $\mathbf{x}(t) \in B_s$ for all $t \in [0, T)$. Moreover, the right side of (23) is negative whenever $\|\mathbf{x}\| > \gamma k_f \|\mathbf{u}\|$. Hence one can easily show that

24 $$V[t, \mathbf{x}(t)] \le \max_{\|\mathbf{x}\| \le \gamma k_f \|\mathbf{u}\|} V(t, \mathbf{x}) \le \beta^2 \gamma^2 k_f^2 \mu^2, \forall t \ge 0.$$

From (24) and (19) it follows that

$$\textbf{25} \quad \|\mathbf{x}\| \le \frac{\beta\gamma k_f \mu}{\alpha} \le \min\{r, s\}.$$

This last observation removes the circularity in the argument, and shows that $\|\mathbf{x}(t)\| \le \min\{r, s\}\ \forall t \ge 0$. Now, from (19) and (23), it follows that

$$\textbf{26} \quad \frac{d}{dt}\{V[t, \mathbf{x}(t)]\} \le -\frac{1}{\beta^2} V[t, \mathbf{x}(t)] + \frac{\gamma k_f}{\alpha}\{V[t, \mathbf{x}(t)]\}^{1/2} \|\mathbf{u}(t)\|.$$

Let $W(t) = V[t, \mathbf{x}(t)]^{1/2}$. Then $W(t)$ is differentiable except when $\mathbf{x}(t) = \mathbf{0}$, and is directionally differentiable even there. Hence the one-sided derivative

$$\textbf{27} \quad \dot{W}_+(t) = \lim_{h \to 0^+} \frac{W(t+h) - W(t)}{h}$$

exists for all $t \ge 0$. For notational convenience the subscript "+" is dropped hereafter. Now $\dot{V} = 2W\dot{W}$; hence it follows from (26) that

$$\textbf{28} \quad 2W\dot{W} \le -\frac{1}{\beta^2} W^2 + \frac{\gamma k_f}{\alpha} W \|\mathbf{u}\|,$$

or

$$\textbf{29} \quad \dot{W}(t) \le -\frac{1}{2\beta^2} W(t) + \frac{\gamma k_f}{2\alpha} \|\mathbf{u}(t)\|, \ \forall t \ge 0.$$

Let $h(t)$ denote the solution of

$$\textbf{30} \quad \dot{h}(t) + \frac{1}{2\beta^2} h(t) = \frac{\gamma k_f}{2\alpha} \|\mathbf{u}(t)\|, \ h(0) = W(0).$$

Then (29) implies that $W(t) \le h(t)\ \forall t \ge 0$. But note that $h(\cdot)$ is just the output of an L_p-stable first-order system with the transfer function

$$\textbf{31} \quad \hat{g}(s) = \frac{\gamma k_f / 2\alpha}{s + 1/2\beta^2}$$

driven by the input $\|\mathbf{u}(t)\|$. Now $\|\mathbf{u}(\cdot)\| \in L_p$ since $\mathbf{u} \in L_p^m$. By Theorem (6.4.30), it follows that $h(\cdot) \in L_p$, which in turn implies that $W(\cdot) \in L_p$. Since

$$\textbf{32} \quad \|\mathbf{x}(t)\| \le \{V[t, \mathbf{x}(t)]\}^{1/2}/\alpha = W(t)/\alpha,$$

it follows that $\mathbf{x}(\cdot) \in L_p^n$. Finally (17) and (8) imply that

$$\textbf{33} \quad \|\mathbf{y}(t)\| = \|\mathbf{g}[t, \mathbf{x}(t), \mathbf{u}(t)]\| \le k_g[\|\mathbf{x}(t)\| + \|\mathbf{u}(t)\|],$$

whence $\mathbf{y}(\cdot) \in L_p^l$.

To complete the proof of small signal L_p-stability, it only remains to demonstrate the existence of a constant γ_p such that (11) holds. For this purpose, note that the inverse Laplace transform of $\hat{g}$ of (31) is

$$\mathbf{34} \quad g(t) = \frac{\gamma k_f}{2\alpha} \exp(-t/2\beta^2) \in L_1.$$

Therefore,

$$\mathbf{35} \quad \| g \|_1 = \gamma k_f / 4\beta^2 \alpha.$$

Hence, by (6.4.31), it follows from (30) that

$$\mathbf{36} \quad \| h \|_p \leq [\gamma k_f / 4\beta^2 \alpha] \| \mathbf{u} \|_p.$$

Since $W(t) \leq h(t)\ \forall t$, (36) and (32) together imply that

$$\mathbf{37} \quad \| \mathbf{x} \|_p \leq \frac{\gamma k_f}{4\alpha^2 \beta^2} \| \mathbf{u} \|_p.$$

Finally, we can conclude from (37) and (33) that

$$\mathbf{38} \quad \| \mathbf{y} \|_p \leq k_g [(\gamma k_f / 4\alpha^2 \beta^2) + 1] \| \mathbf{u} \|_p.$$

The proof of the "global" result is entirely parallel. ■

The next theorem is a nonlinear analog of Theorem (4).

39 Theorem *Suppose the system (7) is autonomous, reachable, and uniformly observable. Under these conditions, if the system is small signal L_2-stable, then* $\mathbf{0}$ *is an attractive equilibrium of (9).*

Proof Since the system (9) is assumed to be small signal L_p-stable, there exist constants r_2 and $\gamma_2 > 0$ such that

$$\mathbf{40} \quad \| \mathbf{u} \|_\infty \leq r_2,\ \mathbf{u} \in L_2^m \Rightarrow \mathbf{y} \in L_2^l \text{ and } \| \mathbf{y} \|_2 \leq \gamma_2 \| \mathbf{u} \|_2.$$

Now, since (7) is reachable, there exist a constant $r > 0$ and a function β of class K satisfying the conditions of Definition (12). Choose $\delta > 0$ such that $\beta(\delta) < r_2$, and let $\mathbf{x}_0 \in B_\delta$ be arbitrary. Then, by definition, there is an input $\mathbf{u}(\cdot)$ with $\| \mathbf{u} \|_\infty \leq \beta(\| \mathbf{x}_0 \|) < r_2$ and a finite time t^* such that $\mathbf{s}(t^*, 0, \mathbf{0}, \mathbf{u}) = \mathbf{x}_0$. Since $\mathbf{s}(t^*, 0, \mathbf{0}, \mathbf{u})$ depends only on the values of $\mathbf{u}(t)$ for $t \in [0, t^*)$, we can assume that $\mathbf{u}(t) = \mathbf{0}$ for $t \geq t^*$, which, together with $\| \mathbf{u} \|_\infty < \beta(\| \mathbf{x}_0 \|)$, means that $\mathbf{u} \in L_2^m$. Now consider the solution trajectory $\mathbf{s}(t, 0, \mathbf{x}_0, \mathbf{0})$ of the unforced system (9). Since (7) is an autonomous system, we see that $\mathbf{s}(t, 0, \mathbf{x}_0, \mathbf{0}) = \mathbf{s}(t^* + t, 0, \mathbf{0}, \mathbf{u})$. By small signal L_2-stability, we know that the corresponding output $\mathbf{y}$ belongs to L_2^l since $\mathbf{u} \in L_2^m$, which implies that

41 $\int_t^\infty \| \mathbf{y}(\tau) \|^2 \, d\tau \to 0$ as $t \to \infty$.

Now, by the definition (13) of uniform observability,

42 $\int_0^\infty \| \mathbf{y}(\tau) \|^2 \, d\tau \geq \alpha(\| \mathbf{x}_0 \|)$,

provided $\mathbf{u}(t) \equiv \mathbf{0}$ for all $t \geq 0$. Again, by using the time-invariance of the system and noting that $\mathbf{u}(t) = \mathbf{0}$ for all $t \geq t^*$, it follows that

43 $\int_t^\infty \| \mathbf{y}(\tau) \|^2 \, d\tau \geq \alpha(\| \mathbf{x}(t) \|), \forall t \geq t^*$.

Now (41) and (43) show that $\alpha(\| \mathbf{x}(t) \|) \to 0$ as $t \to \infty$, which in turn shows that $\mathbf{x}(t) \to \mathbf{0}$ as $t \to \infty$. ■

44 Corollary *Suppose the system (7) is autonomous, globally reachable, and uniformly observable. Under these conditions, if the system is L_2-stable, then $\mathbf{0}$ is a globally attractive equilibrium of (9).*

The proof is analogous to that of Theorem (39) and is thus omitted.

Thus far, attention has been restricted to "open-loop" systems. Next, Theorem (15) and Corollary (44) are used to relate the external and internal stability of the feedback system shown in Figure 6.4, which is the same as Figure 5.15. This system is described by

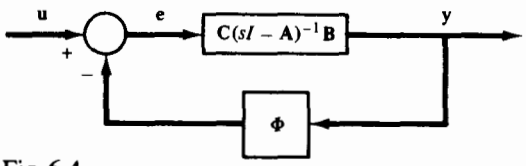


Fig. 6.4

45 $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{e}(t)$, $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)$, $\mathbf{e}(t) = \mathbf{u}(t) - \Phi[t, \mathbf{y}(t)]$,

where $\mathbf{x}(t) \in \mathbf{R}^n$, $\mathbf{u}(t) \in \mathbf{R}^m$, $\mathbf{y}(t) \in \mathbf{R}^l$, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are matrices of compatible dimensions; and $\Phi: \mathbf{R}_+ \times \mathbf{R}^l \to \mathbf{R}^m$ satisfies $\Phi(t, \mathbf{0}) = \mathbf{0}$ $\forall t \geq 0$.

46 Theorem *Consider the system (45), and suppose Φ is globally Lipschitz continuous; i.e., suppose there exists a finite constant μ such that*

47 $\| \Phi(t, \mathbf{y}_1) - \Phi(t, \mathbf{y}_2) \| \leq \mu \| \mathbf{y}_1 - \mathbf{y}_2 \|$, $\forall t \geq 0$, $\forall \mathbf{y}_1, \mathbf{y}_2 \in \mathbf{R}^l$.

Under these conditions, if the unforced system is globally exponentially stable, then the

forced system is L_p-stable wb for all $p \in [1, \infty]$. Suppose the pair $(\mathbf{A}, \mathbf{B})$ *is controllable, and the pair* $(\mathbf{C}, \mathbf{A})$ *is observable. Under these conditions, if the forced system is L_2-stable, then* $\mathbf{x} = \mathbf{0}$ *is a globally attractive equilibrium of the unforced system.*

Proof The system (45) can be written as

$$\mathbf{48} \quad \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) - \mathbf{B}\,\Phi[t, \mathbf{C}\mathbf{x}(t)] + \mathbf{B}\mathbf{u}(t).$$

First, suppose (47) is satisfied. Then the right side of (48) is globally Lipschitz continuous in $\mathbf{x}$ and $\mathbf{u}$. Hence, by Theorem (15), if $\mathbf{x} = \mathbf{0}$ is a globally exponentially stable equilibrium of the unforced system, then the forced system is L_p-stable wb for all $p \in [1, \infty]$.

Next, suppose the forced system (48) is L_2-stable. If it can be shown that the system is reachable and uniformly observable, then the global attractivity of $\mathbf{x} = \mathbf{0}$ would follow from Corollary (44). To show that (48) is reachable, let $\mathbf{x}_0 \in \mathbf{R}^n$ be arbitrary. Then, since $(\mathbf{A}, \mathbf{B})$ is a controllable pair, there exists a finite time t^* and a continuous function $\mathbf{e}(t)$, $t \in [0, t^*]$ such that the resulting solution of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$ starting at $\mathbf{x}(0) = \mathbf{0}$ satisfies $\mathbf{x}(t^*) = \mathbf{x}_0$. Now apply the control signal

$$\mathbf{49} \quad \mathbf{u}(t) = \mathbf{e}(t) - \Phi[t, \mathbf{C}\mathbf{x}(t)]$$

to the system (48). Then it is clear that once again we will have $\mathbf{x}(t^*) = \mathbf{x}_0$. Showing that $\| \mathbf{u} \|_\infty$ is bounded by a function of the form $\beta(\| \mathbf{x}_0 \|)$ is easy and is left as an exercise. To prove that the system (48) is uniformly observable, suppose $\mathbf{u}(t) \equiv \mathbf{0}$. Then, since $(\mathbf{C}, \mathbf{A})$ is an observable pair, there exist a time T and constants $a, b > 0$ such that

$$\mathbf{50} \quad \| \mathbf{x}(0) \|^2 \leq \int_0^T [a \| \mathbf{e}(t) \|^2 + b \| \mathbf{y}(t) \|^2]\, dt.$$

The proof of this statement is not difficult and is left as an exercise. Hence

$$\mathbf{51} \quad a \| \mathbf{e} \|_2^2 + b \| \mathbf{y} \|_2^2 \geq \| \mathbf{x}(0) \|^2.$$

Now note that if $\mathbf{u} \equiv \mathbf{0}$, then $\mathbf{e}(t) = -\Phi[t, \mathbf{y}(t)]$. Since

$$\mathbf{52} \quad \| \Phi[t, \mathbf{y}(t)] \leq \mu \| \mathbf{y}(t) \|^2, \ \forall t \geq 0,$$

it follows that

$$\mathbf{53} \quad \| \mathbf{e} \|_2^2 \leq \mu^2 \| \mathbf{y} \|_2^2.$$

Combining (50) and (53) shows that

$$\mathbf{54} \quad \| \mathbf{x}(0) \|^2 \leq (a\mu^2 + b) \| \mathbf{y} \|_2^2.$$

This shows that (48) is uniformly observable. Now Corollary (44) enables one to conclude that the equilibrium $\mathbf{x} = \mathbf{0}$ is globally attractive. ■

Problem 6.4 Prove Theorem (15) in the special case where **f** and **g** are autonomous, and **f** is C^2, as follows: Define

$$\mathbf{A} := \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{0}}.$$

Using Theorem (5.8.1), show that **A** is a Hurwitz matrix. Define the higher order "remainder term"

$$\mathbf{f}_1(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{A}\mathbf{x}.$$

Then rewrite (7) in the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{f}_1[\mathbf{x}(t)] + \{\mathbf{f}[\mathbf{x}(t), \mathbf{u}(t)] - \mathbf{f}(\mathbf{x}(t), \mathbf{0})\}.$$

Interpret this equation as the exponentially stable system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ driven by an input which itself depends on **x**. Construct an implicit inequality bounding $\mathbf{x}(t)$ and then use Gronwall's inequality [Lemma (5.7.1)] to get an explicit bound on $\mathbf{x}(t)$.

Problem 6.5 Repeat Problem 6.4 for the case where **f** and **g** might be time-varying by using Bellman's inequality instead of the Gronwall inequality. Bellman's inequality (which is an extension of Gronwall's inequality) is as follows: Suppose $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are continuous functions defined on the interval $[0, \infty)$, and that $b(t)$, $c(t)$ are nonnegative for all $t \geq 0$. Suppose that the function $u(\cdot)$ satisfies the implicit inequality

$$u(t) \leq a(t) + b(t) \int_0^t c(\tau) u(\tau) d\tau, \forall t \geq 0.$$

Then

$$u(t) \leq a(t) + b(t) \int_0^t c(\tau) a(\tau) \exp\left[\int_\tau^t b(s) c(s) ds \right] d\tau, \forall t \geq 0.$$

Problem 6.6 Extend Theorem (6.3.46) to time-varying systems.

6.4 OPEN-LOOP STABILITY OF LINEAR SYSTEMS

Before attempting to study the stability of interconnected systems such as in Figure 6.2, it is helpful first to obtain conditions under which the operators G_1 and G_2 represent L_p-stable subsystems. In this section, we concentrate on linear systems and obtain necessary and sufficient conditions for a linear system to be L_p-stable. The term *open-loop stability* refers to the fact that we study the subsystems G_1 and G_2 individually, rather than the overall closed-loop system, which is described by (6.2.24).

6.4.1 Time-Invariant Systems

Throughout most of this subsection, attention is restricted to single-input, single-output (SISO) systems. Once the SISO case is thoroughly analyzed, the results for the MIMO (multi-input, multi-output) case follow easily. Consider a SISO time-invariant system, which is characterized by a scalar transfer function $\hat{h}(s)$. If $\hat{h}(s)$ is a rational function of s (i.e., a ratio of two polynomials in s), then it is well-known that such a system is L_∞-stable (BIBO stable) if and only if

1. $\hat{h}(s)$ is a *proper* rational function (i.e., the degree of the numerator polynomial of $\hat{h}$ is less than or equal to that of the denominator polynomial), and
2. All poles of $\hat{h}$ have negative real parts.

However, the situation is more complicated if $\hat{h}(s)$ is not rational. Such a situation arises whenever $\hat{h}(\cdot)$ is the transfer function of a distributed system, such as an RC transmission line (integrated circuit), an LC transmission line (power line), or if $\hat{h}(\cdot)$ represents a simple delay, etc. In what follows, precise conditions are given under which a scalar $\hat{h}(s)$ (rational or irrational) represents an L_p-stable system. These conditions illustrate one of the chief advantages of the input-output approach to stability, namely, that it places lumped systems [rational $\hat{h}(s)$] and distributed systems [irrational $\hat{h}(s)$] in a unified framework; this is much harder to achieve using Lyapunov theory.

To do this, the sets $\mathbf{A}$ and $\hat{\mathbf{A}}$ are introduced. Basically (as shown later), $\mathbf{A}$ is the set of BIBO stable impulse responses, while $\hat{\mathbf{A}}$ is the set of BIBO stable transfer functions. The precise definitions are given next.

1 Definition *The symbol* $\mathbf{A}$ *denotes the set of generalized functions (distributions)* $f(\cdot)$ *such that* $f(t) = 0$ *when* $t < 0$, *and have the form*

$$\mathbf{2} \qquad f(t) = \sum_{i=0}^{\infty} f_i \, \delta(t - t_i) + f_a(t), \text{ if } t \geq 0.$$

where $\delta(\cdot)$ *denotes the unit delta distribution,* $0 \leq t_0 < t_1 < \cdots$ *are constants,* $f_a(\cdot)$ *is a measurable function, and in addition,*

$$\mathbf{3} \qquad \sum_{i=0}^{\infty} |f_i| < \infty, \int_0^{\infty} |f_a(t)| \, dt < \infty.$$

The norm $\| f \|_{\mathbf{A}}$ *of a distribution in* $\mathbf{A}$ *is defined by*

$$\mathbf{4} \qquad \| f(\cdot) \|_{\mathbf{A}} = \sum_{i=0}^{\infty} |f_i| + \int_0^{\infty} |f_a(t)| \, dt.$$

The convolution of two distributions f *and* g *in* $\mathbf{A}$, *denoted by* $f * g$, *is defined by*

$$5 \qquad (f * g)(t) = \int_0^t f(t-\tau)\, g(\tau)\, d\tau = \int_0^t f(\tau)\, g(t-\tau)\, d\tau.$$

Thus **A** consists of all distributions that vanish for $t < 0$, and for $t \geq 0$ consist of a sum of delayed impulses and a measurable function, with the additional property that the weights of the impulses form an absolutely summable sequence and the measurable function is absolutely integrable. One can think of **A** as the space $L_1[0, \infty)$ augmented by delayed impulses.

Note that, in computing the convolution of two distributions, one should take

$$6 \qquad \delta(t-\tau) * \delta(t-\theta) = \delta(t-\tau-\theta),$$

$$7 \qquad \delta(t-\tau) * f_a(t) = f_a(t-\tau).$$

In other words, the convolution of two delayed unit impulses with delays of τ and θ respectively is another delayed unit impulse with delay $\tau + \theta$, while the convolution of a delayed unit impulse $\delta(t-\tau)$ and a measurable function $f_a(t)$ is the delayed measurable function $f_a(t-\tau)$. Thus, given two elements f, g in **A** of the form

$$8 \qquad f(t) = \sum_{i=0}^{\infty} f_i\, \delta(t - t_i^{(f)}) + f_a(t), \quad g(t) = \sum_{i=0}^{\infty} g_i\, \delta(t - t_i^{(g)}) + g_a(t),$$

their convolution is given by

$$9 \qquad (f * g)(t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f_i g_j\, \delta(t - t_i^{(f)} - t_j^{(g)})$$

$$+ \sum_{i=0}^{\infty} f_i\, g_a(t - t_i^{(f)}) + \sum_{j=0}^{\infty} g_j\, f_a(t - t_j^{(g)}) + \int_0^t f_a(t-\tau)\, g(\tau)\, d\tau.$$

10 Example The function

$$f_1(t) = \exp(-\alpha t)$$

belongs to L_1 and hence to **A**, whenever $\alpha > 0$. The distribution

$$f_2(t) = \sum_{i=0}^{\infty} \frac{1}{(i+1)^2}\, \delta(t - iT), \; T > 0 \text{ given},$$

which is a sequence of evenly spaced delayed impulses, belongs to **A** because the sequence weights $\{1/(i+1)^2\}$ is absolutely summable. However, the distribution

$$f_3(t) = \sum_{i=0}^{\infty} \frac{1}{i+1}\, \delta(t - iT), \; T > 0 \text{ given},$$

does not belong to **A** because the sequence $\{1/(i+1)\}$ is not absolutely summable. The

distribution

$$f_4(t) = \delta(t) + \exp(-t)$$

belongs to **A** and $\|f_4\|_{\mathbf{A}} = 2$.

Remarks

1. Note that L_1 is a subset of **A**; further, if $f \in L_1$, then

11 $\|f\|_{\mathbf{A}} = \|f\|_1$.

2. If $f, g \in \mathbf{A}$ and at least one of them is in L_1 (i.e., does not contain any impulses), then $f * g$ does not contain any impulses. This is clear from (9). It is shown subsequently that if $f \in L_1$, $g \in \mathbf{A}$, then $f * g \in L_1$, i.e., L_1 is an **ideal** in **A**.

As mentioned previously, the set **A** can be interpreted as the set of BIBO stable impulse responses; in other words, a system with an impulse response $h(\cdot)$ is BIBO stable if and only if $h(\cdot) \in \mathbf{A}$. To prove this important result, we first derive some useful properties of **A**. These properties imply that **A** is a Banach algebra with identity, and that it has no zero divisors. [However, a reader who does not know what these terms mean need not worry about it; the terms say nothing more than Lemma (12)].

12 Lemma *The set* **A**, *together with the function* $\|\cdot\|_{\mathbf{A}}$ *and the convolution* $*$, *has the following properties:*

(i) $\|\cdot\|_{\mathbf{A}}$ is a norm on **A**, and **A** is complete under this norm.

(ii) The convolution operation is commutative; i.e.,

13 $f * g = g * f$, $\forall f, g \in \mathbf{A}$.

(iii) The convolution operation is bilinear; i.e.,

14 $f * (\alpha g) = \alpha(f * g)$, $\forall \alpha \in \mathbf{R}$, $\forall f, g \in \mathbf{A}$,

15 $f * (g + h) = f * g + f * h$, $\forall f, g, h \in \mathbf{A}$.

(iv) Whenever $f, g \in \mathbf{A}$, we have that $f * g \in \mathbf{A}$, and in fact

16 $\|f * g\|_{\mathbf{A}} \leq \|f\|_{\mathbf{A}} \cdot \|g\|_{\mathbf{A}}$.

(v) $\delta(\cdot)$ is the unit element of **A**; i.e.,

17 $\delta * f = f * \delta = f$, $\forall f \in \mathbf{A}$.

(vi) **A** has no divisors of the zero element; i.e.,

18 $f * g = 0 \Rightarrow f = 0$ or $g = 0$.

Proof (outline) (i) It is easy to verify that $\|\cdot\|_{\mathbf{A}}$ is indeed a norm on **A**. The completeness of **A** under this norm is more difficult to show, and the fact is stated here without proof.

(ii) and (iii) are obvious.

(iv) Suppose $f, g \in \mathbf{A}$ are of the form (8). Then

$$\textbf{19} \quad (f * g)(t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f_i g_j \, \delta(t - t_i^{(f)} - t_j^{(g)})$$

$$+ \sum_{i=0}^{\infty} f_i \, g_a(t - t_i^{(f)}) + \sum_{j=0}^{\infty} g_j \, f_a(t - t_j^{(g)}) + \int_0^t f_a(t-\tau)\, g(\tau)\, d\tau.$$

The first term on the right side represents the distributional part of $f * g$, while the last three terms represent the measurable part of $f * g$. To compute $\|f * g\|_{\mathbf{A}}$, we take each of the terms separately. First,

$$\textbf{20} \quad \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} |f_i| \cdot |g_j| = \left[\sum_{i=0}^{\infty} |f_i| \right] \cdot \left[\sum_{i=0}^{\infty} |g_j| \right].$$

Next,

$$\textbf{21} \quad \int_0^{\infty} \left| \sum_{i=0}^{\infty} f_i \, g_a(t - t_i^{(f)}) \right| dt \leq \sum_{i=0}^{\infty} |f_i| \int_0^{\infty} |g_a(t - t_i^{(f)})| \, dt$$

$$= \left[\sum_{i=0}^{\infty} |f_i| \right] \cdot \left[\int_0^{\infty} |g_a(t)| \, dt \right].$$

Similarly,

$$\textbf{22} \quad \int_0^{\infty} \left| \sum_{j=0}^{\infty} g_j \, f_a(t - t_j^{(g)}) \right| dt \leq \left[\sum_{j=0}^{\infty} |g_j| \right] \cdot \left[\int_0^{\infty} |f_a(t)| \, dt \right].$$

Finally,

$$\textbf{23}\quad \int_0^\infty \left| \int_0^t |f_a(t-\tau)\, g(\tau)\, d\tau \right| dt \le \int_0^\infty \int_0^t |f_a(t-\tau)\, g_a(\tau)|\, d\tau\, dt$$

$$= \int_0^\infty \int_\tau^\infty |f_a(t-\tau)\, g_a(\tau)|\, dt\, d\tau$$

$$= \int_0^\infty |g_a(\tau)|\, d\tau \cdot \left[\int_\tau^\infty |f_a(t-\tau)|\, dt \right] d\tau$$

$$= \left[\int_0^\infty |g_a(\tau)|\, d\tau \right] \cdot \left[\int_0^\infty |f_a(t)|\, dt \right].$$

Note that the order of integration was interchanged after the first step. Putting together the four bounds (20)-(23) proves (16).

(v) is obvious.

The proof of (vi) is beyond the scope of this book. The reader is instead referred to Hille and Phillips (1957), Theorem 4.18.4. ■

Suppose $f \in \mathbf{A}$. Then, whenever $\operatorname{Re} s \ge 0$, the integral

$$\textbf{24}\quad \hat{f}(s) = \int_0^\infty f(t)\, e^{-st}\, dt = \sum_{i=0}^\infty f_i e^{-st_i} + \hat{f}_a(s)$$

converges and is well-defined. Therefore, all elements of $\mathbf{A}$ are Laplace-transformable, and the region of convergence of the Laplace transform includes the closed right half-plane

$$\textbf{25}\quad C_+ = \{s : \operatorname{Re} s \ge 0\}.$$

Now the set $\hat{\mathbf{A}}$ can be introduced.

26 Definition *The symbol $\hat{\mathbf{A}}$ denotes the set of all functions $\hat{f}: C_+ \to C$ that are Laplace transforms of elements of $\mathbf{A}$.*

Thus, according to Definition (26), "$\hat{f} \in \hat{\mathbf{A}}$" is just another way of saying that the inverse Laplace transform of $\hat{f}$ belongs to $\mathbf{A}$. When we deal with feedback systems, the symbol $\hat{\mathbf{A}}$ comes in handy to keep the notation from proliferating.

27 Lemma *Suppose $\hat{f}(s)$ is a rational function of s. Then $\hat{f} \in \hat{\mathbf{A}}$ if and only if*

(i) $\hat{f}$ *is proper, and*

(ii) *all poles of* $\hat{f}$ *have negative real parts.*

Proof If (i) and (ii) hold, then it is clear that $f(\cdot)$, the inverse Laplace transform of $\hat{f}$, consists of two parts: a measurable function which is bounded by a decaying exponential, and possibly an impulse at time $t = 0$. Hence $f \in \mathbf{A}$, i.e., $\hat{f} \in \hat{\mathbf{A}}$. To prove the necessity of these conditions, suppose (i) does not hold. Then $f(\cdot)$ contains higher order impulses and therefore does not belong to $\mathbf{A}$; if (ii) does not hold, then the measurable part of $f(\cdot)$ is not absolutely integrable. ■

Having defined $\mathbf{A}$, we can define its extension $\mathbf{A}_e$, in exactly the same way that one defines L_{pe} from L_p.

28 Definition *The set* $\mathbf{A}_e$ *consists of all generalized functions* $f(\cdot)$ *which have the property that all truncations* f_T *of* f *belong to* $\mathbf{A}$, *for all* $T \geq 0$, *and is called the* **extension** *of* $\mathbf{A}$.

The set $\mathbf{A}_e$ has some very useful properties. Most physical systems, even those that are "unstable," have impulse responses that belong to $\mathbf{A}_e$ [for example, consider $h(t) = e^t$]. Moreover, it can be shown that if $h(t)$ is any regular measure that vanishes for $t < 0$ and has no singular part, then the corresponding operator H defined by

$$\textbf{29} \quad (Hf)(t) = \int_0^t h(t-\tau) f(\tau)\, d\tau$$

maps L_{pe} into itself for all $p \in [1, \infty]$ if and only if $h \in \mathbf{A}_e$. Thus systems whose impulse responses lie in $\mathbf{A}_e$ are the most general (linear time-invariant) systems that one needs to consider in the present context.

Now the main results of this subsection are stated and proved.

30 Theorem *Consider the operator* H *defined by (29), where* $h \in \mathbf{A}_e$. *Then the following four statements are equivalent:*

(i) H *is* L_1*-stable wb.*

(ii) H *is* L_∞*-stable wb.*

(iii) H *is* L_p*-stable wb for all* $p \in [1, \infty]$.

(iv) $h \in \mathbf{A}$.

Moreover, if $h \in \mathbf{A}$, *then*

$$\textbf{31} \quad \|h * f\|_p \leq \|h\|_{\mathbf{A}} \cdot \|f\|_p, \ \forall p \in [1, \infty].$$

Remarks Theorem (30) brings out fully the importance of the set $\mathbf{A}$. According to this theorem, a necessary and sufficient condition for a system of the form (29) to have any one of various forms of stability is that the impulse response belong to $\mathbf{A}$. This justifies the description of $\mathbf{A}$ as the set of stable impulse responses.

Proof It is first proven that (iv) implies each of (i), (ii) and (iii). Accordingly, suppose $h \in \mathbf{A}$.

(iv) $\Rightarrow$ (i): If $f \in L_1$, then $f \in \mathbf{A}$, and in fact

$$\|f\|_{\mathbf{A}} = \|f\|_1. \tag{32}$$

Hence, by (iv) of Lemma (12), $h * f \in \mathbf{A}$, and

$$\|h * f\|_{\mathbf{A}} \leq \|h\|_{\mathbf{A}} \cdot \|f\|_{\mathbf{A}} = \|h\|_{\mathbf{A}} \cdot \|f\|_1. \tag{33}$$

Next, because f contains no impulses, neither does $h * f$, which means that $h * f \in L_1$, and

$$\|h * f\|_{\mathbf{A}} = \|h * f\|_1. \tag{34}$$

Now (33) and (34) together imply that

$$\|h * f\|_1 \leq \|h\|_{\mathbf{A}} \cdot \|f\|_1, \tag{35}$$

which shows that H is L_1-stable wb.

(iv) $\Rightarrow$ (ii): Suppose $f \in L_\infty$, and write $h \in \mathbf{A}$ in the form

$$h(t) = \sum_{i=0}^{\infty} h_i \, \delta(t - t_i) + h_a(t). \tag{36}$$

Then

$$(h * f)(t) = \sum_{i=0}^{\infty} h_i f(t - t_i) + \int_0^t h_a(\tau) f(t - \tau) \, d\tau, \tag{37}$$

$$\begin{aligned} |(h * f)(t)| &\leq \sum_{i=0}^{\infty} |h_i| \cdot |f(t - t_i)| + \int_0^t |h_a(\tau)| \, |f(t - \tau)| \, d\tau \\ &\leq \operatorname{ess.sup}_t |f(t)| \cdot \left[\sum_{i=0}^{\infty} |h_i| + \int_0^{\infty} |h_a(\tau)| \, d\tau \right] \\ &= \|f\|_\infty \cdot \|h\|_{\mathbf{A}}. \end{aligned} \tag{38}$$

Since (37) holds for (almost) all t, we see that $h * f \in L_\infty$, and that

39 $\|h * f\|_\infty \leq \|h\|_{\mathbf{A}} \cdot \|f\|_\infty$.

This shows that H is L_∞-stable wb.

(iv) $\Rightarrow$ (iii): This part of the proof is omitted as it is a special case of a more general result for linear time-varying systems; see Theorem (75).

Now the reverse implications are proved.

(i) $\Rightarrow$ (iv): Suppose (i) is true. Since H is linear, (i) implies that H is continuous. Now, since every element in $\mathbf{A}$ can be approximated arbitrarily closely in the sense of distributions by an element of L_1, this in turn implies that H maps $\mathbf{A}$ into itself. Now let $f(t) = \delta(t)$; then $h * f = h * \delta = h$, which by assumption belongs to $\mathbf{A}$. This shows that (iv) is true.

(ii) $\Rightarrow$ (iv): This is a special case of Theorem (53).

(iii) $\Rightarrow$ (iv): Suppose (iii) is true, i.e., that H is L_p-stable for all $p \in [1, \infty]$. Then, in particular, H is L_1-stable. As shown above, this implies (iv). ■

40 Theorem *Consider the operator H defined by (29), where $h \in \mathbf{A}$. Then H is L_2-stable, and*

41 $\gamma_2(H) = \sup_\omega |\hat{h}(j\omega)|$.

Remarks The main purpose of Theorem (40) is to prove the bound (41). For an arbitrary $p \in [1, \infty]$, we have the inequality (31), which shows that

42 $\gamma_p(H) \leq \|h\|_{\mathbf{A}}, \ \forall p \in [1, \infty]$.

However, if $p = 2$, then the tighter bound (41) applies.

Proof Let $g = h * f$. Then $\hat{g}(j\omega) = \hat{h}(j\omega)\hat{f}(j\omega)$. Using Parseval's equality gives

$$\textbf{43} \quad \|g\|_2^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\hat{g}(j\omega)|^2 \, d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\hat{h}(j\omega)|^2 \, |\hat{f}(j\omega)|^2 \, d\omega$$

$$= \sup_\omega |\hat{h}(j\omega)|^2 \, \frac{1}{2\pi}\int_{-\infty}^{\infty} |\hat{f}(j\omega)|^2 \, d\omega = \sup_\omega |\hat{h}(j\omega)|^2 \, \|f\|_2^2.$$

Taking the square root of both sides of (43) shows that $\gamma_2(H)$ can be no larger than the right side of (41). Proving that $\gamma_2(H)$ actually equals this quantity is messy but not difficult; the reader is referred to Vidyasagar (1981), Lemma (3.1.69), for the details. ■

The results of this subsection up to now are summarized below:

1. The sets $\mathbf{A}$ and $\hat{\mathbf{A}}$ have been introduced.

2. It has been shown that, in a very precise sense, $\mathbf{A}$ is the set of stable impulse responses and $\hat{\mathbf{A}}$ is the set of stable transfer functions.

3. The useful bounds (41) and (42) have been obtained. Actually, it can be shown that the bound (42) is exact when $p = 1$ and ∞, i.e., $\gamma_1(H) = \gamma_\infty(H) = \|h\|_{\mathbf{A}}$. However, the proofs are a bit involved and are not given here. The interested reader is referred to Vidyasagar (1981), Chapter 3.

This subsection is concluded with a brief discussion of MIMO linear time-invariant systems. The results in this case follow very easily from Theorem (30).

Consider a system with m inputs and l outputs, with the input-output relationship

$$\mathbf{44} \quad (\mathbf{H}f)(t) = \int_0^t H(t-\tau) f(\tau)\, d\tau,$$

where the impulse response matrix $H(\cdot) \in \mathbf{A}_e^{l \times m}$. In analogy with the SISO case, such operators are the most general linear time-invariant operators that one needs to consider in the present context.

The criteria for the L_p-stability of systems of the form (44) are given next.

45 Theorem *Consider an operator* $\mathbf{H}$ *of the form (44), where* $H \in \mathbf{A}_e^{l \times m}$. *Under these conditions, the following four statements are equivalent:*

(i) $\mathbf{H}$ *is* L_1*-stable wb.*

(ii) $\mathbf{H}$ *is* L_∞*-stable wb.*

(iii) $\mathbf{H}$ *is* L_p*-stable wb for all* $p \in [1, \infty]$.

(iv) $H \in \mathbf{A}^{l \times m}$.

The proof is left as an exercise, since it is entirely parallel to that of Theorem (30).

Basically, Theorem (45) states that the set of matrices whose elements all belong to $\mathbf{A}$ is precisely the set of stable MIMO impulse response matrices, whereas the set of matrices whose elements all belong to $\hat{\mathbf{A}}$ is precisely the set of stable MIMO transfer matrices.

Theorem (30) leads to the useful bound (42). A similar bound can be obtained for MIMO systems and is given below without proof. In this connection, it is worthwhile to recall the definition (6.1.27) of the norm on L_p^n.

46 Lemma *Consider an operator of the form (44), where* $H(\cdot) \in \mathbf{A}^{l \times m}$. *Then*

$$\mathbf{47} \quad \gamma_p(\mathbf{H}) \leq \|M_1\|_{i2},$$

where $\|\cdot\|_{i2}$ *denotes the matrix norm induced by the Euclidean vector norm, and* M_1 *is an* $l \times m$ *matrix whose* ij*-th element is* $\|h_{ij}\|_{\mathbf{A}}$. *If* $p = 2$, *then*

48 $\gamma_2(\mathbf{H}) = \| M_2 \|_{i2}$,

where M_2 is the $l \times m$ matrix whose ij-th element is given by

49 $(M_2)_{ij} = \sup_{\omega} | \hat{h}_{ij}(j\omega) |$.

It is easy to verify that (47) reduces to (42), and (48) reduces to (41) in the case of scalar systems.

6.4.2 Time-Varying Systems

In the previous subsection, the focus was on time-invariant systems. In the present subsection, we study a class of operators that represents a natural generalization of those of the form (29). Specifically, we consider operators G of the form

50 $(Gf)(t) = \sum_{i=0}^{\infty} g_i(t) f(t - t_i) + \int_0^t g_a(t, \tau) f(\tau) d\tau$.

Actually, since $f(t) = 0$ whenever $t < 0$, one can rewrite (50) as

51 $(Gf)(t) = \sum_{i \in I(t)} g_i(t) f(t - t_i) + \int_0^t g_a(t, \tau) f(\tau) d\tau$,

where

52 $I(t) = \{i : t_i \leq t\}$.

In other words, (51) is obtained from (50) by taking the summation only over those indices i such that $t_i \leq t$.

Theorem (53) gives necessary and sufficient conditions for an operator G of the form (50) to be L_∞-stable.

53 **Theorem** *Consider an operator G of the form (50), where*

54 $t \mapsto \sum_{i \in I(t)} | g_i(t) | \in L_{\infty e}$,

55 $\tau \mapsto g_a(t, \tau) \in L_1$, $\forall t \geq 0$,

56 $t \mapsto \int_0^t | g_a(t, \tau) | \, d\tau \in L_{\infty e}$.

Then G maps $L_{\infty e}$ into itself. Further, G is L_∞-stable wb if and only if

$$\textbf{57} \quad \sup_t \left\{ \sum_{i \in I(t)} |g_i(t)| + \int_0^t |g_a(t, \tau)| \, d\tau \right\} =: c_\infty < \infty.$$

Remarks Note that conditions (54) to (56) are quite easy to satisfy. For instance, if the set of indices $I(t)$ is finite for each finite t, and if each function $g_i(\cdot)$ is continuous, then (54) is satisfied. Further, the index set $I(t)$ will indeed be finite for each finite t, provided the delay $t_i \to \infty$ as $i \to \infty$. Similarly, if the function g_a is continuous, then (55) and (56) are satisfied.

Proof It is left as a problem to show that if (54) to (56) hold, then G does indeed map $L_{\infty e}$ into itself.

To show that (57) is a necessary and sufficient condition for G to be L_∞-stable wb, we tackle first the sufficiency, since that is much easier.

"If" Suppose (57) holds and that $f \in L_\infty$. Then

$$\textbf{58} \quad |(Gf)(t)| = \left| \sum_{i \in I(t)} g_i(t) f(t - t_i) + \int_0^t g_a(t, \tau) f(\tau) \, d\tau \right|$$

$$\leq \left[\sum_{i \in I(t)} |g_i(t)| + \int_0^t |g_a(t, \tau)| \, d\tau \right] \cdot \|f\|_\infty$$

$$\leq c_\infty \|f\|_\infty.$$

Since the right side of (58) is independent of t, it follows that $Gf \in L_\infty$ and that

$$\textbf{59} \quad \|Gf\|_\infty \leq c_\infty \|f\|_\infty.$$

Hence G is L_∞-stable wb.

"Only if" We show the contrapositive, namely that if (57) does not hold, then the ratio $\|Gf\|_\infty / \|f\|_\infty$ can be made arbitrarily large. To simplify the details, it is assumed that all the delay terms in (50) are zero, i.e., that

$$\textbf{60} \quad (Gf)(t) = \int_0^t g(t, \tau) f(\tau) \, d\tau,$$

where the subscript "a" on g_a is dropped. The proof in the general case is left as an exercise (see Problem 6.13). By assumption, the function

61 $$r(t) = \int_0^t |g(t, \tau)|\, d\tau$$

is unbounded. Let $k < \infty$ be an arbitrary constant; it is shown that there exists a function $f_k \in L_\infty$ of unit norm such that $\|Gf_k\|_\infty \geq k$. Since $r(\cdot)$ is unbounded, there exists a time $t > 0$ such that $r(t) \geq k$. Fix this t, and define

62 $$f_k(\tau) = \text{sign}\, |g(t, \tau)|, \quad \forall \tau \in [0, t],$$

where the sign of 0 is taken as 0. Then

63 $$g(t, \tau) f_k(\tau) = |g(t, \tau)|, \quad \forall \tau \in [0, t], \text{ and}$$

64 $$(Gf_k)(t) = \int_0^t g(t, \tau) f_k(\tau)\, d\tau \geq k.$$

Therefore

65 $$\|Gf_k\|_\infty = \sup_t |(Gf_k)(t)| \geq k.$$

Since this argument can be repeated for any k, it follows that G cannot be L_∞-stable wb. ■

The proof of the "only if" part leaves open the possibility that, if $r(\cdot)$ of (61) is unbounded, then G is L_∞-stable though not L_∞-stable wb. However, it can be shown, using the principle of uniform boundedness, that for operators of the form (50), L_∞-stability and L_∞-stability wb are equivalent properties; see Desoer and Thomasian (1963) or Desoer and Vidyasagar (1975), Theorem (4.7.5).

The next theorem gives necessary and sufficient conditions for G to be L_1-stable.

66 **Theorem** *Consider an operator of the form (50), where*

67 $$t \mapsto \sum_{i \in I(t)} |g_i(t + t_i)| \in L_{\infty e},$$

68 $$t \mapsto g_a(t, \tau) \in L_1, \quad \forall \tau \geq 0,$$

69 $$\tau \mapsto \int_\tau^\infty |g_a(t, \tau)|\, dt \in L_{\infty e}.$$

Under these conditions, G maps L_{1e} into itself. Further, G is L_1-stable wb if and only if

$$\textbf{70} \quad \sup_{\tau}\left[\sum_{i=0}^{\infty}|g_i(\tau+t_i)|+\int_{\tau}^{\infty}|g_a(t,\tau)|\,d\tau\right]=:c_1<\infty.$$

Proof It is left as an exercise to show that if (67) to (69) hold, then G maps L_{1e} into itself.

"If" Suppose (70) holds and that $f\in L_1$. Then

$$\textbf{71} \quad \int_0^{\infty}|(Gf)(t)|\,dt\le\int_0^{\infty}\sum_{i\in I(t)}|g_i(t)|\cdot|f(t-t_i)|\,dt+\int_0^{\infty}\int_0^{t}|g_a(t,\tau)f(\tau)|\,d\tau\,dt$$

$$=\sum_{i=0}^{\infty}\int_0^{\infty}|g_i(t+t_i)|\cdot|f(t)|\,dt+\int_0^{\infty}\int_{\tau}^{\infty}|g_a(t,\tau)|\cdot|f(\tau)|\,dt\,d\tau$$

$$\le\sup_{\tau}\left[\sum_{i=0}^{\infty}|g_i(\tau+t_i)|+\int_{\tau}^{\infty}|g_a(t,\tau)|\,dt\right]\cdot\left[\int_0^{\infty}|f(\tau)|\,d\tau\right]$$

$$=c_1\|f\|_1.$$

This shows that G is L_1-stable wb.

"Only if" Suppose G is L_1-stable wb. Since G is linear, this implies that G is continuous on L_1. Sinnce every distribution in $\mathbf{A}$ can be expressed as a limit, in the sense of distributions, of a sequence of functions in L_1, it follows that G maps $\mathbf{A}$ into $\mathbf{A}$ with finite gain. Let $f(t)=\delta(t-\tau)$, where $\tau\in\mathbf{R}_+$ is a given number. Then

$$\textbf{72} \quad (Gf)(t)=\sum_{i\in I(t)}g_i(t)\,\delta(t-\tau-t_i)+g_a(t,\tau)=\sum_{i=0}^{\infty}g_i(\tau+t_i)\,\delta(t-\tau-t_i)+g_a(t,\tau).$$

Since G maps $\mathbf{A}$ into itself, it follows [upon using the fact that $g_a(t,\tau)=0$ if $t<\tau$] that

$$\textbf{73} \quad \|G\delta(t-\tau)\|_{\mathbf{A}}=\sum_{i=0}^{\infty}|g_i(\tau+t_i)|+\int_{\tau}^{\infty}|g_a(t,\tau)|\,dt<\infty.$$

Now, since $\|\delta(t-\tau)\|_{\mathbf{A}}=1\ \forall\tau\ge 0$ and G maps $\mathbf{A}$ into itself with finite gain, it finally follows that

$$\textbf{74} \quad \sup_{\tau}\|G\delta(t-\tau)\|_{\mathbf{A}}<\infty.$$

But (74) is the same as (70). ■

Finally, it is shown that if G of (50) is both L_1-stable and L_∞-stable, then it is L_p-stable for all $p \in [1, \infty]$. The theorem is stated in full generality, but is only proved in a slightly simplified case; the general case is proved in Willems (1969b).

75 Theorem *Suppose an operator G of the form (50) satisfies both (57) and (70). Then G is L_p-stable for all $p \in [1, \infty]$. Moreover,*

$$\textbf{76} \quad \gamma_p(G) \leq c_1^{1/p} c_\infty^{1/q},$$

where $q = p/(p-1)$ is the conjugate index of p.

Proof The theorem is proved under the additional assumption that $g_i(t) \equiv 0 \ \forall t$, and the subscript "$a$" on the function g_a is dropped for convenience. The proof in the general case is similar [see Problem 6.14 or Willems (1969b)].

If it is assumed that $g_i(t) \equiv 0 \ \forall t$, then (57) and (70) reduce respectively to

$$\textbf{77} \quad \sup_t \int_0^t |g(t, \tau)| \, d\tau =: c_\infty < \infty,$$

$$\textbf{78} \quad \sup_\tau \int_\tau^\infty |g(t, \tau)| \, dt =: c_1 < \infty.$$

Suppose $f \in L_p$. It can be assumed that $1 < p < \infty$, since (76) is clearly true if $p = 1$ or ∞. Now

$$\textbf{79} \quad |(Gf)(t)| \leq \int_0^t |g(t, \tau)| \cdot |f(\tau)| \, d\tau$$

$$= \int_0^t |g(t, \tau)|^{1/q} |g(t, \tau)|^{1/p} |f(\tau)| \, d\tau$$

$$\leq \left[\int_0^t |g(t, \tau)| \, d\tau \right]^{1/q} \cdot \left[\int_0^t |g(t, \tau)| \, |f(\tau)|^p \, d\tau \right]^{1/p}$$

by Hölder's inequality. Next,

$$\textbf{80} \quad |(Gf)(t)|^p \leq \left[\int_0^t |g(t, \tau)| \, d\tau \right]^{p/q} \cdot \left[\int_0^t |g(t, \tau)| \, |f(\tau)|^p \, d\tau \right]$$

$$\leq c_\infty^{p/q} \int_0^t |g(t,\tau)|\,|f(\tau)|^p\,d\tau,$$

81
$$\int_0^\infty |(Gf)(t)|^p\,dt \leq c_\infty^{p/q} \int_0^\infty \int_0^t |g(t,\tau)|\,|f(\tau)|^p\,d\tau\,dt$$

$$= c_\infty^{p/q} \int_0^\infty \int_\tau^\infty |g(t,\tau)|\,|f(\tau)|^p\,dt\,d\tau$$

$$= c_\infty^{p/q} \int_0^\infty \left[\int_\tau^\infty |g(t,\tau)|\,dt \right] \cdot |f(\tau)|^p\,d\tau$$

$$\leq c_\infty^{p/q} c_1 \|f\|_p^p.$$

Raising both sides of (81) to the power $1/p$ gives

82 $\|Gf\|_p \leq c_\infty^{1/q} c_1^{1/p} \|f\|_p.$

This proves that G is L_p-stable wb for all p, and establishes the bound (76). ■

This subsection is concluded by showing that if G is a time-invariant operator, then Theorems (53), (66) and (75) together reduce to Theorem (30). Accordingly, suppose

83 $g_i(t) \equiv h_i,\ \forall t \geq 0,$

84 $g_a(t,\tau) = h_a(t-\tau),\ \forall t, \tau \geq 0,$

so that G corresponds to a time-invariant system with the impulse response

85 $h(t) = \sum_{i=0}^\infty h_i\,\delta(t-t_i) + h_a(t).$

Then the condition (57) for L_∞-stability becomes

86 $\sup_t \left[\sum_{i \in I(t)} |h_i| + \int_0^t |h_a(t-\tau)|\,d\tau \right] < \infty.$

However, as $t \to \infty$, the index set $I(t)$ eventually includes all i. Thus (86) is equivalent to requiring that $h \in \mathbf{A}$. Similarly, the condition (70) for L_1-stability becomes

$$\textbf{87} \quad \sup_{\tau} \left[\sum_{i=0}^{\infty} |h_i| + \int_{\tau}^{\infty} |h_a(t-\tau)|\, dt \right] < \infty,$$

which is also equivalent to requiring that $h \in \mathbf{A}$. Finally, since in the time-invariant case

$$\textbf{88} \quad c_1 = c_\infty = \|h\|_{\mathbf{A}},$$

the bound (76) reduces to (31).

Problem 6.7 Determine whether or not each of the following distributions belongs to $\mathbf{A}$:

$$\text{(a)}\ f(t) = \sum_{i=0}^{\infty} (1/i^2)\, \delta(t - 1 + 1/i).$$

In this case the times at which the distributions occur cluster at the point $t = 1$.

$$\text{(b)}\ f(t) = \sum_{i=1}^{\infty} \exp(-i^2 t).$$

$$\text{(c)}\ f(t) = \delta(t-1) + \exp(-2t) \sin 4t.$$

Problem 6.8 Suppose $f(\cdot) \in \mathbf{A}$. Show that

$$|\hat{f}(s)| \leq \|f(\cdot)\|_{\mathbf{A}}, \ \forall s \in C_+.$$

Problem 6.9 Determine whether each of the functions below belongs to $\hat{\mathbf{A}}$.

$$\text{(a)}\ \hat{f}(s) = e^{-s} \frac{s^2 + 5s + 5}{s^2 + s + 10}.$$

$$\text{(b)}\ \hat{f}(s) = \frac{1}{\cosh \sqrt{s}}.$$

[Hint: Do a partial fraction expansion of $\hat{f}(s)$.]

Problem 6.10 Show that if $\hat{f} \in \hat{\mathbf{A}}$, then the function $s \mapsto \exp(-sT)\hat{f}(s) \in \hat{\mathbf{A}}$ for all $T \geq 0$.

Problem 6.11 Suppose $\hat{f}$ is a rational function, and define the operator $F: x \mapsto f * x$. Show that the following four statements are equivalent: (i) F is L_1-stable wb; (ii) F is L_2-stable wb; (iii) F is L_∞-stable wb; (iv) $\hat{f} \in \hat{\mathbf{A}}$.

Problem 6.12 Determine whether or not each of the operators below is (i) L_1-stable, and (ii) L_∞-stable, using Theorems (66) and (53) respectively.

(a) $(Hu)(t) = u(t-2) + \int_0^t \sin t\, e^{-2(t-\tau)}\, u(\tau)\, d\tau.$

(b) $(Hu)(t) = \int_0^t e^{(-t+2\tau)}\, u(\tau)\, d\tau.$

Problem 6.13 Complete the proof of Theorems (53) and (66) without assuming that all delay terms are zero.

Problem 6.14 Prove Theorem (75) without assuming that all delays are zero.

6.5 LINEAR TIME-INVARIANT FEEDBACK SYSTEMS

In this section, we study conditions under which a feedback interconnection of linear time-invariant subsystems results in a stable system. These results are important and useful in their own right. Moreover, they are a point of departure for the stability analysis of nonlinear and/or time-varying systems.

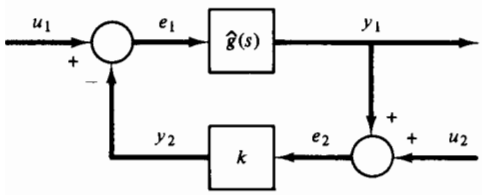


Fig. 6.5

This section is divided into three parts. In the first subsection, the focus is on SISO systems with a constant scalar feedback of the form shown in Figure 6.5. Necessary and sufficient conditions for feedback stability are derived, and a graphical test for verifying these conditions is presented; this graphical test is a generalization of the familiar Nyquist criterion. In the second and third subsections, the attention is on MIMO systems of the form shown in Figure 6.2, where both G_1 and G_2 can represent systems with dynamics (in contrast with the system in Figure 6.5, where the feedback element is non-dynamic). Throughout the section, the emphasis is on the challenge posed by the fact that various subsystems can be distributed. The study of feedback stability in the case where the constituent subsystems are linear, time-invariant, and also lumped, belongs more properly in a book devoted to linear system theory; see, for example, Kailath (1980), Chen (1986), or Vidyasagar (1985).

The following lemma, popularly known as a Paley-Wiener theorem, is the central tool used in this section.

1 Lemma *Suppose* $\hat{f} \in \hat{\mathbf{A}}$. *Then the function* $1/\hat{f}$ *belongs to* $\hat{\mathbf{A}}$ *if and only if*

$$\inf_{\mathrm{Re}\, s \geq 0} |\hat{f}(s)| > 0. \tag{2}$$

The proof of this important result is well beyond the scope of this book, and can be found in Hille and Phillips (1957), p. 150. But the necessity of the condition (2) is quite easy to see. If $1/\hat{f}$ belongs to $\hat{\mathbf{A}}$, then the function $1/\hat{f}(s)$ is bounded over the closed right half-plane C_+ [defined in (6.4.25)]. But this is the same as $|\hat{f}(s)|$ being bounded away from 0 over C_+, which is what (2) says. The sufficiency of (2) is, of course, considerably more difficult to establish, but if $\hat{f}(s)$ is rational, then the sufficiency is easily seen (Problem 6.15).

An element $f \in \mathbf{A}$ is called a *unit* of $\mathbf{A}$ if there exists a $g \in \mathbf{A}$ such that $f * g = \delta(t)$, i.e., if f has a multiplicative inverse in $\mathbf{A}$. In such a case we also say that $\hat{f}$ is a unit of $\hat{\mathbf{A}}$. Now Lemma (1) gives a simple necessary and sufficient condition for a given f to be a unit of A, namely that (2) must hold.

Lemma *Suppose* $\hat{F} \in \hat{\mathbf{A}}^{n \times n}$. *Then the function* $[\hat{F}(\cdot)]^{-1} : s \mapsto [\hat{F}(s)]^{-1}$ *also belongs to* $\hat{\mathbf{A}}^{n \times n}$ *if and only if*

$$\inf_{\mathrm{Re}\, s \geq 0} |\det \hat{F}(s)| > 0. \tag{4}$$

Remark Note that Lemma (3) allows us to determine whether the *matrix-valued* function $[\hat{F}(\cdot)]^{-1}$ belongs to $\hat{\mathbf{A}}^{n \times n}$ by examining the *scalar-valued* function $\det [\hat{F}(\cdot)]$.

Proof Since the determinant of $\hat{F}$ is obtained by forming sums and products of the various components of $\hat{F}$ (all of which belong to $\hat{\mathbf{A}}$), and since $\hat{\mathbf{A}}$ is closed under both of these operations, it follows that $\hat{\Delta} := \det \hat{F} \in \hat{\mathbf{A}}$.

"If" Suppose (4) holds. Then, by Lemma (1), the function $1/\hat{\Delta}$ belongs to $\hat{\mathbf{A}}$. Now write

$$[\hat{F}(s)]^{-1} = \frac{1}{\hat{\Delta}(s)} \operatorname{Adj} [\hat{F}(s)], \tag{5}$$

where $\operatorname{Adj} \hat{F}(s)$ denotes the adjoint matrix of $\hat{F}$, i.e., the matrix of cofactors of $\hat{F}$. Now $\operatorname{Adj} \hat{F} \in \hat{\mathbf{A}}^{n \times n}$, since the components of $\operatorname{Adj} \hat{F}$ are determinants of various submatrices of $\hat{F}$. Hence, if $1/\hat{\Delta} \in \hat{\mathbf{A}}$, then $[\hat{F}(\cdot)]^{-1} \in \hat{\mathbf{A}}^{n \times n}$.

"Only if" Suppose $[\hat{F}(\cdot)]^{-1} \in \hat{\mathbf{A}}^{n \times n}$. Then $\det [\hat{F}(\cdot)] = 1/\hat{\Delta} \in \hat{\mathbf{A}}$. Since $\hat{\Delta} \in \hat{\mathbf{A}}$, Lemma (1) now implies that (4) holds. ■

6.5.1 SISO Systems with Constant Feedback

Consider the system shown in Figure 6.5, where $\hat{g}(s)$ is the transfer function of a linear time-invariant SISO system, and $k \neq 0$ is a constant. There is no loss of generality in assuming that $k \neq 0$, since if $k = 0$ then there is no feedback, and the overall system is stable if and only if $\hat{g} \in \hat{\mathbf{A}}$. Suppose $1 + k\hat{g}(s)$ is not identically zero (which essentially says that the feedback system is *well-posed*); then one can explicitly write down the transfer matrix relating y to u. Indeed, we have

$$\mathbf{6} \qquad \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \frac{1}{1+k\hat{g}} \begin{bmatrix} \hat{g} & -k\hat{g} \\ k\hat{g} & k \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} =: \hat{H} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix}.$$

Let $\hat{H}$ denote the 2×2 transfer matrix in (6). Then the feedback system is (for example) BIBO stable if and only if the output $y \in L_\infty^2$ whenever the input $u \in L_\infty^2$. Since the system is linear and time-invariant, a great many stability notions are equivalent to the requirement that $\hat{H} \in \hat{\mathbf{A}}^{2\times 2}$ [see Theorem (6.4.45)]. Hence, throughout this subsection, feedback stability is taken to mean that $\hat{H} \in \hat{\mathbf{A}}^{2\times 2}$.

7 Lemma *$\hat{H} \in \hat{\mathbf{A}}^{2\times 2}$ if and only if*

$$\mathbf{8} \qquad \frac{1}{1+k\hat{g}} =: \hat{r} \in \hat{\mathbf{A}}.$$

Proof "Only if" Suppose $\hat{H} \in \hat{\mathbf{A}}^{2\times 2}$. Then $\hat{h}_{22} = k\hat{r} \in \hat{\mathbf{A}}$. But since $k \neq 0$, this in turn implies that $\hat{r} \in \hat{\mathbf{A}}$.

"If" Suppose $\hat{r} \in \hat{\mathbf{A}}$. It is shown in turn that each of the four elements of $\hat{H}$ belongs to $\hat{\mathbf{A}}$. First, $\hat{h}_{22} = k\hat{r} \in \hat{\mathbf{A}}$. Next,

$$\mathbf{9} \qquad \hat{h}_{21} = -\hat{h}_{12} = 1 - \hat{r} \in \hat{\mathbf{A}}.$$

Finally,

$$\mathbf{10} \qquad \hat{h}_{11} = \hat{h}_{21}/k \in \hat{\mathbf{A}}.$$

This completes the proof. ■

The quantity $1 + k\hat{g}$ is often referred to as the *return difference*. Thus Lemma (7) states that the feedback system is stable if and only if the reciprocal of the return difference is a stable transfer function. Now the challenge is to find some easily verifiable conditions for ensuring that (8) holds. Suppose $\hat{g} \in \hat{\mathbf{A}}$, i.e., that the system is open-loop stable. Then $1 + k\hat{g}$ also belongs to $\hat{\mathbf{A}}$. Hence, by Lemma (1), it follows that $1/(1 + k\hat{g}) \in \hat{\mathbf{A}}$ (and the feedback system is stable) if and only if

$$\mathbf{11} \qquad \inf_{\mathrm{Re}\, s \geq 0} |1 + k\hat{g}(s)| > 0,$$

i.e., the return difference is bounded away from zero over C_+. However, the condition $\hat{g} \in \hat{\mathbf{A}}$ is very restrictive, and one would like to have a criterion that also applies to systems that are open-loop unstable. Such a result is given next.

12 Theorem *Suppose $\hat{g}$ is of the form*

13 $\hat{g}(s) = \hat{g}_a(s) + \hat{g}_r(s)$,

where $\hat{g}_a \in \hat{\mathbf{A}}$ *and* $\hat{g}_r$ *is rational and strictly proper. Then* $\hat{H}$ *of (6) belongs to* $\hat{\mathbf{A}}^{2\times 2}$ *if and only if (11) holds.*

Remarks Before proving the theorem, some remarks are in order to explain the hypothesis and value of the theorem.

1. The hypothesis on $\hat{g}$ is that it consists of a stable part which could be distributed plus an unstable part which is lumped. In particular, this implies that $\hat{g}$ is meromorphic on the open RHP, and that it has only a finite number of singularities in the open RHP, each of which is a pole of finite order.

2. The theorem is useful because it shows that (11) is a necessary and sufficient condition for the feedback stability of a broad class of systems, not just those that are open-loop stable. With this as the starting point, it is possible to derive a Nyquist-like graphical stability test.

Proof "Only if" Suppose $\hat{H} \in \hat{\mathbf{A}}^{2\times 2}$. Then, as shown in Lemma (7), it follows that $\hat{r} = 1/(1 + k\hat{g}) \in \hat{\mathbf{A}}$, which in turn implies that $\hat{r}$ is bounded over the closed RHP, i.e., that

14 $\sup_{\mathrm{Re}\, s \geq 0} |\hat{r}(s)| < \infty$.

But since $1 + k\hat{g} = 1/\hat{r}$, (14) is equivalent to (11).

"If" Suppose (11) holds. To show that $\hat{H} \in \hat{\mathbf{A}}^{2\times 2}$, it is enough by Lemma (7) to show that $\hat{r} = 1/(1 + k\hat{g}) \in \hat{\mathbf{A}}$. For this purpose, express the rational function $\hat{g}_r(s)$ as $\alpha(s)/\beta(s)$ where α and β are polynomials with no common zeros. Let δ denote the degree of the polynomial β. Since $\hat{g}_r$ is assumed to be strictly proper, the degree of α is at most $\delta - 1$. Now define

15 $\hat{n}(s) = \dfrac{\alpha(s)}{(s+1)^\delta}$, $\hat{d}(s) = \dfrac{\beta(s)}{(s+1)^\delta}$,

and note that

16 $\hat{n}, \hat{d} \in \hat{\mathbf{A}}$, and $\hat{g}_r(s) = \dfrac{\hat{n}(s)}{\hat{d}(s)}$.

Now

17 $\hat{g} = \hat{g}_a + \hat{g}_r = \dfrac{\hat{n} + \hat{d}\hat{g}_a}{\hat{d}}$,

18 $$\hat{r} = \frac{1}{1 + k\hat{g}} = \frac{\hat{d}}{\hat{d} + k(\hat{n} + \hat{d}\hat{g}_a)} = \frac{\hat{d}}{\hat{q}},$$

where

19 $$\hat{q} := \hat{d} + k(\hat{n} + \hat{d}\hat{g}_a).$$

Suppose we could establish, starting from (11), that

20 $$\inf_{\mathrm{Re}\, s \geq 0} |\hat{q}(s)| > 0.$$

Then Lemma (1) would imply that $1/\hat{q} \in \hat{\mathbf{A}}$, which in turn implies, due to (18), that $\hat{r} \in \hat{\mathbf{A}}$, and the stability of H would be proved. Accordingly, the proof is completed by showing that (20) is true. Note that

21 $$\hat{q} = \hat{d}(1 + k\hat{g}).$$

By (11), the quantity $|1 + k\hat{g}(s)|$ is bounded away from zero over the closed RHP C_+. On the other hand, $\hat{d}$ could have some zeros in C_+, namely the the poles of $\hat{g}_r$. Let $\lambda_1, \cdots, \lambda_r$ denote the poles of $\hat{g}_r$ in C_+, and select some open disks $B_1, \cdots, B_r$, with B_i centered at λ_i, such that none of the disks B_i contains a zero of $\hat{n}$. Since $\hat{n}$ and $\hat{d}$ have no common zeros, this can be achieved by making the disks B_i sufficiently small. One other technicality is that if some λ_i is on the $j\omega$-axis, then B_i is chosen to be a half-disk, so that $B_i \subseteq C_+$. Now define B to be the union of the sets B_1 through B_r, and note that, by assumption,

22 $$\inf_{s \in B} |\hat{n}(s)| > 0,$$

since B does not contain any zeros of $\hat{n}$. Similarly it follows that

23 $$\inf_{s \in C_+ - B} |\hat{d}(s)| > 0.$$

So, if (11) holds, then (21) and (23) show that

24 $$\inf_{s \in C_+ - B} |\hat{q}(s)| > 0.$$

What if $s \in B$? At the zeros of $\hat{d}$, we have, from (19), that

25 $$\hat{q}(s) = k\hat{n}(s) \neq 0.$$

Hence, by selecting the disks B_i small enough and making use of (22), one can ensure that

26 $\inf_{s \in B} |\hat{q}(s)| > 0.$

Finally, combining (24) and (26) establishes (20). As shown earlier, this completes the proof. ■

Now let us consider the issue of how one might go about verifying whether the condition (11) holds. If $\hat{g}$ is rational, then the familiar Nyquist criterion of undergraduate control theory allows one to test whether (11) holds by examining only the behavior of the function $\hat{g}(j\omega)$ as ω varies over the real numbers. In attempting to extend the test to distributed systems, the main difficulty one faces is the irrationality of the function $\hat{g}(s)$. Specifically, write $\hat{g}_a(s)$ [where $\hat{g} = \hat{g}_a + \hat{g}_r$; see (13)] in the form

27 $$\hat{g}_a(s) = \sum_{i=0}^{\infty} g_i e^{-st_i} + \hat{g}_m(s),$$

where $\hat{g}_m$ is the Laplace transform of a function $g_1 \in L_1$. Now, by the Riemann-Lebesgue lemma, $|\hat{g}_m(j\omega)| \to 0$ as $|\omega| \to \infty$. On the other hand, the first term

28 $$\hat{g}_{ap}(j\omega) = \sum_{i=0}^{\infty} g_i e^{-j\omega t_i}$$

is an "almost periodic" function of ω. In the practically important case where the delays t_i are all commensurate, i.e.,

29 $t_i = iT$, T given,

the function $\hat{g}_{ap}(j\omega)$ has the form

30 $$\hat{g}_{ap}(j\omega) = \sum_{i=0}^{\infty} g_i (e^{-j\omega T})^i,$$

and is a periodic function of ω with period $2\pi/T$. In either case, it is quite possible that $\hat{g}_a(j\omega)$ has no specific limit as $|\omega| \to \infty$. This difficulty does not arise if $\hat{g}$ is a proper rational function. In spite of this, however, it is nevertheless possible to state a Nyquist-like graphical stability test. The proof in the case of commensurate delays is given by Willems (1969a) and in the noncommensurate case by Callier and Desoer (1972). To avoid technicalities, only the case of commensurate delays is discussed here, and the proof is omitted.

Before proceeding to the graphical stability criterion, it is necessary to introduce two preliminary concepts, namely (i) the indented $j\omega$-axis, and (ii) the argument of the return difference function $1 + k\hat{g}(j\omega)$. If $\hat{g}$ has a pole at some point $j\lambda_i$ on the $j\omega$-axis, then $1 + k\hat{g}(j\lambda_i)$ is undefined. To circumvent this difficulty, the $j\omega$-axis is "indented" around the pole, as shown in Figure 6.6, by going around the pole. Let B_i denote the half-disk shown in Figure 6.6; i.e., let

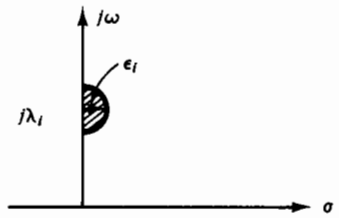


Fig. 6.6

31 $B_i = \{s \in C_+ : |s - j\lambda_i| < \varepsilon_i\}$.

Then, by choosing the radius ε_i sufficiently small, one can ensure that

32 $\inf_{s \in B_i} |1 + k\hat{g}(s)| > 0.$

Hence deleting the half-disk B_i from the closed RHP does not alter whether (11) holds or not. Similar indentation can be performed around all $j\omega$-axis poles of $\hat{g}$.

After the $j\omega$-axis is indented, it is clear that, corresponding to each $\omega \in \mathbf{R}$, there is exactly one point on the indented $j\omega$-axis whose imaginary part is ω, but whose real part may or may not be zero. By a slight abuse of notation, let $\hat{g}(j\omega)$ denote the value of $\hat{g}$ at the point on the indented $j\omega$-axis whose imaginary part is ω; this is illustrated in Figure 6.7.

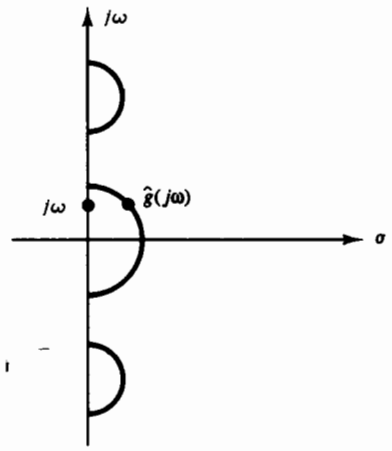


Fig. 6.7

Once the $j\omega$-axis is indented as indicated above and $\hat{g}(j\omega)$ is suitably defined, the quantity $\hat{g}(j\omega)$ is a uniformly continuous function of ω. It can also be assumed that $1+k\hat{g}(j\omega)\neq 0\ \forall\omega$, since otherwise the condition (11) is immediately violated and no further analysis is needed. Thus it is possible to define a continuous function $\phi(j\omega)$ such that

33 $1+k\hat{g}(j\omega)=|1+k\hat{g}(j\omega)|\exp[j\phi(j\omega)]$, and

34 $\phi(j0)=0$ if $1+k\hat{g}(j0)>0$, π if $1+k\hat{g}(j0)<0$.

One can also think of $\phi(j\omega)$ as the argument of the complex number $1+k\hat{g}(j\omega)$; hence we could denote it by the more suggestive notation $\text{Arg}\,[1+k\hat{g}(j\omega)]$.

Now the result can be stated.

35 Theorem (Graphical Stability Test) *Suppose $\hat{g}$ has the form (13), and in addition, suppose that the delays in the distributional part of $\hat{g}_a$ are uniformly spaced, as in (29). Let μ_+ denote the number of poles of $\hat{g}$ with positive real part. Then $\hat{H}$ of (6) belongs to $\hat{\mathbf{A}}^{2\times 2}$ if and only if*

36 $(i)\ \inf_{\omega\in\mathbb{R}}|1+k\hat{g}(j\omega)|>0$, and

37 $(ii)\ \lim_{n\to\infty}[\phi(j2\pi n/T)-\phi(-j2\pi n/T)]=2\pi\mu_+$.

As mentioned earlier, the proof can be found in Willems (1969a).

Theorem (35) has an interpretation quite similar to that of the standard Nyquist criterion. Condition (i) or (36) is equivalent to the following statement: The plot of $\hat{g}(j\omega)$ is *bounded away* from the "critical point" $-1/k$. Note that this is a stronger statement than "The plot of $\hat{g}(j\omega)$ does not pass through the critical point $-1/k$." The latter statement suffices if $\hat{g}$ is a strictly proper rational function, or even if g_a does not contain any delayed impulses, since in this case $\hat{g}(j\omega)$ has a well-defined limit as $|\omega|\to\infty$. But in general it is necessary to use the more precise condition (36). Condition (ii) or (37) is a generalization of the familiar requirement that the plot of $\hat{g}(j\omega)$ encircle the critical point $-1/k$ exactly μ_+ times in the counterclockwise direction. If $\hat{g}$ has the general form (13), then the phrase "encircle" has no meaning since $\phi(j\omega)$ need not have a specific limit as $|\omega|\to\infty$. This is why, in (37), the phase ϕ is evaluated at specially chosen, evenly spaced, frequencies $2\pi n/T$. Of course, if g_a does not contain any delayed impulses, then one can make do with the simpler statement (40) below. The discussion can be summarized as follows:

38 Corollary *Suppose $\hat{g}$ has the form (13), and suppose in addition that $g_a(\cdot)$ does not contain any delayed impulses. Let μ_+ denote the number of poles of $\hat{g}$ with positive real part. Then $\hat{H}$ of (6) belongs to $\hat{\mathbf{A}}^{2\times 2}$ if and only if*

39 $(i)\ 1+k\hat{g}(j\omega)\neq 0,\ \forall\omega\in \mathbf{R}\cup\{\infty\}$, and

40 $(ii)\ \lim_{\omega\to\infty}[\phi(j\omega)-\phi(-j\omega)]=2\pi\mu_+$.

41 **Example** Consider a system of the form shown in Figure 6.5, where

$$\hat{g}(s)=\exp(-s)\,\frac{s^2+4s+2}{s^2-1}.$$

In this case $\hat{g}$ represents a system with the rational transfer function $(s^2+4s+2)/(s^2-1)$ followed (or preceded) by a delay of 1 second. The objective is to determine the range of values of the gain k for which the feedback system is stable.

The first step is to demonstrate that $\hat{g}$ is of the form (13). This turns out to be surprisingly difficult and serves as a motivation for introducing the set $\hat{B}$ in the next subsection. By partial fraction expansion, we have

$$\frac{s^2+4s+2}{s^2-1}=1+\frac{0.5}{s+1}+\frac{3.5}{s-1}.$$

Hence

$$e^{-s}\,\frac{s^2+4s+2}{s^2-1}=e^{-s}\left[1+\frac{0.5}{s+1}\right]+\frac{3.5e^{-s}}{s-1}.$$

Clearly the first term belongs to $\hat{\mathbf{A}}$, since it is the product of two functions, each of which belongs to $\hat{\mathbf{A}}$; in fact, the first term is the Laplace transform of $\delta(t-1)+0.5\exp(-t-1)\,U(t-1)$, where $U(t-1)$ is the delayed unit step function. Now consider the last term, and expand it as

$$\frac{3.5e^{-s}}{s-1}=\frac{3.5e^{-1}}{s-1}+\frac{3.5(e^{-s}-e^{-1})}{s-1}.$$

The first term is rational and unstable, but what is the nature of the second term? Let

$$\hat{f}(s)=\frac{e^{-s}-e^{-1}}{s-1}.$$

Then

$$f(t)=-e^{-1}\cdot e^{t}+e^{(t-1)}\cdot U(t-1)=0,\ \forall t\geq 1.$$

Hence $f\in L_1$, and $\hat{f}\in\hat{\mathbf{A}}$, though $\hat{f}$ is of course not rational. So finally it follows that $\hat{g}=\hat{g}_a+\hat{g}_r$, where

$$\hat{g}_a(s)=e^{-s}\left[1+\frac{0.5}{s+1}\right]+\frac{3.5(e^{-s}-e^{-1})}{s-1},\ \hat{g}_r(s)=\frac{3.5e^{-1}}{s-1}.$$

Thus $\hat{g}$ is of the form (13), so that Theorem (12) is applicable.

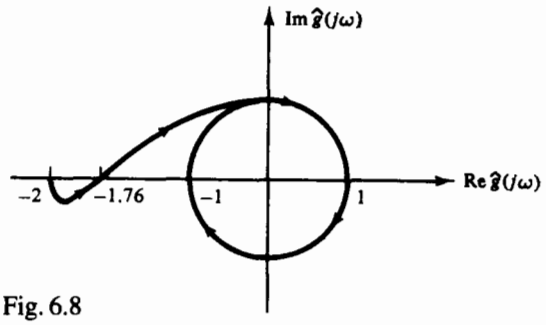


Fig. 6.8

The Nyquist plot of $\hat{g}(j\omega)$ is shown in Figure 6.8. Note that, as $\omega\to\infty$, $\hat{g}(j\omega)$ approaches the periodic function $\exp(-j2\omega)$. Also, μ_+, the number of open RHP poles of $\hat{g}$, equals 1. From Theorem (12) and Figure 6.8, we conclude that the feedback system is stable if and only if

$$-2<-1/k<-1.76, \text{ or } 0.5<k<0.568. \blacksquare$$

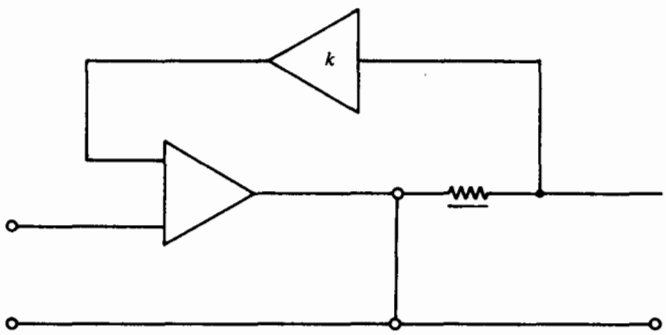


Fig. 6.9

42 Example Consider a uniform RC transmission line with an operational amplifier in the feedback, as shown in Figure 6.9. Suppose the transmission line is driven by a voltage input and has a voltage output. Then the forward transfer function $\hat{g}(s)$ equals the so-called A-parameter of the transmission, which is one of the four chain parameters. For a uniform transmission line of infinite length, it is known (Protonotarios and Wing, 1970) that

$$\hat{g}(s) = \frac{1}{\cosh\sqrt{\lambda s}},$$

where λ is a physical constant. Let us take $\lambda = 1$ for simplicity. Then, as shown by Protonotarios and Wing, $g(\cdot) \in L_1$. Hence Corollary (38) applies, and one can determine the range of values of the gain k for which the system of Figure 6.9 is stable by examining the Nyquist plot of $\hat{g}(j\omega)$. Now

$$\hat{g}(j\omega) = \frac{1}{\cos x \cosh x + \sin x \sinh x},$$

where $x = \sqrt{\omega/2}$. Hence the Nyquist plot more or less spirals into the origin, as shown in Figure 6.10. The range of values of k for which the feedback system is stable is

$$-1 < k < 23.16. \blacksquare$$

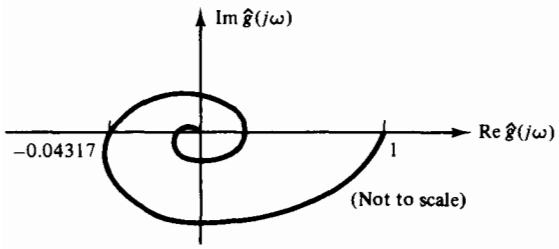


Fig. 6.10

6.5.2 The Set $\hat{B}$

The next two subsections introduce some tools that are useful in analyzing the stability of general feedback systems of the form shown in Figure 6.2, when both G_1 and G_2 are linear and time-invariant but could be multi-input, multi-output (MIMO). The set $\hat{B}$ is introduced in the present subsection, while coprime factorizations are introduced in the next.

Let $\sigma > 0$ be a given number; then the set $\mathbf{A}_\sigma$ consists of all distributions $f(\cdot)$ such that $f(t) = 0$ for $t < 0$, and have the form

$$\textbf{43} \quad f(t) = \sum_{i=0}^{\infty} f_i \delta(t - t_i) + f_a(t) \text{ if } t \geq 0,$$

which satisfy the conditions

$$\textbf{44} \quad \sum_{i=0}^{\infty} |f_i| e^{\sigma t_i} < \infty, \int_0^{\infty} |f_a(t)| e^{\sigma t} dt < \infty.$$

Note that if $\sigma = 0$, then (44) reduces to (6.4.3). Hence the condition (44) is more restrictive than (6.4.3), and as a result, $\mathbf{A}_\sigma$ is a subset of $\mathbf{A}$ for all $\sigma > 0$. Also, if $\sigma > \theta$, then $\mathbf{A}_\sigma$ is a

subset of $\mathbf{A}_\theta$. Thus, if $f, g \in \mathbf{A}_\sigma$, then their convolution $f * g$ can be defined as before, and it is routine to show that $f * g \in \mathbf{A}_\sigma$. More generally, if $f \in \mathbf{A}_\sigma$ and $g \in \mathbf{A}_\theta$, then their convolution $f * g$ belongs to the set $\mathbf{A}_{\min\{\sigma, \theta\}}$. As before, let $\hat{\mathbf{A}}_\sigma$ denote the set of Laplace transforms of distributions in $\mathbf{A}_\sigma$. Now define the sets $\mathbf{A}_-$ and $\hat{\mathbf{A}}_-$ as follows:

$$\mathbf{45} \quad \mathbf{A}_- = \bigcup_{\sigma > 0} \mathbf{A}_\sigma, \hat{\mathbf{A}}_- = \bigcup_{\sigma > 0} \hat{\mathbf{A}}_\sigma.$$

In other words, $\mathbf{A}_-$ consists of all distributions f which belong to $\mathbf{A}_\sigma$ for some $\sigma > 0$, and $\hat{\mathbf{A}}_-$ is the set of Laplace transforms of distributions in $\mathbf{A}_-$. Note that $\mathbf{A}_-$ is a proper subset of $\mathbf{A}$, but is closed under convolution.

Suppose $f \in \mathbf{A}_-$. Then, by definition, there exists a $\sigma > 0$ such that (44) holds. This means that the Laplace transform $\hat{f}$ is analytic over the shifted open half-plane $\{s : \text{Re}\, s > -\sigma\}$. As a consequence, all zeros of $\hat{f}$ in the open half-plane $\{s : \text{Re}\, s > -\sigma\}$ are isolated and of finite multiplicity. In particular, all zeros of $\hat{f}$ in the *closed RHP* $\{s : \text{Re}\, s \geq 0\}$ are isolated and of finite multiplicity. Thus if $\hat{f} \in \hat{\mathbf{A}}_-$, then $\hat{f}$ cannot have a sequence of zeros clustering at some point on the $j\omega$-axis. This is not true in general for an arbitrary function in $\mathbf{A}$, and this is one of the main motivations for introducing the set $\mathbf{A}_-$; see Vidyasagar et al. (1982) for an example of a function in $\hat{\mathbf{A}}$ which has a sequence of zeros clustering at a point on the $j\omega$-axis.

Now we introduce one last concept needed to define $\hat{B}$. Suppose $\hat{f} \in \hat{\mathbf{A}}_-$. Then we say that $\hat{f}$ is *bounded away from zero at infinity* if there exists a constant $r < \infty$ such that

$$\mathbf{46} \quad \inf_{s \in C_+, |s| \geq r} |\hat{f}(s)| > 0.$$

In effect, (46) just states that all zeros of $\hat{f}$ in C_+ lie in some compact subset thereof. Since each of these zeros is isolated, $\hat{f}$ has only finitely many zeros in C_+. It is easy to see that if $\hat{f}, \hat{g} \in \hat{\mathbf{A}}_-$ are each bounded away from zero at infinity, then so is their product $\hat{f}\hat{g}$.

47 Definition *The set $\hat{B}$ consists of all functions $\hat{f} = \hat{a}/\hat{b}$, where $\hat{a}, \hat{b} \in \hat{\mathbf{A}}_-$, and in addition $\hat{b}$ is bounded away from zero at infinity.*

48 Lemma *Suppose $\hat{f} \in \hat{B}$. Then (i) all singularities of $\hat{f}$ in C_+ are poles of finite order, and they are all isolated; (ii) all poles of $\hat{f}$ in C_+ lie in some compact subset thereof, so that $\hat{f}$ has only finitely many poles in C_+; (iii) all zeros of $\hat{f}$ in C_+ are isolated and of finite multiplicity.*

The proof is a ready consequence of earlier observations.

49 Examples Consider again the transfer function of Example (41), namely

$$\hat{g}(s) = e^{-s} \frac{s^2 + 4s + 2}{s^2 - 1}.$$

Recall that we had to work a bit to show that $\hat{g}$ is of the form (13). On the other hand,

showing that $\hat{g} \in \hat{B}$ is quite simple: $\hat{g} = \hat{a}/\hat{b}$, where

$$\hat{a}(s) = e^{-s}\frac{s^2+4s+2}{(s+1)^2}, \hat{b}(s) = \frac{s-1}{s+1}.$$

As another example, consider a uniform RC transmission line whose input is the voltage applied at one end and whose output is the current at the other end. In this case the transfer function is the so-called C parameter among the chain parameters, and equals

$$\hat{f}(s) = \frac{1}{\sqrt{s}\sinh\sqrt{s}}.$$

This transfer function has a pole at the origin, but otherwise all of its poles are on the negative real axis. To show that $\hat{f}$ belongs to $\hat{B}$, write it as $\hat{n}/\hat{d}$, where

$$\hat{n}(s) = \frac{\sqrt{s}}{(s+1)\sinh\sqrt{s}}, \hat{d}(s) = \frac{s}{s+1}.$$

Then $\hat{n} \in \hat{\mathbf{A}}_-$ (though this is perhaps not obvious), while $\hat{d}$ belongs to $\hat{\mathbf{A}}_-$ and is bounded away from zero at infinity.

An example of a transfer function which does not belong to $\hat{B}$ is provided by

$$\hat{h}(s) = \frac{1}{\cosh s},$$

which is the voltage transfer function of a uniform LC transmission line of unit (normalized) length. Since $\hat{h}$ has infinitely many poles on the $j\omega$-axis, it follows from Lemma (48) that h does not belong to $\hat{B}$.

50 Lemma *Suppose $\hat{f}, \hat{g} \in \hat{B}$; then $f \pm g, \hat{f}\hat{g} \in \hat{B}$.*

Proof Write $\hat{f} = \hat{a}/\hat{b}$, $\hat{g} = \hat{c}/\hat{d}$, where $\hat{a}, \hat{b}, \hat{c}, \hat{d} \in \hat{\mathbf{A}}_-$, and in addition $\hat{b}, \hat{d}$ are bounded away from zero at infinity. Then

$$\mathbf{51} \quad \hat{f} \pm \hat{g} = \frac{\hat{a}\hat{d} \pm \hat{b}\hat{c}}{\hat{b}\hat{d}}, \hat{f}\hat{g} = \frac{\hat{a}\hat{c}}{\hat{b}\hat{d}}.$$

In each case, both the numerator and the denominator belong to $\hat{\mathbf{A}}_-$; in addition, since both $\hat{b}$ and $\hat{d}$ are bounded away from zero at infinity, so is $\hat{b}\hat{d}$. ■

The set of *all* ratios $\hat{a}/\hat{b}$ where $\hat{a}, \hat{b} \in \hat{\mathbf{A}}$ and $b \neq 0$ is called the **field of fractions** of $\hat{\mathbf{A}}$. Lemma (50) states that $\hat{B}$ is a *subring* of this field of fractions.

Next, several useful properties of the set $\hat{B}$ are proved. As the proofs are somewhat technical, they may be skipped in the first reading, and the reader can jump ahead to Theorem (67), which is the main result.

52 Lemma *Suppose* $\hat{f} \in \hat{\mathbf{A}}$ *and that* $\hat{f}(\sigma) = 0$ *for some* $\sigma > 0$*; then the function* $s \mapsto \hat{f}(s)/(s-\sigma)$ *is the Laplace transform of a function in* L_1*. Suppose* $\hat{f} \in \hat{\mathbf{A}}$ *and that* $\hat{f}(\sigma + j\omega) = 0$ *for some* $\sigma > 0$ *and some* $\omega \neq 0$*; then the function* $s \mapsto \hat{f}(s)/[(s-\sigma)^2 + \omega^2]$ *is the Laplace transform of a function in* L_1*.*

Proof Express $f(t)$ in the form

$$\mathbf{53} \quad f(t) = \sum_{i=0}^{\infty} f_i \, \delta(t - t_i) + f_a(t) =: f_d(t) + f_a(t),$$

where f_d is the distributional part of f and f_a is the measurable part of f. Suppose $\hat{f}(\sigma) = 0$. Then

$$\mathbf{54} \quad \frac{\hat{f}(s)}{s-\sigma} = \frac{\hat{f}(s) - \hat{f}(\sigma)}{s-\sigma} = \frac{\hat{f}_d(s) - \hat{f}_d(\sigma)}{s-\sigma} + \frac{\hat{f}_a(s) - \hat{f}_a(\sigma)}{s-\sigma}.$$

It is shown that each of the two functions on the right side of (54) is the Laplace transform of a function in L_1. For convenience, let $\hat{L}_1$ denote the set of Laplace transforms of functions in L_1, and observe that $\hat{L}_1 \subseteq \hat{\mathbf{A}}$. First, using the same reasoning as in Example (41), one can show that

$$\mathbf{55} \quad f_i \frac{e^{-st_i} - e^{-\sigma t_i}}{s-\sigma} \in \hat{L}_1, \ \forall i.$$

Note that the inverse Laplace transform of the function in (55) is just

$$\mathbf{56} \quad f_i[-e^{-\sigma t_i} e^{\sigma t} + e^{\sigma(t-t_i)} U(t-t_i)] = 0, \ \forall t \geq t_i,$$

where $U(\cdot)$ denotes the unit step function. Hence the function in (56) has compact support and therefore belongs to L_1, which shows that (55) is true. Moreover, the norm of this function is given by

$$\mathbf{57} \quad \left\| f_i \frac{e^{-st_i} - e^{-\sigma t_i}}{s-\sigma} \right\|_{\hat{\mathbf{A}}} = |f_i| \int_0^{t_i} e^{\sigma(t-t_i)} \, dt = |f_i| \, (1 - e^{-\sigma t_i})/\sigma \leq |f_i|/\sigma, \ \forall t_i.$$

Now consider the function

$$\mathbf{58} \quad \frac{\hat{f}_d(s) - \hat{f}_d(\sigma)}{s-\sigma} = \sum_{i=0}^{\infty} f_i \frac{e^{-st_i} - e^{-\sigma t_i}}{s-\sigma}.$$

From (57) it follows that the right side of (58) is absolutely convergent, i.e.,

$$\textbf{59} \quad \sum_{i=0}^{\infty} \| f_i \frac{e^{-st_i} - e^{-\sigma t_i}}{s-\sigma} \|_{\hat{\mathbf{A}}} \leq \sum_{i=0}^{\infty} |f_i| / \sigma < \infty.$$

Since $\hat{\mathbf{A}}$ is a Banach space, this shows that the summation on the right side of (57) is well-defined and belongs to $\hat{\mathbf{A}}$.

Now consider the second term on the right side of (54). Suppose first that $\hat{f}_a(s)$ is a rational function of s. Then so is the term in question; moreover, it is strictly proper and has no pole at $s = \sigma$. Hence it belongs to $\hat{L}_1$. Now, every element in L_1 can be expressed as a limit of a sum of decaying exponentials (Kammler 1976); equivalently, every function in $\hat{L}_1$ can be expressed as a limit in $\hat{\mathbf{A}}$ of a sequence of stable rational functions. This shows that

$$\textbf{60} \quad \frac{\hat{f}_a(s) - \hat{f}_a(\sigma)}{s-\sigma} \in \hat{L}_1,$$

and the proof is complete. The case of complex zeros follows similarly. ■

61 Corollary *Suppose $\hat{f} \in \hat{\mathbf{A}}$ and that $\hat{f}$ has a zero of multiplicity m at a real $\sigma > 0$; then the function $s \mapsto \hat{f}(s)/(s-\sigma)^m$ belongs to $\hat{L}_1$. Suppose $\hat{f} \in \hat{\mathbf{A}}$ and that $\hat{f}$ has a zero of multiplicity m at a point $\sigma + j\omega$ where $\sigma > 0$ and $\omega \neq 0$. Then the function $s \mapsto \hat{f}(s)/[(s-\sigma)^2 + \omega^2]^m$ belongs to $\hat{L}_1$.*

Proof Apply Lemma (52) repeatedly. ■

62 Corollary *Let $\hat{f} \in \hat{\mathbf{A}}$, and suppose that $\hat{f}(\sigma) = 0$ for some $\sigma > 0$. Then the function $s \mapsto (s+1)\hat{f}(s)/(s-\sigma)$ belongs to $\hat{\mathbf{A}}$. Suppose $\hat{f}(\sigma + j\omega) = 0$ for some $\sigma > 0$ and some $\omega \neq 0$. Then the function $s \mapsto (s+1)^2 \hat{f}(s)/[(s-\sigma)^2 + \omega^2]$ belongs to $\hat{\mathbf{A}}$.*

Remarks Note the contrast between Lemma (52) Corollary (62). In the former, it is shown that $\hat{f}(s)/(s-\sigma)$ belongs to $\hat{L}_1$, whereas here it is claimed that $(s+1)\hat{f}(s)/(s-\sigma)$ belongs to $\hat{\mathbf{A}}$. The difference arises because $1/(s-\sigma)$ is strictly proper, as a result of which the function $\hat{f}(s)/(s-\sigma)$ does not have an impulsive part; in contrast, $(s+1)/(s-\sigma)$ is not strictly proper, and as a result $(s+1)\hat{f}(s)/(s-\sigma)$ could contain an impulsive part.

Proof Suppose $\hat{f}(\sigma) = 0$. Then

$$\textbf{63} \quad \frac{s+1}{s-\sigma}\hat{f}(s) = \hat{f}(s) + \frac{1+\sigma}{s-\sigma}\hat{f}(s) \in \hat{\mathbf{A}}.$$

The case of complex zeros follows similarly. ■

Of course there is nothing special about the term $s+1$, and one can replace it by any other first order polynomial $\alpha s + \beta$.

64 Corollary *Let $\hat{f} \in \hat{\mathbf{A}}$, and suppose $\hat{f}$ has a zero of multiplicity m at $\sigma > 0$; then the function $s \mapsto (s+1)^m \hat{f}(s)/(s-\sigma)^m$ belongs to $\hat{\mathbf{A}}$. Suppose $\hat{f}$ has a zero of multiplicity m at a point $\sigma + j\omega$ where $\sigma > 0$ and $\omega \neq 0$. Then the function $s \mapsto (s+1)^{2m} \hat{f}(s)/[(s-\sigma)^2 + \omega^2]^m$ belongs*

to $\hat{\mathbf{A}}$.

Proof Apply Corollaries (61) and (62) repeatedly. ■

65 Lemma *Suppose* $\hat{f} \in \hat{\mathbf{A}}_-$ *and is bounded away from zero at infinity. Then there exists a unit* $\hat{u}$ *of* $\hat{\mathbf{A}}_-$ *and a proper stable rational function* $\hat{v}$ *such that* $\hat{f} = \hat{u}\hat{v}$.

Proof Since $\hat{f} \in \hat{\mathbf{A}}_-$, there is some $\sigma > 0$ such that $f \in \mathbf{A}_\sigma$. Let $z_1, \cdots, z_k$ denote the distinct C_+-zeros of $\hat{f}$, with multiplicities $m_1, \cdots, m_k$ respectively. Define

$$\textbf{66} \quad \hat{v}(s) = \prod_{i=1}^{k} \left[\frac{s - z_i}{s+1} \right]^{m_i}.$$

Then, from Lemma (52) and Corollary (64), it follows that $\hat{u}(s) := \hat{f}(s)/\hat{v}(s)$ belongs to $\hat{\mathbf{A}}_\theta$ for all $\theta < \sigma$. (To prove this, one uses the fact that $\operatorname{Re} z_i > -\sigma$ even if possibly some of the z_i have zero real parts.) Now $\hat{u}$ is bounded away from zero at infinity, since both $\hat{f}$ and $\hat{v}^{-1}$ have this property. Moreover, $\hat{u}(s)$ has no finite zeros in the half-plane $\{s : \operatorname{Re} s \geq -\theta\}$. Hence, by a slight modification of Lemma (1), $\hat{u}$ is a unit of $\hat{\mathbf{A}}_\theta$ and hence of $\hat{\mathbf{A}}_-$. ■

67 Theorem *Suppose* $\hat{a}, \hat{b} \in \hat{\mathbf{A}}_-$, *and suppose that* $\hat{b}$ *is bounded away from zero at infinity. Then the ratio* $\hat{g} = \hat{a}/\hat{b}$ *is of the form (13). Specifically, if* $p_1, \cdots, p_k$ *are the distinct* C_+*-zeros of* $\hat{b}$, *of multiplicities* $m_1, \cdots, m_k$ *respectively, then*

$$\textbf{68} \quad \hat{g}(s) = \frac{\hat{a}(s)}{\hat{b}(s)} = \hat{g}_a(s) + \hat{g}_r(s),$$

where

$$\textbf{69} \quad \hat{g}_r(s) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{r_{ij}}{(s - p_i)^j}, \text{ and } \hat{g}_a \in \hat{\mathbf{A}}_-,$$

and the constants r_{ij} *are evaluated as*

$$\textbf{70} \quad r_{i,m_i - j} = \frac{1}{j!} \frac{d^j}{ds^j} [(s - p_i)^{m_i} \hat{g}(s)]_{s = p_i}, \; j = 0, \cdots, m_i - 1.$$

Remarks Theorem (67) states two things: (i) If $\hat{b} \in \hat{\mathbf{A}}_-$ and is bounded away from zero at infinity, then for all practical purposes $\hat{b}$ is like a proper stable rational function. To see why, suppose $\hat{g} = \hat{a}/\hat{b}$ and express $\hat{b}$ as $\hat{u}\hat{v}$ where $\hat{v}$ is a proper stable rational function and $\hat{u}$ is a unit of $\hat{\mathbf{A}}_-$. Then the function $\hat{c} = \hat{a}/\hat{u}$ belongs to $\hat{\mathbf{A}}_-$, and moreover

$$\textbf{71} \quad \hat{g} = \frac{\hat{a}}{\hat{b}} = \frac{\hat{a}}{\hat{u}\hat{v}} = \frac{\hat{c}}{\hat{v}},$$

where the denominator $\hat{v}$ now has a very simple form. (ii) Suppose we start with (71) and carry out a partial fraction expansion as in (69). Then the part $\hat{g}_a$ which is "left over" belongs

to $\hat{\mathbf{A}}_-$. This was the calculation we had to do in Example (41), and Theorem (67) makes it precise. Note that the formula (70) is the familiar expression for the coefficients in a partial fraction expansion.

Proof With all the preliminary work, the proof of the theorem is actually quite easy. First, write $\hat{g}$ in the form (71), where

$$\mathbf{72} \quad \hat{v}(s) = \prod_{i=1}^{k} \left[\frac{s - p_i}{s + 1} \right]^{m_i},$$

and expand $1/\hat{v}(s)$ as a partial fraction sum

$$\mathbf{73} \quad \frac{1}{\hat{v}(s)} = \prod_{i=1}^{k} \left[\frac{s+1}{s-p_i} \right]^{m_i} = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{q_{ij}}{(s-p_i)^j}.$$

To prove that $\hat{g} = \hat{c}/\hat{v}$ is of the form (13), it is enough to show that each term $\hat{c}(s)/(s-p_i)^j$ is of the form (13), since a finite sum of functions of the form (13) is again of this form. (Note that both $\hat{\mathbf{A}}$ and the set of strictly proper rational functions are closed under addition.) Accordingly, consider the ratio $\hat{c}(s)/(s-p_i)^j$. Since $\hat{c} \in \hat{\mathbf{A}}_-$, there exists a $\sigma > 0$ such that $\hat{c} \in \hat{\mathbf{A}}_\sigma$, and of course $\operatorname{Re} p_i \geq 0 > -\sigma$. Now write

$$\mathbf{74} \quad \frac{\hat{c}(s)}{(s-p_i)^j} = \frac{\hat{c}(s) - \sum_{l=0}^{j-1} (s-p_i)^l \hat{c}^{(l)}(p_i)/l!}{(s-p_i)^j} + \frac{\sum_{l=0}^{j-1} (s-p_i)^l \hat{c}^{(l)}(p_i)/l!}{(s-p_i)^j}.$$

The second term on the right side is a strictly proper rational function. As for the first term, its numerator has a zero of multiplicity j at $s = p_i$. Hence, by Corollary (61), the first term belongs to $\hat{\mathbf{A}}_\sigma$ and hence to $\hat{\mathbf{A}}_-$. This shows that the overall function $\hat{c}(s)/\hat{v}(s)$ is of the form (13). Now the formula (70) for the constants r_{ij} follows in the usual fashion. ■

This subsection is concluded with an obvious result.

75 Lemma *Suppose $\hat{F} \in \hat{B}^{l \times m}$. Then $\hat{F}$ can be decomposed as*

$$\mathbf{76} \quad \hat{F}(s) = \hat{F}_a(s) + \hat{F}_r(s),$$

where $\hat{F}_a \in \hat{\mathbf{A}}_-^{l \times m}$ and $\hat{F}_r(s)$ is an $l \times m$ matrix whose elements are strictly proper rational functions of s.

In summary, the set $\hat{B}$ is very useful for at least two reasons: First, every element of $\hat{B}$ has the form (13); as a consequence, the stability of feedback systems whose forward-path element belongs to $\hat{B}$ and which have a constant gain in the feedback path, can be easily determined using a comprehensive and physically meaningful graphical stability test [Theorem (35)]. Second, determining whether or not a given function belongs to $\hat{B}$ is easier than determining whether or not a given function has the form (13); compare Examples (41)

and (49). At this point, it is natural to ask: Is every function of the form (13) a member of the set $\hat{B}$? The answer is no, as is easily shown. Recall that in (13) the function $\hat{g}_a$ belongs to the set $\hat{\mathbf{A}}$, whereas, as shown in Theorem (67), the function $\hat{g}_a$ belongs to $\hat{\mathbf{A}}_-$. Thus, if one chooses a function $\hat{g}_a$ which belongs to $\hat{\mathbf{A}}$ but not to $\hat{\mathbf{A}}_-$ [for example, a function which has a sequence of zeros clustering on the $j\omega$-axis; see the example in Vidyasagar et al. (1982)], then this function is of the form (13) but does not belong to $\hat{B}$.

6.5.3 Coprime Factorizations

In the final subsection of this section, the notion of coprime factorizations is introduced, and it is shown how one may analyze the feedback stability of distributed systems using this notion.

77 Definition *Two elements $a, b \in \mathbf{A}$ are said to be* **coprime** *if there exist elements $x, y \in \mathbf{A}$ such that*

78 $$x * a + y * b = \delta(t),$$

or equivalently

79 $$\hat{x}(s)\,\hat{a}(s) + \hat{y}(s)\,\hat{b}(s) = 1, \quad \forall s \in C_+.$$

In this case we also say that $\hat{a}, \hat{b} \in \hat{\mathbf{A}}$ are coprime.

80 Example Let

$$\hat{a}(s) = 2e^{-s}, \quad \hat{b}(s) = 1 + e^{-2s}.$$

Then $\hat{a}, \hat{b}$ are coprime, since (79) is satisfied with

$$\hat{x}(s) = -0.5\,e^{-s}, \quad \hat{y}(s) = 1. \blacksquare$$

In order to show that two elements $a, b \in \mathbf{A}$ are coprime using Definition (77), it is necessary to display explicitly a solution pair (x, y) satisfying the relation (78), which is often referred to as the *Bezout identity*. The next result gives a criterion for coprimeness that is much easier to verify.

81 Lemma *Two elements $a, b \in \mathbf{A}$ are coprime if and only if*

82 $$\inf_{\mathrm{Re}\, s \geq 0} \max\{|\hat{a}(s)|, |\hat{b}(s)|\} > 0.$$

Proof (Partial) "Only if" Suppose a, b are coprime, and select $x, y \in \mathbf{A}$ such that (78) is satisfied. Then, since $\hat{x}, \hat{y} \in \hat{\mathbf{A}}$, the quantities $\hat{x}(s), \hat{y}(s)$ are bounded as s varies over the closed right half-plane C_+. Define

83 $$\mu = \sup_{\mathrm{Re}\, s \geq 0} [|\hat{x}(s)| + |\hat{y}(s)|].$$

Now it is easy to show that, for each fixed $s \in C_+$, we have

84 $$1 = |\hat{x}(s)\hat{a}(s) + \hat{y}(s)\hat{b}(s)| \leq [|\hat{x}(s)| + |\hat{y}(s)|] \max\{|\hat{a}(s)|, |\hat{b}(s)|\}.$$

Hence (83) and (84) imply that

85 $$\inf_{\mathrm{Re}\, s \geq 0} \max\{|\hat{a}(s)|, |\hat{b}(s)|\} \geq \frac{1}{\mu} > 0.$$

"If" This part of the proof is given in Callier and Desoer (1978); see also Vidyasagar (1985), p. 342. ■

The coprimeness condition given in Lemma (81) can be restated in another equivalent form.

86 Lemma *Two elements $a, b \in \mathbf{A}$ are coprime if and only if there does not exist a sequence $\{s_i\}$ in C_+ such that $\hat{a}(s_i) \to 0$, $\hat{b}(s_i) \to 0$ as $i \to \infty$.*

Proof It is shown that the "no common zeroing sequence" condition given above is equivalent to (82). First, suppose that (82) is true; then obviously no sequence $\{s_i\}$ in C_+ can be found such that $|\hat{a}(s_i)|$, $|\hat{b}(s_i)|$ both approach zero. Conversely, suppose (82) is false, and select a sequence $\{s_i\}$ in C_+ such that

87 $$\max\{|\hat{a}(s_i)|, |\hat{b}(s_i)|\} \to 0 \text{ as } i \to \infty.$$

Then it is immediate that $|\hat{a}(s_i)|, |\hat{b}(s_i)| \to 0$ as $i \to \infty$. ■

The condition for coprimeness given in Lemma (86) says, in effect, that $\hat{a}$ and $\hat{b}$ have no common zeros in C_+. However, since the region C_+ is unbounded, a little care is needed. It is possible that $\hat{a}(s)$ and $\hat{b}(s)$ never both vanish at a common point $s \in C_+$, but nevertheless approach zero along some common (unbounded) sequence $\{s_i\}$ in C_+. Lemma (86) says that, in order for a and b to be coprime, this cannot happen either.

88 Definition *Suppose $\hat{p}(s) = \hat{a}(s)/\hat{b}(s)$, where $\hat{a}, \hat{b} \in \hat{\mathbf{A}}$. Then the ordered pair $(\hat{a}, \hat{b})$ is said to be a* **fractional representation of** $\hat{p}$. *We also say that the pair (a, b) is a fractional representation of p. If, in addition, $\hat{a}$ and $\hat{b}$ are coprime, then the ordered pair $(\hat{a}, \hat{b})$ is called a* **coprime factorization of** $\hat{p}$. *We also say that the ordered pair (a, b) is a coprime factorization of p.*

89 Example Consider the transfer function

$$\hat{h}(s) = \frac{1}{\cosh s}$$

which was first introduced in Example (49), representing the transfer function of an LC transmission line. As was shown in Example (49), this transfer function does not belong to

the set $\hat{B}$, so as of now we have no means of analyzing the stability of a system obtained by placing even a constant feedback gain around h (as in Figure 6.5). Now it is possible to rewrite $\hat{h}(s)$ as

$$\hat{h}(s) = \frac{2e^{-s}}{1+e^{-2s}} =: \frac{\hat{a}(s)}{\hat{b}(s)}.$$

As shown in Example (80), $\hat{a}$ and $\hat{b}$ are coprime; therefore the ordered pair $(2e^{-s}, 1+e^{-2s})$ is a coprime factorization of $\hat{h}$. As we shall see, this will enable us to analyze feedback systems involving $\hat{h}$.

90 Lemma *Every rational function has a coprime factorization.*

Proof Suppose $\hat{p}(s)$ is a rational function, and express $\hat{p}(s)$ as $\alpha(s)/\beta(s)$ where α and β are polynomials with no common zeros in C_+. Let k equal the larger of the degrees of α and β, and define

$$\textbf{91} \quad \hat{a}(s) = \frac{\alpha(s)}{(s+1)^k}, \hat{b}(s) = \frac{\beta(s)}{(s+1)^k}.$$

Then $\hat{a}, \hat{b} \in \hat{\mathbf{A}}$, and clearly $\hat{p}(s) = \hat{a}(s)/\hat{b}(s)$. Moreover, $\hat{a}$ and $\hat{b}$ are coprime, since they satisfy the criterion of Lemma (86). Hence $(\hat{a}, \hat{b})$ is a coprime factorization of $\hat{p}$. ■

If $\hat{p}$ is irrational, it need not have a coprime factorization. An example of such a function is constructed in Vidyasagar et al. (1982). The idea is to choose two irrational functions $\hat{a}, \hat{b} \in \hat{\mathbf{A}}$ such that each function has an infinite sequence of zeros in C_+, and such that the two zero sequences have a common accumulation point on the $j\omega$-axis. In this case, the ratio $\hat{p} = \hat{a}/\hat{b}$ does not have a coprime factorization. But the next result shows that a large number of irrational functions do have coprime factorizations, and brings out the importance of the set $\hat{B}$.

92 Lemma *Suppose $\hat{q}$ has a coprime factorization and that $\hat{r} \in \hat{\mathbf{A}}$; then $\hat{p} := \hat{q} + \hat{r}$ has a coprime factorization. In particular, every function in $\hat{B}$ has a coprime factorization.*

Remark In effect, this lemma shows that a function has a coprime factorization if its "unstable part" has a coprime factorization.

Proof Suppose $(\hat{a}, \hat{b})$ is a coprime factorization of $\hat{q}$, and choose $\hat{x}, \hat{y} \in \hat{\mathbf{A}}$ such that (79) holds. Then, since $\hat{q} = \hat{a}/\hat{b}$, elementary algebra shows that

$$\textbf{93} \quad \hat{p} = \frac{\hat{a}}{\hat{b}} + \hat{r} = \frac{\hat{a} + \hat{b}\hat{r}}{\hat{b}}.$$

Now $\hat{a} + \hat{r}\hat{b}$ and $\hat{b}$ belong to $\hat{\mathbf{A}}$; they are also coprime, since

94 $\hat{x}(\hat{a}+\hat{r}\hat{b})+(\hat{y}-\hat{x}\hat{r})\hat{b}=1, \forall s\in C_+.$

Hence the ordered pair $(\hat{a}+\hat{r}\hat{b}, \hat{b})$ is a coprime factorization of $\hat{p}$. The last sentence of the lemma follows because, from Theorem (67), every function in $\hat{B}$ is the sum of a rational function [which has a coprime factorization by Lemma (90)], and a function in $\hat{\mathbf{A}}$. ■

The extension of these ideas to MIMO systems is straight-forward. For this purpose, a little notation is introduced. Let the symbol $\hat{F}$ denote the set of all ratios $\hat{a}/\hat{b}$ where $\hat{a}, \hat{b}\in\hat{\mathbf{A}}$ and $\hat{b}\neq 0$, and in addition two ratios $\hat{a}/\hat{b}$, $\hat{c}/\hat{d}$ are deemed to be equal if $\hat{a}\hat{d}=\hat{b}\hat{c}$. Then $\hat{F}$ is called the **field of fractions** associated with $\hat{\mathbf{A}}$. One can think of $\hat{F}$ as the set of all transfer functions that have a fractional representation over $\hat{\mathbf{A}}$. Next, the symbol $M(S)$ is denotes the set of all matrices, of whatever order, whose components all belong to the set S. Typically, the set S will be one of $\mathbf{A}$, $\hat{\mathbf{A}}$, $\hat{F}$, $\hat{B}$, and so on. The reason for introducing the generic symbol M, which denotes "matrix," is that often the precise dimensions of the various matrices encountered below are immaterial to the main argument.

95 Definition *Suppose A, $B\in M(\mathbf{A})$ have the same number of columns; then A and B are said to be* **right-coprime** *if there exist X, $Y\in M(\mathbf{A})$ such that the identity*

96 $X*A+Y*B=I\delta(t)$

holds, or equivalently,

97 $\hat{X}(s)\hat{A}(s)+\hat{Y}(s)\hat{B}(s)=I, \forall s\in C_+.$

In this case we also say that $\hat{A}$, $\hat{B}\in M(\hat{\mathbf{A}})$ are right-coprime.

98 Definition *Suppose A, $B\in M(\mathbf{A})$ have the same number of rows; then A and B are said to be* **left-coprime** *if there exist X, $Y\in M(\mathbf{A})$ such that*

99 $A*X+B*Y=I\delta(t),$

or equivalently,

100 $\hat{A}(s)\hat{X}(s)+\hat{B}(s)\hat{Y}(s)=I, \forall s\in C_+.$

In this case we also say that $\hat{A}$, $\hat{B}\in M(\hat{\mathbf{A}})$ are left-coprime.

Thus the notion of coprimeness for scalar-valued functions introduced in Definition (77) has *two* distinct generalizations to the matrix case, namely left-coprimeness and right-coprimeness. This is not altogether surprising, since matrix multiplication is not commutative. Note that A and B are right-coprime if and only if A' and B' are left-coprime. With the aid of this observation, all of the results given below for right-coprimeness can be readily translated into results for left-coprimeness.

A necessary and sufficient condition for two scalar functions to be coprime is given in Lemma (81). Again, there are two generalizations of this result to the matrix case, one for right-coprimeness and another for left-coprimeness. Only the right-coprimeness result is stated here, and the reader can easily infer the corresponding result for left-coprimeness.

101 Lemma *Suppose $A, B \in M(\mathbf{A})$ have the same number of columns. Then A and B are right-coprime if and only if*

$$\mathbf{102} \quad \inf_{\mathrm{Re}\, s \geq 0} |\det [\hat{M}'(s)\hat{M}(s)]| > 0,$$

where

$$\mathbf{103} \quad \hat{M}(s) = \begin{bmatrix} \hat{A}(s) \\ \hat{B}(s) \end{bmatrix}.$$

Note that the condition (102) is a little stronger than: $\hat{M}(s)$ has full column rank for all $s \in C_+$. If $\hat{M}(s)$ has full column rank for all $s \in C_+$, then certainly $\det [\hat{M}'(s)\hat{M}(s)] > 0$ for all $s \in C_+$. But this is not all (102) says: it says something more, namely that the quantity is *bounded away from zero.*

104 Definition *Suppose $\hat{P} \in M(\hat{F})$. Then an ordered pair $(\hat{N}, \hat{D})$ is a* **right-coprime factorization (rcf)** *of $\hat{P}$ if (i) $\hat{P} = \hat{N}\hat{D}^{-1}$, and (ii) $\hat{N}$ and $\hat{D}$ are right-coprime. In this case we also say that (N, D) is a right-coprime factorization of P.*

The concept of a left-coprime factorization (lcf) is defined analogously.

The next result is a matrix analog of Lemma (90).

105 Lemma *If $\hat{P}(s)$ is a matrix of rational functions of s, then $\hat{P}$ has an rcf and an lcf.*

The proof of this result may be found in Vidyasagar (1978b).

The next result is a matrix analog of Lemma (92). Its proof is left as an exercise to the reader.

106 Lemma *If $\hat{P}$ has an rcf (respectively, an lcf), and if $\hat{Q} \in M(\hat{\mathbf{A}})$, then $\hat{P} + \hat{Q}$ has an rcf (respectively, an lcf). Every matrix in $M(B)$ has both an rcf and an lcf.*

Up to now we have had a barrage of definitions and lemmas. Finally we come to the *pièce de resistance* of this section, which is a set of necessary and sufficient conditions for the feedback stability of systems of the form shown in Figure 6.2, in the case where G_1 and G_2 are linear time-invariant systems. Let $\hat{G}_i$, $i = 1,2$ denote the transfer matrices of these two systems, and note that $\hat{G}_1$ and $\hat{G}_2$ have complementary dimensions. In other words, if $\hat{G}_1$ has dimensions $l \times m$, then $\hat{G}_2$ has dimensions $m \times l$. As a consequence, both $\hat{G}_1\hat{G}_2$ and $\hat{G}_2\hat{G}_1$ are square matrices, though possibly of different dimensions. A standard identity in matrix theory can now be used to show that, for each fixed s, we have that $\det [I + \hat{G}_1(s)\hat{G}_2(s)] = \det [I + \hat{G}_2(s)\hat{G}_1(s)]$. If this determinant is not identically zero as a

function of s, then we say that the feedback system is **well-posed.** In this case it is possible to solve the system equations (6.2.27) in the form

$$\textbf{107}\quad \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \end{bmatrix} = \hat{H} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix},$$

where

$$\textbf{108}\quad \hat{H} = \begin{bmatrix} (I+\hat{G}_2\hat{G}_1)^{-1} & -\hat{G}_2(I+\hat{G}_1\hat{G}_2)^{-1} \\ \hat{G}_1(I+\hat{G}_2\hat{G}_1)^{-1} & (I+\hat{G}_1\hat{G}_2)^{-1} \end{bmatrix}.$$

We say that the feedback system is *stable* if $\hat{H} \in M(\hat{\mathbf{A}})$. From Theorem (6.4.45), it follows that $\hat{H} \in M(\hat{\mathbf{A}})$ is a necessary and sufficient condition for various forms of stability. The next theorem can therefore be thought of as providing necessary and sufficient conditions for several types of stability at once.

109 Theorem *Suppose* $(\hat{N}_i, \hat{D}_i)$ *is an rcf of* $\hat{G}_i$, *and* $(\tilde{N}_i, \tilde{D}_i)$ *is an lcf of* $\hat{G}_i$, *for* $i = 1, 2$. *Then the following statements are equivalent.*

(i) $\hat{H} \in M(\hat{\mathbf{A}})$.

$$\textbf{110}\quad \text{(ii)} \inf_{\text{Re}\, s \geq 0} |\det \hat{\Delta}(s)| > 0,$$

where

$$\textbf{111}\quad \hat{\Delta} = \tilde{N}_1\hat{N}_2 + \tilde{D}_1\hat{D}_2.$$

$$\textbf{112}\quad \text{(iii)} \inf_{\text{Re}\, s \geq 0} |\det \tilde{\Delta}(s)| > 0,$$

where

$$\textbf{113}\quad \tilde{\Delta} = \tilde{N}_2\hat{N}_1 + \tilde{D}_2\hat{D}_1.$$

Remark Note that conditions (110) and (112) can be verified using graphical criteria analogous to Theorem (35) and Corollary (38). More on this later.

Proof Using the matrix identities

$$\textbf{114}\quad A(I+BA)^{-1} = (I+AB)^{-1}A,$$

$$\textbf{115}\quad (I+BA)^{-1} = I - B(I+AB)^{-1}A,$$

one can obtain two equivalent expressions for $\hat{H}$ from (108), one of which involves only $(I+\hat{G}_1\hat{G}_2)^{-1}$ and the other of which involves only $(I+\hat{G}_2\hat{G}_1)^{-1}$. They are

$$\textbf{116}\quad \hat{H}=\begin{bmatrix} I-\hat{G}_2(I+\hat{G}_1\hat{G}_2)^{-1}\hat{G}_1 & -\hat{G}_2(I+\hat{G}_1\hat{G}_2)^{-1} \\ (I+\hat{G}_1\hat{G}_2)^{-1}\hat{G}_1 & (I+\hat{G}_1\hat{G}_2)^{-1} \end{bmatrix},$$

$$\textbf{117}\quad \hat{H}=\begin{bmatrix} (I+\hat{G}_2\hat{G}_1)^{-1} & -(I+\hat{G}_2\hat{G}_1)^{-1}\hat{G}_2 \\ \hat{G}_1(I+\hat{G}_2\hat{G}_1)^{-1} & I-\hat{G}_1(I+\hat{G}_2\hat{G}_1)^{-1}\hat{G}_2 \end{bmatrix}.$$

It is now shown, using the expression (116), that $\hat{H}\in M(\hat{\mathbf{A}})$ if and only if (110) holds. The proof that $\hat{H}\in M(\hat{\mathbf{A}})$ if and only if (112) holds is entirely analogous, and proceeds from the expression (117) for $\hat{H}$.

Substitute

$$\textbf{118}\quad \hat{G}_1=\tilde{D}_1^{-1}\tilde{N}_1,\ \hat{G}_2=\hat{N}_2\hat{D}_2^{-1}$$

in (116) and clear fractions. This gives

$$\textbf{119}\quad \hat{H}=\begin{bmatrix} I-\hat{N}_2\hat{\Delta}^{-1}\tilde{N}_1 & -\hat{N}_2\hat{\Delta}^{-1}\tilde{D}_1 \\ \hat{D}_2\hat{\Delta}^{-1}\tilde{N}_1 & \hat{D}_2\hat{\Delta}^{-1}\tilde{D}_1 \end{bmatrix}.$$

First, suppose (110) holds. Then, by Lemma (3), it follows that $\hat{\Delta}^{-1}\in M(\hat{\mathbf{A}})$. Since all the matrices on the right side of (119) belong to $M(\hat{\mathbf{A}})$, it follows that $\hat{H}\in M(\hat{\mathbf{A}})$. To prove the converse, suppose $\hat{H}\in M(\hat{\mathbf{A}})$. Then since $I\in M(\hat{\mathbf{A}})$, it follows from (119) that

$$\textbf{120}\quad \begin{bmatrix} \hat{N}_2\hat{\Delta}^{-1}\tilde{N}_1 & \hat{N}_2\hat{\Delta}^{-1}\tilde{D}_1 \\ \hat{D}_2\hat{\Delta}^{-1}\tilde{N}_1 & \hat{D}_2\hat{\Delta}^{-1}\tilde{D}_1 \end{bmatrix}=\begin{bmatrix} \hat{N}_2 \\ \hat{D}_2 \end{bmatrix}\hat{\Delta}^{-1}[\tilde{N}_1\ \tilde{D}_1]=:\hat{M}\in M(\hat{\mathbf{A}}).$$

Now select matrices $\tilde{X}_1, \tilde{Y}_1, \hat{X}_2, \hat{Y}_2\in M(\hat{\mathbf{A}})$ such that

$$\textbf{121}\quad \tilde{N}_1\tilde{X}_1+\tilde{D}_1\tilde{Y}_1=I,\ \hat{X}_2\hat{N}_2+\hat{Y}_2\hat{D}_2=I.$$

Such matrices exist because $\tilde{N}_1, \tilde{D}_1$ are left-coprime, and $\hat{N}_2, \hat{D}_2$ are right-coprime. Now from (120) it follows that

$$\textbf{122}\quad \hat{\Delta}^{-1}=[\hat{X}_2\ \hat{Y}_2]\hat{M}\begin{bmatrix} \hat{X}_1 \\ \hat{Y}_1 \end{bmatrix}\in M(\hat{\mathbf{A}}).$$

Since $\hat{\Delta}\in M(\hat{\mathbf{A}})$, Lemma (3) now shows that (110) holds. ■

123 Example As an illustration of Theorem (109), consider a feedback system of the form shown in Figure 6.5, where

$$\hat{g}(s) = \frac{1}{\cosh s}$$

and k is a constant. As shown in Example (89), a coprime factorization of $\hat{g}$ is given by the ordered pair $(\hat{n}(s), \hat{d}(s)) := (2e^{-s}, 1+e^{-2s})$. Since k is just a constant, it has the obvious coprime factorization $(k, 1)$. Since the system is SISO, we need not bother about lcf's and rcf's, as both notions coincide. Hence the function $\hat{\Delta}$ of (111) is in this instance given by

$$\hat{\Delta}(s) = \hat{d}(s) + k\hat{n}(s) = 1 + e^{-2s} + ke^{-s}.$$

It is now shown that, for every $k \in \mathbf{R}$, there is an $s_0 \in C_+$ such that $\hat{\Delta}(s_0) = 0$. This means that (110) is violated, and hence the feedback system is unstable. Consider the equation

$$1 + e^{-2s} + ke^{-s} = 0.$$

This is a quadratic equation in $x := e^{-s}$. Let α, β denote the two roots of the equation

$$1 + x^2 + kx = 0.$$

Then clearly $\alpha\beta = 1$, which shows that both roots are nonzero and that at least one root has magnitude less than or equal to 1. Let α denote a root such that $0 < |\alpha| \leq 1$, and select $s_0 \in C_+$ such that $e^{-s_0} = \alpha$. Then $\hat{\Delta}(s_0) = 0$, which means that (110) does not hold. Thus we can conclude that the feedback system is unstable for all real k.

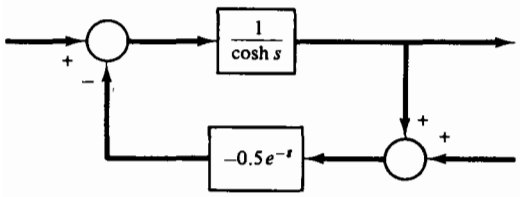


Fig. 6.11

Now suppose the feedback operator is changed from a constant k to a pure delay of 1 second and a gain of -0.5, as shown in Figure 6.11. In this case, since the feedback operator $-0.5\exp(-s)$ belongs to $\hat{\mathbf{A}}$, it has the coprime factorization $(-0.5\exp(-s), 1)$. Now the function $\hat{\Delta}$ becomes

$$\hat{\Delta}(s) = \hat{d}(s) + \hat{n}(s)\cdot(-0.5e^{-s}) = 1 + e^{-2s} - e^{-2s} = 1.$$

Hence (110) is satisfied, and we conclude that the system of Figure 6.11 is (BIBO) stable.

124 Example As another illustration of Theorem (109), consider again the system of Figure 6.5, with

$$\hat{g}(s) = \frac{1}{\tanh s}, k = 1.$$

Now $\hat{g}(s)$ has the coprime factorization $(\hat{n}, \hat{d})$ with

$$\hat{n}(s) = 1 + e^{-2s}, \hat{d}(s) = 1 - e^{-2s}.$$

It is easy to see that $\hat{g} = \hat{n}/\hat{d}$. Also, $\hat{n}$ and $\hat{d}$ are coprime since (79) is satisfied with $\hat{x} = \hat{y} = 1$. Hence

$$\hat{\Delta}(s) = \hat{d}(s) + k\hat{n}(s) = \hat{d}(s) + \hat{n}(s) = 1.$$

Hence (110) holds, and the feedback system is stable.

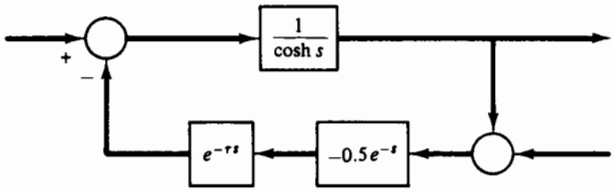


Fig. 6.12

125 Example In this example, we analyze whether the stable feedback systems of Examples (123) and (124) remain stable when a small delay is inserted into the loop. Consider first the system of Figure 6.12, which is the same as that in Figure 6.11 except for an additional delay in the feedback. It is now shown that, for every $\varepsilon > 0$, there is a $\tau < \varepsilon$ such that the system of Figure 6.12 is unstable. In other words, though the system of Figure 6.11 is stable, it can be destabilized by the insertion of an arbitrarily small delay in the loop.

The feedback operator of the system of Figure 6.12 is $-0.5 \exp[-(\tau + 1)s]$, which still belongs to $\hat{\mathbf{A}}$. Hence

$$\hat{\Delta}(s) = 1 + e^{-2s} + (2e^{-s}) \cdot (-0.5e^{-(\tau+1)s}) = 1 + e^{-2s} - e^{-(2+\tau)s}.$$

Choose $\tau = 2/m$ where m is an integer, and define $x = \exp(-2s/m)$. Then

$$\hat{\Delta} = 1 + x^m - x^{m+1}.$$

Now consider the polynomial equation

$$0 = 1 + x^m - x^{m+1}.$$

The product of all the roots of this equation is 1. This implies that all roots are nonzero, and that at least one root (call it α) has magnitude no larger than 1. Choose $s_0 \in C_+$ such that $\exp(-2s_0/m) = \alpha$. Then $\hat{\Delta}(s_0) = 0$, which shows that (110) is violated and that the feedback system is unstable whenever $\tau = 2/m$, m an integer. By choosing m sufficiently large, we can

make τ as small as we wish. This shows that the system in Figure 6.12 is unstable for arbitrarily small choices of τ.

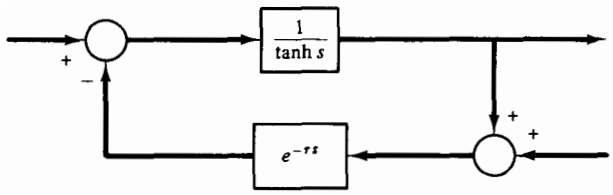


Fig. 6.13

Similarly, consider the feedback system of Figure 6.13. In this case

$$\hat{\Delta} = 1 - e^{-2s} + (1 + e^{-2s}) e^{-\tau s} = 1 + e^{-\tau s} - e^{-2s} + e^{-(2+\tau)s}.$$

As before, let $\tau = 2/m$, $x = \exp(-\tau s)$, and consider the polynomial equation

$$0 = 1 + x - x^m + x^{m+1}.$$

Earlier reasoning applies here, and one can conclude that there is at least one root α such that $0 < |\alpha| \leq 1$. If $s_0 \in C_+$ is chosen such that $\exp(-2s_0/m) = \alpha$, then $\hat{\Delta}(s_0) = 0$, which implies that (110) is violated and that the feedback system is unstable.

As a final remark, note that both $1/\cosh s$ and $1/\tanh s$ have infinitely many poles in C_+. It is difficult, if not impossible, to find stabilizing controllers for such systems which continue to maintain closed-loop stability even when a small delay is inserted into the loop. ■

This section is concluded with a graphical stability criterion for MIMO systems which is a generalization, to distributed systems, of a result in Rosenbrock (1970).

126 Theorem *Consider the system of Figure 6.2. Suppose G_1 and G_2 are linear and time-invariant, with transfer matrices $\hat{G}_1$ and $\hat{G}_2$ respectively. Suppose that each $\hat{G}_i$ is of the form*

127 $\hat{G}_i(s) = \hat{G}_{ia}(s) + \hat{G}_{ir}(s)$, $i = 1, 2$,

where $\hat{G}_{ia} \in M(\hat{\mathbf{A}})$ and $\hat{G}_{ir}$ is a matrix of strictly proper rational functions. Suppose the delays of both $\hat{G}_{1a}$ and $\hat{G}_{2a}$ are rationally related, i.e., suppose there exists a $T > 0$ such that

128 $G_{ia}(t) = \sum_{j=0}^{\infty} G_{ij}\, \delta(t - iT) + G_{im}(t)$, $G_{im} \in M(L_1)$.

Finally, let μ_i denote the McMillan degree of the unstable part of $\hat{G}_{ir}$. Then the transfer function $\hat{H}$ of (108) belongs to $M(\hat{\mathbf{A}})$ if and only if the following two conditions hold:

129 $(i)\ \inf_{\omega \in \mathbb{R}} |\det [I + \hat{G}_1(j\omega)\hat{G}_2(j\omega)]| > 0$, and

130 $(ii)\ \lim_{n \to \infty} \phi(j2\pi n/T) - \phi(-j2\pi n/T) = 2\pi(\mu_1 + \mu_2)$,

where

131 $\phi(j\omega) = \text{Arg det}\,[I + \hat{G}_1(j\omega)\hat{G}_2(j\omega)]$.

Remarks

1. Suppose $\hat{G}$ is a matrix of rational functions and that p is a pole of $\hat{g}$. Then the **McMillan degree** of p as a pole of $\hat{G}$ is the highest order it has as a pole of any minor of $\hat{G}$. The McMillan degree of the rational matrix $\hat{G}$ is the sum of the McMillan degrees of all of its poles. Thus μ_1 is the sum of the McMillan degrees of all the unstable poles of $\hat{G}_1$. Since the $j\omega$-axis is indented to go around the purely imaginary poles, only those poles with *positive* real parts should be counted in computing μ_1. Similar remarks apply to μ_2.
2. Note that the hypothesis of Theorem (126) requires that the all delays in both G_{1a} and G_{2a} must be rationally related.
3. If the system is SISO and $\hat{G}_2$ is just a constant k (so that $\mu_2 = 0$), then Theorem (126) reduces to Theorem (35).

Problem 6.15 Prove Lemma (1) for rational functions $\hat{f}$.

Problem 6.16 Using Theorem (35), determine the range of constant feedback gains k which can be placed around the following transfer function in order to produce a stable closed-loop system:

$$\hat{g}(s) = \left[e^{-s} + \frac{20}{s-10}\right]\left[e^{-s} - \frac{20}{s+50}\right]^2$$

Problem 6.17 Using Theorem (109), determine whether or not each of the following feedback systems is stable.

(a) $\hat{g}_1(s) = \dfrac{1}{\cosh s}, \hat{g}_2(s) = \dfrac{e^{-s}}{s+10}$.

(b) $\hat{g}_1(s) = \dfrac{1}{\tanh s}, \hat{g}_2(s) = \dfrac{s+1}{s+2}$.

6.6 TIME-VARYING AND/OR NONLINEAR SYSTEMS

The previous section was addressed to the stability of linear time-invariant feedback systems. In the present section we study time-varying and/or nonlinear systems. Two general methods are presented for the analysis of such systems. The first, known as the small gain approach, can be used to study L_p-stability for all values of $p \in [1, \infty]$, whereas the second method, known as the passivity approach, can be used to study L_2-stability. Using the relationships between input-output and Lyapunov stability derived in Section 6.3, both approaches can also be used to analyze the Lyapunov stability of nonlinear feedback systems. In particular, the small gain approach leads to the circle criterion while the passivity approach leads to the Popov criterion.

6.6.1 The Small Gain Approach

The starting point of the small gain approach is the following result:

1 Theorem *Consider the system in Figure 6.2, and suppose* $p \in [1, \infty]$ *is specified. Suppose in addition that both* G_1 *and* G_2 *are causal and* L_p*-stable wb, and let* $\gamma_{1p} = \gamma_p(G_1)$, $\gamma_{2p} = \gamma_p(G_2)$. *Under these conditions, the system of Figure 6.2 is* L_p*-stable if*

2 $\gamma_{1p}\gamma_{2p} < 1$.

Remarks

1. The inequality (2) is often called the **small gain condition,** and the name of the approach derives from this.
2. Theorem (1) can be interpreted as a perturbational result. Suppose we begin with two subsystems G_1 and G_2 which are stable in themselves, and then we interconnect them in the fashion shown in Figure 6.2. Then the resulting system is also stable provided the "loop gain" is less than one.
3. Theorem (1) is also valid with "wb" replaced throughout by "wfg." The proof of the amended version is left as an exercise (Problem 6.18).

Proof To streamline the proof, let us introduce some notation. If $\mathbf{a}, \mathbf{b} \in \mathbf{R}^n$, then the statement "$\mathbf{a} \leq \mathbf{b}$" is equivalent to "$a_i \leq b_i \; \forall i$," and to "$\mathbf{b} - \mathbf{a} \in \mathbf{R}_+^n$." Note that if $\mathbf{a}, \mathbf{b} \in \mathbf{R}^n$, $\mathbf{a} \leq \mathbf{b}$, and $\mathbf{A} \in \mathbf{R}_+^{n \times n}$, then $\mathbf{Aa} \leq A\mathbf{b}$. This is the matrix generalization of the fact that multiplying both sides of a scalar inequality by a nonnegative number preserves the inequality.

Suppose the system equations (6.2.24) are satisfied, so that

$$\mathbf{3} \qquad \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} - \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

$$\mathbf{4} \qquad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}.$$

It is shown that if $u_i \in L_p^{n_i}$ for $i = 1, 2$, then $e_1, y_2 \in L_p^{n_1}$, and $e_2, y_1 \in L_p^{n_2}$. Since G_i is causal and L_p-stable wb, it follows from (4) that

$$\mathbf{5} \qquad \begin{bmatrix} \|y_1\|_{Tp} \\ \|y_2\|_{Tp} \end{bmatrix} \leq \begin{bmatrix} \gamma_{1p} & 0 \\ 0 & \gamma_{2p} \end{bmatrix} \begin{bmatrix} \|e_1\|_{Tp} \\ \|e_2\|_{Tp} \end{bmatrix}, \forall T \geq 0.$$

Taking norms in (3) gives

$$\mathbf{6} \qquad \begin{bmatrix} \|e_1\|_{Tp} \\ \|e_2\|_{Tp} \end{bmatrix} \leq \begin{bmatrix} \|u_1\|_{Tp} \\ \|u_2\|_{Tp} \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \|y_1\|_{Tp} \\ \|y_2\|_{Tp} \end{bmatrix}, \forall T \geq 0.$$

Substituting from (5) into (6) gives

$$\mathbf{7} \qquad \begin{bmatrix} \|e_1\|_{Tp} \\ \|e_2\|_{Tp} \end{bmatrix} \leq \begin{bmatrix} \|u_1\|_{Tp} \\ \|u_2\|_{Tp} \end{bmatrix} + \begin{bmatrix} 0 & \gamma_{2p} \\ \gamma_{1p} & 0 \end{bmatrix} \begin{bmatrix} \|e_1\|_{Tp} \\ \|e_2\|_{Tp} \end{bmatrix}, \forall T \geq 0,$$

or

$$\mathbf{8} \qquad \begin{bmatrix} 1 & -\gamma_{2p} \\ -\gamma_{1p} & 1 \end{bmatrix} \begin{bmatrix} \|e_1\|_{Tp} \\ \|e_2\|_{Tp} \end{bmatrix} \leq \begin{bmatrix} \|u_1\|_{Tp} \\ \|u_2\|_{Tp} \end{bmatrix}, \forall T \geq 0.$$

Now examine the matrix

$$\mathbf{9} \qquad \mathbf{M} := \begin{bmatrix} 1 & -\gamma_{2p} \\ -\gamma_{1p} & 1 \end{bmatrix} \in \mathbf{R}^{2\times 2}.$$

If $\gamma_{1p}\gamma_{2p} < 1$, then $\mathbf{M}$ is nonsingular, and

$$\mathbf{10} \qquad \mathbf{M}^{-1} = \frac{1}{1-\gamma_{1p}\gamma_{2p}} \begin{bmatrix} 1 & \gamma_{2p} \\ \gamma_{1p} & 1 \end{bmatrix} \in \mathbf{R}_+^{2\times 2}.$$

Hence we can multiply both sides of (8) by $\mathbf{M}^{-1}$ and the inequality still holds. This results in

$$\mathbf{11} \qquad \begin{bmatrix} \|e_1\|_{Tp} \\ \|e_2\|_{Tp} \end{bmatrix} \leq \frac{1}{1-\gamma_{1p}\gamma_{2p}} \begin{bmatrix} 1 & \gamma_{2p} \\ \gamma_{1p} & 1 \end{bmatrix} \begin{bmatrix} \|u_1\|_{Tp} \\ \|u_2\|_{Tp} \end{bmatrix}, \forall T \geq 0.$$

If $u_i \in L_p^{n_i}$ for $i = 1, 2$, then

$$\textbf{12} \quad \begin{bmatrix} \|u_1\|_{Tp} \\ \|u_2\|_{Tp} \end{bmatrix} \leq \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix}, \forall T \geq 0.$$

Now (11) and (12) imply that

$$\textbf{13} \quad \begin{bmatrix} \|e_1\|_{Tp} \\ \|e_2\|_{Tp} \end{bmatrix} \leq \frac{1}{1-\gamma_{1p}\gamma_{2p}} \begin{bmatrix} 1 & \gamma_{2p} \\ \gamma_{1p} & 1 \end{bmatrix} \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix}, \forall T \geq 0.$$

However, since the right side of (13) is independent of T, it follows from Lemma (6.1.24) that $e_i \in L_p^{n_i}$ for $i = 1, 2$. Finally, combining (13) and (5) gives the bound

$$\textbf{14} \quad \begin{bmatrix} \|y_1\|_p \\ \|y_2\|_p \end{bmatrix} \leq \frac{1}{1-\gamma_{1p}\gamma_{2p}} \begin{bmatrix} \gamma_{1p} & \gamma_{1p}\gamma_{2p} \\ \gamma_{1p}\gamma_{2p} & \gamma_{2p} \end{bmatrix} \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix}, \forall T \geq 0.$$

This shows that $y_1 \in L_p^{n_2}$, $y_2 \in L_p^{n_1}$. The inequalities (13) and (14) show that the system is L_p-stable wb. ■

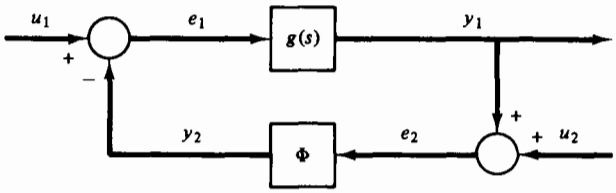


Fig. 6.14

As an application of Theorem (1), consider the SISO feedback system shown in Figure 6.14, where the forward path element is linear, time-invariant, and has the transfer function $\hat{g}(s)$; and the feedback element is a memoryless, possibly time-varying nonlinearity. Specifically, suppose $\phi: \mathbf{R}_+ \times \mathbf{R} \to \mathbf{R}$ is a given continuous function, and define a corresponding operator Φ on L_{1e} by

$$\textbf{15} \quad (\Phi x)(t) = \phi[t, x(t)], \forall t \geq 0.$$

We say that ϕ (or Φ) *belongs to the sector* $[a, b]$ if it is true that

$$\textbf{16} \quad \phi(t, 0) = 0, \text{ and } a \leq \frac{\phi(t, \sigma)}{\sigma} \leq b, \forall \sigma \neq 0, \forall t \geq 0,$$

or equivalently

$$\textbf{17} \quad a\sigma^2 \leq \sigma\phi(t, \sigma) \leq b\sigma^2, \forall \sigma \in \mathbf{R}, \forall t \geq 0.$$

Note that (17) is the scalar version of Definition (5.6.9). Now a direct application of Theorem (1) leads to a simple sufficient condition for the system of Figure 6.14.

18 Lemma *Consider the system of Figure 6.14, where* $\hat{g} \in \hat{\mathbf{A}}$, *and* ϕ *belongs to the sector* $[-r, r]$. *Then the system is* L_2*-stable wb provided*

19 $\sup_{\omega \in \mathbb{R}} |\hat{g}(j\omega)| < r^{-1}$.

Proof Apply Theorem (1) with $G_1 = \hat{g}$ (or, more precisely, let $G_1 : x \mapsto g * x$), let $G_2 = \Phi$, and let $p = 2$. Then both G_1 and G_2 are causal and L_2-stable; in addition,

20 $\gamma_2(G_1) = \sup_{\omega} |\hat{g}(j\omega)|, \gamma_2(G_2) \leq r,$

where the first inequality follows from Theorem (6.4.40). Now the condition (2) for L_2-stability becomes

21 $\sup_{\omega} |\hat{g}(j\omega)| \cdot r < 1,$

which is the same as (19). ■

Remarks In the proof of Lemma (18), the fact that Φ is a memoryless nonlinearity is not used; the only relevant property of Φ is that $\gamma_2(\Phi) \leq r^{-1}$.

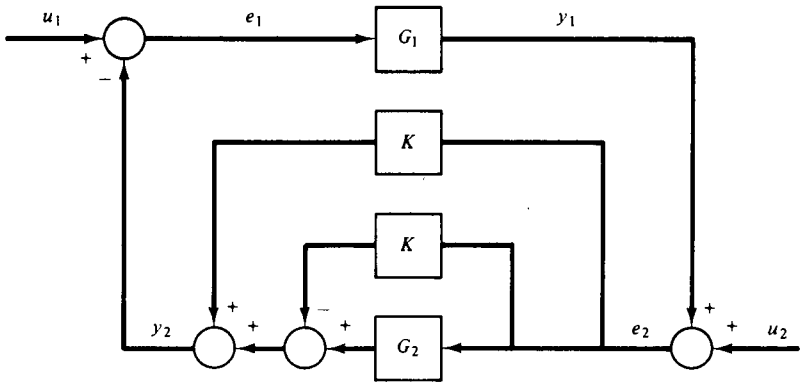


Fig. 6.15

By making a suitable transformation of the system in Figure 6.2, one can significantly expand the range of applicability of Theorem (1). The idea is to introduce an additional L_p-stable linear operator K which is first subtracted and then added to G_2, as shown in Figure 6.15. Then, through block diagram manipulations, the system is redrawn as shown in Figure 6.16, where K now acts as a feedback around G_1 and a feed-forward around G_2. Note that the first external input is changed from u_2 to $u_1 - Ku_2$ as a result of these manipulations. Now the system of Figure 6.16 can be interpreted as that in Figure 6.17. If this system is L_p-stable, then so is the original one in Figure 6.2 (and conversely; see Problem 6.19). These ideas are formalized in the next result.

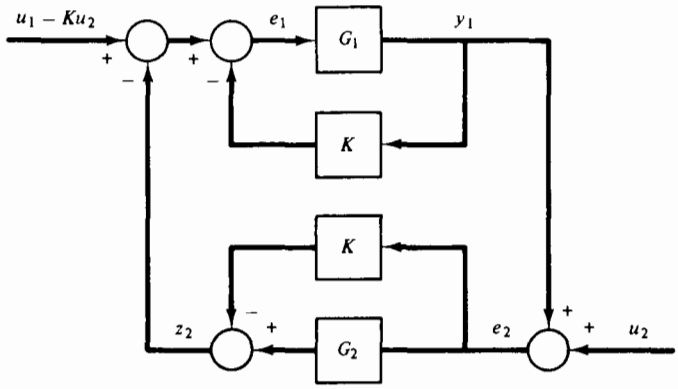


Fig. 6.16

22 Theorem (Loop Transformation) *Consider the system shown in Figure 6.2, and suppose* $p \in [1, \infty]$ *is specified. Suppose* G_2 *is causal and* L_p*-stable wb. Under these conditions, the system is* L_p*-stable wb if there exists a causal linear operator* K *which is* L_p*-stable wb such that (i)* $G_1(I + KG_1)^{-1}$ *is causal and* L_p*-stable wb, and (ii)*

23 $\gamma_p[G_1(I+KG_1)^{-1}]\gamma_p(G_2-K) < 1.$

Proof The system of Figure 6.2 is described by the familiar equations

24 $e_1 = u_1 - y_2, e_2 = u_2 + y_1, y_1 = G_1 e_1, y_2 = G_2 e_2.$

Now define a new output

25 $z_2 = y_2 - Ke_2 = (G_2 - K)e_2,$

and eliminate y_2 from (24). This gives

26 $e_1 = u_1 - y_2 = u_1 - z_2 - Ke_2 = u_1 - z_2 - K(u_2 + y_1).$

Using the fact that K is linear, (26) can be rewritten as

27 $e_1 = (u_1 - Ku_2) - z_2 - Ky_1.$

The other equations in (24) now become

28 $e_2 = u_2 + y_1, y_1 = G_1 e_1, z_2 = (G_2 - K)e_2.$

Clearly (27) and (28) describe the system of Figure 6.16. Now this system can be rearranged as in Figure 6.17, with

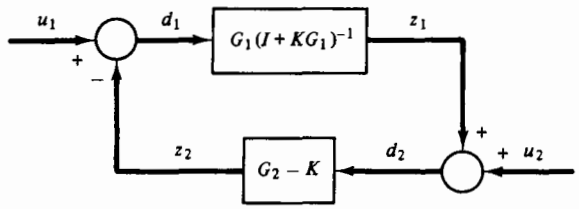


Fig. 6.17

29 $v_1 = u_1 - Ku_2,\ v_2 = u_2,\ z_1 = y_1,\ z_2 = y_2 - Ke_2,\ d_1 = e_1 + Ky_1,\ d_2 = e_2.$

By applying Theorem (1) to the system of Figure 6.17, one can conclude that this system is L_p-stable wb if (i) $G_1(I+KG_1)^{-1}$ and $G_2 - K$ are causal and L_p-stable wb, and (ii) (23) holds. To complete the proof, it only remains to show that the *original* system (24) is also L_p-stable wb.

For this purpose, introduce the notation

30 $\gamma_a = \gamma_p[G_1(I+KG_1)^{-1}],\ \gamma_b = \gamma_p(G_2 - K),$

and note that $\gamma_a\gamma_b < 1$ from (23). Now, from (29), we have

31 $$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} I & -K \\ 0 & I \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

Hence, whenever $(u_1, u_2) \in L_p^n$, it follows that $(v_1, v_2) \in L_p^n$, and moreover,

32 $$\begin{bmatrix} \|v_1\|_p \\ \|v_2\|_p \end{bmatrix} \leq \begin{bmatrix} 1 & k_p \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix},$$

where k_p is the induced norm of the linear operator K viewed as a map from $L_p^{n_1}$ into $L_p^{n_2}$. Now the fact that $\gamma_a\gamma_b < 1$ implies that the system of Figure 6.17 is L_p-stable wb. Hence, whenever $(u_1, u_2) \in L_p^n$, [which in turn implies that $(v_1, v_2) \in L_p^n$], it follows that $(d_1, d_2) \in L_p^n$, $(z_1, z_2) \in L_p^n$. Moreover, in analogy with (13) and (14) one obtains the bounds

33 $$\begin{bmatrix} \|d_1\|_p \\ \|d_2\|_p \end{bmatrix} \leq \frac{1}{1-\gamma_a\gamma_b} \begin{bmatrix} 1 & \gamma_b \\ \gamma_a & 1 \end{bmatrix} \begin{bmatrix} \|v_1\|_p \\ \|v_2\|_p \end{bmatrix},$$

$$\begin{bmatrix} \|z_1\|_p \\ \|z_2\|_p \end{bmatrix} \leq \frac{1}{1-\gamma_a\gamma_b} \begin{bmatrix} \gamma_a & \gamma_a\gamma_b \\ \gamma_a\gamma_b & \gamma_a \end{bmatrix} \begin{bmatrix} \|v_1\|_p \\ \|v_2\|_p \end{bmatrix}.$$

Substituting from (32) into (33) gives

$$\textbf{34}\quad \begin{bmatrix} \|d_1\|_p \\ \|d_2\|_p \end{bmatrix} \le \frac{1}{1-\gamma_a\gamma_b} \begin{bmatrix} 1 & \gamma_b + k_p \\ \gamma_a & \gamma_a k_p + 1 \end{bmatrix} \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix},$$

$$\begin{bmatrix} \|z_1\|_p \\ \|z_2\|_p \end{bmatrix} \le \frac{1}{1-\gamma_a\gamma_b} \begin{bmatrix} \gamma_a & \gamma_a(\gamma_b + k_p) \\ \gamma_a\gamma_b & \gamma_b(\gamma_a k_p + 1) \end{bmatrix} \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix}.$$

To obtain bounds on the L_p-norms of e_1, e_2, y_1, y_2, notice that the system equations (3) and (29) imply that

$$\textbf{35}\quad y_1 = z_1,\ e_2 = d_2,\ y_2 = z_2 + Ke_2,\ e_1 = d_1 + Ky_1.$$

Hence

$$\textbf{36}\quad \|y_1\|_p = \|z_1\|_p,\ \|e_2\|_p = \|d_2\|_p,$$

$$\|y_2\|_p \le \|z_2\|_p + k_p\|e_2\|_p,\ \|e_1\|_p \le \|d_1\|_p + k_p\|y_1\|_p.$$

Substituting from (32), (33) and (34) into (36), and performing a certain amount of routine algebra, yields

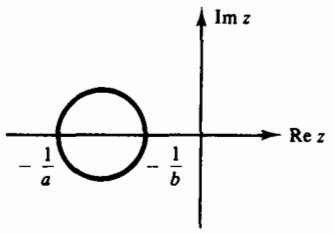
$$\textbf{37}\quad \begin{bmatrix} \|e_1\|_p \\ \|e_2\|_p \end{bmatrix} \le \frac{1}{1-\gamma_a\gamma_b} \begin{bmatrix} \gamma_a k_p + 1 & (\gamma_a k_p + 1)(\gamma_b + k_p) \\ \gamma_a & \gamma_a k_p + 1 \end{bmatrix} \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix},$$

$$\textbf{38}\quad \begin{bmatrix} \|y_1\|_p \\ \|y_2\|_p \end{bmatrix} \le \frac{1}{1-\gamma_a\gamma_b} \begin{bmatrix} \gamma_a & \gamma_a(\gamma_b + k_p) \\ \gamma_a(\gamma_b + k_p) & (\gamma_a k_p + 1)(\gamma_b + k_p) \end{bmatrix} \begin{bmatrix} \|u_1\|_p \\ \|u_2\|_p \end{bmatrix}.$$

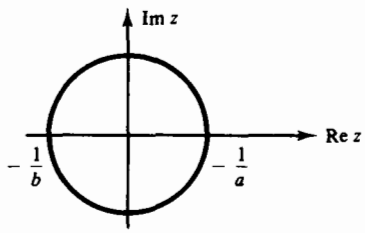
Hence the system of (3) is L_p-stable wb. ■

In the original version of the small gain theorem, the two subsystems G_1 and G_2 are not distinguished, in the sense that both the stability condition (2) and the bounds (13), (14) remain the same if the indices 1 and 2 are interchanged throughout. However, this symmetry is no longer present in Theorem (22), for the obvious reason that now the two subsystems are treated differently: One has a feedback placed around it while the other has a feedforward placed around it.

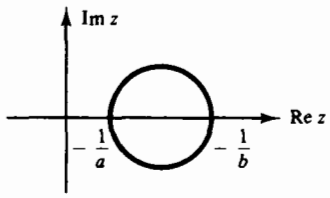
Combining Theorem (22) with the Nyquist stability criterion [Theorem (6.5.35)] leads to a very widely applicable result known as the circle criterion. One bit of notation is introduced to facilitate the statement of the theorem. Suppose a and b are nonzero real numbers with $a < b$. Then the symbol $D(a, b)$ denotes the disk in the complex plane which is centered on the real axis and whose circumference passes through the two points $-1/a$ and $-1/b$ (see Figure 6.18). Equivalently,



(a) $D(a, b)$ when $0 < a < b$.



(b) $D(a, b)$ when $a < 0 < b$.



(c) $D(a, b)$ when $a < b < 0$.

Fig. 6.18

39 $$D(a, b) = \left\{ z \in C : \left| z + \frac{b+a}{2ba} \right| \leq \left| \frac{b-a}{2ba} \right| \right\}.$$

40 Theorem (Circle Criterion) *Consider the system of Figure 6.14, where the nonlinearity Φ is memoryless and belongs to the sector $[a, b]$, and the transfer function $\hat{g}$ has the form*

41 $$\hat{g}(s) = \hat{g}_a(s) + \hat{g}_r(s),$$

where $\hat{g}_r$ is rational and strictly proper, $\hat{g}_a \in \hat{\mathbf{A}}$, and there exists a $T > 0$ such that

42 $$g_a(t) = \sum_{i=0}^{\infty} g_i \delta(t - iT) + g_m(t), \; g_m \in L_1.$$

Under these conditions, the feedback system is L_2-stable wb if one of the following conditions, as appropriate, holds:

Case (1). $ab > 0$: (i) With the $j\omega$-axis indented around the purely imaginary poles of $\hat{g}$ as in Section 6.5.1, the plot of $\hat{g}(j\omega)$ is bounded away from the disk $D(a, b)$, i.e.,

43 $$\inf_{\omega \in \mathbb{R},\, z \in D(a,\, b)} |\hat{g}(j\omega) - z| > 0.$$

(ii) Let μ_+ denote the number of poles of $\hat{g}$ with positive real part; then

44 $$\lim_{n \to \infty} \text{Arg}\,[\hat{g}(j2\pi n/T) - z] - \text{Arg}\,[\hat{g}(-j2\pi n/T) - z] = 2\pi\mu_+, \quad \forall z \in D(a, b).$$

Case (2). $0 = a < b$: (i) $\hat{g} \in \hat{\mathbf{A}}$, and (ii)

45 $$\inf_{\omega \in \mathbb{R}} \text{Re}\,\hat{g}(j\omega) > -\frac{1}{b}.$$

Case (3). $a < 0 < b$: (i) $\hat{g} \in \hat{\mathbf{A}}$, and (ii) the plot of $\hat{g}(j\omega)$ is contained in the disk $D(a, b)$, and is bounded away from the circumference of the disk.

Remarks

1. Note that, in (42), all the delays in the impulsive part of g_a are commensurate.
2. If g_a in (42) has no delayed impulses, then (44) is equivalent to the simpler condition: The plot of $\hat{g}(j\omega)$ encircles the disk $D(a, b)$ in the counterclockwise direction exactly μ_+ times as ω increases from $-\infty$ to ∞.
3. Recall that Theorem (6.3.46) relates the L_2-stability of a system to the global attractivity of the unforced equilibrium; now Theorem (40) gives a sufficient condition for L_2-stability. By combining the two theorems, one can recover the circle criterion of Lyapunov stability [Theorem (5.6.37)]. However, the input-output version of the circle criterion is more general than the Lyapunov stability version, since the former applies even to distributed systems, delay systems, etc.
4. If $b - a \to 0$ and a, b both approach a constant $k \neq 0$, then Theorem (40) reduces to the sufficiency part of the graphical Nyquist criterion [Theorem (6.5.35)]. It is shown in Theorem (126) later in this section that, in a certain sense, the circle criterion also gives a necessary condition.
5. As stated, Theorem (40) requires the feedback element Φ to be a *memoryless* nonlinearity. However, this fact is not used in the proof. It is shown in Theorem (126) that the circle criterion guarantees L_2-stability even if Φ is a *dynamic* nonlinear map belonging to the sector $[a, b]$.

Proof Define

46 $$r = \frac{b-a}{2}, \quad k = \frac{b+a}{2}.$$

Then $b = k + r$, $a = k - r$, and the map $\sigma \mapsto \phi(t, \sigma) - k\sigma$ belongs to the sector $[-r, r]$. Now apply Theorem (22) (the loop transformation theorem) with $K = kI$, $p = 2$, and combine with Lemma (18). Since the map $\Phi - K$ belongs to the sector $[-r, r]$, its L_2-gain is at most r. Hence the system is L_2-stable wb provided (i) $\hat{g}/(1 + k\hat{g}) \in \hat{\mathbf{A}}$, and (ii)

$$\textbf{47} \quad \sup_{\omega} \left| \frac{\hat{g}(j\omega)}{1+k\hat{g}(j\omega)} \right| \cdot r < 1.$$

Now it is shown that the hypotheses of the theorem enable us to deduce that the above two conditions hold.

Consider first case (1). Since the point $-1/k$ belongs to the disk $D(a, b)$, (43) and (44) show that the hypotheses of the Nyquist criterion [Theorem (6.5.35)] are satisfied. Hence $\hat{g}/(1+k\hat{g}) \in \hat{\mathbf{A}}$. To establish (47), we again make use of (43). From (39), only elementary algebra is needed to show that

$$\textbf{48} \quad \left| \frac{z}{1+kz} \right| < \frac{1}{r} \text{ iff } z \notin D(a, b).$$

Since the plot of $\hat{g}(j\omega)$ is bounded away from the disk $D(a, b)$, (47) follows, and the system is L_2-stable wb.

Next, consider Case (2). In this case $k = b/2$, and $-1/k = -2/b$. The bound (45) states that the Nyquist plot of $\hat{g}(j\omega)$ is confined to the half-plane $\{z : \text{Re}\, z > -1/b\}$, and of course $-1/k < -1/b$. Hence

$$\textbf{49} \quad \text{Re}\, \hat{g}(j\omega) + 1/k > 0, \forall \omega,$$

and as a consequence,

$$\textbf{50} \quad \text{Arg}\, [\hat{g}(j\omega) + 1/k] \in (-\pi/2, \pi/2), \forall \omega.$$

In particular,

$$\textbf{51} \quad \text{Arg}\, [\hat{g}(j 2\pi n/T) + 1/k] - \text{Arg}\, [\hat{g}(-j 2\pi n/T) + 1/k] \in (-\pi, \pi), \forall \omega.$$

Consider the limit of the quantity in (51) as $n \to \infty$. This limit must be an integer multiple of 2π. Now (51) implies that the limit must therefore be zero. Since $\hat{g} \in \hat{\mathbf{A}}$, it has no poles with positive real parts. Hence by Theorem (6.5.35) it follows that $\hat{g}/(1+k\hat{g}) \in \hat{\mathbf{A}}$. Finally, it is routine to show that

$$\textbf{52} \quad \left| \frac{z}{1+(b/2)z} \right| < \frac{2}{b} \text{ iff } \text{Re}\, z > -\frac{1}{b}.$$

Therefore (45) implies (47), and L_2-stability wb now follows from Theorem (22).

Finally consider Case (3). If $k = 0$, then $a = -r$, $b = r$, and the L_2-stability wb of the system follows from Lemma (18), so it can be assumed that $k \neq 0$. The new feature is that $ab < 0$, so that some inequalities get reversed when both sides are multiplied by ab. As a consequence we get

53 $\left| \dfrac{z}{1+kz} \right| \leq \dfrac{1}{r}$ iff $z \in D(a, b)$.

Compare (53) and (48). Now (53) shows that the point $-1/k$ lies outside $D(a, b)$. Indeed, if $k > 0$ then $-1/k < -1/b$, or else if $k < 0$ then $-1/k > -1/a$. In other words, the disk $D(a, b)$ lies entirely to one side of the point $-1/k$. If $k > 0$, then the fact that $\hat{g}(j\omega) \in D(a, b)\ \forall\omega$ implies that

54 $\mathrm{Re}\,\hat{g}(j\omega) + 1/k > 0, \forall\omega.$

As in the proof for Case (2), (54) implies that $\hat{g}/(1+k\hat{g}) \in \hat{\mathbf{A}}$. Similarly, if $k < 0$, then

55 $\mathrm{Re}\,\hat{g}(j\omega) + 1/k < 0, \forall\omega,$

and once again $\hat{g}/(1+k\hat{g}) \in \hat{\mathbf{A}}$. Now the hypotheses in Case (3) show that (47) is also satisfied. Hence the L_2-stability of the system now follows from Theorem (22). ■

56 Example Consider the system of Figure 6.14, with

$$\hat{g}(s) = \left[e^{-0.1s} + \frac{2}{s-1} \right] \cdot \left[1 + e^{-0.1s} - \frac{2}{s+4} \right].$$

Then $\hat{g}$ is a product of two functions, each of which belongs to $\hat{\mathbf{B}}$; hence $\hat{g} \in \hat{\mathbf{B}}$. By Theorem (6.5.67), it follows that $\hat{g}$ is of the form (41), and it is evident that the delays in the impulsive part of $g(\cdot)$ are commensurate. Hence Theorem (40) applies to $\hat{g}$.

Now $\hat{g}$ is unstable, and $\mu_+ = 1$. So Cases (2) and (3) of Theorem (40) are inapplicable, and only Case (1) may possibly apply. The Nyquist plot of $\hat{g}(j\omega)$ is shown in Figure 6.19. From the figure one can see that if

$$-1.5 < -\frac{1}{a} < -\frac{1}{b} < -1.05, \text{ i.e., } 0.667 < a < b < 0.95,$$

then the hypotheses of Case (1) are satisfied. Hence the system of Figure 6.14 is L_2-stable whenever Φ belongs to the sector $[a, b]$ with $[a, b]$ a subset of $(0.667, 0.95)$. Another way of saying the same thing is that the feedback system is L_2-stable whenever Φ belongs to the sector $[0.667 + \varepsilon, 0.95 - \varepsilon]$ for some $\varepsilon > 0$.

57 Example Consider the system of Figure 6.14, with

$$\hat{g}(s) = e^{-0.1s} \left[\frac{s+2}{s+1} + e^{-0.2s} \frac{s+4}{s+2} \right].$$

Then $\hat{g} \in \hat{\mathbf{A}}$ and thus falls within the scope of Theorem (40).

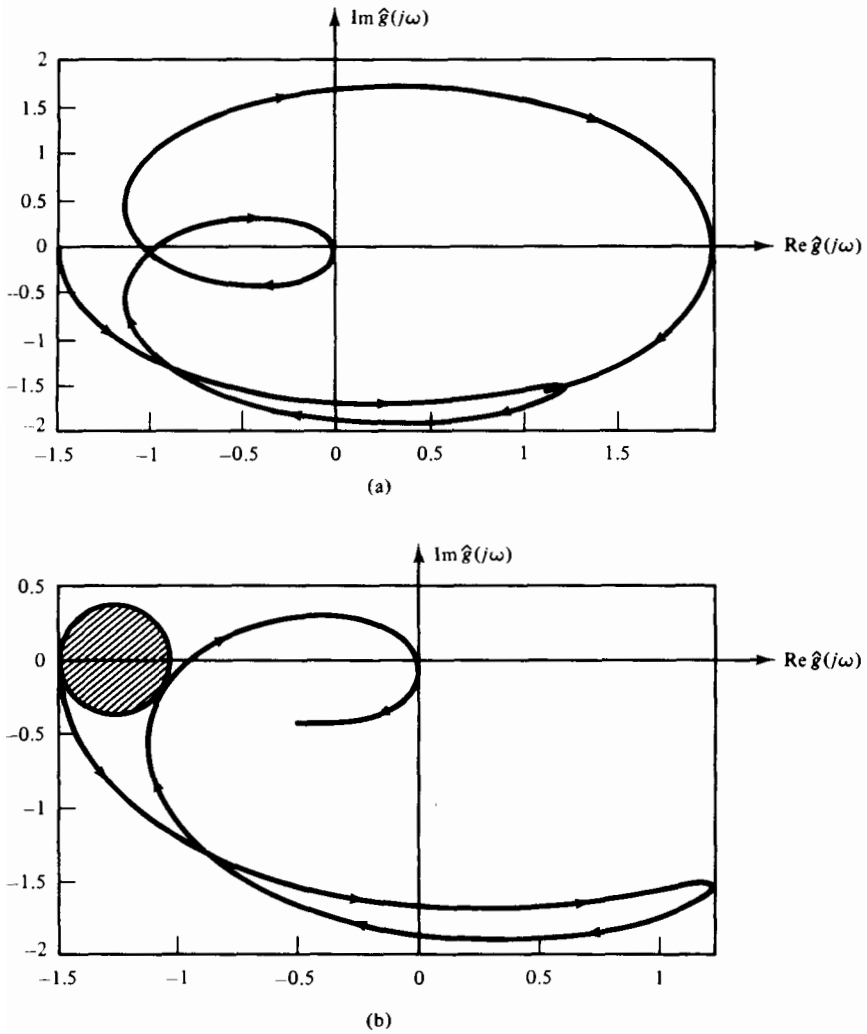


Fig. 6.19

The Nyquist plot of $\hat{g}$ is shown in Figure 6.20, from which one can see that

$$\gamma_2(G) = \sup_{\omega} |\hat{g}(j\omega)| = 4.$$

Suppose we apply the small gain theorem directly without bothering with loop transformations. Then Theorem (1) tells us that the system under study is L_2-stable whenever

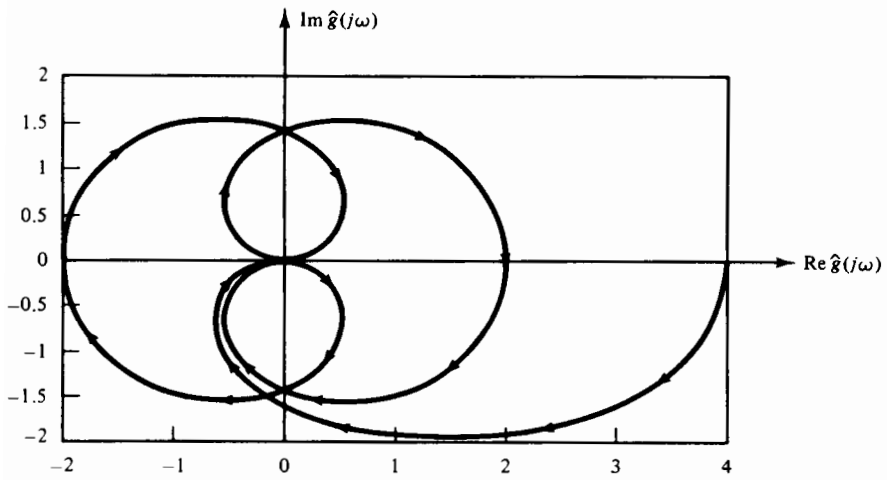


Fig. 6.20

$$\gamma_2(\Phi) < \frac{1}{\gamma_2(G)} = 0.25.$$

In particular, whenever Φ is a memoryless nonlinear element in the sector $[-0.25+\varepsilon, 0.25-\varepsilon]$, the system is L_2-stable.

Now let us apply Case (3) of Theorem (40). From Figure 6.20 one can see that the Nyquist plot of $\hat{g}(j\omega)$ is always in the interior of the disk $D(a, b)$ provided

$$-\frac{1}{b} < -2, \; 4 < -\frac{1}{a}, \text{ i.e., } -0.25 < a < b < 0.5.$$

Thus the feedback system is L_2-stable for all memoryless nonlinearities in the sector $[-0.25+\varepsilon, 0.5-\varepsilon]$ for some $\varepsilon > 0$. By using Theorem (40) instead of Theorem (1), we are able to extend the upper limit from $0.25-\varepsilon$ to $0.5-\varepsilon$.

Next, let us apply Case (1) of Theorem (40). Since $\mu_+ = 0$ in this case, it is desired that the Nyquist plot of $\hat{g}(j\omega)$ neither intersect nor encircle the disk $D(a, b)$. From Figure 6.20 one can see that these conditions are satisfied provided the Nyquist plot lies entirely to one side of the disk $D(a, b)$. This happens provided either

$$-\frac{1}{a} < -\frac{1}{b} < -2, \text{ i.e., } 0 < a < b < 0.5,$$

or

$$4 < -\frac{1}{a} < -\frac{1}{b}, \text{ i.e., } -0.25 < a < b < 0.$$

Combined with the fact that $\hat{g} \in \hat{\mathbf{A}}$, this once again shows that the feedback system is L_2-

stable whenever Φ belongs to the sector $[-0.25+\varepsilon, 0.5-\varepsilon]$. Hence there is no advantage to applying Case (1) in this instance.

6.6.2 The Passivity Approach

In this subsection, an alternative approach to L_2-stability is presented, known as the passivity approach. The Popov criterion is among the useful stability criteria that can be obtained using this approach. In contrast with the small gain approach which can be used to analyze L_p-stability for all values of $p \in [1, \infty]$, the passivity approach is naturally geared to analyzing L_2-stability; it is, however, possible to analyze L_∞-stability using passivity methods, but this is not discussed in this book.

The next result, though not the most general of its kind, is adequate for the present purposes; it is taken from Vidyasagar (1977). A still more general result, based on the so-called dissipativity approach, can be found in Moylan and Hill (1978).

58 Theorem *Consider the feedback system of Figure 6.2. Suppose there exist constants* ε_i, δ_i, $i = 1, 2$, *such that*

$$\mathbf{59} \quad \langle x, G_i x \rangle_T \geq \varepsilon_i \| x \|_{T2}^2 + \delta_i \| G_i x \|_{T2}^2, \ \forall T \geq 0, \ \forall x \in L_{2e}, \ i = 1, 2.$$

Then the system is L_2*-stable wb if*

$$\mathbf{60} \quad \delta_1 + \varepsilon_2 > 0, \ \delta_2 + \varepsilon_1 > 0.$$

Proof The system under study is described by the equations

$$\mathbf{61} \quad e_1 = u_1 - y_2, \ e_2 = u_2 + y_1, \ y_1 = G_1 e_1, \ y_2 = G_2 e_2.$$

As a consequence, it readily follows that

$$\mathbf{62} \quad \langle y_1, e_1 \rangle_T + \langle y_2, e_2 \rangle_T = \langle y_1, u_1 \rangle_T + \langle y_2, u_2 \rangle_T, \ \forall T \geq 0.$$

Now, from (59), it follows that

$$\mathbf{63} \quad \langle y_i, e_i \rangle_T \geq \varepsilon_i \| e_i \|_{T2}^2 + \delta_i \| y_i \|_{T2}^2, \ i = 1, 2.$$

Hence

$$\mathbf{64} \quad \langle y_1, e_1 \rangle_T + \langle y_2, e_2 \rangle_T \geq \varepsilon_1 \| e_1 \|_{T2}^2 + \delta_1 \| y_1 \|_{T2}^2 + \varepsilon_2 \| e_2 \|_{T2}^2 + \delta_2 \| y_2 \|_{T2}^2.$$

Now note that $\| e_i \|_{T2}^2 = \langle e_i, e_i \rangle_T$, and substitute for e_1, e_2 from (61). This gives, after routine computations,

$$\mathbf{65} \quad \varepsilon_1 \| e_1 \|_{T2}^2 + \varepsilon_2 \| e_2 \|_{T2}^2 = \varepsilon_1 [\| y_2 \|_{T2}^2 - 2\langle y_2, u_1 \rangle_T + \| u_1 \|_{T2}^2]$$

$$+\varepsilon_2[\|y_1\|_{T2}^2+2\langle y_1, u_2\rangle_T+\|u_2\|_{T2}^2].$$

Combining (62), (64), and (65), and rearranging gives an implicit inequality, namely

66 $$(\delta_1+\varepsilon_2)\|y_1\|_{T2}^2+(\delta_2+\varepsilon_1)\|y_2\|_{T2}^2$$
$$\leq \langle y_1, (u_1-2\varepsilon_2 u_2)\rangle_T+\langle y_2, (u_2+2\varepsilon_1 u_1)\rangle_T-\varepsilon_1\|u_1\|_{T2}^2-\varepsilon_2\|u_2\|_{T2}^2.$$

Using Schwarz' inequality and the triangle inequality on the right side of (66) gives

67 $$[\|y_1\|_{T2} \quad \|y_2\|_{T2}]\begin{bmatrix}\delta_1+\varepsilon_2 & 0\\ 0 & \delta_2+\varepsilon_1\end{bmatrix}\begin{bmatrix}\|y_1\|_{T2}\\ \|y_2\|_{T2}\end{bmatrix}$$
$$\leq[\|y_1\|_{T2} \quad \|y_2\|_{T2}]\begin{bmatrix}1 & 2|\varepsilon_2|\\ 2|\varepsilon_1| & 1\end{bmatrix}\begin{bmatrix}\|u_1\|_{T2}\\ \|u_2\|_{T2}\end{bmatrix}$$
$$+[\|u_1\|_{T2} \quad \|u_2\|_{T2}]\begin{bmatrix}|\varepsilon_1| & 0\\ 0 & |\varepsilon_2|\end{bmatrix}\begin{bmatrix}\|u_1\|_{T2}\\ \|u_2\|_{T2}\end{bmatrix}.$$

This vector inequality is of the form

68 $$\mathbf{x}'\mathbf{A}\mathbf{x}\leq\mathbf{x}'\mathbf{B}\mathbf{z}+\mathbf{z}'\mathbf{C}\mathbf{z},$$

where

69 $$\mathbf{x}=\begin{bmatrix}\|y_1\|_{T2}\\ \|y_2\|_{T2}\end{bmatrix},\ \mathbf{z}=\begin{bmatrix}\|u_1\|_{T2}\\ \|u_2\|_{T2}\end{bmatrix},$$

and the definitions the matrices **A**, **B**, **C** are obvious. Now **A** is positive definite, and hence has a symmetric square root **S**. By "completing the square," (68) can be rewritten as

70 $$[\mathbf{S}\mathbf{x}-\frac{1}{2}\mathbf{S}^{-1}\mathbf{B}\mathbf{z}]'\,[\mathbf{S}\mathbf{x}-\frac{1}{2}\mathbf{S}^{-1}\mathbf{B}\mathbf{z}]\leq\mathbf{z}'\,(\frac{1}{4}\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}+\mathbf{C})\,\mathbf{z},$$

and therefore

71 $$\|\mathbf{S}\mathbf{x}-\frac{1}{2}\mathbf{S}^{-1}\mathbf{B}\mathbf{z}\|\leq\|\mathbf{M}\mathbf{z}\|,$$

where **M** is the symmetric square root of the matrix $\mathbf{C}+(\mathbf{B}'\mathbf{A}\mathbf{B})/4$ and $\|\cdot\|$ denotes the Euclidean norm. Finally,

72 $$\|\mathbf{S}\mathbf{x}\|\leq[\|\mathbf{M}\|_i+(1/2)\|\mathbf{S}^{-1}\mathbf{B}\|_i]\,\|\mathbf{z}\|,$$

where $\|\cdot\|_i$ denotes the matrix norm induced by the Euclidean norm. Since **S** is nonsingular, one can deduce from (72) that

73 $\|\mathbf{x}\| \le (1/\|\mathbf{S}^{-1}\|_i)[\|\mathbf{M}\|_i + (1/2)\|\mathbf{S}^{-1}\mathbf{B}\|_i]\|\mathbf{z}\|.$

This shows that the relation between u and y is L_2-stable wb. By Lemma (6.2.33), this shows that the system as a whole is L_2-stable wb. ■

Remark Theorem (58) is stated for SISO systems, but it is easy to see that this fact is not really used in the proof. Thus the theorem is applicable to MIMO systems as well, but an essential restriction is that both G_1 and G_2 are "square," i.e., have an equal number of inputs and outputs; otherwise quantities such as $\langle y_1, e_1 \rangle_T$ do not make sense. On the other hand, the so-called dissipativity approach does not have any such restrictions.

Several useful results can now be obtained as corollaries of Theorem (58). To state them in their original historical form, two terms are introduced.

74 **Definition** *An operator* $G: L_{2e} \to L_{2e}$ *is said to be* **passive** *if*

75 $\langle x, Gx \rangle_T \ge 0, \forall T \ge 0, \forall x \in L_{2e},$

and is **strictly passive** *if there exists a constant* $\varepsilon > 0$ *such that*

76 $\langle x, Gx \rangle_T \ge \varepsilon \|x\|_{T2}^2, \forall T \ge 0, \forall x \in L_{2e}.$

77 **Corollary** *The feedback system of Figure 6.2 is* L_2*-stable wb if both* G_1 *and* G_2 *are strictly passive.*

Proof In this case (59) holds with $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, and $\delta_1 = \delta_2 = 0$. Hence (60) is satisfied and the result follows from Theorem (58). ■

78 **Corollary** *The feedback system of Figure 6.2 is* L_2*-stable wb if either (i)* G_1 *is strictly passive and has finite gain, and* G_2 *is passive, or (ii)* G_2 *is strictly passive and has finite gain, and* G_1 *is passive.*

Proof Suppose (i) is true. Select constants $\varepsilon > 0$ and $\gamma < \infty$ such that

79 $\langle x, G_1 x \rangle_T \ge \varepsilon \|x\|_{T2}^2, \forall T \ge 0, \forall x \in L_{2e},$

80 $\|G_1 x\|_{T2} \le \gamma \|x\|_{T2}, \forall T \ge 0, \forall x \in L_{2e},$

and observe that

81 $\langle x, G_2 x \rangle_T \ge 0, \forall T \ge 0, \forall x \in L_{2e}.$

Thus G_2 satisfies (59) with $\varepsilon_2 = \delta_2 = 0$. Now (80) implies that

82 $\|x\|_{T2} \ge \dfrac{1}{\gamma}\|G_1 x\|_{T2}, \forall T \ge 0, \forall x \in L_{2e}.$

Pick any $\alpha \in (0, \varepsilon)$, and note from (79) and (82) that

83 $$<x, G_1x>_T \geq (\varepsilon-\alpha)\|x\|_{T2}^2 + \alpha\|x\|_{T2}^2$$

$$\geq (\varepsilon-\alpha)\|x\|_{T2}^2 + \frac{\alpha}{\gamma^2}\|G_1x\|_{T2}^2, \forall T \geq 0, \forall x \in L_{2e}.$$

Hence G_1 satisfies (59) with

84 $\varepsilon_1 = \varepsilon - \alpha, \delta_1 = \alpha/\gamma^2.$

Since both ε_1 and δ_1 are positive, (60) is satisfied and the result follows from Theorem (60). If (ii) holds, simply interchange the indices 1 and 2 throughout. ■

85 Corollary *Consider the feedback system of Figure 6.2, and suppose there exist real constants ε and δ and a positive finite constant γ such that*

86 $<x, G_1x>_T \geq \varepsilon\|x\|_{T2}^2, \forall T \geq 0, \forall x \in L_{2e},$

87 $\|G_1x\|_{T2} \leq \gamma\|x\|_{T2}, \forall T \geq 0, \forall x \in L_{2e},$

88 $<x, G_2x>_T \geq \delta\|G_2x\|_{T2}^2, \forall T \geq 0, \forall x \in L_{2e}.$

Under these conditions, the system is L_2-stable wb if

89 $\varepsilon + \delta > 0.$

Proof As in the proof of Corollary (78), (86) and (87) together imply (83). Hence G_1 satisfies (59) with ε_1 and δ_1 defined in (84). Now (88) states that G_2 satisfies (59) with $\varepsilon_2 = 0, \delta_2 = \delta$. Hence, for sufficiently small α, we have

90 $$\delta_1 + \varepsilon_2 = \frac{\alpha}{\gamma^2} > 0, \delta_2 + \varepsilon_1 = \varepsilon + \delta - \alpha > 0.$$

Hence (60) is satisfied and the result follows from Theorem (58). ■

The well-known Popov criterion can now be obtained as an application of Corollary (85). A preliminary result, which is of independent interest, is stated first.

91 Lemma *Suppose $\hat{g} \in \hat{\mathbf{A}}$, and define $G: L_{2e} \to L_{2e}$ by $Gx = g * x$. Define*

92 $$\varepsilon = \inf_{\omega \in \mathbb{R}} \operatorname{Re} \hat{g}(j\omega).$$

Then (79) holds (with G_1 replaced by G).

Proof Since G is causal, it follows that $(Gx)_T = (Gx_T)_T$, for all $T \geq 0$ and for all $x \in L_{2e}$. Also,

$$\mathbf{93}\quad <x, Gx>_T = \int_0^T x(t)\,(Gx)(t)\,dt = \int_0^\infty x_T(t)\,(Gx)_T(t)\,dt = \int_0^\infty x_T(t)\,(Gx_T)(t)\,dt.$$

Now $x_T \in L_2$ whenever $x \in L_{2e}$, and therefore x_T has a Fourier transform; denote it by $\hat{x}_T(j\omega)$. Also, since $\hat{g} \in \hat{\mathbf{A}}$, the function Gx_T belongs to L_2, and its Fourier transform is $\hat{g}(j\omega)\,x_T(j\omega)$. By Parseval's theorem,

$$\mathbf{94}\quad <x, Gx>_T = \frac{1}{2\pi}\,\mathrm{Re}\int_{-\infty}^{\infty} [\hat{x}_T(j\omega)]^*\,\hat{g}(j\omega)\,\hat{x}_T(j\omega)\,d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{Re}\,\hat{g}(j\omega)\,|\hat{x}_T(j\omega)|^2\,d\omega$$

$$\geq \frac{1}{2\pi}\int_{-\infty}^{\infty} \varepsilon\,|\hat{x}_T(j\omega)|^2\,d\omega$$

$$= \varepsilon\,\|x_T\|_2^2 = \varepsilon\,\|x\|_{T2}^2.$$

Since (94) is true for every $T \geq 0$ and every $x \in L_{2e}$, the lemma is proved. ■

95 Theorem (Popov Criterion) *Consider the feedback system shown in Figure 6.14, and suppose the following conditions hold: (i) $g(\cdot)$ has a distributional derivative, and $g, \dot{g} \in \mathbf{A}$. (ii) Φ is a memoryless time-invariant nonlinearity of the form*

$$\mathbf{96}\quad (\Phi x)(t) = \phi[x(t)],$$

where $\phi: \mathbf{R} \to \mathbf{R}$ is continuous and belongs to the sector $[0, b]$ where b could be infinite. Finally, suppose there exists a constant $q \geq 0$ such that

$$\mathbf{97}\quad \inf_{\omega \in \mathbb{R}} \mathrm{Re}\,[(1 + j\omega q)\,\hat{g}(j\omega)] + \frac{1}{b} =: \beta > 0.$$

Under these conditions, the functions e_1, e_2, y_1, y_2 belong to L_2 whenever u_1, u_2 and $\dot{u}_2$ belong to L_2; moreover, there exists a constant $\gamma < \infty$ such that

$$\mathbf{98}\quad \|e_i\|_2,\ \|y_i\|_2 \leq \gamma(\|u_1\|_2 + \|u_2\|_2 + \|\dot{u}_2\|_2),\ i = 1, 2.$$

Remarks

1. Popov's criterion applies only to *time-invariant* systems, since $\phi(\cdot)$ in (96) does not depend explicitly on t.

2. The conclusions of the theorem do not quite say that the system under study is L_2-stable. Rather, the second input u_2 *and its derivative* $\dot{u}_2$ must both belong to L_2 in order for the signals e_1, e_2, y_1, y_2 to belong to L_2.

3. If $q = 0$, then (97) reduces to (45), and the Popov criterion reduces to Case (2) of the circle criterion. But in this case we already know from Theorem (40) that (i) the nonlinearity ϕ could be time-varying, and (ii) true L_2-stability can be concluded in that there is no restriction on $\dot{u}_2$. Hence the full power of the Popov criterion comes into play only when (97) is satisfied for some $q > 0$, since in this case the inequality (97) is weaker than (45).

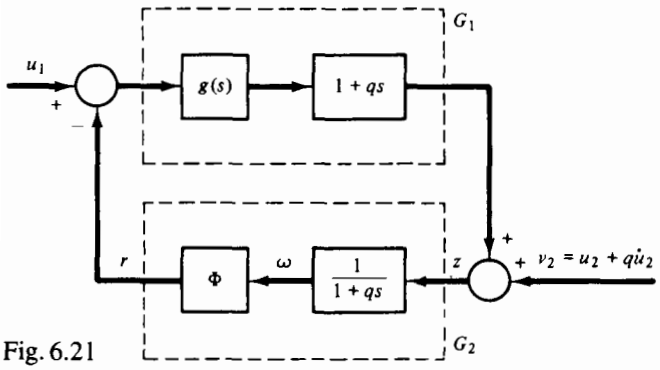


Fig. 6.21

Proof Rearrange the system of Figure 6.14 as shown in Figure 6.21 by introducing the "multiplier" $1 + qs$. In the process the input u_2 is modified to $v_2 = u_2 + q\dot{u}_2$, which belongs to L_2 by assumption. Now define G_1 and G_2 as shown in Figure 6.21 and apply Corollary (85). The assumption that $g, \dot{g} \in \mathbf{A}$ imply that $(1 + qs)\hat{g}(s) \in \hat{\mathbf{A}}$. Hence G_1 has finite L_2-gain wb and satisfies (87). Next, from Lemma (91), G_1 satisfies (86) with

99 $$\varepsilon := \inf_{\omega \in \mathbb{R}} \operatorname{Re}[(1 + j\omega q)\hat{g}(j\omega)].$$

It is now claimed that G_2 satisfies (88) with $\delta = 1/b$. If this can be shown, then the theorem follows, because $\varepsilon + \delta > 0$ by virtue of (97).

Thus, to complete the proof of the theorem, it only remains to show that

100 $$<z, G_2 z>_T = <z, r>_T \geq \frac{1}{b}\|r\|_{T2}^2, \forall T \geq 0, \forall z \in L_{2e}.$$

Now from Figure 6.21, we see that

101 $$r(t) = \phi[w(t)], z(t) = w(t) + q\dot{w}(t).$$

From Remark (3) above, it can be supposed that $q > 0$, since if $q = 0$ the theorem follows

from the circle criterion. If $q > 0$, then w is the convolution of z and the function $(1/q)\exp(-t/q)$, which shows that $w(0) = 0$. Now

$$\mathbf{102} \quad <z, r>_T = \int_0^T \phi[w(t)]\, w(t)\, dt + q \int_0^T \phi[w(t)]\, \dot{w}(t)\, dt.$$

However,

$$\mathbf{103} \quad \int_0^T \phi[w(t)]\, \dot{w}(t)\, dt = \int_{w(0)}^{w(T)} \phi(\sigma)\, d\sigma = \int_0^{w(T)} \phi(\sigma)\, d\sigma \geq 0, \ \forall T \geq 0,$$

since the graph of $\phi(\cdot)$ always lies in the first or third quadrant. Hence (102) and (103) together imply that

$$\mathbf{104} \quad <z, r>_T \geq \int_0^T \phi[w(t)]\, w(t)\, dt.$$

Now, since ϕ belongs to the sector $[0, b]$, it follows that

$$\mathbf{105} \quad 0 \leq \frac{\phi(\sigma)}{\sigma} \leq b, \ \forall \sigma \neq 0.$$

Hence

$$\mathbf{106} \quad \sigma\phi(\sigma) \geq \frac{1}{b}[\phi(\sigma)]^2, \ \forall \sigma,$$

and as a result,

$$\mathbf{107} \quad \phi[w(t)]\, w(t) \geq \frac{1}{b}\{\phi[w(t)]\}^2 = \frac{1}{b}[r(t)]^2, \ \forall t \in [0, T].$$

Combining (104) and (107) gives

$$\mathbf{108} \quad <z, r>_T \geq \frac{1}{b} \| r \|_{T2}^2,$$

thus establishing (100) and completing the proof. ■

The inequality (97) can be given a graphical interpretation, which makes it useful in practice. Suppose we plot $\mathrm{Re}\, \hat{g}(j\omega)$ versus $\omega\, \mathrm{Im}\, \hat{g}(j\omega)$ as ω varies from 0 to ∞. This graph is called the **Popov plot**, in contrast to the Nyquist plot in which one plots $\mathrm{Re}\, \hat{g}(j\omega)$ versus $\mathrm{Im}\, \hat{g}(j\omega)$. Since both $\mathrm{Re}\, \hat{g}(j\omega)$ and $\omega\, \mathrm{Im}\, \hat{g}(j\omega)$ are even functions of ω, it is only necessary to draw the plot for $\omega \geq 0$. The inequality (97) means that one can draw a straight line through the point $-1/b + j0$ with a slope $1/q \geq 0$ such that the Popov plot lies to the right of the line and does not touch it; such a line is called a **Popov line.**

109 Example Let

$$\hat{g}(s) = \frac{1}{s^2 + 4s + 4}.$$

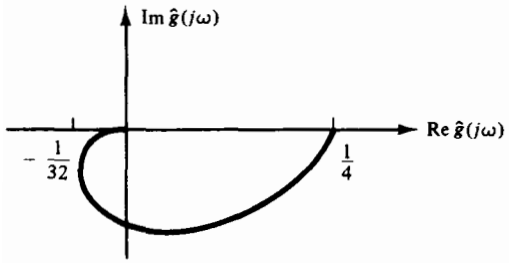


Fig. 6.22

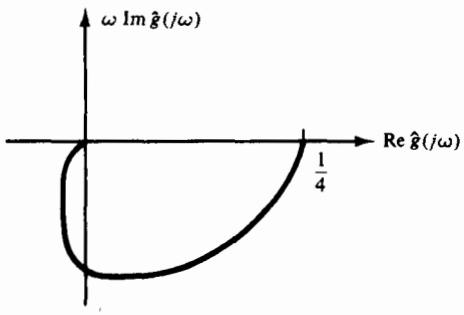


Fig. 6.23

The Nyquist plot and the Popov plot of $\hat{g}$ are shown in Figures 6.22 and 6.23, respectively. From Figure 6.23, one can see that no matter how small $1/b$ is (i.e., no matter how large b is), it is always possible to draw a suitable Popov line, as indicated. Hence the system of Figure 6.14 is L_2-stable for all *time-invariant* nonlinearities $\phi(\cdot)$ in the sector $[0, b]$ for all finite b. On the other hand,

$$\inf_{\omega \in \mathbb{R}} \operatorname{Re} \hat{g}(j\omega) = -\frac{1}{32}.$$

So (45) is satisfied whenever $b < 32$. Thus, by applying Case (2) of the circle criterion, one sees that the system of Figure 6.14 is L_2-stable for all *possibly time-varying* nonlinearities in the sector $[0, b]$ for $b < 32$. Hence, by restricting the memoryless element to be time-varying, we are able to infer stability for a larger class of nonlinearities.

Application: Aizerman's Conjecture

Aizerman's conjecture was stated in Section 5.6. Now consider the input-output version of Aizerman's conjecture, stated next:

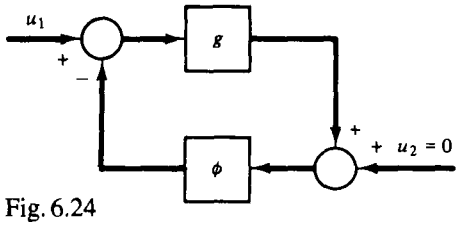


Fig. 6.24

110 Conjecture *Consider the system of Figure 6.24, and suppose that* $g, \dot{g} \in \mathbf{A}$. *Suppose in addition that* $\hat{g}/(1+k\hat{g}) \in \hat{\mathbf{A}}$ *for all* $k \in [0, b]$. *Then the system in* L_2*-stable for all nonlinearities* ϕ *in the sector* $[0, b]$.

Essentially Aizerman's conjecture (input-output version) states that if the system of Figure 6.24 is L_2-stable whenever $\phi(\cdot)$ is a constant gain of value $k \in [0, b]$, then it is also L_2-stable for all (memoryless time-invariant) nonlinearities in the sector $[0, b]$. In general, Aizerman's conjecture is false; see Willems (1971) for a class of counterexamples. However, Popov's criterion provides a means of identifying a large class of transfer functions $\hat{g}(\cdot)$ for which Aizerman's conjecture is true.

Suppose $g, \dot{g} \in \mathbf{A}$. If $g(\cdot)$ contains any impulses, then $\dot{g}$ would contain higher order impulses and thus would not belong to $\mathbf{A}$. Thus the assumptions $g \in \mathbf{A}$, $\dot{g} \in \mathbf{A}$ imply that g does not contain any impulses, i.e., that $g \in L_1$. Because $g \in L_1$, it follows from the Riemann-Lebesgue lemma that $\hat{g}(j\omega)$ has a definite limit (namely 0) as $\omega \to \infty$. Since $\hat{g}/(1+k\hat{g}) \in \hat{\mathbf{A}}$ $\forall k \in [0, b]$ by assumption, the graphical stability criterion of Theorem (6.5.35) implies that the Nyquist plot of $\hat{g}(j\omega)$ neither intersects nor encircles the half-line segment $(-\infty, -1/b]$. Because the only difference between the Nyquist plot and the Popov plot is in the vertical axis, the same is true of the Popov plot as well. Now, suppose the Popov plot of $\hat{g}$ has the shape shown shown in Figure 6.25. Then the L_2-stability wb of the nonlinear feedback system is assured for all time-invariant nonlinearities in the sector $[0, b]$, because of the Popov criterion. By Theorem (6.3.46), the state-space version of Aizerman's conjecture is satisfied by such systems. On the other hand, suppose the Popov plot of $\hat{g}$ has the appearance shown in Figure 6.26. In this case, Popov's criterion is not satisfied. However, since the Popov criterion is only a sufficient condition for stability, it still does not follow that Aizerman's conjecture is false for such a $\hat{g}$. Thus, in summary, the Popov criterion provides a readily verifiable sufficient condition for determining whether Aizerman's conjecture is valid for a particular transfer function $\hat{g}(\cdot)$.

6.6.3 Necessity of the Circle Criterion

This section is concluded by showing that, in a sense to be made precise below, the circle criterion provides a necessary as well as sufficient condition for absolute stability. Hence in a sense the circle criterion is *not* overly conservative. To minimize the technical

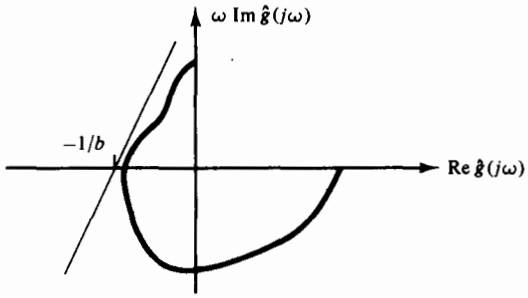


Fig. 6.25

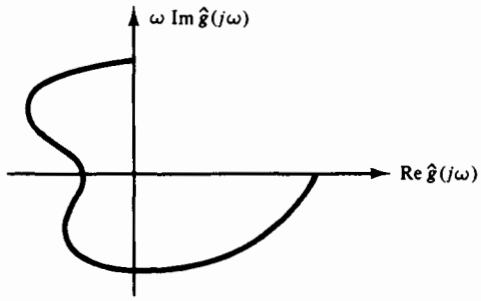


Fig. 6.26

details, attention is restricted to systems of the form shown in Figure 6.27, where the forward path consists of a *lumped* linear time-invariant system, and the feedback path consists of a causal, but not necessarily memoryless, nonlinear element $\Phi: L_{2e} \to L_{2e}$. Let a, b be given real numbers with $a < b$. Then we say that Φ *belongs to the sector* $[a, b]$ if it is true that

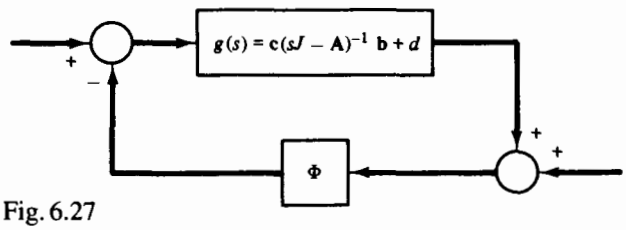


Fig. 6.27

111 $\|\Phi x - [(b+a)/2]x\|_{T2} \leq (b-a)/2 \|x\|_{T2}, \forall T \geq 0, \forall x \in L_{2e}$.

Note that the above definition of the sector $[a, b]$ is consistent with the earlier one if Φ happens to be a memoryless nonlinearity of the form (15). However, the present definition is broader since it is not restricted to memoryless operators. We say that Φ belongs to the

sector (a, b) if it belongs to the sector $[a+\varepsilon, b-\varepsilon]$ for some $\varepsilon > 0$. Before presenting the main result of this subsection, which is Theorem (126), an important special case is treated.

112 Lemma *Consider the system of Figure 6.27, and let $r > 0$ be a given real number. Then the following two statements are equivalent:*

(I) *The system is L_2-stable wb for every Φ belonging to the sector $(-r, r)$.*

(II) $\hat{g} \in \hat{\mathbf{A}}$ *and*

$$\textbf{113} \quad \sup_{\omega} |\hat{g}(j\omega)| \leq \frac{1}{r}.$$

Proof "(II) $\Rightarrow$ (I)" Suppose (II) is true. Then the operator $G: x \mapsto g * x$ is L_2-stable, and by Theorem (6.4.40),

$$\textbf{114} \quad \gamma_2(G) = \sup_{\omega} |\hat{g}(j\omega)| \leq \frac{1}{r}.$$

Now suppose $\Phi: L_{2e} \to L_{2e}$ belongs to the sector $(-r, r)$. Then there exists an $\varepsilon > 0$ such that Φ belongs to the sector $[-r+\varepsilon, r-\varepsilon]$. Hence Φ is L_2-stable wb, and

$$\textbf{115} \quad \gamma_2(\Phi) \leq r - \varepsilon < r.$$

Hence

$$\textbf{116} \quad \gamma_2(G) \cdot \gamma_2(\Phi) < 1,$$

and the L_2-stability wb of the system follows from Theorem (1).

"(I) $\Rightarrow$ (II)" Suppose (I) is true. Then, in particular, the system is L_2-stable with $\Phi = 0$, since the zero operator belongs to the sector $(-r, r)$. Since $\hat{g}$ is rational, L_2-stability is equivalent to L_∞-stability, and $\hat{g} \in \hat{\mathbf{A}}$. To prove (113), it is shown that if (113) is false then (I) is false. Accordingly, suppose (113) is false, i.e., that

$$\textbf{117} \quad \sup_{\omega} |\hat{g}(j\omega)| > \frac{1}{r},$$

and select an ω_0 such that

$$\textbf{118} \quad |\hat{g}(j\omega_0)| > \frac{1}{r}.$$

To be precise, suppose

$$\textbf{119} \quad \hat{g}(j\omega_0) = \frac{1}{r-\varepsilon} \exp(j\theta),$$

where ε lies in the interval $(0, r)$ and $\theta \in [0, 2\pi)$. Now let

120 $\tau = \theta/\omega_0$,

and define $\Phi: L_{2e} \to L_{2e}$ by

121 $(\Phi x)(t) = -(r-\varepsilon)\, x(t-\tau)$.

In other words, Φ is a gain of $-(r-\varepsilon)$ cascaded with a delay of τ; note that Φ is *not* memoryless. Now Φ is also linear and time-invariant, and its transfer function is

122 $\hat{\phi}(s) = -(r-\varepsilon)\exp(-\tau s)$.

From the construction it is clear that

123 $1 + \hat{g}(j\omega_0)\hat{\phi}(j\omega_0) = 0$.

Hence the function $s \mapsto 1/[1+\hat{g}(s)\hat{\phi}(s)]$ is unbounded over C_+; as a result, the transfer function $1/(1+\hat{g}\hat{\phi})$ cannot represent an L_2-stable system. Thus the system under study is L_2-unstable for the particular choice of Φ in (121). Since this Φ belongs to the sector $[-r+\varepsilon, r-\varepsilon]$ and hence to the sector $(-r, r)$, it follows that (I) is false. ■

Suppose we try to modify Lemma (112) by restricting Φ to the "closed" sector $[-r, r]$ and changing (113) to

124 $\sup_\omega |\hat{g}(j\omega)| < \dfrac{1}{r}$.

Then a subtle difficulty arises in the proof. If (124) is violated, i.e., if

125 $\sup_\omega |\hat{g}(j\omega)| \geq \dfrac{1}{r}$,

then there need not exist any *finite* ω_0 such that $|\hat{g}(j\omega_0)| \geq 1/r$. This is the reason for stating Lemma (112) in that particular form, with Φ restricted to the "open" sector $(-r, r)$ and stating (113) with a non-strict inequality.

Lemma (112) is an example of a result for so-called **absolute stability.** This term refers to the study of the stability of *an entire family* of systems, instead of just specific systems. The main idea in absolute stability theory is to deduce the stability of an entire family of systems by studying only some of its members. By examining the proof of Lemma (112), the reader can easily prove the following result: If the feedback system of Figure 6.28 is L_2-stable wb for all real constants $k \in (-r, r)$ and all delays $\tau \geq 0$, then the system of Figure 6.27 is L_2-stable wb for all nonlinearities Φ in the sector $(-r, r)$.

Now for the main result.

126 Theorem *Consider the system of Figure 6.27, and suppose a, b are two given real numbers with $a < b$. Then the following two statements are equivalent:*

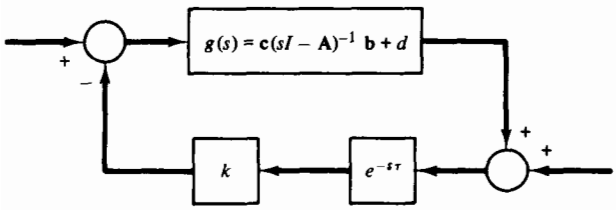


Fig. 6.28

(I) The feedback system is L_2-stable wb for every Φ belonging to the sector (a, b).

(II) The transfer function $\hat{g}$ satisfies one of the following conditions as appropriate: (1) If $ab > 0$, then the Nyquist plot of $\hat{g}(j\omega)$ does not intersect the interior of the disk $D(a, b)$ defined in (39), and encircles the interior of the disk $D(a, b)$ exactly μ_+ times in the counterclockwise direction, where μ_+ is the number of poles of $\hat{g}$ with positive real part. (2) If $a = 0$, then $\hat{g}$ has no poles with positive real part, and

127 $$\operatorname{Re} \hat{g}(j\omega) \geq -\frac{1}{b}, \forall \omega.$$

(3) If $ab < 0$, then $\hat{g} \in \hat{\mathbf{A}}$, and the Nyquist plot of $\hat{g}(j\omega)$ lies inside the disk $D(a, b)$ for all ω.

Remarks There are some slight differences between Statement (II) above and the conditions in Theorem (40). These differences arise because here the nonlinear element Φ is assumed to lie in the "open" sector (a, b), while in Theorem (40) the nonlinearity is restricted to lie in the "closed" sector $[a, b]$. As a consequence, the region in the complex plane in which the Nyquist plot of $\hat{g}(j\omega)$ is required to lie is a closed set here, while it is an open set in Theorem (40). The most notable difference arises in Case (2), i.e., $a = 0$. If Φ is permitted to lie in the sector $[0, b]$, then obviously $\hat{g}$ itself must be stable, because $\Phi = 0$ is a permissible choice. But if Φ can only belong to the *open* sector $(0, b)$, then $\Phi = 0$ is *not* a permissible choice, and $\hat{g}$ itself need not be stable. The differences in the other two cases are quite minor. Note that it is quite routine to modify Theorem (40) to provide for the case where Φ belongs to the open sector (a, b). Thus Theorem (126) shows that the circle criterion [suitably modified for the fact that the sector (a, b) is open] is in fact a necessary as well as sufficient condition for absolute stability, provided the forward path element is lumped. Of course, the power of Theorem (40) lies in that it is applicable to distributed systems as well.

Proof "(II) $\Rightarrow$ (I)" This follows from Theorem (40), after adjusting for the fact that Φ belongs to the sector (a, b) rather than the sector $[a, b]$. The details are left as an exercise. (See Problem 6.22.)

"(I) $\Rightarrow$ (II)" Each of the three cases is handled separately.

Case (3). $ab < 0$: In this case $\Phi = 0$ belongs to the sector (a, b). Since the system is L_2-stable wb when $\Phi = 0$, it follows that $\hat{g}$ itself is L_2-stable wb, and since $\hat{g}$ is rational, that $\hat{g} \in \hat{\mathbf{A}}$. Now define, as before,

128 $k = \dfrac{b+a}{2}, r = \dfrac{b-a}{2}.$

Again, $\Phi = kI$ is a permissible choice, and the L_2-stability wb of the system with this particular choice of Φ shows that

129 $\hat{g}_t := \dfrac{\hat{g}}{1+k\hat{g}} \in \hat{\mathbf{A}}.$

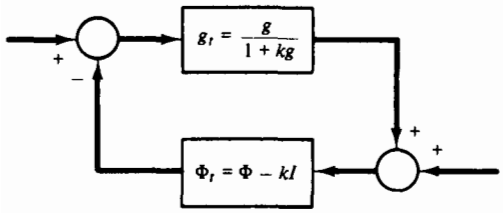


Fig. 6.29

Now redraw the system of Figure 6.27 as in Figure 6.29, where $\hat{g}_t$ is defined above, and

130 $\Phi_t = \Phi - kI \in \text{sector}\,(-r, r).$

Of course the inputs in Figure 6.29 are not the same as those in Figure 6.27, but as shown in the proof of Theorem (22), the system of Figure 6.27 is L_2-stable wb if the system of Figure 6.29 is L_2-stable wb, and the converse follows easily (see Problem 6.19). Hence the hypothesis that (I) is true implies that the system of Figure 6.29 is L_2-stable wb for every Φ in the sector $(-r, r)$. Now apply Lemma (112). This shows that

131 $\sup_{\omega} |\hat{g}_t(j\omega)| \leq \dfrac{1}{r},$

or, equivalently,

132 $\sup_{\omega} \left| \dfrac{\hat{g}(j\omega)}{1+[(b+a)/2]\hat{g}(j\omega)} \right| \leq \dfrac{2}{b-a}.$

As shown in the proof of Theorem (40), this implies [after allowing for the non-strict inequality in (132)] that $\hat{g}(j\omega) \in D(a, b)\ \forall \omega$.

Case (2). $a = 0$: The reasoning is much the same as in Case (3). In this case $k = r = b/2$, and (131) becomes

133 $\sup_{\omega} \left| \dfrac{\hat{g}(j\omega)}{1+(b/2)\hat{g}(j\omega)} \right| \leq \dfrac{2}{b},$

which can be easily shown to be equivalent to (127). Also, since $-2/b < -1/b$, (127) implies

that

134 $\text{Arg}\,[\hat{g}(j\omega)+(2/b)]\in(-\pi/2,\,\pi/2),\ \forall\omega.$

As a consequence,

135 $\lim_{\omega\to\infty}\text{Arg}\,[\hat{g}(j\omega)+(2/b)]-\text{Arg}\,[\hat{g}(-j\omega)+(2/b)]=0,$

since this quantity must be an integer multiple of 2π. Now, by assumption, the feedback system is L_2-stable if $\Phi=kI=(b/2)I$. Comparing (135) with the argument condition in Theorem (6.5.35), one sees that $\hat{g}$ cannot have any poles with positive real part.

Case (1). $ab>0$: Define k and r as in (128), and redraw the system as in Figure 6.29. Then (129) and (131) follow as before, and (131) implies, as in the proof of Theorem (40), that $\hat{g}(j\omega)$ does not intersect the interior of the disk $D(a,\,b)$. Since the feedback system is stable with $\Phi=kI$, the Nyquist plot must encircle the point $-1/k$ exactly μ_+ times in the counterclockwise direction. Since the plot of $\hat{g}(j\omega)$ does not intersect the interior of the disk $D(a,\,b)$, the same encirclement condition applies to every point in the interior of $D(a,\,b)$. ■

Problem 6.18 Modify the statement of Theorem (1) to the case where the operators G_1 and G_2 are L_p-stable wfg but not necessarily L_p-stable wb, and prove the resulting statement. Obtain estimates corresponding to (13) and (14).

Problem 6.19 Consider the feedback system of Figure 6.2 and its transformed version in Figure 6.16. Suppose K is a linear operator and that K is L_p-stable wb. Show that the system of Figure 6.2 is L_p-stable wb if and only if the system of Figure 6.16 is L_p-stable wb.

Problem 6.20 Prove (48).

Problem 6.21 Obtain explicit bounds on the norms $\|e\|_{T2}$ and $\|y\|_{T2}$ from (73).

Problem 6.22 In Theorem (126), show that (II) implies (I).

Problem 6.23 Show that, in Case (2) of Theorem (126), not only is $\hat{g}$ not permitted to have any poles in the open right half-plane, but it is also not permitted to have any *repeated* poles on the $j\omega$-axis.

Problem 6.24 Using Theorem (40), find some possible choices of the sector $[a,\,b]$ such the feedback system of Figure 6.14 is L_2-stable when $\hat{g}(s)$ is as in Problem 6.16.

Problem 6.25 Using Theorem (95), find the possible values of the constant $b>0$ such that the feedback system of Figure 6.14 is stable with

$$\hat{g}(s)=\frac{s-1}{(s+1)(s+2)(s+5)(s+10)}\ ,$$

and Φ a time-invariant and memoryless nonlinearity belonging to the sector $[0,\,b]$.

6.7 DISCRETE-TIME CONTROL SYSTEMS

In this brief section, the discrete-time analogs of the contents of the first six sections are presented, mostly without proof. Since in most cases the details of the discrete-time results are virtually identical to those in the continuous-time case, only the differences are highlighted.

6.7.1 Stability Definitions

Let S denote the linear space of all sequences $\{x_i\}_{i\geq 0}$. For $p\in[1,\infty)$ define

$$\mathbf{1}\qquad l_p=\{x\in S:\sum_{i=0}^{\infty}|x_i|^p<\infty\},$$

and define

$$\mathbf{2}\qquad l_\infty=\{x\in S: x \text{ is a bounded sequence }\}.$$

Then l_p is a subspace of S for each $p\in[1,\infty]$. If we define the norms

$$\mathbf{3}\qquad \|x\|_p=[\sum_{i=0}^{\infty}|x_i|^p]^{1/p},\ \forall x\in l_p,$$

$$\mathbf{4}\qquad \|x\|_i=\sup_i|x_i|,\ \forall x\in l_\infty,$$

then the pair $(l_p, \|\cdot\|_p)$ is a Banach space for each $p\in[1,\infty]$. Note that $l_p\subseteq l_q$ if $p<q$.

A sequence $x\in S$ is said to **have finite support** if there is an integer N such that $x_i=0\ \forall i>N$. Clearly, if x has finite support then $x\in l_p\ \forall p\in[1,\infty]$. Hence the set S acts as the "extension" of l_p for each $p\in[1,\infty]$, and there is no need for a symbol such as l_{pe}.

A binary relation on S is defined just as in Section 6.2, namely as a subset of S^2.

5 **Definition** *Suppose R is a binary relation on S. Then R is* l_p**-stable** *if*

$$\mathbf{6}\qquad (x,y)\in R,\ x\in l_p\Rightarrow y\in l_p.$$

R is l_p**-stable with finite gain (wfg)** *if it is l_p-stable, and in addition there exist finite constants γ_p and b_p such that*

$$\mathbf{7}\qquad (x,y)\in R,\ x\in l_p\Rightarrow \|y\|_p\leq\gamma_p\|x\|_p+b_p.$$

R is l_p**-stable with finite gain and zero bias (wb)** *if it is l_p-stable, and in addition there exists a finite constant γ_p such that*

$$\mathbf{8}\qquad (x,y)\in R,\ x\in l_p\Rightarrow \|y\|_p\leq\gamma_p\|x\|_p.$$

The definitions of feedback stability are now entirely analogous to Definition (6.2.32), with L_p replaced by l_p.

It is left to the reader to state and prove discrete-time analogs of the contents of Section 6.3.

6.7.2 Linear Time-Invariant Systems

Suppose $x, y \in S$. Then their convolution $x * y \in S$ is defined by

$$\textbf{9} \qquad (x * y)_i = \sum_{j=0}^{i} x_{i-j}\, y_j = \sum_{j=0}^{i} x_j\, y_{i-j}.$$

The set l_1 plays the same role in discrete-time systems that the set **A** does in continuous-time systems.

10 Lemma *Suppose $x, y \in l_1$. Then $x * y \in l_1$, and*

$$\textbf{11} \qquad \|x * y\|_1 \leq \|x\|_1 \cdot \|y\|_1.$$

Every linear time-invariant operator $A : S \to S$ has the representation $Ax = a * x$, where the sequence $a \in S$ is called the *unit pulse response of A*.

12 Theorem *Suppose $A : S \to S$ is linear and time-invariant, and let a denote the unit pulse response of A. Then the following four statements are equivalent:*

(i) *A is l_1-stable wb.*

(ii) *A is l_∞-stable wb.*

(iii) *A is l_p-stable wb for all $p \in [1, \infty]$.*

(iv) $a \in l_1$.

Moreover, if $a \in l_1$, then

$$\textbf{13} \qquad \|a * x\|_p \leq \|a\|_1 \cdot \|x\|_p, \quad \forall x \in l_p, \quad \forall p \in [1, \infty].$$

Every sequence $f \in l_1$ has a z-transform defined by

$$\textbf{14} \qquad \tilde{f}(z) = \sum_{i=0}^{\infty} f_i z^i,$$

which converges whenever $|z| \leq 1$. Note that z is raised to positive powers in (14), not negative powers. The symbol $\tilde{l}_1$ denotes the set of z-transforms of sequences in l_1. If $f \in l_1$, then $\tilde{f}$ is analytic on the open unit disk, continuous on the closed unit disk; moreover, any zeros of $\tilde{f}$ in the open unit disk are isolated.

Theorem (12) is the discrete-time analog of Theorem (6.4.30). The next result is the analog of Theorem (6.4.40).

15 Theorem *Suppose $A: S \to S$ is linear and time-invariant, and that its unit pulse response a belongs to l_1. Then*

$$\gamma_2(A) = \max_{\theta \in [0, 2\pi]} |\tilde{a}(e^{j\theta})| . \tag{16}$$

Note that $\tilde{a}(e^{j\theta})$ a continuous function of θ, and that the interval $[0, 2\pi]$ is compact; hence in (16) we can say "max" instead of "sup".

The stability results for linear time-varying systems can be obtained from Theorems (6.4.53), (6.4.66), and (6.4.75) simply by replacing all integrals by summations. In fact, it is possible to place all of these results into a unified framework by replacing $\mathbf{R}$ (which plays the role of the "time set" in continuous-time systems) by an arbitrary locally compact Abelian group; see Vidyasagar and Bose (1975) for details.

A causal linear discrete-time system G has an input-output representation of the form

$$(Gx)_i = \sum_{j=0}^{i} g_{ij} x_j, \ \forall i \geq 0. \tag{17}$$

18 Theorem *The operator $G: S \to S$ defined by (17) is l_∞-stable wb if and only if*

$$\sup_i \sum_{j=0}^{i} |g_{ij}| =: c_\infty < \infty. \tag{19}$$

G is l_1-stable wb if and only if

$$\sup_j \sum_{i=j}^{\infty} |g_{ij}| =: c_1 < \infty. \tag{20}$$

If G satisfies both (19) and (20), then G is l_p-stable for all $p \in [1, \infty]$, and

$$\gamma_p(G) \leq c_1^{1/p} \, c_\infty^{1/q}, \tag{21}$$

where $q = p/(p-1)$.

Now we come to feedback stability. The analog of Lemma (6.5.1) is the following.

22 Lemma *Suppose $\tilde{f} \in \tilde{l}_1$. Then $1/\tilde{f} \in \tilde{l}_1$ if and only if*

$$\tilde{f}(z) \neq 0 \ \forall z \text{ with } |z| \leq 1. \tag{23}$$

In many ways the discrete-time theory is much simpler than the continuous-time theory, since there are fewer technicalities. For instance, since the closed unit disk is a compact set and since $\tilde{f}(\cdot)$ is continuous on the closed unit disk, (23) is equivalent to

24 $\inf_{|z|\le 1} |\tilde{f}(z)| > 0.$

Getting a graphical criterion to test (23) is also easy. Suppose $\tilde{f} \in \tilde{l}_1$, and suppose one plots $\tilde{f}(e^{j\theta})$ as θ increases from 0 to 2π. If $\tilde{f}(e^{j\theta}) = 0$ for some θ, then (23) fails at once, and no further testing is necessary. If $\tilde{f}(e^{j\theta}) \neq 0 \; \forall \theta \in [0, 2\pi]$, then it follows from the fact that all zeros of $\tilde{f}(\cdot)$ in the open unit disk are isolated that f has only a finite number of zeros in the closed (and open) unit disk. By the principle of the argument, the number of zeros of $\tilde{f}(\cdot)$ in the closed unit disk is precisely the number of times that the plot of $\tilde{f}(e^{j\theta})$ encircles the origin in the counterclockwise direction as θ increases from 0 to 2π. The discussion can be summarized as follows:

25 **Lemma** *Suppose* $\tilde{f} \in \tilde{l}_1$. *Then* $1/\tilde{f} \in \tilde{l}_1$ *if and only if the following two conditions hold:*

26 (*i*) $\tilde{f}(e^{j\theta}) \neq 0, \; \forall \theta \in [0, 2\pi]$, and

27 (*ii*) $\operatorname{Arg} \tilde{f}(e^{j2\pi}) - \operatorname{Arg} \tilde{f}(e^{j0}) = 0.$

Consider now the linear time-invariant feedback system of Figure 6.5 [with $\hat{g}(s)$ changed to $\tilde{g}(z)$, of course]. Assume that $\tilde{g}(z)$ has the form

28 $\tilde{g}(z) = \tilde{g}_s(z) + \tilde{g}_r(z),$

where $\tilde{g}_s \in \tilde{l}_1$, and $\tilde{g}_r$ is rational but analytic at $z = 0$. (The analyticity assumption is to ensure that $\tilde{g}$ corresponds to a causal system.) If $\tilde{g}_r$ has some poles on the unit circle, then the unit circle should be "indented" so that these poles do not lie inside the region enclosed by the indented circle (see Figure 6.30). By a slight abuse of notation, let $\tilde{g}(e^{j\theta})$ denote the value of $\tilde{g}(z)$ at the unique point z on the indented unit circle whose argument is θ. Then we have the following result, which is an analog of Theorem (6.5.35).

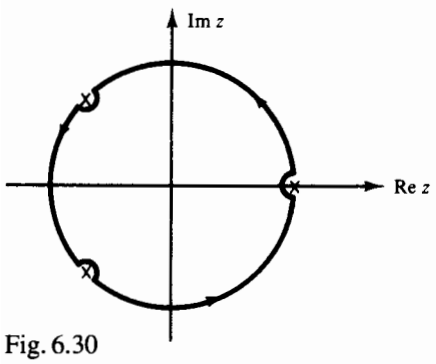


Fig. 6.30

29 Theorem *Consider the system of Figure 6.5, with $\hat{g}(s)$ replaced by $\tilde{g}(z)$, and suppose $\tilde{g}(z)$ has the form (28). Then the feedback system is stable if and only if (i)*

$$\textbf{30} \quad \tilde{g}(e^{j\theta}) \neq -\frac{1}{k}, \forall \theta \in [0, 2\pi],$$

and (ii) the plot of $\tilde{g}(e^{j\theta})$ encircles the point $-1/k$ exactly μ_+ times in the clockwise direction as θ increases from 0 to 2π, where μ_+ is the number of poles of $\tilde{g}_r$ in the open unit disk.

Notice an important difference between Theorem (29) and Theorem (6.5.35) [or Corollary (6.5.38)]. In the latter case, the encirclements are required to be in the *counterclockwise* direction, whereas in the present case they are required to be in the *clockwise* direction. The difference arises because of the orientation of the contour in the two cases. In the discrete-time case, the "unstable" region, i.e., the region where the poles are forbidden to lie in order to have a stable system, lies to the *left* of the unit circle. Hence, as shown in Figure 6.30, the unit circle is a positively oriented curve. In contrast, in the continuous-time case, the unstable region lies to the *right* of the $j\omega$-axis as ω increases from $-\infty$ to ∞.

The discrete-time analog of the set $\hat{B}$ is straight-forward. Given a $\rho > 1$, define

$$\textbf{31} \quad l_{1\rho} = \{x \in S: \sum_{i=0}^{\infty} |x_i| \rho^i < \infty\},$$

and define

$$\textbf{32} \quad l_{1+} = \bigcup_{\rho > 1} l_{1\rho}.$$

Then $l_{1+} \subseteq l_1$. Let $\tilde{l}_{1+}$ denote the set of z-transforms of sequences in l_{1+}.

33 Definition *The set $\hat{B}_d$ consists of all ratios $\tilde{n}(z)/\tilde{d}(z)$ where $\tilde{n}, \tilde{d} \in \hat{B}_d$, and in addition $\tilde{d}(0) \neq 0$.*

It is left to the reader to state and prove properties of the set $\hat{B}_d$ analogous to those of $\hat{B}$ as stated in Section 6.5.2. In particular, every $\tilde{f} \in \hat{B}_d$ has the form (28).

The extension of the contents of Section 6.5.3 to discrete-time systems is totally straight-forward. For a unified treatment of the coprime factorization approach, see Vidyasagar (1985).

6.7.3 Nonlinear Feedback Systems

Now let us turn our attention to nonlinear feedback systems. The small gain theorem [Theorem (6.6.1)], the passivity theorem [Theorem (6.6.58)] and its various corollaries all apply, with the obvious modifications, to discrete-time systems.

Consider now the feedback system of Figure 6.14, where the forward path is linear and time-invariant with transfer function $\tilde{g}(z)$ [instead of $\hat{g}(s)$], and the feedback element Φ is memoryless and of the form

34 $(\Phi x)_i = \phi(i, x_i)$

for some continuous function $\phi: Z_+ \times \mathbf{R} \to \mathbf{R}$, where

35 $Z_+ = \{0, 1, 2, \cdots\}$

denotes the set of nonnegative integers. In analogy with (6.6.17), we say that "ϕ or Φ belongs to the sector $[a, b]$ if

36 $a\sigma^2 \leq \sigma\phi(i, \sigma) \leq b\sigma^2, \forall \sigma \neq 0, \forall i \in Z_+.$

It is easy to see what is meant by "ϕ or Φ belongs to the sector $[a, b]$" by making the obvious changes in (6.6.17).

As in the continuous-time case, the circle criterion is obtained by combining the small gain theorem and the graphical stability criterion for linear time-invariant systems. Note that the disk $D(a, b)$ is defined as in (6.6.39).

37 Theorem *Consider the system of Figure 6.14. Suppose $\tilde{g}$ has the form (28), and that Φ is memoryless and belongs to the sector $[a, b]$. Then the feedback system is l_2-stable wb if one of the following conditions, as appropriate, holds:*

Case (1). $ab > 0$: The plot of $\tilde{g}(e^{j\theta})$ as θ increases from 0 to 2π does not intersect the disk $D(a, b)$, and encircles it exactly μ_+ times in the clockwise direction, where μ_+ is the number of poles of $\tilde{g}$ in the open unit disk.

Case (2). $a = 0$: $\tilde{g} \in \tilde{l}_1$, and

38 $\operatorname{Re} \tilde{g}(e^{j\theta}) > -\dfrac{1}{b}, \forall \theta \in [0, 2\pi].$

Case (3). $ab < 0$: $\tilde{g} \in \tilde{l}_1$, and $\tilde{g}(e^{j\theta})$ lies in the interior of the disk $D(a, b)$ for all $\theta \in [0, 2\pi]$.

The "Popov criterion" for discrete-time systems is, however, noticeably different from its continuous-time counterpart, and is given next. In proving this theorem, it is helpful to notice that if $\tilde{x}(z)$ is the z-transform of the sequence $\{x_0, x_1, x_2, \cdots\}$, then $z\tilde{x}(z)$ is the z-transform of the delayed sequence $\{0, x_0, x_1, \cdots\}$.

39 Theorem *Consider the feedback system of Figure 6.14 with $\hat{g}(s)$ replaced by $\tilde{g}(z)$. Suppose $\tilde{g} \in \tilde{l}_1$ and that Φ is a memoryless nonlinearity belonging to the sector $[0, b]$. Then the feedback system is l_2-stable if there exists a $q \geq 0$ such that*

40 $$\text{Re}\,\{[1-q(e^{j\theta}-1)]\,\tilde{g}(e^{j\theta})\} > -\frac{1}{b}+\frac{qb\,(1+2q)^2\gamma^2}{4},\ \forall\theta\in[0,\,2\pi],$$

where

41 $$\gamma := \max_{\theta\in[0,\,2\pi]} |\tilde{g}(e^{j\theta})|.$$

Remarks

1. If $q=0$, then (40) reduces to (38). So Theorem (39) gives a less restrictive stability condition than the circle condition. However, it is not clear how one would test systematically whether there exists a q satisfying the hypotheses of the theorem.

2. Comparing Theorem (39) with the Popov criterion [Theorem (6.6.95)], one can see several significant differences. First and foremost, the continuous-time criterion is only applicable to time-invariant systems, whereas no such restriction applies to the present case. Next, in the continuous-time case, both $\hat{g}(s)$ and $s\,\hat{g}(s)$ are assumed to belong to $\hat{\mathbf{A}}$. Such an assumption is not needed in the discrete-time case, since if $\tilde{g}(z)$ belongs to $\tilde{l}_1$, then so does the function $z\,\tilde{g}(z)$. [Note that the inverse transform of the function $z\,\tilde{g}(z)$ is the delayed sequence $\{0, g_0, g_1, \cdots\}$ which belongs to l_1 if $\tilde{g}\in\tilde{l}_1$.] Similarly, in the continuous-time case, it is assumed that u_2 and its derivative $\dot{u}_2$ both belong to L_2. Again, such an assumption is not needed in the discrete-time case, since if $u_2\in l_2$, then so does the delayed sequence $\{0, u_{20}, u_{21}, \cdots\}$, which is the inverse transform of $z\,\tilde{u}_2(z)$. Hence in the discrete-time case one has "true" l_2-stability if (40) holds. In view of these differences, perhaps one should not even refer to Theorem (39) as a discrete-time version of the Popov criterion.

Proof Given sequences $x,\ y\in S$, define their truncated inner product $\langle x,\ y\rangle_N$ as

42 $$\langle x,\ y\rangle_N = \sum_{i=0}^{N} x_i\,y_i.$$

Now the discrete-time analog of Theorem (6.6.58) is as follows: The system of Figure 6.2 is l_2-stable if there exist constants $\varepsilon_1,\ \delta_1,\ \varepsilon_2,\ \delta_2$ such that

43 $$\langle x,\ G_i x\rangle_N \geq \varepsilon_i\,\|x\|_{N2}^2 + \delta_i\,\|G_i x\|_{N2}^2,\ \forall N\in Z_+,\ \forall x\in S,\ \text{for } i=1,\,2,$$

where

44 $$\|x\|_{N2} = \langle x,\ x\rangle_N^{1/2} = \left[\sum_{i=0}^{N} x_i^2\right]^{1/2},$$

and such that

45 $\delta_1 + \varepsilon_2 > 0, \delta_2 + \varepsilon_1 > 0.$

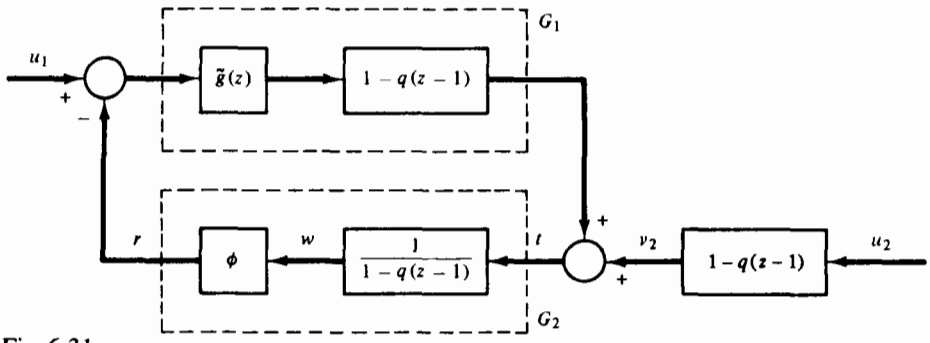


Fig. 6.31

Redraw the system under study as shown in Figure 6.31, and note that $v_2 \in l_2$ whenever $u_2 \in l_2$. Also, the transfer function $1/[1-q(z-1)] \in l_1$ for all $q \geq 0$, and

46 $$\max_{\theta \in [0, 2\pi]} \left| \frac{1}{1-q(e^{j\theta}-1)} \right| = 1.$$

Let us first analyze the operator G_1. Define

47 $$\beta := \min_{\theta \in [0, 2\pi]} \operatorname{Re} \{[1-q(e^{j\theta}-1)]\tilde{g}(e^{j\theta})\},$$

48 $$\mu := \max_{\theta \in [0, 2\pi]} |[1-q(e^{j\theta}-1)]\tilde{g}(e^{j\theta})|.$$

Then (40) implies that

49 $$\beta > -\frac{1}{b} + \frac{qb(1+2q)^2\gamma^2}{4}.$$

Next, observe that

50 $$\mu \leq \max_{\theta \in [0, 2\pi]} |1-q(e^{j\theta}-1)| \cdot \max_{\theta \in [0, 2\pi]} |\tilde{g}(e^{j\theta})| = (1+2q)\gamma.$$

Thus (49) and (50) together imply that

51 $$\beta > -\frac{1}{b} + \frac{qb\mu^2}{4}.$$

Now, by analogy with Lemma (6.6.91), it follows that

52 $\langle x, G_1 x\rangle_N \geq \beta \| x \|_{N2}^2, \forall N \in Z_+, \forall x \in S.$

Also, it follows from Theorem (15) that

53 $\| G_1 x \|_{N2} \leq \mu \| x \|_{N2}, \forall N \in Z_+, \forall x \in S.$

Combining (52) and (53) shows [cf. (6.6.83)] that, for all $\alpha \geq 0$,

54 $\langle x, G_1 x\rangle_N \geq (\beta - \alpha) \| x \|_{N2}^2 + \dfrac{\alpha}{\mu^2} \| G_1 x \|_{N2}^2, \forall N \in Z_+, \forall x \in S.$

Now let us study the operator G_2. With the various quantities defined as in Figure 6.31, we have

55 $r_i = \phi_i(w_i), \forall i,$ and

56 $\tilde{t}(z) = [1 - q(z-1)]\, \tilde{w}(z),$ i.e.,

57 $t_i = w_i - q(w_{i-1} - w_i), \forall i,$

where w_{-1} is taken as zero. Thus

58 $\langle t, G_2 t\rangle_N = \sum_{i=0}^{N} r_i t_i = \sum_{i=0}^{N} \phi_i(w_i) w_i - q \sum_{i=0}^{N} \phi_i(w_i)(w_{i-1} - w_i).$

Let us examine the second summation on the right side, keeping in mind that $0 \leq \phi_i(w_i)/w_i \leq b, \forall i.$ Now

59
$$\begin{aligned} \phi_i(w_i)(w_{i-1} - w_i) &= \frac{\phi_i(w_i)}{w_i}(w_{i-1} w_i - w_i^2) \\ &= -\frac{\phi_i(w_i)}{w_i}(w_i - 0.5 w_{i-1})^2 + \frac{\phi_i(w_i)}{4 w_i} w_{i-1}^2 \\ &\leq \frac{\phi_i(w_i)}{4 w_i} w_{i-1}^2 \leq \frac{b}{4} w_{i-1}^2. \end{aligned}$$

Hence

60 $\sum_{i=0}^{N} \phi_i(w_i)(w_{i-1} - w_i) \leq \dfrac{b}{4} \sum_{i=0}^{N} w_{i-1}^2 \leq \dfrac{b}{4} \| w \|_{N2}^2,$

61 $-q \sum_{i=0}^{N} \phi_i(w_i)(w_{i-1} - w_i) \geq -\dfrac{qb}{4} \| w \|_{N2}^2.$

Applying Theorem (15) and Equation (46) to

62 $$\tilde{w}(z) = \frac{\tilde{t}(z)}{1 - q(z-1)}$$

shows that

63 $$\| w \|_{N2} \leq \| t \|_{N2}, \forall N \in Z_+.$$

Substituting (63) into (61) gives

64 $$-q \sum_{i=0}^{N} \phi_i(w_i)(w_{i-1} - w_i) \geq -\frac{qb}{4} \| t \|_{N2}^2.$$

Let us return to the first summation on the right side of (58). Since Φ belongs to the sector $[0, b]$, it follows that

65 $$\sum_{i=0}^{N} \phi_i(w_i) w_i \geq \frac{1}{b} \sum_{i=0}^{N} [\phi_i(w_i)]^2 = \frac{1}{b} \| G_2 t \|_{N2}^2.$$

Finally, substituting from (64) and (65) into (58) gives

66 $$<t, G_2 t>_N \geq \frac{1}{b} \| G_2 t \|_{N2}^2 - \frac{qb}{4} \| t \|_{N2}^2.$$

Now we are in a position to complete the proof. Equations (54) and (66) show that (43) holds with

67 $$\varepsilon_1 = \beta - \alpha, \delta_1 = \frac{\alpha}{\mu^2}, \varepsilon_2 = -\frac{qb}{4}, \delta_2 = \frac{1}{b}.$$

Note that the constant α is not yet specified. Thus the system under study is l_2-stable if it is possible to choose $\alpha \geq 0$ such that

68 $$\frac{\alpha}{\mu^2} - \frac{qb}{4} > 0, \beta - \alpha + \frac{1}{b} > 0.$$

The first inequality in (68) leads to

69 $$\alpha > \frac{qb\mu^2}{4},$$

while the second inequality in (68) leads to

79 $$\beta > \alpha - \frac{1}{b}.$$

Hence, if (51) holds, then one can choose $\alpha > 0$ such that both (69) and (70) hold, thus establishing the l_2-stability of the system. ■

Note that actually (51) is a sufficient condition for l_2-stability, and (40) is just a more conversative version of this condition with μ replaced by the bound (50).

Notes and References

The foundations of feedback stability theory were laid in the early 1960's by Sandberg and Zames. Three papers by Sandberg (1964a, 1964b, 1965) and two papers by Zames (1966a, 1966b) are cited here, but many more could have been cited. A more complete list of the early references can be found in the book by Desoer and Vidyasagar (1975). The material in Section 6.3 on the relationship between input-output and Lyapunov stability is based on Vidyasagar and Vannelli (1982); see Hill and Moylan (1980) for related results. The analysis in Section 6.4 on open-loop stability is due to several authors, but Theorem (6.4.75) is due to Willems (1969b). The graphical stability test of Theorem (6.5.35) is due to Willems (1969a); see Callier and Desoer (1972) for a more general version. The algebra $\hat{B}$ of Section 6.5.2 was introduced by Callier and Desoer (1978). The idea of coprime factorizations over $\hat{\mathbf{A}}$ is due to Vidyasagar (1972, 1975, 1978). See the book by Vidyasagar (1985) for a discussion of the stability of linear distributed feedback systems in an abstract setting. The material in Section 6.6 owes its genesis to the early work of Sandberg and Zames referred to earlier. The form of the Popov criterion given in Theorem (6.6.95) is due to Desoer (1965). The form of the passivity theorem given in Theorem (6.6.58) is due to Vidyasagar (1977); see Moylan and Hill (1978) for a more general version.

7. DIFFERENTIAL GEOMETRIC METHODS

In this chapter, we give a brief introduction to some results in nonlinear system theory obtained using methods from differential geometry. In the preceding chapters, the attention has been on the stability issue, which can broadly be described as a concern as to whether or not a given system is well-behaved in some sense. The emphasis was on avoiding detailed calculations of solutions to system equations, and on obtaining broad and nonspecific conclusions. In contrast, in the present chapter the emphasis is much more on the detailed behavior of a system: Is it *reachable*, in the sense that, starting from a given initial state, one is able to steer the system to all nearby states? Is it *observable*, in the sense that, for each state, there exists at least one corresponding input which permits us to discriminate between this state and all nearby states, by measuring only the corresponding output? Is the system *feedback linearizable*, in the sense that, by making an appropriate change of coordinates and applying a nonlinear state feedback, the equations describing the system can be made to look linear? Differential geometric methods provide a powerful means to address these and other questions. Chronologically, differential geometric methods are of even more recent vintage than input-output methods, and most of the results presented in this chapter date back no more than ten years. As such, these methods are still the subject of current research.

Most of the results presented in this chapter pertain to nonlinear systems described by a set of equations of the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t)] + \sum_{i=0}^{m} u_i(t)\, \mathbf{g}_i[\mathbf{x}(t)],$$

where $\mathbf{x}(t) \in \mathbf{R}^n$, and $\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m$ are vector fields on $\mathbf{R}^n$. (This term is defined in Section 7.1 below.) Two features of this system can be observed at once. First, the system is *time-invariant*, in that there is no explicit dependence on time. The class of systems studied in Chapter 6 can either be time-invariant or time-varying, and the relative efficacy of the analysis methods presented therein remains pretty much the same. In contrast, differential geometric methods are much more efficient when applied to time-invariant systems than to time-varying systems. Second, the system above is *linear in the control*, in contrast to a general nonlinear (time-invariant) system with n states and m inputs, which is described by

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

Again, the analysis methods of Chapter 6 apply quite well to the more general nonlinear system, whereas the methods presented in this chapter apply only to the less general class of systems. This is not a fundamental restriction of differential geometric methods. In fact, it

is possible to use differential geometric methods to study the more general class of nonlinear systems as well. However, the increase in complexity is enormous. Hence the attention in this chapter is focused on the less general, "linear in the control" class of systems, so as to keep the technicalities to a minimum.

While the class of systems studied in this chapter is more restrictive than that in Chapter 6, the analysis is more thorough. The questions of reachability, observability, and feedback linearizability are all answered. As a prelude to this, some background material in differential geometry is presented.

It should be emphasized that the contents of the present chapter represent only a small part of the results that can be found in differential geometric control theory. Also, to keep the treatment at an elementary level, a great many simplifications are made. First, it is assumed that the differential equations describing the system under study are defined on some open subset of $\mathbf{R}^n$, rather than on an abstract n-dimensional manifold. Second, all definitions and computations are carried out using so-called "local coordinates." There is a price to be paid for doing things in this way. The assumption that all the action takes place on some open subset of $\mathbf{R}^n$ rules out many interesting situations, such as systems defined on the circle or a torus. Using local coordinates to compute everything obscures one of the major motivations of modern differential geometry, which is to obtain a "coordinate-free" description of various entities. A student desirous of a more general treatment of the differential geometric approach to nonlinear control systems should consult an advanced text, such as Isidori (1989) or Nijmeijer and van der Schaft (1990).

7.1 BASICS OF DIFFERENTIAL GEOMETRY

In this section, we begin by recalling a well-known theorem from advanced calculus, namely the inverse function theorem. Then the notions of a vector field, a form, and various types of Lie derivatives are introduced.

Suppose $\mathbf{f}: \mathbf{R}^n \to \mathbf{R}^m$, and suppose each component of $\mathbf{f}$ is continuously differentiable with respect to each of its arguments. (In other words, suppose that $\mathbf{f}$ is C^1.) Then the $m \times n$ matrix whose ij-th entry is $\partial f_i / \partial x_j$ is called the **Jacobian matrix** of $\mathbf{f}$ and is denoted by $\partial \mathbf{f} / \partial \mathbf{x}$. We say that $\mathbf{f}$ is **smooth** if every component of $\mathbf{f}$ has continuous derivatives of all orders with respect to all combinations of its arguments. Suppose U, V are open subsets of $\mathbf{R}^n$ and that $\mathbf{f}: U \to V$ is C^1. Then we say that $\mathbf{f}$ is a **diffeomorphism** of U onto V if (i) $\mathbf{f}(U) = V$, (ii) $\mathbf{f}$ is one-to-one, and (iii) the inverse function $\mathbf{f}^{-1}: V \to U$ is also C^1. $\mathbf{f}$ is called a **smooth diffeomorphism** if both $\mathbf{f}$ and $\mathbf{f}^{-1}$ are smooth functions.

Theorem (1) can be found in most standard texts on advanced calculus, e.g., Royden(1963)

1 Theorem (Inverse Function Theorem) *Suppose* $\mathbf{f}: \mathbf{R}^n \to \mathbf{R}^n$ *is* C^1 *at* $\mathbf{x}_0 \in \mathbf{R}^n$, *and let* $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$. *Suppose* $[\partial \mathbf{f} / \partial \mathbf{x}]_{\mathbf{x} = \mathbf{x}_0}$ *is nonsingular. Then there exist open sets* $U \subseteq \mathbf{R}^n$ *containing* $\mathbf{x}_0$ *and* $V \subseteq \mathbf{R}^n$ *containing* $\mathbf{y}_0$ *such that* $\mathbf{f}$ *is a diffeomorphism of* U *onto* V. *If, in addition,* $\mathbf{f}$ *is smooth, then* $\mathbf{f}^{-1}$ *is also smooth, i.e.,* $\mathbf{f}$ *is a smooth diffeomorphism.*

Since smooth diffeomorphisms are the only kind of diffeomorphisms used in this book, the adjective "smooth" is often omitted.

Throughout the remainder of the chapter, X denotes an open subset of $\mathbf{R}^n$, where n is a fixed integer (known as the order of the system under study).

2 Definition *A* **vector field** *on X is a smooth function mapping X into $\mathbf{R}^n$. The set of all vector fields on X is denoted by $V(X)$. The set of all smooth real-valued functions mapping X into $\mathbf{R}$ is denoted by $S(X)$.*

Note that if $a, b \in S(X)$, then the functions $a \pm b$ and the product function ab defined by $ab: \mathbf{x} \mapsto a(\mathbf{x})b(\mathbf{x})$ are also smooth functions and belong to $S(X)$. Hence $S(X)$ is a **ring** under the usual definitions of addition and multiplication. As for the set $V(X)$, obviously it is a linear vector space over the real field. But much more is true. Suppose $a \in S(X)$ and that $\mathbf{f} \in V(X)$ (i.e., suppose a is a smooth real-valued function and that $\mathbf{f}$ is a vector field on X). Then the function mapping $\mathbf{x}$ into $a(\mathbf{x})\mathbf{f}(\mathbf{x})$ is also a vector field on X. Moreover, one can easily verify the following properties: For each $a, b \in S(X), \mathbf{f}, \mathbf{g} \in V(X)$, we have

3 $a(\mathbf{f}+\mathbf{g}) = a\mathbf{f} + a\mathbf{g}$,

$(a+b)\mathbf{f} = a\mathbf{f} + b\mathbf{f}$,

$(a \cdot b)\mathbf{f} = a \cdot (b\mathbf{f})$.

These show that $V(X)$ is a **module** over the ring $S(X)$, but this terminology is not used further in this book.

4 Definition *A* **form** *on X is a smooth function mapping X into $(\mathbf{R}^n)^*$, which is the set of $1 \times n$ row vectors. The set of all forms on X is denoted by $F(X)$.*

Note that it is customary to write vector fields as column vectors, and forms as row vectors.

5 Examples Let $X = \mathbf{R}^2$. Then the following are vector fields on X:

$$\begin{bmatrix} x_1^2 - 2x_2 \\ x_1 + x_2^3 \end{bmatrix}, \begin{bmatrix} \cosh x_1 \\ \exp(x_1 + x_2^2) \end{bmatrix}.$$

The following is *not* a (smooth) vector field, since it does not have continuous derivatives of *all* orders. (Note that the first component has continuous derivatives of order two or less, but the third derivative with respect to x_2 is not continuous at $x_2 = 0$.)

$$\begin{bmatrix} x_1 - 2x_2^3 \\ x_2 \end{bmatrix} \text{ if } x_2 \geq 0, \begin{bmatrix} x_1 - x_2^4 \\ x_2 \end{bmatrix} \text{ if } x_2 < 0.$$

The following are examples of forms: Note that a form is written as a row vector whereas a vector field is written as a column vector.

$$[\sin(x_1 - 3x_2^3) \quad x_1^2 + x_2],\ [\exp(x_2 - 2x_1^2) \quad x_1 - \tanh x_2].$$

Next we define various operations involving vector fields, forms, and smooth real-valued functions.

Suppose $\mathbf{x}_0 \in X$ is given. A **curve** in X passing through $\mathbf{x}_0$ is a smooth function $\mathbf{c}$ mapping some open interval $(-\alpha, \beta)$ containing 0 into X, such that $\mathbf{c}(0) = \mathbf{x}_0$. Suppose $\mathbf{f}$ is a vector field on X and that $\mathbf{x}_0 \in X$ is given. Then by the contents of Chapter 2, especially Corollary (2.4.22), we know that there exists a unique solution of the differential equation

$$\mathbf{6} \qquad \frac{d}{dt}\mathbf{x}(t) = \mathbf{f}[\mathbf{x}(t)],\ \mathbf{x}(0) = \mathbf{x}_0$$

for sufficiently small values of t. Viewed as a function of t, the solution $\mathbf{x}(\cdot)$ defines a curve passing through $\mathbf{x}_0$; it is called the **integral curve** of $\mathbf{f}$ passing through $\mathbf{x}_0$, and is denoted by $\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)$. (Note that this is a minor variation of the notation employed in Chapter 5.) Note that, for each fixed t, $\mathbf{s}_{\mathbf{f},t}$ maps X into itself. Moreover, by Theorem (2.4.57), $\mathbf{s}_{\mathbf{f},t}$ is locally a (smooth) diffeomorphism.

For later purposes, it is useful to define the transformation of a vector field under a change of coordinates. Suppose $\mathbf{f} \in V(X)$, that $\mathbf{x}_0 \in X$ is given, and T is a (smooth) diffeomorphism in some neighbourhood of $\mathbf{x}_0$. Now suppose we make a change of coordinates $\mathbf{y} = T(\mathbf{x})$. What do the integral curves of $\mathbf{f}$ through $\mathbf{x}_0$ look like in the new coordinates? Suppose $\mathbf{x}(t)$ satisfies (6). Then, letting $\mathbf{J}$ denote the Jacobian matrix of T, $\mathbf{y}(t) = T[\mathbf{x}(t)]$ satisfies

$$\mathbf{7}' \qquad \dot{\mathbf{y}}(t) = \mathbf{J}[\mathbf{x}(t)]\,\mathbf{f}[\mathbf{x}(t)] = [\mathbf{Jf}][T^{-1}\mathbf{y}(t)].$$

Thus in the new coordinates the vector field $\mathbf{f}$ looks like

$$\mathbf{8} \qquad \mathbf{f}_T(\mathbf{y}) = \mathbf{J}[T^{-1}(\mathbf{y})]\,\mathbf{f}[T^{-1}(\mathbf{y})].$$

This can be compactly expressed as

$$\mathbf{9} \qquad \mathbf{f}_T = (\mathbf{Jf}) \cdot T^{-1}.$$

This notation means: Given an argument $\mathbf{y}$, first find its preimage $T^{-1}\mathbf{y}$, and then evaluate the function $\mathbf{Jf}$ at $T^{-1}\mathbf{y}$. Thus changing coordinates from $\mathbf{x}$ to $\mathbf{y} = T(\mathbf{x})$ does not merely result in $\mathbf{f}(\mathbf{x})$ getting replaced by $\mathbf{f}[T^{-1}(\mathbf{y})]$; it is also necessary to premultiply by $\mathbf{J}[T^{-1}(\mathbf{y})]$ so as to ensure that the integral curve of $\mathbf{f}_T$ in the $\mathbf{y}$ coordinates corresponds exactly to the integral curve of $\mathbf{f}$ in the $\mathbf{x}$ coordinates. An equivalent way of expressing this is as follows:

$$\mathbf{10} \qquad \mathbf{s}_{\mathbf{f},t} = T^{-1}\mathbf{s}_{\mathbf{f}_T,t}T.$$

This equation means: Choose an $\mathbf{x}_0 \in X$; first apply T (thus changing to the $\mathbf{y}$ coordinates); then follow the integral curve of the transformed vector field $\mathbf{f}_T$ passing through $T(\mathbf{x}_0)$; finally apply T^{-1} (thus changing back to the $\mathbf{x}$ coordinates). The answer is the same as following the integral curve of $\mathbf{f}$ through $\mathbf{x}_0$. Finally, note that since T is a diffeomorphism, one

can also write (10) as

11 $T\mathbf{s}_{\mathbf{f},t} = \mathbf{s}_{\mathbf{f}_T,t} T.$

12 Example A scalar example is enough to make this point clear. Suppose $X = \mathbf{R}$ and let f be the vector field

$$f(x) = x^2 + 1.$$

Suppose $x_0 = 1$. Then the integral curve of f passing through x_0 is given by

13 $x(t) = \tan(t + \pi/4).$

Of course, $x(t)$ is well-defined only for sufficiently small t. Now suppose we make the change of coordinates

$$y = x^2 + 1 =: T(x),$$

which is a local diffeomorphism around x_0 (but not a global diffeomorphism). Then, in terms of y we have

$$f(x) = x^2 + 1 = y.$$

Also, $T(x_0) = 2$. Now, if we solve the equation

$$\dot{y}(t) = y(t),\ y(0) = 2,$$

we get $y(t) = 2\exp(t)$, which does not match with (13). On the other hand, let us define the transformed vector field $f_T(y)$ in accordance with (9). This gives

$$J(x) = 2x,\ T^{-1}(y) = (1-y)^{1/2},$$

$$f_T(y) = 2x \cdot (x^2 + 1)\,|_{x=(1-y)^{1/2}} = 2y\,(1-y)^{1/2}.$$

If we solve the equation

$$\dot{y}(t) = 2y(t)\,[1 - y(t)]^{1/2},\ y(0) = 2,$$

we get

$$y(t) = \sec^2(t + \pi/4),$$

which is just the solution (13) transformed into the y coordinate.

14 Example As a more familiar illustration of the vector field transformation (9), suppose both $\mathbf{f}$ and T are *linear*, i.e., suppose

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x},\ T(\mathbf{x}) = \mathbf{M}\mathbf{x},$$

where $\mathbf{A}, \mathbf{M}$ are both $n \times n$ matrices, and $\mathbf{M}$ is nonsingular. Then (9) becomes

$$\mathbf{f}_T(\mathbf{x}) = \mathbf{M}\mathbf{A}\mathbf{M}^{-1}\mathbf{x},$$

which is just the usual similarity transformation. ■

In this book we have made a deliberate choice to avoid the "modern" coordinate-free approach to differential geometry, and to compute everything explicitly. But this is a good place to point out one of the advantages of the modern approach. In it, both $\mathbf{f}$ and $\mathbf{f}_T$ represent the *same* vector field, but expressed with respect to different coordinate system; thus one would make a distinction between an abstract vector field, and its concrete representation with respect to a specific coordinate system. This is similar to the distinction made in linear algebra between an abstract linear map and its concrete (matrix) representation with respect to a specific basis [cf. Example (14)]. In the long run, the modern approach is cleaner conceptually, though it does require a higher level of abstraction.

Suppose $a \in S(X)$, i.e., a is a smooth real-valued function on X. Then its gradient, denoted by ∇a or $\mathbf{d}a$, is defined as the row vector $[\partial a/\partial x_1 \cdots \partial a/\partial x_n]$. Note that $\mathbf{d}a$ is a form on X. Now suppose $\mathbf{f} \in V(X)$. Then the map

$$\mathbf{15} \quad \mathbf{x} \mapsto \mathbf{d}a(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) : X \to \mathbf{R}$$

is smooth; it is called the **Lie derivative** of the function a with respect to the vector field $\mathbf{f}$, and is denoted by $L_{\mathbf{f}}a$. Note that $L_{\mathbf{f}}a \in S(X)$. Now suppose $\mathbf{h} \in F(X)$, i.e., $\mathbf{h}$ is a form on X. Then the map

$$\mathbf{16} \quad \mathbf{x} \mapsto \mathbf{h}(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x})$$

is smooth and real-valued. It is denoted by $<\mathbf{h}, \mathbf{f}>$ and belongs to $S(X)$.

The Lie derivative $L_{\mathbf{f}}a$ can be interpreted as the derivative of a along integral curves of the vector field $\mathbf{f}$. Notice that

$$\mathbf{17} \quad (L_{\mathbf{f}}a)(\mathbf{x}) = \lim_{t \to 0} \frac{1}{t} \{a[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)] - a(\mathbf{x}_0)\}.$$

We have already encountered this concept in Chapter 5 in connection with taking the derivative of a Lyapunov function candidate along the solution trajectories of a particular differential equation. Note that, given the differential equation (6) and a Lyapunov function candidate a, the derivative $\dot{a}$ defined according to Definition (5.2.23) is precisely $L_{\mathbf{f}}a$. Hence there is no need to give examples of the computation of the Lie derivative of a smooth function. As for the quantity $<\mathbf{h}, \mathbf{f}>$, this is called the inner product of the form $\mathbf{h}$ and the vector field $\mathbf{f}$. It is nothing more than the scalar product of the row vector $\mathbf{h}$ and the column vector $\mathbf{f}$.

A form $\mathbf{h}$ is called **exact** if there exists a smooth function $a \in S(X)$ such that $\mathbf{h} = \mathbf{d}a$. In an abstract setting, it is not always easy to determine whether a given form is exact or not. This is true even if X is restricted to be an open subset of $\mathbf{R}^n$, as is done here. However, if X is an *open ball* in $\mathbf{R}^n$, i.e., a set of the form $\{\mathbf{x}: \|\mathbf{x} - \mathbf{x}_0\| < \varepsilon\}$ for some $\mathbf{x}_0$ and ε, then it is easy to determine whether a given form $\mathbf{h} \in F(X)$ is exact: Form the $n \times n$ matrix $J_{\mathbf{h}}$ whose ij-th element is $\partial h_j / \partial x_i$; then $\mathbf{h}$ is exact if and only if $J_{\mathbf{h}}$ is a symmetric matrix for all $\mathbf{x} \in X$.

Next we define a very important concept.

18 Definition *Suppose* $\mathbf{f}, \mathbf{g} \in V(X)$. *Then the* **Lie Bracket** *of* $\mathbf{f}$ *and* $\mathbf{g}$ *is denoted by* $[\mathbf{f}, \mathbf{g}]$, *and is the vector field defined by*

$$\mathbf{19} \quad [\mathbf{f}, \mathbf{g}] = \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \cdot \mathbf{f} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \cdot \mathbf{g}.$$

20 Example Let $X = \mathbf{R}^2$ as before, and suppose $\mathbf{f}, \mathbf{g} \in V(X)$ are given by

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_1 - x_2^2 \\ x_1 x_2 \end{bmatrix}, \quad \mathbf{g}(\mathbf{x}) = \begin{bmatrix} \sin(x_1 + x_2) \\ \cos(x_1 - x_2) \end{bmatrix}.$$

Then routine computations show that

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} 1 & -2x_2 \\ x_2 & x_1 \end{bmatrix}, \quad \frac{\partial \mathbf{g}}{\partial \mathbf{x}} = \begin{bmatrix} \cos(x_1 + x_2) & \cos(x_1 + x_2) \\ -\sin(x_1 - x_2) & \sin(x_1 - x_2) \end{bmatrix},$$

and

$$\begin{aligned} [\mathbf{f}, \mathbf{g}] &= \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \cdot \mathbf{f} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \cdot \mathbf{g} \\ &= \begin{bmatrix} (x_1 + x_1 x_2 - x_2^2)\cos(x_1 + x_2) - \sin(x_1 + x_2) + 2x_2 \cos(x_1 - x_2) \\ (-x_1 + x_1 x_2 + x_2^2)\sin(x_1 - x_2) - x_2 \sin(x_1 + x_2) - x_1 \cos(x_1 - x_2) \end{bmatrix}. \end{aligned}$$

The next several lemmas bring out several useful geometric interpretations of the Lie bracket.

21 Lemma *Suppose* $\mathbf{f}, \mathbf{g}$ *are vector fields on* X, *and let* $\mathbf{x}_0 \in X$. *Then*

$$\mathbf{22} \quad [\mathbf{f}, \mathbf{g}](\mathbf{x}_0) = \lim_{t \to 0} \frac{1}{t} \left\{ \frac{\partial}{\partial \mathbf{x}} [\mathbf{s}_{\mathbf{f},-t}(\mathbf{x})]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)} \cdot \mathbf{g}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)] - \mathbf{g}(\mathbf{x}_0) \right\}.$$

Remarks Unfortunately Lemma (21) makes less sense in the "local coordinates" that we are using than it does in a global or coordinate-free setting. In essence, Lemma (21) states that the Lie bracket $[\mathbf{f}, \mathbf{g}]$ can be thought of as a directional derivative of the vector field $\mathbf{g}$ along integral curves of $\mathbf{f}$, in much the same way that $L_{\mathbf{f}} a$ is the directional derivative of the real-valued function a along the integral curves of $\mathbf{f}$ [see (17)]. However, in the case

of an abstract manifold, one cannot just subtract the vectors $\mathbf{g}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)]$ and $\mathbf{g}(\mathbf{x}_0)$, since they "live" in distinct tangent spaces. The extra factor $\frac{\partial}{\partial \mathbf{x}}[\mathbf{s}_{\mathbf{f},-t}(\mathbf{x})]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)}$ "pulls back" the vector $\mathbf{g}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)]$ into the tangent space of X at $\mathbf{x}_0$; the need for this factor can only be appreciated in a more abstract setting.

Proof Since we are taking limits as $t \to 0$, it is only necessary to compute the various quantities inside the braces up to a first order term in t. For small t, we have

$$\mathbf{23} \quad \mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0) = \mathbf{x}_0 + t\mathbf{f}(\mathbf{x}_0) + \mathbf{o}(t).$$

Hence

$$\mathbf{24} \quad \mathbf{g}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)] = \mathbf{g}[\mathbf{x}_0 + t\mathbf{f}(\mathbf{x}_0)] + \mathbf{o}(t) = \mathbf{g}(\mathbf{x}_0) + t\left[\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right]_{\mathbf{x}_0} \cdot \mathbf{f}(\mathbf{x}_0) + \mathbf{o}(t).$$

Again, from (23),

$$\mathbf{25} \quad \mathbf{s}_{\mathbf{f},-t}(\mathbf{x}) = \mathbf{x} - t\mathbf{f}(\mathbf{x}) + \mathbf{o}(t),$$

so that

$$\mathbf{26} \quad \frac{\partial \mathbf{s}_{\mathbf{f},-t}(\mathbf{x})}{\partial \mathbf{x}} = I - t\frac{\partial \mathbf{f}}{\partial \mathbf{x}} + \mathbf{O}(t).$$

Hence

$$\mathbf{27} \quad \left[\frac{\partial \mathbf{s}_{\mathbf{f},-t}(\mathbf{x})}{\partial \mathbf{x}}\right]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)} = I - t\left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)} + \mathbf{O}(t) = I - t\left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{x}_0} + \mathbf{O}(t),$$

since $\mathbf{s}_{\mathbf{f},-t}(\mathbf{x}_0) = \mathbf{x}_0$ to zeroth order in t. Substituting all this in (22) and gathering terms shows that the right side of (22) equals

$$\mathbf{28} \quad \lim_{t \to 0} \frac{1}{t}\left\{[I - t\frac{\partial \mathbf{f}}{\partial \mathbf{x}}]\cdot[\mathbf{g} + t\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f}] - \mathbf{g}\right\} = \frac{\partial \mathbf{g}}{\partial \mathbf{x}}\,\mathbf{f} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\,\mathbf{g} = [\mathbf{f}, \mathbf{g}],$$

where the argument $\mathbf{x}_0$ has been suppressed for clarity. ■

29 Lemma *Suppose* $\mathbf{f}, \mathbf{g} \in V(X)$, *and let* $\mathbf{x}_0 \in X$. *Then*

$$\mathbf{30} \quad [\mathbf{f}, \mathbf{g}](\mathbf{x}_0) = \lim_{t \to 0} \frac{1}{t^2}\{[\mathbf{s}_{\mathbf{g},-t} \cdot \mathbf{s}_{\mathbf{f},-t} \cdot \mathbf{s}_{\mathbf{g},t} \cdot \mathbf{s}_{\mathbf{f},t}](\mathbf{x}_0) - \mathbf{x}_0\}$$

Remarks Lemma (29) gives yet another interpretation of the Lie bracket. Suppose we start at a point $\mathbf{x}_0$, and follow the integral curve of the vector field $\mathbf{f}$ for a very short time t; then, from that point, follow the integral curve of $\mathbf{g}$ for a duration t; then, from that point,

follow the integral curve of $\mathbf{f}$ *backwards in time* for a duration t; and finally, from that point, follow the integral curve of $\mathbf{g}$ *backwards in time* for a duration t. Where do we end up? Well, to first order in t, we get back to the point $\mathbf{x}_0$. However, to *second* order in t, we end up at $\mathbf{x}_0 + t^2[\mathbf{f}, \mathbf{g}](\mathbf{x}_0)$. This is just what Lemma (29) states. Note that the map $\mathbf{s}_{\mathbf{f},-t}$ is the inverse of the map $\mathbf{s}_{\mathbf{f},t}$, and similarly $\mathbf{s}_{\mathbf{g},-t}$ is the inverse of the map $\mathbf{s}_{\mathbf{g},t}$. Hence, if the maps $\mathbf{s}_{\mathbf{f},t}$ and $\mathbf{s}_{\mathbf{g},t}$ commute (i.e., $\mathbf{s}_{\mathbf{f},t}\mathbf{s}_{\mathbf{g},t} = \mathbf{s}_{\mathbf{g},t}\mathbf{s}_{\mathbf{f},t}$ for all sufficiently small t), then the limit in (30) is zero. Thus one can think of the Lie bracket $[\mathbf{f}, \mathbf{g}]$ as a measure of the extent to which the "solution" maps $\mathbf{s}_{\mathbf{f},t}$ and $\mathbf{s}_{\mathbf{g},t}$ fail to commute. This point is developed further in Lemma (40) below.

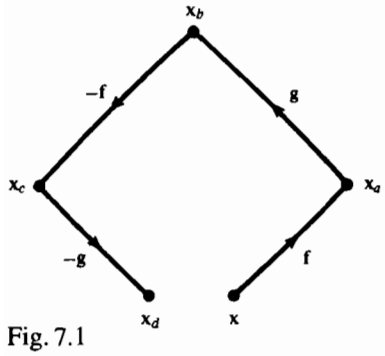


Fig. 7.1

Proof The proof is quite routine, though tedious. To follow it, refer to Figure 7.1. To compute the limit in (30), it is necessary to compute the various quantities up to second order in t. By definition,

$$\mathbf{31} \quad \frac{d}{dt}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)] = \mathbf{f}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)].$$

Hence

$$\mathbf{32} \quad \frac{d}{dt}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)]_{t=0} = \mathbf{f}(\mathbf{x}_0).$$

Next, (31) implies that

$$\mathbf{33} \quad \frac{d^2}{dt^2}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)]_{t=0} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \cdot \mathbf{f}(\mathbf{x}_0).$$

Now, by Taylor series,

$$\mathbf{34} \quad \mathbf{x}_a = \mathbf{x}_0 + t\mathbf{f}(\mathbf{x}_0) + \frac{t^2}{2}\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \cdot \mathbf{f}(\mathbf{x}_0) + \mathbf{o}(t^2).$$

Next, in analogy with (34),

$$\mathbf{35}\quad \mathbf{x}_b = \mathbf{x}_a + t\mathbf{g}(\mathbf{x}_a) + \frac{t^2}{2}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\cdot\mathbf{g}(\mathbf{x}_a) + \mathbf{o}(t^2).$$

Let us now substitute for $\mathbf{x}_a$ from (34) and simplify by neglecting all terms which are of order higher than t^2. Thus it is only necessary to estimate $\mathbf{g}(\mathbf{x}_a)$ to first order in t since it is multiplied by t, and it is safe to replace $[\partial \mathbf{g}/\partial \mathbf{x}](\mathbf{x}_a)\cdot\mathbf{g}(\mathbf{x}_a)$ by $[\partial \mathbf{g}/\partial \mathbf{x}](\mathbf{x}_0)\cdot\mathbf{g}(\mathbf{x}_0)$, since this term is multiplied by $t^2/2$. Since

$$\mathbf{36}\quad \mathbf{g}(\mathbf{x}_a) = \mathbf{g}(\mathbf{x}_0) + t\frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}_0)\,\mathbf{f}(\mathbf{x}_0) + \mathbf{o}(t),$$

(35) and (36) imply that

$$\mathbf{37}\quad \mathbf{x}_b = \mathbf{x}_0 + t[\mathbf{f}(\mathbf{x}_0) + \mathbf{g}(\mathbf{x}_0)] + \mathbf{o}(t^2)$$
$$+ t^2\left[\frac{1}{2}\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_0)\cdot\mathbf{f}(\mathbf{x}_0) + \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}_0)\cdot\mathbf{f}(\mathbf{x}_0) + \frac{1}{2}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}_0)\cdot\mathbf{g}(\mathbf{x}_0)\right].$$

The process is now repeated, and the results are shown below; the argument is $\mathbf{x}_0$ unless indicated otherwise.

$$\mathbf{38}\quad \mathbf{x}_c = \mathbf{x}_b - t\mathbf{f}(\mathbf{x}_b) + \frac{t^2}{2}\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_b)\cdot\mathbf{f}(\mathbf{x}_b) + \mathbf{o}(t^2)$$
$$= \mathbf{x}_0 + t\mathbf{g}(\mathbf{x}_0) + t^2\left[\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\cdot\mathbf{f} + \frac{1}{2}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\cdot\mathbf{g} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\cdot\mathbf{g}\right] + \mathbf{o}(t^2),$$

$$\mathbf{39}\quad \mathbf{x}_d = \mathbf{x}_c - t\mathbf{g}(\mathbf{x}_c) + \frac{t^2}{2}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}_c)\cdot\mathbf{g}(\mathbf{x}_c) + \mathbf{o}(t^2)$$
$$= \mathbf{x}_0 + t^2\left[\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\mathbf{g}\right] + \mathbf{o}(t^2).$$

This completes the proof. ■

In the remarks prior to the proof of Lemma (29), it was stated that $[\mathbf{f}, \mathbf{g}]$ is a measure of the extent to which the solution or "flow" maps $\mathbf{s}_{\mathbf{f},t}$ and $\mathbf{s}_{\mathbf{g},t}$ fail to commute. The next result sheds more light on the relationship.

40 Lemma *Suppose* $\mathbf{f}, \mathbf{g} \in V(X)$. *Then*

$$\mathbf{41}\quad [\mathbf{f}, \mathbf{g}] = \mathbf{0} \text{ iff } \mathbf{s}_{\mathbf{f},t}\,\mathbf{s}_{\mathbf{g},\tau} = \mathbf{s}_{\mathbf{g},\tau}\,\mathbf{s}_{\mathbf{f},t},\ \forall t, \tau \text{ sufficiently small.}$$

Lemma (40) states that the Lie bracket of two vector fields $\mathbf{f}$ and $\mathbf{g}$ is *identically* zero if and only if the solution maps $\mathbf{s}_{\mathbf{f},t}$ and $\mathbf{s}_{\mathbf{g},\tau}$ commute for all sufficiently small t, τ. Actually, it is easy to see that if the commutativity relationship (41) holds for all sufficiently small t, τ, then in fact the same relationship holds for *all* t, τ for which the solution maps are defined.

Proof "If" Follows directly from Lemma (29).

"Only if" Let $\mathbf{x}_0 \in X$ be arbitrary, and define

$$\mathbf{42} \quad \mathbf{c}(t) = \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t}(\mathbf{x})\, \mathbf{g}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)}.$$

Then $\mathbf{c}$ is also a vector field. In fact, comparing with (9), we see that it is the transformed version of $\mathbf{g}$ under the diffeomorphism $\mathbf{s}_{\mathbf{f},-t}$. Let us compute the time derivative of $\mathbf{c}(t)$. By definition,

$$\mathbf{43} \quad \dot{\mathbf{c}}(t) = \lim_{\tau \to 0} \frac{1}{\tau} [\mathbf{c}(t+\tau) - \mathbf{c}(t)]$$

$$= \lim_{\tau \to 0} \left\{ \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t-\tau}(\mathbf{x})\, \mathbf{g}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},t+\tau}(\mathbf{x}_0)} - \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t}(\mathbf{x})\, \mathbf{g}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)} \right\}.$$

Now let us observe that

$$\mathbf{44} \quad \mathbf{s}_{\mathbf{f},-t-\tau} = \mathbf{s}_{\mathbf{f},-t}\, \mathbf{s}_{\mathbf{f},-\tau}.$$

In other words, $\mathbf{s}_{\mathbf{f},-t-\tau}$ is the composition of $\mathbf{s}_{\mathbf{f},-\tau}$ followed by $\mathbf{s}_{\mathbf{f},-t}$. Hence, by the chain rule,

$$\mathbf{45} \quad \frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t-\tau}(\mathbf{x}) = \left[\frac{\partial}{\partial \mathbf{y}} \mathbf{s}_{\mathbf{f},-t}(\mathbf{y}) \right]_{\mathbf{s}_{\mathbf{f},-\tau}(\mathbf{x})} \frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-\tau}(\mathbf{x}).$$

Now define

$$\mathbf{46} \quad \mathbf{z} = \mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0),$$

and apply (45) in (43). Using the fact that

$$\mathbf{47} \quad \mathbf{s}_{\mathbf{f},t+\tau}(\mathbf{x}_0) = \mathbf{s}_{\mathbf{f},\tau}(\mathbf{z})$$

in (45) gives

$$\mathbf{48} \quad \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t-\tau}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},t+\tau}(\mathbf{x}_0)} = \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t}(\mathbf{x}) \right]_{\mathbf{z}} \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-\tau}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},\tau}(\mathbf{z})}.$$

Similarly

$$\mathbf{49} \quad \mathbf{g}[\mathbf{s}_{\mathbf{f},t+\tau}(\mathbf{x}_0)] = \mathbf{g}[\mathbf{s}_{\mathbf{f},\tau}(\mathbf{z})].$$

Substituting from (48) and (49) into (43) gives

$$\mathbf{50} \quad \dot{\mathbf{c}}(t) = \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t}(\mathbf{x}) \right]_{\mathbf{z}} \cdot \lim_{\tau \to 0} \frac{\mathbf{d}(\tau)}{\tau},$$

where

$$\mathbf{51} \quad \mathbf{d}(\tau) = \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-\tau}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},\tau}(\mathbf{z})} \mathbf{g}[\mathbf{s}_{\mathbf{f},\tau}(\mathbf{z})] - \mathbf{g}(\mathbf{z}).$$

But by Lemma (21),

$$\mathbf{52} \quad \lim_{\tau \to 0} \frac{\mathbf{d}(\tau)}{\tau} = [\mathbf{f}, \mathbf{g}](\mathbf{z}) = \mathbf{0}.$$

Hence $\dot{\mathbf{c}} \equiv \mathbf{0}$, which means that $\mathbf{c}(t) = \mathbf{c}(0)$ for all t (when it is defined).

To complete the proof, note that $\mathbf{s}_{\mathbf{f},0}(\mathbf{x}) = \mathbf{x}$. Hence

$$\mathbf{53} \quad \frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},0}(\mathbf{x}) = I, \ \forall \mathbf{x}.$$

As a consequence, from (42), $\mathbf{c}(0) = \mathbf{g}(\mathbf{x}_0)$. Coupled with the fact that $\dot{\mathbf{c}} \equiv \mathbf{0}$, this implies that $\mathbf{c}(t) \equiv \mathbf{g}(\mathbf{x}_0)\ \forall t$. From the definition (42) of $\mathbf{c}(t)$, this means that

$$\mathbf{54} \quad \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},-t}(\mathbf{x})\, \mathbf{g}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)} = \mathbf{g}(\mathbf{x}_0), \ \forall t.$$

Now we make use of the vector field transformation formula (9). Fix $t \in \mathbf{R}$, and note that $\mathbf{s}_{\mathbf{f},-t}$ is a local diffeomorphism around $\mathbf{x}_0$. Apply the formula (9), with $\mathbf{f}$ replaced by $\mathbf{g}$, and T replaced by $\mathbf{s}_{\mathbf{f},-t}$ [and note that $(\mathbf{s}_{\mathbf{f},-t})^{-1} = \mathbf{s}_{\mathbf{f},t}$]. Then (9) shows that

$$\mathbf{55} \quad \mathbf{g}_T(\mathbf{x}) = \left[\mathbf{s}_{\mathbf{f},-t}(\mathbf{x})\, \mathbf{g}(\mathbf{x}) \right]_{\mathbf{s}_{\mathbf{f},t}(\mathbf{x})} = \mathbf{g}(\mathbf{x}),$$

where the last step follows from (54). (Recall that $\mathbf{x}_0$ is arbitrary.) This means that the vector field $\mathbf{g}$ remains *invariant* under the diffeomorphism $\mathbf{s}_{\mathbf{f},t}$. To put it another way, if $[\mathbf{f}, \mathbf{g}] = \mathbf{0}$, then the vector field $\mathbf{g}$ remains invariant under the flow of the vector field $\mathbf{f}$. Now apply (11) with $\mathbf{f}$ and $\mathbf{f}_T$ replaced by $\mathbf{g}$, T replaced by $\mathbf{s}_{\mathbf{f},-t}$, and the time variable t replaced by τ. This gives

$$\mathbf{56} \quad \mathbf{s}_{\mathbf{f},-t}\, \mathbf{s}_{\mathbf{g},\tau} = \mathbf{s}_{\mathbf{g},\tau}\, \mathbf{s}_{\mathbf{f},-t}.$$

Pre- and post-multiplying both sides by $\mathbf{s}_{\mathbf{f},t}$ gives

$$\mathbf{57} \quad \mathbf{s}_{\mathbf{g},\tau}\, \mathbf{s}_{\mathbf{f},t} = \mathbf{s}_{\mathbf{f},t}\, \mathbf{s}_{\mathbf{g},\tau}.$$

This completes the proof of the "if" part. ■

One last question before we move on to other topics. What happens to Lie brackets when we change coordinates? Suppose $\mathbf{f}(\mathbf{x})$, $\mathbf{g}(\mathbf{x})$ are two given vector fields, and we make a coordinate change $\mathbf{y} = T(\mathbf{x})$. Then, as discussed above, $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ get transformed into $\mathbf{f}_T(\mathbf{y})$ and $\mathbf{g}_T(\mathbf{y})$ respectively. Now one can compute the Lie bracket of the vector fields either before or after the coordinate change. Do both procedures give the same answer? In other words, is it true that

$$\mathbf{58} \quad \{\mathbf{J}(\mathbf{x})[\mathbf{f}, \mathbf{g}](\mathbf{x})\}_{\mathbf{x}=T^{-1}(\mathbf{y})} = \frac{\partial \mathbf{g}_T(\mathbf{y})}{\partial \mathbf{y}} \mathbf{f}_T(\mathbf{y}) - \frac{\partial \mathbf{f}_T(\mathbf{y})}{\partial \mathbf{y}} \mathbf{g}_T(\mathbf{y})?$$

The reader should not be surprised to learn that the answer is yes. One can of course verify (58) directly by substituting for the various quantities. But a more "modern" reason for believing (58) is to note that the Lie bracket of two vector fields is defined in terms of the behavior of certain integral curves [see Lemma (29)], and the transformations of the vector fields, from $\mathbf{f}$ and $\mathbf{g}$ to $\mathbf{f}_T$ and $\mathbf{g}_T$ respectively, are intended precisely to ensure that the integral curves match in the two coordinate systems.

Recall that if $\mathbf{f}$ is a vector field on X and a is a smooth real-valued function, then $L_{\mathbf{f}}a$ is also a smooth real-valued function defined by (15). The next lemma relates repeated Lie derivatives to the Lie bracket.

59 Lemma *Suppose* $a \in S(X)$ *and* $\mathbf{f}, \mathbf{g} \in V(X)$. *Then*

$$\mathbf{60} \quad L_{[\mathbf{f}, \mathbf{g}]}a = L_{\mathbf{f}}(L_{\mathbf{g}}a) - L_{\mathbf{g}}(L_{\mathbf{f}}a).$$

Proof The result is established via routine though lengthy computations. By definition,

$$\mathbf{61} \quad (L_{\mathbf{f}}L_{\mathbf{g}}a)(\mathbf{x}) = \nabla(L_{\mathbf{g}}a)(\mathbf{x})\,\mathbf{f}(\mathbf{x}).$$

Hence, it is useful to compute $\nabla(L_{\mathbf{g}}a)$. Now

$$\mathbf{62} \quad (L_{\mathbf{g}}a)(\mathbf{x}) = \nabla a(\mathbf{x})\,\mathbf{g}(\mathbf{x}) = \sum_{j=1}^{n} \frac{\partial a}{\partial x_j} \cdot g_j.$$

Therefore

$$\mathbf{63} \quad \nabla(L_{\mathbf{g}}a)_i = \frac{\partial}{\partial x_i}(L_{\mathbf{g}}a)(\mathbf{x}) = \sum_{j=1}^{n} \left[\frac{\partial^2 a}{\partial x_i \partial x_j} g_j + \frac{\partial a}{\partial x_j} \frac{\partial g_j}{\partial x_i} \right].$$

This can be concisely expressed. Define $\nabla^2 a$ to be the $n \times n$ matrix whose ij-th element is $\partial^2 a / \partial x_i \partial x_j$. This matrix $\nabla^2 a$ is called the **Hessian matrix** of a. Note that $\nabla^2 a$ is symmetric. Now (63) can be expressed as

64 $$\nabla(L_{\mathbf{g}}a) = \mathbf{g}'\,\nabla^2 a + \nabla a\,\frac{\partial \mathbf{g}}{\partial \mathbf{x}}.$$

Therefore

65 $$L_{\mathbf{f}}L_{\mathbf{g}}a = \nabla(L_{\mathbf{g}}a)\,\mathbf{f} = \mathbf{g}'\,\nabla^2 a\,\mathbf{f} + \nabla a\,\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\,\mathbf{f},$$ and

66 $$L_{\mathbf{f}}L_{\mathbf{g}}a - L_{\mathbf{g}}L_{\mathbf{f}}a = \mathbf{g}'\,\nabla^2 a\,\mathbf{f} - \mathbf{f}'\,\nabla^2 a\,\mathbf{g} + \nabla a\,(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\mathbf{g}).$$

However, since $\nabla^2 a$ is a symmetric matrix, we have

67 $$\mathbf{g}'\,\nabla^2 a\,\mathbf{f} = \mathbf{f}'\,\nabla^2 a\,\mathbf{g},$$

and so

68 $$L_{\mathbf{f}}L_{\mathbf{g}}a - L_{\mathbf{g}}L_{\mathbf{f}}a = \nabla a\,(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\mathbf{g}) = L_{[\mathbf{f},\,\mathbf{g}]}a.$$

This completes the proof. ■

It is possible to prove a more general result than Lemma (59), using the concept of the Lie derivative of a form with respect to a vector field.

69 **Definition** *Suppose* $\mathbf{f} \in V(X)$ *and that* $\mathbf{h} \in F(X)$. *Then the* **Lie derivative** *of* $\mathbf{h}$ *with respect to* $\mathbf{f}$ *is also a form, and is defined by*

70 $$L_{\mathbf{f}}\mathbf{h} = \mathbf{f}'\left[\frac{\partial \mathbf{h}'}{\partial \mathbf{x}}\right]' + \mathbf{h}\frac{\partial \mathbf{f}}{\partial \mathbf{x}}.$$

Note that $\mathbf{h}'$ is a column vector, and that $\partial \mathbf{h}'/\partial \mathbf{x}$ is just the usual Jacobian matrix of $\mathbf{h}'$. So far we have defined three types of Lie derivatives: Suppose $\mathbf{f}, \mathbf{g} \in V(X)$, $a \in S(X)$, and $\mathbf{h} \in F(X)$. Then the Lie derivative of the vector field $\mathbf{g}$ with respect to $\mathbf{f}$ is just the Lie bracket $[\mathbf{f}, \mathbf{g}]$. The Lie derivative of the smooth function a with respect to $\mathbf{f}$ is defined in (15) as $\nabla a\mathbf{f}$. The Lie derivative of the form $\mathbf{h}$ with respect to $\mathbf{f}$ is given by (70). Note that the Lie derivatives of a vector field, a real-valued function and a form are again respectively a vector field, a real-valued function and a form. These derivatives can be related via a Leibniz type of product formula.

71 **Lemma** *Suppose* $\mathbf{f}, \mathbf{g} \in V(X)$ *and* $\mathbf{h} \in F(X)$. *Then*

72 $$L_{\mathbf{f}}\langle \mathbf{h}, \mathbf{g}\rangle = \langle L_{\mathbf{f}}\mathbf{h}, \mathbf{g}\rangle + \langle \mathbf{h}, L_{\mathbf{f}}\mathbf{g}\rangle.$$

Proof As usual the proof follows just by substituting for the various expressions and clearing terms. Note that $\langle \mathbf{h}, \mathbf{g}\rangle(\mathbf{x})$ is just $\mathbf{h}(\mathbf{x})\,\mathbf{g}(\mathbf{x})$. Thus

$$\mathbf{73}\quad [\nabla\langle \mathbf{h}, \mathbf{g}\rangle]_i = \frac{\partial}{\partial x_i}[\mathbf{h}(\mathbf{x})\,\mathbf{g}(\mathbf{x})] = \sum_{j=1}^{n}\left(\frac{\partial h_j}{\partial x_i} g_j + h_j \frac{\partial g_j}{\partial x_i}\right),$$

or, in other words, [cf. Definition (69)],

$$\mathbf{74}\quad \nabla\langle \mathbf{h}, \mathbf{g}\rangle = \mathbf{g}'\left(\frac{\partial \mathbf{h}'}{\partial \mathbf{x}}\right) + \mathbf{h}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}.$$

Hence

$$\mathbf{75}\quad L_{\mathbf{f}}\langle \mathbf{h}, \mathbf{g}\rangle = \nabla\langle \mathbf{h}, \mathbf{g}\rangle \mathbf{f} = \mathbf{g}'\left(\frac{\partial \mathbf{h}'}{\partial \mathbf{x}}\right)\mathbf{f} + \mathbf{h}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f},$$

while

$$\begin{aligned}\mathbf{76}\quad \langle L_{\mathbf{f}}\mathbf{h}, \mathbf{g}\rangle + \langle \mathbf{h}, L_{\mathbf{f}}\mathbf{g}\rangle &= \langle L_{\mathbf{f}}\mathbf{h}, \mathbf{g}\rangle + \langle \mathbf{h}, [\mathbf{f}, \mathbf{g}]\rangle \\ &= \mathbf{f}'\left(\frac{\partial \mathbf{h}'}{\partial \mathbf{x}}\right)'\mathbf{g} + \mathbf{h}\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\mathbf{g} + \mathbf{h}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f} - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\mathbf{g}\right) \\ &= \mathbf{f}'\left(\frac{\partial \mathbf{h}'}{\partial \mathbf{x}}\right)'\mathbf{g} + \mathbf{h}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f}.\end{aligned}$$

The equality of the two quantities in (75) and (76) follows upon noting that $\mathbf{f}'(\partial \mathbf{h}'/\partial \mathbf{x})'\mathbf{g}$ is just a scalar and therefore equals its transpose. ■

Some other properties of the Lie bracket are ready consequences of the definition.

77 **Lemma** *Suppose* $\mathbf{f}, \mathbf{g}, \mathbf{h} \in V(X)$, $a \in S(X)$, *and* $\alpha, \beta \in \mathbf{R}$. *Then*

78 $[\mathbf{f}, \alpha\mathbf{g} + \beta\mathbf{h}] = \alpha[\mathbf{f}, \mathbf{g}] + \beta[\mathbf{f}, \mathbf{h}]$,

79 $[\mathbf{f}, \mathbf{g}] = -[\mathbf{g}, \mathbf{f}]$,

80 $[\mathbf{f}, [\mathbf{g}, \mathbf{h}]] + [\mathbf{g}, [\mathbf{h}, \mathbf{f}]] + [\mathbf{h}, [\mathbf{f}, \mathbf{g}]] = \mathbf{0}$,

81 $[\mathbf{f}, a\mathbf{g}] = a[\mathbf{f}, \mathbf{g}] + (L_{\mathbf{f}}a)\mathbf{g}$.

Remarks Equation (79) displays the *anti-symmetry* of the Lie bracket. Together (78) and (79) show the the *bilinearity* of the Lie bracket. Equation (80) is known as the *Jacobi identity*. Equation (81) is a type of product rule. In fact, if we replace the Lie bracket $[\mathbf{f}, \mathbf{g}]$ by the Lie derivative symbol $L_{\mathbf{f}}\mathbf{g}$, then (81) can be rewritten as

82 $L_{\mathbf{f}}(a\mathbf{g}) = a\,L_{\mathbf{f}}\mathbf{g} + (L_{\mathbf{f}}a)\,\mathbf{g},$

which looks just like a product rule.

Proof Both (78) and (79) are ready consequences of Definition (18). The formulas (80) and (81) can be established through routine computations; the details are left as an exercise. ■

Suppose $\mathbf{f}_1, \cdots, \mathbf{f}_k \in V(X)$, and $\mathbf{x} \in X$. Then we say that the vector fields $\mathbf{f}_1, \cdots, \mathbf{f}_k$ are **linearly independent** at $\mathbf{x}$ if the column vectors $\mathbf{f}_1(\mathbf{x}), \cdots, \mathbf{f}_k(\mathbf{x})$ are linearly independent (over the field of real numbers). Linear independence of forms is defined analogously. It is clear, by virtue of continuity, that if $\mathbf{f}_1, \cdots, \mathbf{f}_k$ are linearly independent at $\mathbf{x}$, then they are in fact linearly independent at all points in some neighborhood of $\mathbf{x}$, i.e., in open set containing $\mathbf{x}$.

This section is concluded with one last bit of notation. To denote repeated Lie brackets of vector fields, it is convenient to introduce the "ad" symbol. Given $\mathbf{f}, \mathbf{g} \in V(X)$, we define

83 $\mathrm{ad}_{\mathbf{f}}^0\mathbf{g} = \mathbf{g},\ \mathrm{ad}_{\mathbf{f}}^{i+1}\mathbf{g} = [\mathbf{f}, \mathrm{ad}_{\mathbf{f}}^i\mathbf{g}].$

Thus

84 $\mathrm{ad}_{\mathbf{f}}^1\mathbf{g} = [\mathbf{f}, \mathbf{g}],\ \mathrm{ad}_{\mathbf{f}}^2\mathbf{g} = [\mathbf{f}, [\mathbf{f}, \mathbf{g}]],$

and so on.

Problem 7.1 Compute the Lie brackets of the various vector fields defined in Examples (5) and (20).

Problem 7.2 Prove the following alternate version of Lemma (29): If $\mathbf{f}, \mathbf{g} \in V(X)$, then

$$[\mathbf{f},\mathbf{g}](\mathbf{x}_0) = \lim_{t \to 0^+} \frac{1}{t^2} \left\{ [\mathbf{s}_{-\mathbf{g},t}\, \mathbf{s}_{-\mathbf{f},t}\, \mathbf{s}_{\mathbf{g},t}\, \mathbf{s}_{\mathbf{f},t}](\mathbf{x}_0) - \mathbf{x}_0 \right\}.$$

Problem 7.3 Suppose $\mathbf{f}, \mathbf{g} \in V(X)$ are constant vector fields. Show that $[\mathbf{f},\mathbf{g}] = \mathbf{0}$.

Problem 7.4 Let *Aff* denote the set of *affine* vector fields on $\mathbf{R}^n$, i.e., the set of vector fields of the form

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b},\ \mathbf{A} \in \mathbf{R}^{n \times n},\ \mathbf{b} \in \mathbf{R}^n.$$

Show that the set *Aff* is closed under the Lie bracket, i.e., that $[\mathbf{f},\mathbf{g}] \in Aff$ whenever $\mathbf{f}, \mathbf{g} \in Aff$.

Problem 7.5 Suppose $\mathbf{f} \in V(X)$, and $a, b \in S(X)$; i.e., suppose $\mathbf{f}$ is a vector field and a, b are smooth functions. Prove the Leibniz-type product formula

$$L_{\mathbf{f}}(ab) = a\,L_{\mathbf{f}}b + b\,L_{\mathbf{f}}a.$$

Problem 7.6 Suppose $a \in S(X)$, $\mathbf{h} \in F(X)$, and $\mathbf{f} \in V(X)$. Using Definition (47), prove the product-type formula

$$L_{\mathbf{f}}(a\mathbf{h}) = (L_{\mathbf{f}}a)\mathbf{h} + aL_{\mathbf{f}}\mathbf{h}.$$

Problem 7.7 Using the Jacobi identity (60), prove that if $\mathbf{f}, \mathbf{g}, \mathbf{h} \in V(X)$, then

$$L_{\mathbf{f}}[\mathbf{g},\mathbf{h}] = [L_{\mathbf{f}}\mathbf{g},\mathbf{h}] + [\mathbf{g},L_{\mathbf{f}}\mathbf{h}].$$

7.2 DISTRIBUTIONS, FROBENIUS THEOREM

In this section, we present a useful tool in differential geometry, namely the Frobenius theorem. Along the way we introduce important concepts such as submanifolds, distributions, and involutivity.

1 Definition *A subset $M \subseteq X$ is a k-***dimensional submanifold** *$(k < n)$ of X if it possesses the following property: For each $\mathbf{x}_0 \in M$, there exists an open set $U \subseteq X$ containing $\mathbf{x}_0$ and smooth functions $\phi_{k+1}, \cdots, \phi_n \in S(X)$ such that (i) $\{\mathbf{d}\phi_i(x), i = k+1, \cdots, n\}$ is a linearly independent set of row vectors for all $\mathbf{x} \in U$, and (ii)*

2 $U \cap M = \{\mathbf{x} \in U : \phi_i(\mathbf{x}) = 0 \text{ for } i = k+1, \cdots, n\}$.

This definition states that locally M looks like a k-dimensional surface in X defined by the $n-k$ independent equations $\phi_i(\mathbf{x}) = 0$ for $i = k+1, \cdots, n$. Note that some authors use the term **embedded submanifold** for what is called just a submanifold here; they reserve the term "submanifold" for something more general. However, we shall not require this more general concept.

3 Example Let $X = \mathbf{R}^2$, and let M be the circle of radius 1 centered at the origin, i.e., let

$$M = \{(x_1, x_2) \in \mathbf{R}^2 : x_1^2 + x_2^2 = 1\}.$$

Then, by defining

$$\phi_2(x_1, x_2) = x_1^2 + x_2^2 - 1,$$

one can see that M is a one-dimensional submanifold of $\mathbf{R}^2$; it is usually denoted by S^1.

More generally, let $X = \mathbf{R}^{n+1}$, and define

$$S^n = \{\mathbf{x} \in \mathbf{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = 1\}.$$

Then, by defining

$$\phi_{n+1}(\mathbf{x}) = \sum_{i=1}^{n+1} (x_i^2) - 1,$$

we see that S^n is an n-dimensional submanifold of $\mathbf{R}^{n+1}$; it is called the n-sphere. ■

Suppose M is a k-dimensional submanifold of X. Then, by Definition (1), *there exist* functions $\phi_{k+1}, \cdots, \phi_n \in S(X)$ satisfying (2). However, in general, these functions are *not unique*. As a simple illustration, let $X = \mathbf{R}^3$, and define

4 $$\phi_2 = x_2 - x_1^2 - x_3^2, \phi_3 = x_3 - x_1^2 - x_2^2.$$

Then $\phi_2(\mathbf{0}) = \phi_3(\mathbf{0}) = 0$, and it is easy to see that $\mathbf{d}\phi_2(\mathbf{0}) = [0\ 1\ 0]$ and $\mathbf{d}\phi_3(\mathbf{0}) = [0\ 0\ 1]$ are linearly independent. Hence there exists a neighborhood U of $\mathbf{0}$ such that $\mathbf{d}\phi_2(\mathbf{x})$ and $\mathbf{d}\phi_3(\mathbf{x})$ are linearly independent for all $\mathbf{x} \in U$. It follows that the set

5 $$M = \{\mathbf{x} \in U: \phi_2(\mathbf{x}) = \phi_3(\mathbf{x}) = 0\}$$

is a one-dimensional submanifold of $\mathbf{R}^3$. But if we define

6 $$\psi_2 = \phi_2 + \phi_3, \psi_3 = \phi_2 - \phi_3,$$

then the set

7 $$\{\mathbf{x} \in U: \psi_2(\mathbf{x}) = \psi_3(\mathbf{x}) = 0\}$$

is also equal to M. However, though the functions defining M are not unique, the following statement is easy to establish:

8 Lemma *Suppose M is a k-dimensional submanifold of X. Suppose $\mathbf{x}_0 \in M$, and that there exist open sets $U, V \subseteq X$, each containing $\mathbf{x}_0$, and smooth functions $\phi_{k+1}, \cdots, \phi_n \in S(X)$, $\psi_{k+1}, \cdots, \psi_n \in S(X)$, such that (i) the set $\{\mathbf{d}\phi_{k+1}(\mathbf{x}), \cdots, \mathbf{d}\phi_n(\mathbf{x})\}$ is linearly independent for all $\mathbf{x} \in U$, (ii) the set $\{\mathbf{d}\psi_{k+1}(\mathbf{x}), \cdots, \mathbf{d}\psi_n(\mathbf{x})\}$ is linearly independent for all $\mathbf{x} \in V$, (iii) (2) holds, and (iv)*

9 $$V \cap M = \{\mathbf{x} \in V: \psi_i(\mathbf{x}) = 0 \text{ for } i = k+1, \cdots, n\}.$$

Under these conditions, the following statement is true for each $x \in U \cap V$: The $(n-k)$-dimensional subspace of $(\mathbf{R}^n)^*$ spanned by the row vectors $\{\mathbf{d}\phi_{k+1}(\mathbf{x}), \cdots, \mathbf{d}\phi_n(\mathbf{x})\}$ is the same as the $(n-k)$-dimensional subspace of $(\mathbf{R}^n)^*$ spanned by the row vectors $\{\mathbf{d}\psi_{k+1}(\mathbf{x}), \cdots, \mathbf{d}\psi_n(\mathbf{x})\}$.

The proof is left as an exercise.

10 Definition *Suppose M is a k-dimensional submanifold of X, and choose smooth functions $\phi_{k+1}, \cdots, \phi_n$ such that the conditions of Definition (1) are satisfied. Then the* **tangent space** *of M at $\mathbf{x} \in M$ is the k-dimensional subspace of $\mathbf{R}^n$ defined by*

11 $TM_{\mathbf{x}} = \{\mathbf{v} \in \mathbf{R}^n : \langle \mathbf{d}\phi_i(\mathbf{x}), \mathbf{v} \rangle = 0, \text{ for } i = k+1, \cdots, n\}$.

A vector field $\mathbf{f} \in V(X)$ *is said to be* **tangent to** M **at** $\mathbf{x}$ *if* $\mathbf{f}(\mathbf{x}) \in TM_{\mathbf{x}}$.

It follows readily from Lemma (8) that the above definition of $TM_{\mathbf{x}}$ is **intrinsic**, i.e., does not depend on the particular choice of functions ϕ_i used to represent M in the vicinity of $\mathbf{x}$. In other words $TM_{\mathbf{x}}$ is just the k-dimensional subspace of column vectors that are annihilated by each of the $n-k$ row vectors $\mathbf{d}\phi_{k+1}(\mathbf{x}), \cdots, \mathbf{d}\phi_n(\mathbf{x})$ (or, to be more precise, the subspace spanned by these $n-k$ row vectors).

There is another way of looking at submanifolds with which it is relatively easy to compute. Suppose M is a k-dimensional submanifold of X, and that $\mathbf{x}_0 \in M$. Then, according to Definition (1), there exist an open neighborhood $U \subseteq X$ of $\mathbf{x}_0$ and smooth functions $\phi_{k+1}, \cdots, \phi_n$ such that $\mathbf{d}\phi_{k+1}(\mathbf{x}), \cdots, \mathbf{d}\phi_n(\mathbf{x})$ are linearly independent at all $\mathbf{x} \in U$, and such that (2) holds. Now pick smooth functions $\phi_1, \cdots, \phi_k$ such that $\phi_i(\mathbf{x}_0) = 0$ for $i = 1, \cdots, k$ and such that $\{\mathbf{d}\phi_i(\mathbf{x}_0), i = 1, \cdots, n\}$ is a (row) basis for $\mathbf{R}^n$. This is actually quite easy to do. One could even just choose $\phi_i(\mathbf{x}) = \mathbf{v}_i(\mathbf{x} - \mathbf{x}_0)$, where $\mathbf{v}_1, \cdots, \mathbf{v}_k$ is a set of (constant) row vectors chosen such that $\{\mathbf{v}_1, \cdots, \mathbf{v}_k, \mathbf{d}\phi_{k+1}(\mathbf{x}_0), \cdots, \mathbf{d}\phi_n(\mathbf{x}_0)\}$ is a row basis for $\mathbf{R}^n$. Now define a map $T: \mathbf{R}^n \to \mathbf{R}^n$ by

12 $y_i = (T\mathbf{x})_i = \phi_i(\mathbf{x}), \ 1 \le i \le n.$

By construction, the Jacobian $\partial T/\partial \mathbf{x}$ evaluated at $\mathbf{x}_0$ is nonsingular. Thus, by the inverse function theorem [Theorem (7.1.1)], T is locally a diffeomorphism, say on $U_0 \subseteq U$. One can think of $y_1, \cdots, y_n$ as a new set of coordinates on U_0. What does the submanifold M look like in terms of these new coordinates? Comparing (2), we see that

13 $$M \cap U_0 = \left\{ \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{0} \end{bmatrix} : \mathbf{y}_1 \in N \subseteq \mathbf{R}^k \right\},$$

where N is some neighborhood of $\mathbf{0}$ in $\mathbf{R}^k$. In other words, a k-dimensional submanifold of X is a set which, after a suitable smooth change of coordinates, looks like a k-dimensional "slice" of X.

The tangent space of M also has a simple form in the new coordinates. If we perform the coordinate transformation (12), then in terms of the $\mathbf{y}$ coordinates, ϕ_i is just the i-th coordinate of $\mathbf{y}$. Hence $\mathbf{d}\phi_i$ is just the i-th elementary row vector, with a 1 in the i-th position and zeros elsewhere. To compute the tangent space at $\mathbf{y}_0$ we can apply the formula (11) and observe that a column vector $\mathbf{v}$ is annihilated by each of the elementary row vectors with a 1 in positions $k+1, \cdots, n$ respectively if and only if the last $n-k$ components of $\mathbf{v}$ are zero. Thus, if $\mathbf{y}_0 = T(\mathbf{x}_0)$, then

14 $TM_{\mathbf{y}_0} = \{\mathbf{v} \in \mathbf{R}^n : \mathbf{v} = [\mathbf{v}_a' \ \ \mathbf{0}]', \ \mathbf{v}_a \in \mathbf{R}^k\}$,

$$= \{\mathbf{v} \in \mathbf{R}^n : v_i = 0 \text{ for } i = k+1, \cdots, n\}.$$

A vector field $\mathbf{f}$ is tangent to M at a point $\mathbf{x}_0$ if and only if the coordinate-transformed vector field $\mathbf{f}_T$ [defined in (7.1.9)] has the form

$$\mathbf{15} \quad \mathbf{f}_T(\mathbf{y}) = \begin{bmatrix} \mathbf{f}_a \\ \mathbf{0}_{n-k} \end{bmatrix}.$$

The next definition introduces a very important concept.

16 Definition *A k-dimensional* **distribution** Δ *on X is a map which assigns, to each* $\mathbf{x} \in X$, *a k-dimensional subspace of* $\mathbf{R}^n$ *such that the following smoothness condition is satisfied: For each* $\mathbf{x}_0 \in X$ *there exist an open set* $U \subseteq X$ *containing* $\mathbf{x}_0$ *and k vector fields* $\mathbf{f}_1, \cdots, \mathbf{f}_k$ *such that (i)* $\{\mathbf{f}_1(\mathbf{x}), \cdots, \mathbf{f}_k(\mathbf{x})\}$ *is a linearly independent set for each* $\mathbf{x} \in U$, *and (ii)*

$$\mathbf{17} \quad \Delta(\mathbf{x}) = \text{span}\,\{\mathbf{f}_1(\mathbf{x}), \cdots, \mathbf{f}_k(\mathbf{x})\}, \ \forall \mathbf{x} \in U.$$

It is often convenient to describe a distribution in terms of the vector fields that generate it. Thus, given k vector fields $\mathbf{f}_1, \cdots, \mathbf{f}_k$, we often *define* Δ by the formula (17). But it is important to remember that $\Delta(\mathbf{x})$ is a subspace, and that $\{\mathbf{f}_1(\mathbf{x}), \cdots, \mathbf{f}_k(\mathbf{x})\}$ is *a*, not *the*, basis for it. To amplify this point further, suppose $\mathbf{f}_1$ and $\mathbf{f}_2$ are two vector fields on X with the property that $\mathbf{f}_1(\mathbf{x})$ and $\mathbf{f}_2(\mathbf{x})$ are linearly independent at all $\mathbf{x} \in X$. Then span $\{\mathbf{f}_1, \mathbf{f}_2\}$ and span $\{\mathbf{f}_1 + \mathbf{f}_2, \mathbf{f}_1 - \mathbf{f}_2\}$ describe exactly the same distribution. Indeed, in order to define a k-dimensional distribution, it is often convenient to define it as the span of k *or more* vector fields, of which some k are linearly independent at each point.

In attempting to describe a distribution by means of a set of vector fields that generate it, one can get into the following difficulty: Suppose $\mathbf{f}_1, \cdots, \mathbf{f}_k \in V(X)$ are given, and we define Δ by (17). Then it can happen that the rank of the matrix $[\mathbf{f}_1(\mathbf{x}) \cdots \mathbf{f}_k(\mathbf{x})]$ is not constant as $\mathbf{x}$ varies. To get around this difficulty, it is possible to define a k-dimensional distribution as a map which assigns to each $\mathbf{x} \in X$ a subspace of dimension *no more than k,* and then require that each open set U contain at least one point $\mathbf{y}$ such that the dim $\Delta(\mathbf{y})$ exactly equals k. If this definition is accepted, a great deal of verbal awkwardness is avoided because, in subsequent sections, distributions are invariably defined in terms of some generating set of vector fields. In such a case, if $\Delta(\mathbf{x})$ is actually a k-dimensional subspace of $\mathbf{R}^n$, then we say that $\mathbf{x}$ is a **regular** point of Δ. Alternatively, $\mathbf{x}$ is a regular point of the distribution Δ if there is a neighborhood U of $\mathbf{x}$ such that the dimension of $\Delta(\mathbf{y})$ is the same for all $\mathbf{y} \in U$. Finally, if Δ is a given distribution and $\mathbf{f} \in V(X)$, then we say that $\mathbf{f}$ **belongs to** Δ if $\mathbf{f}(\mathbf{x}) \in \Delta(\mathbf{x})$ $\forall \mathbf{x} \in X$, and denote it by $\mathbf{f} \in \Delta$.

Suppose Δ is a k-dimensional distribution, that $U \subseteq X$ is an open set, and that $\mathbf{f}_1, \cdots, \mathbf{f}_m \in V(X)$, $m \geq k$, are vector fields that span Δ on U. Now suppose $\mathbf{f} \in \Delta$. Then, using Definition (16), one can show that there exist smooth functions $\alpha_1, \cdots, \alpha_m \in S(X)$ such that

$$\mathbf{18}\quad \mathbf{f}(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i(\mathbf{x})\,\mathbf{f}_i(\mathbf{x}),\ \forall \mathbf{x} \in U.$$

It is essential to note that the "coefficients" in (18) are *not constants*, but are smooth functions.

An interesting question in differential geometry is the following: (Unfortunately the brevity of the present treatment does not permit us to explore fully just *why* this question is interesting.) Suppose Δ is a given k-dimensional, everywhere regular, distribution on X; for each $\mathbf{x} \in X$, does there exist a k-dimensional submanifold $M_{\mathbf{x}}$ of X containing $\mathbf{x}$ such that every vector field $\mathbf{f} \in \Delta$ is tangent to $M_{\mathbf{x}}$ at $\mathbf{x}$ [i.e., $TM_{\mathbf{x}} = \Delta(\mathbf{x})$]? If such a submanifold $M_{\mathbf{x}}$ exists for each $\mathbf{x} \in X$, then the distribution Δ is said to be **completely integrable**, and $M_{\mathbf{x}}$ is said to be the **integral manifold** of Δ passing though $\mathbf{x}$. But the question is: When is a distribution completely integrable?

An elegant answer to this question is provided by the next result, commonly known as the Frobenius theorem.

19 Definition *A distribution Δ is* **involutive** *if $[\mathbf{f}, \mathbf{g}] \in \Delta$ whenever $\mathbf{f}, \mathbf{g} \in \Delta$.*

In other words, a distribution is involutive if it is closed under the Lie bracket.

20 Theorem (Frobenius) *A distribution is completely integrable if and only if it is involutive.*

It turns out that the "only if" part is quite easy to prove, and the "if" part is really the substantive part of the theorem.

Proof "Only if" Suppose Δ is a completely integrable k-dimensional distribution. Then, for each $\mathbf{x} \in X$, there corresponds an integral manifold $M_{\mathbf{x}}$ of dimension k. By Definition (1), this means that, for each $\mathbf{x}_0 \in X$, there exist a neighborhood U of $\mathbf{x}_0$ and smooth functions $\phi_{k+1}, \cdots, \phi_n \in S(X)$ such that the differentials of the ϕ_i are linearly independent row vectors at $\mathbf{x}_0$, and such that each function ϕ_i has constant value at all points in $M_{\mathbf{x}_0}$. Now, if we select smooth functions $\phi_1, \cdots, \phi_k \in S(X)$ such that $\{\mathbf{d}\phi_i(\mathbf{x}_0), i = 1, \cdots, n\}$ is a row basis, then the map $T: U \to \mathbf{R}^n$ defined by (12) is a diffeomorphism over some neighborhood $U_0 \subseteq U$ of $\mathbf{x}_0$. Moreover, since each $\mathbf{f} \in \Delta$ is tangent to $M_{\mathbf{x}}$ *at each* $\mathbf{x} \in U_0$, it follows that in the new coordinate system each $\mathbf{f}_T \in \Delta_T$ has the form

$$\mathbf{21}\quad \mathbf{f}_T(\mathbf{y}) = \begin{bmatrix} \mathbf{f}_a(\mathbf{y}) \\ \mathbf{0}_{n-k} \end{bmatrix},\ \forall \mathbf{y} \in T(U_0).$$

The key point is to note here is that the tangency relationship holds *at all points in some neighborhood.* Moreover, since the tangent space $TM_{\mathbf{x}}$ is precisely $\Delta(\mathbf{x})$ for all $\mathbf{x} \in U_0$, it follows that Δ_T is precisely the set of vector fields of the form (21). Now suppose $\mathbf{f}_T, \mathbf{g}_T \in \Delta_T$, i.e., that

$$\mathbf{22}\quad \mathbf{f}_T(\mathbf{y}) = \begin{bmatrix} \mathbf{f}_a(\mathbf{y}) \\ \mathbf{0}_{n-k} \end{bmatrix}, \mathbf{g}_T(\mathbf{y}) = \begin{bmatrix} \mathbf{g}_a(\mathbf{y}) \\ \mathbf{0}_{n-k} \end{bmatrix}.$$

Then it readily follows from the formula (7.1.19) that the bottom $n-k$ elements of $[\mathbf{f}_T, \mathbf{g}_T]$ are also identically zero. Thus $[\mathbf{f}_T, \mathbf{g}_T] \in \Delta_T$ and hence Δ_T (i.e., Δ) is an involutive distribution.

"If" See Appendix C. ■

The "if" part of the proof is by induction on the integer k (the dimension of the distribution Δ). If $k = 1$, then Δ is just span $\{\mathbf{f}\}$ where $\mathbf{f}$ is a nonzero vector field. Now a one-dimensional distribution is automatically involutive, since $[a\mathbf{f}, b\mathbf{f}] \in \text{span}\,\mathbf{f} = \Delta$ whenever $a, b \in S(X)$. In this case, the integral manifold $M_{\mathbf{x}}$ of Δ passing through $\mathbf{x}$ is just the integral curve of the vector field $\mathbf{f}$ passing through $\mathbf{x}$. For $k \geq 2$, the proof is more complicated, and is given in Appendix C.

Every distribution contains an *infinite* number of vector fields; for example, if $\mathbf{f} \in \Delta$, then $a\mathbf{f} \in \Delta$ for all $a \in S(X)$. However, in order to check whether a distribution is involutive or not, it is only necessary to compute a *finite* number of Lie brackets. Suppose $\mathbf{f}_1, \cdots, \mathbf{f}_m \in V(X)$ span the $k(\leq m)$-dimensional distribution Δ over some open set $U \subseteq X$. Then every $\mathbf{f} \in \Delta$ has an expansion of the form (18). Hence, using the bilinearity of the Lie bracket and the various product formulas in Lemma (7.1.77), one can easily establish the following statement: Δ is involutive if and only if $[\mathbf{f}_i, \mathbf{f}_j] \in \Delta$ for each i, j; or, in other words, there exist smooth functions $\alpha_{ijl} \in S(U)$, $1 \leq i, j, l \leq m$, such that

$$\mathbf{23}\quad [\mathbf{f}_i, \mathbf{f}_j](\mathbf{x}) = \sum_{l=1}^{m} \alpha_{ijl}(\mathbf{x})\, \mathbf{f}_l(\mathbf{x}), \ \forall \mathbf{x} \in U.$$

This leads to an alternate form of the Frobenius theorem, which is the form actually used in later sections.

24 Theorem (Alternate Frobenius) *Suppose $\mathbf{f}_1, \cdots, \mathbf{f}_m \in V(X)$, $N \subseteq X$ is an open set, $\mathbf{x}_0 \in N$, and that the set $\{\mathbf{f}_1(\mathbf{x}), \cdots, \mathbf{f}_m(\mathbf{x})\}$ contains k linearly independent vectors at each $\mathbf{x} \in N$. Then there exist functions $\phi_{k+1}, \cdots, \phi_n \in S(X)$ such that (i) $\mathbf{d}\phi_{k+1}(\mathbf{x}_0), \cdots, \mathbf{d}\phi_n(\mathbf{x}_0)$ are linearly independent, and (ii) there exists a neighborhood $V \subseteq N$ of $\mathbf{x}_0$ such that*

$$\mathbf{25}\quad \langle d\phi_i, \mathbf{f}_j \rangle(\mathbf{x}) = 0, \ \forall \mathbf{x} \in V, \text{ for } k+1 \leq i \leq n, \ 1 \leq j \leq m,$$

if and only if the distribution spanned by $\mathbf{f}_1, \cdots, \mathbf{f}_m$ is involutive, i.e., there exist smooth functions α_{ijl} and a neighborhood $U \subseteq N$ of $\mathbf{x}_0$ such that (23) holds.

Remark Suppose the vector fields $\mathbf{f}_1, \cdots, \mathbf{f}_k$ are given by

$$\mathbf{26}\quad \mathbf{f}_1(\mathbf{x})=\begin{bmatrix}1\\0\\\cdot\\\cdot\\\cdot\\0\end{bmatrix},\cdots,\mathbf{f}_k(\mathbf{x})=\begin{bmatrix}0\\\cdot\\\cdot\\\cdot\\0\\1\\\cdot\\\cdot\\\cdot\\0\end{bmatrix},\ \forall \mathbf{x}\in X.$$

In other words, $\mathbf{f}_i$ is a *constant* vector field with a "1" in the i-th position and zeros elsewhere. Define $\Delta = \text{span}\{\mathbf{f}_1, \cdots, \mathbf{f}_k\}$. Then Δ is a k-dimensional distribution consisting of all vector fields of the form

$$\mathbf{27}\quad \mathbf{f}(\mathbf{x})=\begin{bmatrix}\mathbf{f}_a(\mathbf{x})\\\mathbf{0}_{n-k}\end{bmatrix}.$$

It is clear that Δ is completely integrable. Indeed, given $\mathbf{x}_0\in X$, the corresponding integral manifold $M_{\mathbf{x}_0}$ is just the set

$$\mathbf{28}\quad \{\mathbf{x}\in\mathbf{R}^n: x_i = x_{0i}, \text{ for } k+1\le i\le n\},$$

and functions $\phi_{k+1}, \cdots, \phi_n\in S(X)$ satisfying (25) are given by

$$\mathbf{29}\quad \phi_i(\mathbf{x}) = x_i,\ i = k+1, \cdots, n.$$

The point of the Frobenius theorem is that *all* completely integrable distributions can be made to look like this after a suitable change of coordinates. Of course, the theorem is non-constructive in the sense that it tells us that a suitable coordinate transformation *exists* — it does not tell us how to find it. Nevertheless the Frobenius theorem is a very useful result.

30 Example Suppose X is an open neighborhood of the point $\mathbf{x}_0 = [1\ 1\ 1]'$ in $\mathbf{R}^3$, and consider the vector fields

$$\mathbf{f}(\mathbf{x})=\begin{bmatrix}2x_1^2\\x_2^2+x_2x_3\\-2x_3^3\end{bmatrix},\ \mathbf{g}(\mathbf{x})=\begin{bmatrix}0\\x_3\\0\end{bmatrix}.$$

Then it is routine to verify that

$$[\mathbf{f}, \mathbf{g}](\mathbf{x}) = \begin{bmatrix} 0 \\ x_2^2 - (x_2 + x_3)x_3 \\ 0 \end{bmatrix} = \alpha(\mathbf{x})\,\mathbf{g}(\mathbf{x}),$$

where

$$\alpha(\mathbf{x}) = \frac{x_2^2}{x_3} - (x_2 + x_3)$$

is a smooth function of $\mathbf{x}$ in any neighborhood of $\mathbf{x}_0$ that does not intersect the surface $x_3 = 0$. Hence the two-dimensional distribution Δ spanned by $\mathbf{f}$ and $\mathbf{g}$ is involutive if we define X to be the open ball of radius 1 in $\mathbf{R}^3$ centered at $\mathbf{x}_0$. Now the Frobenius theorem assures us that there is a smooth function $h \in S(X)$ such that

$$<\mathbf{d}h, \mathbf{f}>(\mathbf{x}) = 0, \quad <\mathbf{d}h, \mathbf{g}>(\mathbf{x}) = 0, \quad \forall \mathbf{x} \in X.$$

These are two simultaneous partial differential equations that the function h needs to satisfy. In general, it is not easy to solve these equations and find such a function, though the Frobenius theorem guarantees that such a function exists. In the present instance,

$$h(\mathbf{x}) = 2x_1^{-1} + x_3^{-2}$$

is a possible solution. This means that, if we look at the surface in X defined by

$$2x_1^{-1} + x_3^{-2} = \text{constant},$$

then at each point $\mathbf{x}$ on the surface, the tangent plane to the surface at $\mathbf{x}$ is the span of the two vectors $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$.

7.3 REACHABILITY AND OBSERVABILITY

This section has two parts. In the first part of the section, we study the reachability of nonlinear control systems of the form

$$\mathbf{1} \qquad \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \sum_{i=1}^{m} u_i\,\mathbf{g}_i(\mathbf{x}),$$

where $\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m$ are given vector fields in $V(X)$ and $X \subseteq \mathbf{R}^n$ is a given open set. The questions studied are the following: Are there simple necessary and sufficient conditions for the system (1) to be reachable? If the system (1) is *not* reachable, can one carry out a change of state variables in such a way that the "maximally unreachable" part is explicitly displayed? Finally, do these conditions reduce to the familiar conditions for the linear system

2 $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$

to be reachable, so that the nonlinear results are true generalizations of the corresponding results for linear systems?

In the second part of the section, we add an output or measurement equation of the form

3 $\mathbf{y} = \mathbf{h}(\mathbf{x})$

to the system description (1), and ask when such a system is observable. Equation (3) is the nonlinear analog of the output equation

4 $\mathbf{y} = \mathbf{C}\mathbf{x}$

for linear systems. The questions studied are the following: Are there simple necessary and sufficient conditions for the system (1) – (3) to be locally observable? If the system is *not* locally observable, can one carry out a change of state variables such that the system is decomposed into an observable part and an unobservable part? Finally, is it possible to decompose a nonlinear system of the form (1) – (3) into *four* subsystems, which are respectively reachable and observable, reachable and unobservable, unreachable and observable, and last, unreachable and unobservable?

7.3.1 Reachability

We begin with a definition of reachability.

5 **Definition** *The system (1) is said to be (locally)* **reachable** *around a state* $\mathbf{x}_0 \in X$ *if there exists a neighborhood* U *of* $\mathbf{x}_0$ *such that, for each* $\mathbf{x}_f \in U$, *there exist a time* $T > 0$ *and a set of control inputs* $\{u_i(t), t \in [0, T], 1 \le i \le m\}$ *such that, if the system starts in the state* $\mathbf{x}_0$ *at time* 0, *then it reaches the state* $\mathbf{x}_f$ *at time* T.

Note that the above definition is purely local: The system (1) is reachable around $\mathbf{x}_0$ if every state *sufficiently close to* $\mathbf{x}_0$ can be reached from $\mathbf{x}_0$. It is very difficult to analyze the "global" reachability of (1) since nonlinear systems do not satisfy superposition in general. Actually, one can argue that local reachability is all that one can prove, and that the equivalence of local and global reachability is a property peculiar to linear systems alone. Also, for linear systems, there is complete equivalence between *reachability* (the ability to reach any desired final state from a given initial state) and *controllability* (the ability to reach a fixed final state from any given initial state). This is no longer the case for nonlinear systems. Accordingly, the discussion here is restricted to reachability alone. For a thorough discussion of various nuances of this topic, see Isidori (1989) or Nijmeijer and van der Schaft (1990).

The discussion on reachability encompasses several topics. First, the notion of invariant distributions is introduced, and it is shown that the notion is a natural generalization, to nonlinear systems, of the notion of $(\mathbf{A}, \mathbf{B})$-invariant subspaces for linear systems [see e.g., Wonham (1979)]. Then it is shown that every system can be transformed locally to a

reachable part plus a "totally unreachable" part. This extends the canonical decomposition of linear time-invariant systems into a controllable part plus a completely uncontrollable part, introduced by Kalman (1963), to nonlinear systems. Finally, the familiar rank test for the reachability of linear systems is extended to nonlinear systems. All in all, the message is that, with appropriate tools, most of the familiar results for linear time-invariant systems can be extended in a natural way to nonlinear systems.

To motivate the various ideas, we begin with a brief discussion of linear (time-invariant) systems. Consider the linear differential equation

$$\mathbf{6} \qquad \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t).$$

where $\mathbf{x}(t) \in \mathbf{R}^n$, $\mathbf{u}(t) \in \mathbf{R}^m$, $\mathbf{A} \in \mathbf{R}^{n \times n}$ and $\mathbf{B} \in \mathbf{R}^{n \times m}$. Then a subspace M of $\mathbf{R}^n$ is said to be **A-invariant** if

$$\mathbf{7} \qquad \mathbf{A}M \subseteq M.$$

If, in addition,

$$\mathbf{8} \qquad \mathbf{B}(\mathbf{R}^m) \subseteq M,$$

then M is said to be **(A, B)-invariant.** Such subspaces help us to analyze the system (6) in an abstract way. For instance, suppose $\mathbf{x}_0 \in M$; then it is clear from (6) that $\mathbf{x}(t) \in M \ \forall t \geq 0$. Hence, if we express $\mathbf{R}^n$ as a direct sum $M \dotplus N$, then the system equation (6) must have the form

$$\mathbf{9} \qquad \dot{\mathbf{x}}_1 = \mathbf{A}_{11}\mathbf{x}_1 + \mathbf{A}_{12}\mathbf{x}_2 + \mathbf{B}_1\mathbf{u},$$

$$\dot{\mathbf{x}}_2 = \mathbf{A}_{22}\mathbf{x}_2,$$

where $\mathbf{x}_1 + \mathbf{x}_2$ is the decomposition of the vector $\mathbf{x}$. Hence, it is worthwhile to make M as small (i.e., as low-dimensional) as possible. Also, a necessary and sufficient condition for the system (6) to be reachable is that there is no nontrivial (**A**, **B**)-invariant subspace of $\mathbf{R}^n$, i.e., the only (**A**, **B**)-invariant subspace of $\mathbf{R}^n$ is $\mathbf{R}^n$ itself.

To extend these ideas to nonlinear systems of the form (1), a few preliminary definitions and results are needed.

10 Definition *Suppose* Δ *is a distribution on* X, *and* $\mathbf{f} \in V(X)$. *Then* Δ *is said to be* **invariant under f,** *or* **f-invariant,** *if* $[\mathbf{f}, \mathbf{h}] \in \Delta \ \forall \mathbf{h} \in \Delta$.

This definition generalizes the notion of an **A**-invariant subspace for linear vector fields. Let $X = \mathbf{R}^n$, M a k-dimensional subspace of $\mathbf{R}^n$, and let $\mathbf{A} \in \mathbf{R}^{n \times n}$. Let $\mathbf{f}(\mathbf{x})$ be the linear vector field $\mathbf{A}\mathbf{x}$, and let Δ be the distribution generated by the *constant* vector fields $\mathbf{v}_1, \cdots, \mathbf{v}_k$, where $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a basis for M. It is now shown that the distribution Δ is **f**-invariant if and only if $\mathbf{A}M \subseteq M$, i.e., M is an **A**-invariant subspace. Note that every $\mathbf{h} \in \Delta$ has the form

$$\mathbf{11} \quad \mathbf{h}(\mathbf{x}) = \sum_{i=1}^{k} h_i(\mathbf{x})\, \mathbf{v}_i,\ h_i \in S(X).$$

To show that Δ is an $\mathbf{f}$-invariant distribution if and only if $\mathbf{A}M \subseteq M$, consider first the "only if" part of the statement. Suppose Δ is $\mathbf{f}$-invariant. Then certainly $[\mathbf{f}, \mathbf{v}_i] \in \Delta$ for each i. But

$$\mathbf{12} \quad [\mathbf{f}, \mathbf{v}_i] = [\mathbf{Ax}, \mathbf{v}_i] = \mathbf{A}\mathbf{v}_i = \text{constant},\ \forall \mathbf{x}.$$

Hence $[\mathbf{f}, \mathbf{v}_i] \in \Delta$ implies that $\mathbf{A}\mathbf{v}_i \in M$ for all i. Since the $\mathbf{v}_i$'s form a basis for M, it follows that $\mathbf{A}M \subseteq M$. Conversely, suppose that $\mathbf{A}M \subseteq M$, and that $\mathbf{h}$ is of the form (11); it must be shown that $[\mathbf{f}, \mathbf{h}] \in \Delta$. Since the Lie bracket is bilinear, it is enough to show that

$$\mathbf{13} \quad [\mathbf{Ax}, h_i(\mathbf{x})\, \mathbf{v}_i] \in \Delta,\ \forall i.$$

It will then follow that

$$\mathbf{14} \quad [\mathbf{f}, \mathbf{h}] = \sum_{i=1}^{k} [\mathbf{Ax}, h_i(\mathbf{x})\, \mathbf{v}_i] \in \Delta.$$

To establish (13), apply the definition of the Lie bracket. This gives

$$\mathbf{15} \quad [\mathbf{Ax}, h_i(\mathbf{x})\, \mathbf{v}_i] = \mathbf{v}_i \frac{\partial h_i(\mathbf{x})}{\partial \mathbf{x}} \mathbf{Ax} - \mathbf{A}\, h_i(\mathbf{x})\, \mathbf{v}_i = c_i(\mathbf{x})\, \mathbf{v}_i - h_i(\mathbf{x})\, \mathbf{A}\mathbf{v}_i \in \Delta,$$

where

$$\mathbf{16} \quad c_i(\mathbf{x}) = \frac{\partial h_i(\mathbf{x})}{\partial \mathbf{x}} \mathbf{Ax} \in S(X).$$

Hence Δ is $\mathbf{f}$-invariant.

17 Lemma *Suppose $\mathbf{x}_0 \in X$, $\mathbf{f}$ is a vector field on X, and Δ is a k-dimensional distribution on X. Suppose there is a neighborhood U of $\mathbf{x}_0$ such that, when restricted to U, Δ is involutive and $\mathbf{f}$-invariant. Then there exist a neighborhood U_0 of $\mathbf{x}_0$ and a diffeomorphism T on U_0 such that, in terms of the new coordinates $\mathbf{y} = T(\mathbf{x})$, $\mathbf{f}_T$ has the form*

$$\mathbf{18} \quad \mathbf{f}_T(\mathbf{y}) = \begin{bmatrix} \mathbf{f}_a(\mathbf{y}_a, \mathbf{y}_b) \\ \mathbf{f}_b(\mathbf{y}_b) \end{bmatrix},\ \forall \mathbf{y} = \begin{bmatrix} \mathbf{y}_a \\ \mathbf{y}_b \end{bmatrix} \in T(U_0),$$

where $\mathbf{f}_a, \mathbf{y}_a \in \mathbf{R}^k$ and $\mathbf{f}_b, \mathbf{y}_b \in \mathbf{R}^{n-k}$.

Remarks The point of the lemma is that, after a suitable change of coordinates, $\mathbf{f}$ has the "triangular" form shown in (18).

Proof Since Δ is involutive when restricted to U, it follows from the Frobenius theorem that there exist a neighborhood $U_0 \subseteq U$ of $\mathbf{x}_0$ and a diffeomorphism T on U_0 such that, in terms of the new coordinates,

$$\Delta_T = \{\mathbf{h} \in V(X): \mathbf{h}_i(\mathbf{y}) = 0,\ k+1 \le i \le n,\ \forall \mathbf{y} \in T(U_0)\}$$

$$= \{\mathbf{h} \in V(X): \mathbf{h}_b(\mathbf{y}) = \mathbf{0}\ \forall \mathbf{y} \in T(U_0)\}. \tag{19}$$

Now by assumption Δ is $\mathbf{f}$-invariant, or equivalently, Δ_T is $\mathbf{f}_T$-invariant. Let $\mathbf{h}_T \in \Delta_T$ be arbitrary. Then by (7.1.58),

$$[\mathbf{f}, \mathbf{h}]_T = [\mathbf{f}_T, \mathbf{h}_T] = \frac{\partial \mathbf{h}_T}{\partial \mathbf{x}} \mathbf{f}_T - \frac{\partial \mathbf{f}_T}{\partial \mathbf{x}} \mathbf{h}_T. \tag{20}$$

Now, since Δ_T is $\mathbf{f}_T$-invariant, it follows that $[\mathbf{f}_T, \mathbf{h}_T] \in \Delta_T$, which means that the bottom $n-k$ elements of $[\mathbf{f}_T, \mathbf{h}_T]$ must be identically zero. Partition $\mathbf{f}_T$, $\mathbf{h}_T$, $\mathbf{y}$ as

$$\mathbf{f}_T = \begin{bmatrix} \mathbf{f}_a \\ \mathbf{f}_b \end{bmatrix}, \mathbf{h}_T = \begin{bmatrix} \mathbf{h}_a \\ \mathbf{0} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \mathbf{y}_a \\ \mathbf{y}_b \end{bmatrix}, \tag{21}$$

where all vectors with the subscript a have k rows, and all vectors with the subscript b have $n-k$ rows. Similarly, partition $\partial \mathbf{f}_T / \partial \mathbf{y}$ and $\partial \mathbf{h}_T / \partial \mathbf{y}$ as

$$\frac{\partial \mathbf{f}_T}{\partial \mathbf{y}} = \begin{bmatrix} \dfrac{\partial \mathbf{f}_a}{\partial \mathbf{y}_a} & \dfrac{\partial \mathbf{f}_a}{\partial \mathbf{y}_b} \\ \dfrac{\partial \mathbf{f}_b}{\partial \mathbf{y}_a} & \dfrac{\partial \mathbf{f}_b}{\partial \mathbf{y}_b} \end{bmatrix}, \frac{\partial \mathbf{h}_T}{\partial \mathbf{y}} = \begin{bmatrix} \dfrac{\partial \mathbf{h}_a}{\partial \mathbf{y}_a} & \dfrac{\partial \mathbf{h}_a}{\partial \mathbf{y}_b} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{22}$$

Now the fact that $[\mathbf{f}_T, \mathbf{h}_T] \in \Delta_T$ implies that $[\mathbf{f}_T, \mathbf{h}_T]_b \equiv \mathbf{0}$, or, from (20), that

$$\frac{\partial \mathbf{f}_b}{\partial \mathbf{y}_a} \mathbf{h}_a \equiv \mathbf{0}. \tag{23}$$

But since (23) holds for *all* $\mathbf{h}_a$, it follows that

$$\frac{\partial \mathbf{f}_b}{\partial \mathbf{y}_a} \equiv \mathbf{0}, \tag{24}$$

i.e., $\mathbf{f}_b$ is independent of $\mathbf{y}_a$. This is exactly what (18) states. ■

Theorem *Consider the system (1), and suppose an initial state $\mathbf{x}_0 \in X$ is specified. Suppose there exist a neighborhood U of $\mathbf{x}_0$ and a k-dimensional distribution Δ on U such that*

(i) *Δ is involutive, and $\mathbf{x}_0$ is a regular point of Δ.*

(ii) *$\mathbf{g}_i(\mathbf{x}) \in \Delta\ \forall \mathbf{x} \in U$, for $i = 1, \cdots, m$.*

(iii) Δ *is invariant under* $\mathbf{f}$.

Then there exist a neighborhood $U_0 \subseteq U$ *of* $\mathbf{x}_0$ *and a diffeomorphism* T *on* U_0 *such that*

$$\mathbf{26} \quad \mathbf{f}_T(\mathbf{y}) = \begin{bmatrix} \mathbf{f}_a(\mathbf{y}_a, \mathbf{y}_b) \\ \mathbf{f}_b(\mathbf{y}_b) \end{bmatrix}, \mathbf{g}_{iT}(\mathbf{y}) = \begin{bmatrix} \mathbf{g}_{ia}(\mathbf{y}_a, \mathbf{y}_b) \\ \mathbf{0} \end{bmatrix}, i = 1, \cdots, m,$$

where

$$\mathbf{27} \quad \begin{bmatrix} \mathbf{y}_a \\ \mathbf{y}_b \end{bmatrix} = \mathbf{y} = T(\mathbf{x})$$

is a partition of $\mathbf{y}$, *all vectors with the subscript a have k rows, and all vectors with the subscript b have* $n - k$ *rows.*

Proof The form of $\mathbf{f}$ follows from Lemma (17). The form of $\mathbf{g}_i$ follows from the fact that $\mathbf{g}_i \in \Delta$ for all i. ■

Theorem (25) is an important result because it gives a nonlinear analog of the well-known canonical decomposition of unreachable linear systems as given in (9). As one might expect, the result is only local, in the sense that it applies only in some neighborhood of $\mathbf{x}_0$. To apply this theorem effectively, it is desirable to make the integer k as small as possible, so as to make the distribution Δ as small as possible. For this purpose, the following procedure is proposed.

28 Procedure *Step 0:* Define $i = 0$ and set

$$\mathbf{29} \quad \Delta_0 = \text{span}\{\mathbf{g}_1, \cdots, \mathbf{g}_m\}.$$

Step 1: Let $\{\mathbf{h}_1^{(i)}, \cdots, \mathbf{h}_{k_i}^{(i)}\}$ be a set of vector fields that generate Δ_i. Check if Δ_i is involutive in some neighborhood of $\mathbf{x}_0$ by checking whether each Lie bracket $[\mathbf{h}_j^{(i)}, \mathbf{h}_l^{(i)}]$ belongs to Δ_i for $1 \le j, l \le k_i$. Check if Δ_i is invariant under $\mathbf{f}$ by checking whether $[\mathbf{f}, \mathbf{h}_j^{(i)}] \in \Delta_i$ for all j. If Δ_i is both involutive and invariant under $\mathbf{f}$, STOP.

Step 2: If Δ_i is either not involutive or not invariant under $\mathbf{f}$, then set $i = i + 1$, and define Δ_{i+1} as

$$\mathbf{30} \quad \Delta_{i+1} = \text{span}\{\mathbf{h}_j^{(i)}, 1 \le j \le k_i\} \cup \{[\mathbf{h}_j^{(i)}, \mathbf{h}_l^{(i)}], 1 \le j, l \le k_i\}$$

$$\cup \{[\mathbf{f}, \mathbf{h}_j^{(i)}], 1 \le j \le k_i\}.$$

Return to Step 1. Note: In (30), it is only necessary to add those Lie brackets $[\mathbf{h}_j^{(i)}, \mathbf{h}_l^{(i)}]$ and $[\mathbf{f}, \mathbf{h}_j^{(i)}]$ which do not already belong to Δ_i.

This procedure generates a sequence of distributions $\{\Delta_i\}$ such that $\Delta_i(\mathbf{x}) \subseteq \Delta_{i+1}(\mathbf{x})$ for all $\mathbf{x} \in X$. If $\mathbf{x}_0$ is a *regular* point of each Δ_i, which it need not be in general, then $\dim \Delta_{i+1}$ is strictly larger than $\dim \Delta_i$. Since $\dim \Delta_i \le n$ for all i, in such a case the process cannot continue more than n times. When the procedure terminates we will have found a distribution

Δ_c which is both involutive as well as invariant under $\mathbf{f}$ and all $\mathbf{g}_l$. Because of the manner in which Δ_c is generated, it is clear that Δ_c is the *smallest* distribution with these two properties. Lemma (31) below makes this precise.

31 Lemma *Let Δ_c be the distribution generated by Procedure (28), and suppose Δ is another distribution which contains all $\mathbf{g}_i$, is involutive, and invariant under $\mathbf{f}$ over some neighborhood of $\mathbf{x}_0$. Then $\Delta_c \subseteq \Delta$.*

The proof is easy and is left as an exercise.

32 Remark Actually, it can be shown that the above algorithm can be simplified. In Step 1 it is only necessary to check if Δ_i is invariant under $\mathbf{f}$ and all $\mathbf{g}_l$ — it is *not* necessary to check that Δ_i is involutive, because that follows automatically; see Isidori (1989), Lemma 8.7, p. 62. Thus, in Step 2, it is *not* necessary to add Lie brackets of the form $[\mathbf{h}_j^{(i)}, \mathbf{h}_l^{(i)}]$ to the base of Δ_i; it is only necessary to add Lie brackets of the form $[\mathbf{f}, \mathbf{h}_j^{(i)}]$ and $[\mathbf{g}_l, \mathbf{h}_j^{(i)}]$ that do not already belong to Δ_i. In other words, (30) can be replaced by

$$\Delta_{i+1} = \text{span}\,\{\mathbf{h}_j^{(i)},\ 1 \le j \le k_i\} \cup \{[\mathbf{f}, \mathbf{h}_j^{(i)}],\ [\mathbf{g}_l, \mathbf{h}_j^{(i)}],\ 1 \le j \le k_i,\ 1 \le l \le m\}. \tag{33}$$

The intermediate distributions might be affected by this change, but not the final answer Δ_c.

There is yet another way of simplifying the implementation of Procedure (28). As mentioned in the discussion after Procedure (28), if $\mathbf{x}_0$ is a regular point of each Δ_i, then the algorithm will stop after at most $n-1$ repetitions of Step 2. Now suppose the procedure generates, after $i < n-1$ steps, a distribution Δ_i which contains all the vector fields $\mathbf{g}_1, \cdots, \mathbf{g}_m$; is invariant under $\mathbf{f}$; and is involutive. Suppose that, instead of stopping the procedure, we compute Δ_{i+1} using the expression (33). Then it is clear that $\Delta_{i+1} = \Delta_i$, in view of the assumptions on Δ_i. In fact, more is true, namely

$$\Delta_i = \Delta_{i+1} = \cdots = \Delta_{n-1} = \Delta_c. \tag{34}$$

Thus, in principle, one could simply keep applying the expression (33) $n-1$ times, *without* verifying whether Δ_i is invariant or involutive, verifying only that $\mathbf{x}_0$ is a regular point of each Δ_i. In such a case Δ_{n-1} will automatically equal Δ_c.

35 Example As an illustration of Theorem (25), consider the third-order, single-input, bilinear system described by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + u\mathbf{B}\mathbf{x},$$

in a neighborhood of the point $\mathbf{x}_0 = [1\ 1\ 1]'$, where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & -14 \\ 0 & 0 & 0 \\ 0 & 0 & -19 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

This system is of the form (1) with

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x},\ \mathbf{g}(\mathbf{x}) = \mathbf{B}\mathbf{x}.$$

Now let us apply Procedure (28) as amended by Remark (32) to this system to determine the smallest distribution that is (i) involutive, and (ii) invariant under both **f** and **g**. In computing this distribution, the simplification suggested by Remark (32) can be used. The procedure generates the following sequence:

$i = 0$. We begin by setting $\Delta_0 = \text{span}\{\mathbf{B}\mathbf{x}\}$. Then it is necessary to form the Lie brackets $[\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}]$ and $[\mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{x}]$ to see if Δ_0 is invariant under both $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$. Of course $[\mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{x}] = \mathbf{0}$, while

$$[\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}] = (\mathbf{B}\mathbf{A} - \mathbf{A}\mathbf{B})\mathbf{x} =: \mathbf{C}\mathbf{x},$$

where

$$\mathbf{C} = \mathbf{B}\mathbf{A} - \mathbf{A}\mathbf{B} = \begin{bmatrix} 0 & 0 & -48 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence set $\Delta_1 = \text{span}\{\mathbf{B}\mathbf{x}, \mathbf{C}\mathbf{x}\}$.

$i = 1$. To check if Δ_1 is invariant under both $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$, it is necessary to form the Lie brackets $[\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}]$, $[\mathbf{A}\mathbf{x}, \mathbf{C}\mathbf{x}]$, $[\mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{x}]$, and $[\mathbf{B}\mathbf{x}, \mathbf{C}\mathbf{x}]$. Now $[\mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{x}]$ is of course zero, while $[\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}] = \mathbf{C}\mathbf{x} \in \Delta_1$. Next, it is routine to verify that

$$[\mathbf{A}\mathbf{x}, \mathbf{C}\mathbf{x}] = (\mathbf{C}\mathbf{A} - \mathbf{A}\mathbf{C})\mathbf{x} = -19\mathbf{C}\mathbf{x},\ [\mathbf{B}\mathbf{x}, \mathbf{C}\mathbf{x}] = (\mathbf{C}\mathbf{B} - \mathbf{B}\mathbf{C})\mathbf{x} = 3\mathbf{C}\mathbf{x}.$$

Hence Δ_1 is invariant under both $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$. The second equation above also shows that Δ_1 is involutive.

At this stage the reader may ask: Why restrict the analysis to a neighborhood of $\mathbf{x}_0$? Why not consider all of $\mathbf{R}^3$? The difficulty is that, if we consider a neighborhood of the origin for example, then at $\mathbf{x} = \mathbf{0}$ both $\mathbf{C}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$ equal the zero vector; hence, at $\mathbf{x} = \mathbf{0}$, Δ_1 spans a zero-dimensional subspace of $\mathbf{R}^3$ (i.e., a single point). On the other hand, every neighborhood of $\mathbf{0}$ contains a vector $\mathbf{x}$ at which $\mathbf{B}\mathbf{x}$ and $\mathbf{C}\mathbf{x}$ are linearly independent vectors; for example, take $\mathbf{x} = \varepsilon[0\ 0\ 1]'$ for sufficiently small ε. Hence, over any neighborhood of $\mathbf{0}$, both Δ_0 and Δ_1 are *singular* distributions, i.e., they do not have a constant dimension. But this problem does not arise if the analysis is restricted to a sufficiently small neighborhood of $\mathbf{x}_0$, because $\mathbf{B}\mathbf{x}$ and $\mathbf{C}\mathbf{x}$ are linearly independent.

Now back to the example. Since Δ_1 is involutive, the Frobenius theorem guarantees that there exist a smooth function $h(\mathbf{x})$ and a neighborhood U of $\mathbf{x}_0$ such that

$$\langle \mathbf{d}h, \mathbf{B}\mathbf{x} \rangle \equiv 0,\ \langle \mathbf{d}h, \mathbf{C}\mathbf{x} \rangle \equiv 0,\ \forall \mathbf{x} \in U.$$

Thus h satisfies the two partial differential equations

$$h_1(x_1 + 2x_2 + 4x_3) + 2h_2x_2 + 3h_3x_3 = 0,\ h_1x_3 = 0,$$

where $h_i = \partial h/\partial x_i$, and in the second equation the extraneous factor -48 has been omitted. A solution of the above pair of partial differential equations is given by

$$h(\mathbf{x}) = x_2^{-3}x_3^2.$$

To put the system in the form (26), it is necessary to find a new set of variables $\mathbf{y}$ such that $y_3 = h(\mathbf{x})$, and such that the transformation from $\mathbf{x}$ to $\mathbf{y}$ is a local diffeomorphism. An easy way to do this is the following: Compute

$$\mathbf{d}h(\mathbf{x}_0) = [0\ \ -3\ \ 2].$$

Choose two other row vectors $\mathbf{a}$ and $\mathbf{b}$ such that the set $\{\mathbf{a}, \mathbf{b}, \mathbf{d}h(\mathbf{x}_0)\}$ is a row basis for $\mathbf{R}^3$, and define

$$y_1 = \mathbf{a}\mathbf{x},\ y_2 = \mathbf{b}\mathbf{x},\ y_3 = h(\mathbf{x}).$$

Then, since the Jacobian matrix of this transformation is nonsingular at $\mathbf{x}_0$, it follows from the inverse function theorem [Theorem (7.1.1)] that the above map is a local diffeomorphism in some neighborhood of $\mathbf{x}_0$. In the present case, a simple choice is

$\mathbf{a} = [1\ \ 0\ \ 0]'$, $\mathbf{b} = [0\ \ 1\ \ 0]'$, and

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} =: T(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_2^{-3}x_3^2 \end{bmatrix}.$$

Then the inverse transformation is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = T^{-1}(\mathbf{y}) = \begin{bmatrix} y_1 \\ y_2 \\ y_2^{3/2}y_3^{1/2} \end{bmatrix}.$$

The system description in terms of the new variables can be computed directly, or by using the definition (7.1.9) for transforming vector fields. Let us adopt the latter procedure. The Jacobian matrix of the transformation T is given by

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3x_2^{-4}x_3^2 & 2x_2^{-3}x_3 \end{bmatrix}.$$

So the vector field $\mathbf{A}\mathbf{x} =: \mathbf{f}(\mathbf{x})$ gets transformed into

$$\mathbf{f}_T(\mathbf{y}) = \mathbf{J}(\mathbf{x})\mathbf{A}\mathbf{x}\,|_{\mathbf{x}=T^{-1}(\mathbf{y})}$$

$$= \begin{bmatrix} -14x_3 \\ 0 \\ -2x_2^{-3}x_3 \cdot 19x_3 \end{bmatrix}_{\mathbf{x}=T^{-1}(\mathbf{y})} = \begin{bmatrix} -14y_2^{3/2}y_3^{1/2} \\ 0 \\ -38y_3 \end{bmatrix}.$$

Similarly the vector field $\mathbf{Bx} =: \mathbf{g}(\mathbf{x})$ gets transformed into

$$\mathbf{g}_T(\mathbf{y}) = \begin{bmatrix} x_1 + 2x_2 + 4x_3 \\ 2x_2 \\ (-3x_2^{-4}x_3^2)\cdot 2x_2 + (2x_2^{-3}x_3)\cdot 3x_3 \end{bmatrix}_{\mathbf{x}=T^{-1}(\mathbf{y})} = \begin{bmatrix} y_1 + 2y_2 + 4y_2^{3/2}y_3^{1/2} \\ 2y_2 \\ 0 \end{bmatrix}.$$

Hence, in terms of the new variables, the system equations are

$$\dot{y}_1 = -14y_2^{3/2}y_3^{1/2} + u\,(y_1 + 2y_2 + 4y_2^{3/2}y_3^{1/2}),$$

$$\dot{y}_2 = 2uy_2,$$

$$\dot{y}_3 = -38y_3.$$

So the variable y_3 is uncoupled from the other variables y_1 and y_2 as well as from the input u. ■

Next, we explore the relationship between the distribution Δ_c produced by Procedure (28) and the familiar matrix

36 $\mathbf{W} = [\mathbf{B}\ \mathbf{AB}\ \cdots\ \mathbf{A}^{n-1}\mathbf{B}]$

associated with the linear control system (6). In particular, it is shown that, if one views each column vector of $\mathbf{W}$ as defining a constant vector field on $\mathbf{R}^n$, then the distribution Δ_c produced by Procedure (28) is precisely the span of the column vectors of $\mathbf{W}$. Now the system (6) is of the form (1) with

37 $\mathbf{f}(\mathbf{x}) = \mathbf{Ax}$, $\mathbf{g}_i = \mathbf{b}_i$, $i = 1, \cdots, m$,

where $\mathbf{b}_i$ is the i-th column of the matrix $\mathbf{B}$. Applying Procedure (28) gives

38 $\Delta_0 = \text{span}\,\{\mathbf{b}_1, \cdots, \mathbf{b}_m\} = \text{span}\,\mathbf{B}$.

Since each of the vector fields generating Δ_0 is constant, Δ_0 has the same rank at all $\mathbf{x}_0 \in \mathbf{R}^n$, i.e., every $\mathbf{x}_0 \in \mathbf{R}^n$ is a regular point of Δ_0. Now it is claimed that

39 $\Delta_i = \text{span}\{\mathbf{A}^l\mathbf{b}_j,\ 1 \le j \le m,\ 0 \le l \le i\}$.

The proof of (39) follows quite easily by induction on i. Obviously (39) is true when $i = 0$. Now suppose (39) is true for a particular value of i, and that Δ_{i+1} is computed according to (33). Since the Lie bracket to two constant vector fields is zero, and since $[\mathbf{Ax}, \mathbf{A}^l\mathbf{b}_j] = -\mathbf{A}^{l+1}\mathbf{b}_j$, it follows that

40 $\Delta_{i+1} = \text{span}\{\mathbf{A}^l\mathbf{b}_j,\ 1 \le j \le m,\ 0 \le l \le i+1\}$.

In particular, we see that $\Delta_c = \Delta_{n-1} = \text{span}\,\mathbf{W}$. In other words, Δ_c is spanned by the constant vector fields comprising the columns of the matrix $\mathbf{W}$.

Thus far we have introduced the distribution Δ_c via Procedure (28), and have shown that if Δ_c has dimension less than n, then the system (1) can be put into the "triangular" form (26). Clearly, the system (26) is not locally reachable, because the time evolution of the vector $\mathbf{y}_b$ is completely unaffected by the input $\mathbf{u}$. Hence, if $\dim \Delta_c < n$, then the system (1) is not reachable. But is the converse true: If Δ_c has dimension n, does this mean that the system (1) is locally reachable in the sense of Definition (5)? The answer is yes, though the proof is beyond the scope of the book. The interested reader is referred to Isidori (1989), Theorem 8.13, p. 69. The next theorem states the result formally.

41 Theorem *For the system (1), the following statements are equivalent:*

(i) *The system is locally reachable around* $\mathbf{x}_0 \in \mathbf{R}^n$ *in the sense of Definition (5).*

(ii) *There is a neighborhood* U *of* $\mathbf{x}_0$ *such that the distribution* Δ_c *constructed in accordance with Procedure (28) has dimension* n *at all* $\mathbf{x} \in U$.

Note that, when specialized to the linear system (6), Theorem (41) reduces to the familiar statement that the system (6) is reachable if and only if the matrix $\mathbf{W}$ of (36) has rank n. But a naive generalization of this rank test to nonlinear systems is not valid, as brought out in Corollary (42) and Example (45) below, which incidentally illustrate an important difference between linear systems and nonlinear systems.

42 Corollary *Given the system (1), consider the distribution*

43 $\bar{\Delta}_{n-1} = \text{span}\{\text{ad}_{\mathbf{f}}^i\mathbf{g}_j,\ 1 \le j \le m,\ 0 \le i \le n-1\}$.

If $\bar{\Delta}_{n-1}(\mathbf{x}_0)$ *has dimension* n, *then the system is locally reachable around* $\mathbf{x}_0$ *in the sense of Definition (5).*

Proof Clearly $\bar{\Delta}_{n-1}$ is a subset of Δ_c as constructed by (34). Hence, if $\bar{\Delta}_{n-1}(\mathbf{x}_0)$ has dimension n, then so does Δ_c, whence the system is locally reachable by virtue of Theorem (41). ■

Note that $\bar{\Delta}_{n-1}$ can be formed iteratively according to the following procedure: Define $\bar{\Delta}_0 = \Delta_0$ as in (29), and define

44 $\bar{\Delta}_{i+1} = \text{span}\,\{\bar{\Delta}_i, [\mathbf{f}, \mathbf{r}_j^{(i)}], 1 \le j \le l_i\}$,

where l_i is the dimension of $\bar{\Delta}_i$, and $\mathbf{r}_1^{(i)}, \cdots, \mathbf{r}_{l_i}^{(i)}$ is a set of generating vector fields for $\bar{\Delta}_i$. Comparing (44) and (33) shows the difference between $\bar{\Delta}_{n-1}$ and Δ_c. In constructing the former, at each stage we do *not* take Lie brackets of the form $[\mathbf{g}_l, \mathbf{r}_j^{(i)}]$, as we do when constructing Δ_c. In the case of a *linear* system of the form (1) this makes no difference, because in this case each Δ_i is the span of a set of *constant* vector fields, and the Lie bracket of two constant vector fields is zero. Hence $\bar{\Delta}_i = \Delta_i$ for all i, and $\bar{\Delta}_{n-1} = \Delta_c$. So the converse of Corollary (42) is true for the linear system (6). But the next example shows that the converse of Corollary (42) is false in general — $\dim \bar{\Delta}_{n-1} = n$ is *not* a necessary condition for local reachability.

45 Example Consider once again the bilinear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + u\mathbf{B}\mathbf{x},$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & -2 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Suppose $\mathbf{x}_0 = [1 \;\; 1 \;\; 1]'$. Let U be a ball centered at $\mathbf{x}_0$, with a radius small enough that the origin does not belong to U. Then

$$\Delta_0 = \bar{\Delta}_0 = \text{span}\,\{\mathbf{g}\} = \text{span}\,\{\mathbf{B}\mathbf{x}\}.$$

$$\Delta_1 = \text{span}\,\{\mathbf{B}\mathbf{x}, [\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}], [\mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{x}]\} = \text{span}\,\{\mathbf{B}\mathbf{x}, [\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}]\}$$

$$= \bar{\Delta}_1 = \text{span}\,\{\mathbf{B}\mathbf{x}, \mathbf{C}\mathbf{x}\},$$

where

$$\mathbf{C} = \mathbf{B}\mathbf{A} - \mathbf{A}\mathbf{B} = \begin{bmatrix} 0 & 0 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Next,

$$\bar{\Delta}_2 = \text{span}\,\{\bar{\Delta}_1, [\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}], [\mathbf{A}\mathbf{x}, \mathbf{C}\mathbf{x}]\}.$$

But $[\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}] = \mathbf{C}\mathbf{x} \in \bar{\Delta}_1$, while $[\mathbf{A}\mathbf{x}, \mathbf{C}\mathbf{x}] = \mathbf{D}\mathbf{x}$, where

$$\mathbf{D} = \mathbf{CA} - \mathbf{AC} = \begin{bmatrix} 0 & 0 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = -2\mathbf{C}.$$

Hence $[\mathbf{Ax}, \mathbf{Cx}] \in \bar{\Delta}_1$, and as a result $\bar{\Delta}_2 = \bar{\Delta}_1$. Of course $\dim \bar{\Delta}_2 = 2$. However,

$$\Delta_2 = \text{span}\,\{\Delta_1, [\mathbf{Ax}, \mathbf{Bx}], [\mathbf{Ax}, \mathbf{Cx}], [\mathbf{Bx}, \mathbf{Bx}], [\mathbf{Bx}, \mathbf{Cx}]\}.$$

Now $[\mathbf{Bx}, \mathbf{Bx}] = \mathbf{0}$, and we have already seen that

$$\text{span}\,\{\Delta_1, [\mathbf{Ax}, \mathbf{Bx}], [\mathbf{Ax}, \mathbf{Cx}]\} = \Delta_1.$$

But $[\mathbf{Bx}, \mathbf{Cx}] = \mathbf{Ex}$, where

$$\mathbf{E} = \mathbf{CB} - \mathbf{BC} = \begin{bmatrix} 0 & 0 & -4 \\ 0 & 0 & 4 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence

$$\Delta_2 = \text{span}\,\{\mathbf{Bx}, \mathbf{Cx}, \mathbf{Ex}\}.$$

At $\mathbf{x} = \mathbf{x}_0$, we have

$$\Delta_2(\mathbf{x}_0) = \text{span}\left\{\begin{bmatrix} 7 \\ 4 \\ 3 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 4 \\ 0 \end{bmatrix}\right\}.$$

It is easy to verify that $\dim \Delta_2(\mathbf{x}_0) = 3$. It now follows from Theorem (41) that the system is locally reachable. ■

The next theorem gives a simple *sufficient* condition for the system (1) to be locally reachable *around an equilibrium*, i.e., a vector $\mathbf{x}_0 \in X$ such that $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$.

46 Theorem *Consider the system (1), and suppose* $\mathbf{x}_0 \in X$ *satisfies* $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$. *Define the matrix* $\mathbf{A}_0 \in \mathbf{R}^{n \times n}$ *and the vectors* $\mathbf{b}_{i0} \in \mathbf{R}^n$, $i = 1, \cdots, m$, *by*

$$\mathbf{47} \qquad \mathbf{A}_0 = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{x} = \mathbf{x}_0}, \quad \mathbf{b}_{i0} = \mathbf{g}_i(\mathbf{x}_0).$$

Consider the linearized system

$$\mathbf{48} \quad \dot{\mathbf{z}} = \mathbf{A}_0\mathbf{z} + \sum_{i=1}^{m} \mathbf{b}_{i0} v_i.$$

Then the system (1) is locally reachable if the system (48) is reachable, i.e., if the n×nm matrix

$$\mathbf{49} \quad \mathbf{W}_0 = [\mathbf{B}_0 \ \mathbf{A}_0\mathbf{B}_0 \cdots \mathbf{A}_0^{n-1}\mathbf{B}_0]$$

has rank n, where

$$\mathbf{50} \quad \mathbf{B}_0 = [\mathbf{b}_{10} \cdots \mathbf{b}_{m0}].$$

Proof In fact we prove something more, namely: For *every* sufficiently small $T > 0$, the set of states reachable from $\mathbf{x}_0$ at time T contains a neighborhood of $\mathbf{x}_0$. Accordingly, suppose $T > 0$ is specified. Let $\mathbf{e}_1, \cdots, \mathbf{e}_n$ denote the elementary unit vectors in $\mathbf{R}^n$; that is, $\mathbf{e}_i$ contains a "1" in the i-th row and zeros elsewhere. Now by the assumption that the linearized system (48) is reachable, there exists a control input $\mathbf{v}^i(\cdot)$ (which is an m-vector valued function) that steers the system (48) from the state $\mathbf{z} = \mathbf{0}$ at time 0 to the state $\mathbf{z} = \mathbf{e}_i$ at time T. Indeed there are n such input functions $\mathbf{v}^1(\cdot), \cdots, \mathbf{v}^n(\cdot)$. Now, for any $\mathbf{r} = [r_1 \cdots r_n]' \in \mathbf{R}^n$, define the control input

$$\mathbf{51} \quad \mathbf{u}_\mathbf{r} = \sum_{j=1}^{n} r_j \mathbf{v}^j.$$

Let $\mathbf{x}(t, \mathbf{r})$ denote the solution of (1) starting at the initial condition $\mathbf{x}(0) = \mathbf{x}_0$, and with the input $\mathbf{u}_\mathbf{r}$. Then $\mathbf{x}(T, \mathbf{r})$ is well-defined whenever $\|\mathbf{r}\|$ is sufficiently small, say $\|\mathbf{r}\| < \varepsilon$. Let B_ε denote the ball in $\mathbf{R}^n$ of radius ε and centered at $\mathbf{0}$, and define $\mathbf{h}: B_\varepsilon \to \mathbf{R}^n$ by

$$\mathbf{52} \quad \mathbf{h}(\mathbf{r}) = \mathbf{x}(T, \mathbf{r}).$$

Note that $\mathbf{h}(\mathbf{0}) = \mathbf{x}_0$; this follows from the fact that $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$. It is now shown that $[\partial\mathbf{h}/\partial\mathbf{r}](\mathbf{x}_0)$ is nonsingular. It then follows from the inverse function theorem [Theorem (7.1.1)] that locally $\mathbf{h}$ is a diffeomorphism around $\mathbf{x}_0$. Thus, given any $\mathbf{x}_f$ sufficiently close to $\mathbf{x}_0$, there exists an $\mathbf{r} \in \mathbf{R}^n$ such that

$$\mathbf{53} \quad \mathbf{x}_f = \mathbf{h}(\mathbf{r}) = \mathbf{x}(T, \mathbf{r}).$$

Hence the input $\mathbf{u}_\mathbf{r}$ defined by (51) steers the system from $\mathbf{x}_0$ to $\mathbf{x}_f$. This shows that the proof is complete once it is established that $[\partial\mathbf{h}/\partial\mathbf{r}](\mathbf{x}_0)$ is nonsingular.

For this purpose, note that $\mathbf{x}(\cdot, \mathbf{r})$ satisfies the differential equation

$$\mathbf{54} \quad \dot{\mathbf{x}}(t, \mathbf{r}) = \mathbf{f}[\mathbf{x}(t, \mathbf{r})] + \sum_{i=1}^{m} (\mathbf{u}_\mathbf{r})_i \, \mathbf{g}_i[\mathbf{x}(t, \mathbf{r})] = \mathbf{f}[\mathbf{x}(t, \mathbf{r})] + \sum_{i=1}^{m} \sum_{j=1}^{m} r_j \, (\mathbf{v}^j)_i(t) \, \mathbf{g}_i[\mathbf{x}(t, \mathbf{r})].$$

Interchanging the order of summation and differentiating with respect to r_j gives

$$55 \quad \frac{d}{dt}\frac{\partial \mathbf{x}(t, \mathbf{r})}{\partial r_j} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{x}(t, \mathbf{r})} \frac{\partial \mathbf{x}(t, \mathbf{r})}{\partial r_j} + \sum_{i=1}^{m} (\mathbf{v}^j)_i(t)\, \mathbf{g}_i[\mathbf{x}(t, \mathbf{r})] + \sum_{i=1}^{m} (\mathbf{u}_\mathbf{r})_i \left[\frac{\partial \mathbf{g}_i}{\partial \mathbf{x}}\right]_{\mathbf{x}(t, \mathbf{r})} \frac{\partial \mathbf{x}}{\partial \mathbf{r}}(t, \mathbf{r}).$$

Define

$$56 \quad \mathbf{m}_j(t) = \left[\frac{\partial \mathbf{x}(t, \mathbf{r})}{\partial r_j}\right]_{\mathbf{r}=\mathbf{0}}.$$

A differential equation governing $\mathbf{m}_j(\cdot)$ can be obtained from (55) by substituting $\mathbf{r} = \mathbf{0}$. In this case $\mathbf{x}(t, \mathbf{0}) = \mathbf{x}_0$,

$$57 \quad \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{x}(t, \mathbf{0})} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}_0} = \mathbf{A}_0, \text{ and}$$

$$58 \quad \mathbf{g}_i[\mathbf{x}(t, \mathbf{0})] = \mathbf{g}_i(\mathbf{x}_0) = \mathbf{b}_{i0}.$$

Also, if $\mathbf{r} = \mathbf{0}$, then the last term on the right side of (55) drops out, since $(\mathbf{u}_\mathbf{r})_{\mathbf{r}=\mathbf{0}} = \mathbf{0}$. Hence

$$59 \quad \dot{\mathbf{m}}_j(t) = \mathbf{A}_0\, \mathbf{m}_j(t) + \sum_{i=1}^{m} (\mathbf{v}^j)_i(t)\, \mathbf{b}_{i0} = \mathbf{A}_0\, \mathbf{m}_j(t) + \mathbf{B}_0\, \mathbf{v}^j(t).$$

By the manner in which the functions $\mathbf{v}_j(\cdot)$ were chosen, we know that

$$60 \quad \mathbf{m}_j(T) = \mathbf{e}_j.$$

Hence

$$61 \quad \frac{\partial \mathbf{h}(\mathbf{r})}{\partial \mathbf{r}} = \frac{\partial \mathbf{x}(T, \mathbf{r})}{\partial \mathbf{r}} = I,$$

which is obviously nonsingular. ■

62 Example Consider a mass constrained by a nonlinear spring and a nonlinear viscous damper, and driven by an external force. Such a system is described by

$$m\ddot{r} + d(\dot{r}) + k(r) = u,$$

where the various quantities are defined as follows:

m Mass of the object

r Position of the object

d Frictional damping force

k Restoring force of the spring

Using the natural state vector

$$\mathbf{x} = \begin{bmatrix} r \\ \dot{r} \end{bmatrix},$$

one can represent the system dynamics in the form (1), with

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2 \\ -d(x_2) - h(x_1) \end{bmatrix}, \mathbf{g}(\mathbf{x}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Now let $\mathbf{x}_0 = \mathbf{0}$, and note that $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. The matrices defined in (48) become

$$\mathbf{A}_0 = \begin{bmatrix} 0 & 1 \\ -k'(0) & -d'(0) \end{bmatrix}, \mathbf{b}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

where k', d' denote the derivatives of k and d respectively. It is easy to see that the linearized system is reachable, whatever be the constants $k'(0)$ and $d'(0)$. Hence, by Theorem (46), the original nonlinear system is also locally reachable around $\mathbf{0}$.

7.3.2 Observability

Now let us study the observability of systems described by the state equation (1) and the output equation (3). First, a few concepts are introduced.

63 Definition *Consider a system described by (1) and (3). Two states* $\mathbf{x}_0$ *and* $\mathbf{x}_1$ *are said to be* **distinguishable** *if there exists an input function* $\mathbf{u}(\cdot)$ *such that*

64 $\mathbf{y}(\cdot, \mathbf{x}_0, \mathbf{u}) \neq \mathbf{y}(\cdot, \mathbf{x}_1, \mathbf{u})$,

where $\mathbf{y}(\cdot, \mathbf{x}_i, \mathbf{u})$, $i = 1, 2$ *is the output function of the system (1) – (3) corresponding to the input function* $\mathbf{u}(\cdot)$ *and the initial condition* $\mathbf{x}(0) = \mathbf{x}_i$. *The system is said to be* **(locally) observable at** $\mathbf{x}_0 \in X$ *if there exists a neighborhood* N *of* $\mathbf{x}_0$ *such that every* $\mathbf{x} \in N$ *other than* $\mathbf{x}_0$ *is distinguishable from* $\mathbf{x}_0$. *Finally, the system is said be* **(locally) observable** *if it is locally observable at each* $\mathbf{x}_0 \in X$.

As is the case with reachability, there are several subtleties in the observability of nonlinear systems that have no analog in the case of linear systems. These are illustrated through several examples.

65 Example According to Definition (63), two states $\mathbf{x}_0$ and $\mathbf{x}_1$ are distinguishable if (64) holds for *some* choice of input function $\mathbf{u}(\cdot)$; there is no requirement that (64) hold for *all* inputs $\mathbf{u}(\cdot)$. Now for a linear system described by (2) and (4), it is easy to show that the following three statements are all equivalent: (i) (64) holds for *some* input $\mathbf{u}(\cdot)$; (ii) (64) holds for *all* inputs $\mathbf{u}(\cdot)$; (iii) (64) holds with $\mathbf{u} \equiv \mathbf{0}$. This is because, in the case of a linear system described by (2) and (4), the output $\mathbf{y}$ is the *sum* of the zero-input response and the zero-state response. Since such a decomposition is not possible in general for a nonlinear system, the above three statements are *not* equivalent in the case of a general nonlinear system.

To illustrate this point, consider the bilinear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + u\mathbf{B}\mathbf{x}, \; y = \mathbf{c}\mathbf{x},$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{c} = [0 \;\; 1 \;\; 0].$$

Let $\mathbf{x}_0 \equiv \mathbf{0}$. Suppose we set $u \equiv 0$. Then a routine calculation shows that the pair $(\mathbf{c}, \mathbf{A})$ is not observable. In particular, if we let $\mathbf{x}_1 = [1 \;\; 0 \;\; 0]'$, then

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{cA} \\ \mathbf{cA}^2 \end{bmatrix} \mathbf{x}_1 = \mathbf{0}.$$

By familiar arguments in linear system theory, this implies that the states $\mathbf{0}$ and $\mathbf{x}_1$ cannot be distinguished with zero input, i.e.,

$$y(\cdot, \mathbf{0}, 0) = y(\cdot, \mathbf{x}_1, 0).$$

However, suppose we choose a constant input $u(t) \equiv 1 \; \forall t$. Then the system equations become

$$\dot{\mathbf{x}} = (\mathbf{A} + \mathbf{B})\mathbf{x}, \; \mathbf{y} = \mathbf{c}\mathbf{x}.$$

Now, as can be easily verified, the pair $(\mathbf{c}, \mathbf{A} + \mathbf{B})$ is not observable either, *but*

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{c}(\mathbf{A}+\mathbf{B}) \\ \mathbf{c}(\mathbf{A}+\mathbf{B})^2 \end{bmatrix} \mathbf{x}_1 \neq \mathbf{0}.$$

This means that

$$y(\cdot, \mathbf{0}, 1) \neq y(\cdot, \mathbf{x}_1, 1).$$

Hence the states $\mathbf{0}$ and $\mathbf{x}_1$ are distinguishable [by the constant input $\mathbf{u}(t) \equiv 1 \ \forall t$].

Let us carry on the argument. Suppose we apply a constant input $u(t) \equiv k \ \forall t$. Then the state equation becomes

$$\dot{\mathbf{x}} = (\mathbf{A} + k\mathbf{B})\mathbf{x}.$$

Now

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{c}(\mathbf{A}+k\mathbf{B}) \\ \mathbf{c}(\mathbf{A}+k\mathbf{B})^2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ k & 0 & 1 \\ 0 & k & 0 \end{bmatrix} =: \mathbf{M}, \text{ say.}$$

Clearly $\mathbf{M}$ is singular. Thus the pair $(\mathbf{c}, \mathbf{A} + k\mathbf{B})$ is *unobservable* for each fixed k. This means that *there is no constant input* that would permit us to distinguish *all* nonzero states from $\mathbf{0}$. However, let us fix k, and ask: What are the states that cannot be distinguished from $\mathbf{0}$? These are precisely the (nonzero) states that produce an identically zero output, i.e., the states $\mathbf{x}$ such that $\mathbf{Mx} = \mathbf{0}$. An easy calculation shows that the states that cannot be distinguished from $\mathbf{0}$ with $u(t) \equiv k$ are

$$\{\alpha[1 \ \ 0 \ \ -k]', \alpha \neq 0\}.$$

Now comes the important point. Given any $\mathbf{x} \neq 0$, *there exists a choice of* k such that

$$\mathbf{x} \neq \alpha[1 \ \ 0 \ \ -k]' \ \forall \alpha \in \mathbf{R}.$$

Hence, *with this particular choice of input,* the states $\mathbf{0}$ and $\mathbf{x}$ can be distinguished. So we see that the system under study *is* (locally) observable.

66 Example It is possible to define a system to be **globally observable** if *every* pair of states $(\mathbf{x}_0, \mathbf{x}_1)$ with $\mathbf{x}_0 \neq \mathbf{x}_1$ is distinguishable. However, this concept is much stronger than local observability. Consider the system

$$\dot{x} = u, \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos x \\ \sin x \end{bmatrix}.$$

Given y_1 and y_2, one can uniquely determine x to within a multiple of 2π. Hence each x is distinguishable from all other *nearby* states, and the system is (locally) observable. However, since x and $x + 2\pi$ cannot be distinguished, the system is *not* globally observable. ■

Now let us derive necessary and sufficient conditions for the system (1) – (3) to be locally observable. For this purpose, it is useful to recall how the standard observability rank test for linear systems is derived. Consider a linear system described by (2) and (4). Suppose we know $\mathbf{u}$ and can measure $\mathbf{y}$; assume for the sake of convenience that $\mathbf{u}(t)$ is a smooth function of t, i.e., has derivatives of all order. Then successive differentiation of the

output equation (4) gives

67 $\mathbf{y}(t) = \mathbf{C}\,\mathbf{x}(t)$,

$\dot{\mathbf{y}}(t) = \mathbf{C}\,\dot{\mathbf{x}}(t) = \mathbf{CA}\,\mathbf{x}(t) + \mathbf{CB}\,\mathbf{u}(t)$,

$\ddot{\mathbf{y}}(t) = \mathbf{CA}^2\,\mathbf{x}(t) + \mathbf{CAB}\,\mathbf{u}(t) + \mathbf{CB}\,\dot{\mathbf{u}}(t)$, $\cdots$

Hence, by successively differentiating $\mathbf{y}$, we can *infer* the quantities

68 $\mathbf{C}\,\mathbf{x}(t)$, $\mathbf{CA}\,\mathbf{x}(t)$, $\mathbf{CA}^2\,\mathbf{x}(t)$, $\cdots$

after *subtracting* the known quantities $\mathbf{CB}\,\mathbf{u}(t)$, $\mathbf{CAB}\,\mathbf{u}(t)$, $\mathbf{CB}\,\dot{\mathbf{u}}(t)$, etc. Now (68) shows that if the matrix

69 $$\mathbf{W}_0 = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{CA}^{n-1} \end{bmatrix}$$

has rank n, then it is possible to determine $\mathbf{x}(t)$ uniquely. (Of course, there is no need to go beyond $\mathbf{A}^{n-1}$ because of the Cayley-Hamilton theorem.)

For nonlinear systems the idea is pretty much the same. Let l denote the number of outputs, and let y_j, $h_j(\mathbf{x})$ denote respectively the j-th components of $\mathbf{y}$ and $\mathbf{h}(\mathbf{x})$. Then

70 $y_j = h_j(\mathbf{x})$,

71 $$\dot{y}_j = \mathbf{d}h_j\,\dot{\mathbf{x}} = \mathbf{d}h_j\,\mathbf{f}(\mathbf{x}) + \sum_{i=1}^{m} u_i\,\mathbf{d}h_j\,\mathbf{g}_i(\mathbf{x}) = (L_{\mathbf{f}}h_j)(\mathbf{x}) + \sum_{i=1}^{m} u_i\,(L_{\mathbf{g}_i}h_j)(\mathbf{x}),$$

where the Lie derivatives $L_{\mathbf{f}}h_j$ and $L_{\mathbf{g}_i}h_j$ are defined in accordance with (7.1.15), and the explicit dependence on t is not displayed in the interests of clarity. Differentiating one more time gives

72 $$\ddot{y}_j = (L_{\mathbf{f}}^2 h_j)(\mathbf{x}) + \sum_{i=1}^{m} u_i\,(L_{\mathbf{g}_i}L_{\mathbf{f}}h_j)(\mathbf{x}) + \sum_{i=1}^{m} \dot{u}_i\,(L_{\mathbf{g}_i}h_j)(\mathbf{x})$$

$$+ \sum_{i=1}^{m} u_i\,(L_{\mathbf{f}}L_{\mathbf{g}_i}h_j)(\mathbf{x}) + \sum_{i=1}^{m}\sum_{s=1}^{m} u_i u_s\,(L_{\mathbf{g}_s}L_{\mathbf{g}_i}h_j)(\mathbf{x}).$$

Expressions for higher derivatives of y_j get progressively nastier, but the pattern is clear

enough. The quantity $y_j^{(k)}$ is a "linear" combination of terms of the form $(L_{\mathbf{z}_s} L_{\mathbf{z}_{s-1}} \cdots L_{\mathbf{z}_1} h_j)(\mathbf{x})$, where $1 \le s \le k$, and each of the vector fields $\mathbf{z}_1, \cdots, \mathbf{z}_s$ is from the set $\{\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m\}$.

In view of the foregoing observation, Theorem (73) below seems quite plausible. However, a little effort is needed to prove the theorem.

73 Theorem (Sufficient Condition for Local Observability) *Consider the system described by (1) and (3), and suppose $\mathbf{x}_0 \in X$ is given. Consider the forms*

$$\mathbf{74} \quad (\mathbf{d}L_{\mathbf{z}_s} L_{\mathbf{z}_{s-1}} \cdots L_{\mathbf{z}_1} h_j)(\mathbf{x}_0),\ s \ge 0,\ \mathbf{z}_i \in \{\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m\},$$

evaluated at $\mathbf{x}_0$. Suppose there are n linearly independent row vectors in this set. Then the system is locally observable around $\mathbf{x}_0$.

Remarks The proof of Theorem (73) is based on several preliminary lemmas and is given by and by. But first it is shown that, when specialized to the linear system described by (2) and (4), the condition of Theorem (73) reduces to the familiar condition that the matrix $\mathbf{W}_o$ of (69) have rank n. Now (2) – (4) is of the form (1) – (3) with

$$\mathbf{75} \quad \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x},\ \mathbf{g}_i(\mathbf{x}) = \mathbf{b}_i,\ h_j(\mathbf{x}) = \mathbf{c}_j\mathbf{x},$$

where $\mathbf{b}_i$ denotes the i-th column of the matrix $\mathbf{B}$ and $\mathbf{c}_j$ denotes the j-th row of the matrix $\mathbf{C}$. Hence, with $s = 0$ in (74), we have

$$\mathbf{76} \quad (\mathbf{d}h_j)(\mathbf{x}) = \mathbf{c}_j,\ j = 1, \cdots, l.$$

Next,

$$\mathbf{77} \quad (L_{\mathbf{f}} h_j)(\mathbf{x}) = \mathbf{c}_j\mathbf{A}\mathbf{x},\ (L_{\mathbf{g}_i} h_j)(\mathbf{x}) = \mathbf{c}_j\mathbf{b}_i.$$

Therefore, the only *nonconstant* functions are $L_{\mathbf{f}} h_j$. Since $\mathbf{d}a = \mathbf{0}$ if a is a constant function, the only *nonzero* vectors of the form (74) with $s = 1$ are

$$\mathbf{78} \quad (\mathbf{d}L_{\mathbf{f}} h_j)(\mathbf{x}) = \mathbf{c}_j\mathbf{A},\ j = 1, \cdots, l.$$

When we take repeated Lie derivatives as in (74), the constant functions do not contribute anything. In fact, the only *nonzero* vectors of the form (74) are

$$\mathbf{79} \quad \mathbf{c}_j\mathbf{A}^k,\ k \ge 0,\ j = 1, \cdots, l.$$

Theorem (73) states that if this set contains n linearly independent row vectors, then the system is (locally) observable. Finally, the Cayley-Hamilton theorem permits one to conclude that the span of the vectors in (79) is exactly the same as the span of the set

80 $\mathbf{c}_j\mathbf{A}^k, k=0, \cdots, n-1, j=1, \cdots, l.$

So the sufficient condition for observability becomes

$$\textbf{81} \quad \operatorname{rank}\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{CA}^{n-1} \end{bmatrix} = n,$$

which is of course the familiar condition.

To prove Theorem (73), a preliminary concept is introduced.

82 Definition *Given the system described by (1) and (3), the* **observation space O** *of the system is the linear space of functions over the field* $\mathbf{R}$ *spanned by all functions of the form*

83 $L_{\mathbf{z}_s} \cdots L_{\mathbf{z}_1} h_j, s \geq 0, \mathbf{z}_1, \cdots, \mathbf{z}_s \in \{\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m\}, 1 \leq j \leq l.$

It is important to note that the observation space consists of all linear combinations of functions of the form (83) with *constant real* coefficients — *not* functions of $\mathbf{x}$. Also, if $s=0$ in (82), then the "zeroth-order" Lie derivative of h_j is to be interpreted as h_j itself.

The next lemma gives an alternative and useful interpretation of the observation space.

84 Lemma *For the system described by (1) and (3), let* $\mathbf{J}$ *denote the linear space of functions over the field* $\mathbf{R}$ *spanned by all functions of the form*

85 $L_{\mathbf{v}_s} \cdots L_{\mathbf{v}_1} h_j, s \geq 0, 1 \leq j \leq l,$

where $\mathbf{v}_1, \cdots, \mathbf{v}_s$ *are vector fields of the form*

$$\textbf{86} \quad \mathbf{v} = \mathbf{f} + \sum_{i=1}^{m} u_i \mathbf{g}_i$$

for some choice of real numbers $u_1, \cdots, u_m \in \mathbf{R}$. *Then* $\mathbf{J} = \mathbf{O}$.

Proof Note that if $\mathbf{v}$, $\mathbf{w}$ are vector fields, we have

87 $L_{\mathbf{v}+\mathbf{w}} h_j = L_{\mathbf{v}} h_j + L_{\mathbf{w}} h_j.$

Now note that (i) each vector field of the form (86) is a linear combination over $\mathbf{R}$ of the vector fields $\{\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m\}$, and (ii) conversely, each vector field in the set $\{\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m\}$ is a linear combination over $\mathbf{R}$ of vector fields of the form (86). It is obvious that (i) is true. To see (ii), observe first that $\mathbf{f}$ is of the form (86) — just set $u_i = 0 \; \forall i$. Next, we have

88 $\mathbf{g}_i = (\mathbf{f}+\mathbf{g}_i)-\mathbf{f},\ i = 1, \cdots, m.$

Hence $\mathbf{g}_i$ is also a linear combination of vector fields of the form (86). It follows that the span of the vector fields $\{\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m\}$ with coefficients in $\mathbf{R}$ is the same as the span of all vector fields of the form (86) with coefficients in $\mathbf{R}$. That $\mathbf{O} = \mathbf{J}$ now follows from repeated applications of (87). ■

Lemma (89) presents another technical result needed in the proof of Theorem (73).

89 Lemma *Let* $\mathbf{u}_1, \cdots, \mathbf{u}_s \in \mathbf{R}^m$, *and consider the piecewise-constant input*

90
$$\mathbf{u}(t) = \begin{cases} \mathbf{u}_1, & 0 \le t < t_1 \\ \mathbf{u}_2, & t_1 \le t < t_1 + t_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \mathbf{u}_s, & \sum_{k=1}^{s-1} t_k \le t \le \sum_{k=1}^{s} t_k \end{cases}$$

Let $y_j(\mathbf{x}_0) = y_j(\mathbf{x}_0, t_1, \cdots, t_s)$ *denote the j-th component of the system (1) – (3) corresponding to the control input* $\mathbf{u}(\cdot)$ *and the initial state* $\mathbf{x}_0$. *Then*

91
$$\left[\frac{\partial}{\partial t_1} \cdots \frac{\partial}{\partial t_s}\mathbf{y}\right]_{t_k = 0\ \forall k} = (\mathbf{d}L_{\mathbf{v}_s} \cdots L_{\mathbf{v}_1} h_j)(\mathbf{x}_0),$$

where $\mathbf{v}_k$ *is the vector field*

92
$$\mathbf{v}_k = \mathbf{f} + \sum_{i=1}^{m} u_{ki}\, \mathbf{g}_i,\ k = 1, \cdots, s.$$

Remarks The lemma says that if we apply a piecewise-constant control of the form (90) and then let the duration of the "pulses" shrink to 0, then the quantity in (91) is equal to a particular repeated Lie derivative.

Lemma (89) can be proved quite easily by induction on k; the proof is left as an exercise.

Proof of Theorem (73) Let $\mathbf{O}$ denote the observation space of the system, and consider the set of row vectors $\mathbf{d}\alpha(\mathbf{x}_0)$ as α varies over $\mathbf{O}$. This is a subspace of $(\mathbf{R}^n)^*$, the set of $1 \times n$ row vectors. Moreover, this subspace is precisely the span of the various row vectors in (74), and hence has dimension n by hypothesis. Now let $\mathbf{J}$ be as in Lemma (84). Since $\mathbf{J} = \mathbf{O}$ by Lemma (84), the hypothesis implies that there exist n linearly independent row vectors of the form $\mathbf{d}\beta_1(\mathbf{x}_0), \cdots, \mathbf{d}\beta_n(\mathbf{x}_0)$, where each β_i is a function of the form (85). Hence, by the inverse function theorem [Theorem (7.1.1)], it follows that the map

$$\mathbf{93} \quad T(\mathbf{x}) = \begin{bmatrix} \beta_1(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \beta_n(\mathbf{x}) \end{bmatrix}$$

is locally a diffeomorphism around $\mathbf{x}_0$.

Choose a neighborhood N of $\mathbf{x}_0$ such that $T: N \to T(N)$ is a diffeomorphism. Suppose $\mathbf{x}_1 \in N$ is *indistinguishable* from $\mathbf{x}_0$; it is shown that $\mathbf{x}_1 = \mathbf{x}_0$. In turn, this implies that every $\mathbf{x} \in N$ other than $\mathbf{x}_0$ is distinguishable from $\mathbf{x}_0$, i.e., that the system is observable at $\mathbf{x}_0$, which is the desired conclusion.

Accordingly, suppose $\mathbf{x}_1 \in N$ is indistinguishable from $\mathbf{x}_0$. This means that $\mathbf{y}(\cdot, \mathbf{x}_0, \mathbf{u}) = \mathbf{y}(\cdot, \mathbf{x}_1, \mathbf{u})$ for *all* inputs $\mathbf{u}(\cdot)$. In particular, let $\mathbf{u}$ be a piecewise-constant input of the form (90). Then $y_j(\mathbf{x}_0) = y_j(\mathbf{x}_1)$. Letting all $t_s \to 0$ and applying (91) shows that

$$\mathbf{94} \quad (\mathbf{d}L_{\mathbf{v}_s} \cdots \mathbf{d}L_{\mathbf{v}_1} h_j)(\mathbf{x}_0) = (\mathbf{d}L_{\mathbf{v}_s} \cdots \mathbf{d}L_{\mathbf{v}_1} h_j)(\mathbf{x}_1), \; j = 1, \cdots, l.$$

This relationship holds for *all* vector fields $\mathbf{v}$ of the form (86), and for *all* integers $s \geq 0$. Now comes the main point. Each function β_i in (93) is of the form (85) for a suitable choice of vector fields $\mathbf{v}_1, \cdots, \mathbf{v}_s$. Hence

$$\mathbf{95} \quad \beta_i(\mathbf{x}_0) = \beta_i(\mathbf{x}_1), \; i = 1, \cdots, n,$$

i.e., $T(\mathbf{x}_0) = T(\mathbf{x}_1)$. But since T is locally a diffeomorphism, this implies that $\mathbf{x}_0 = \mathbf{x}_1$. ■

96 Example Consider once again the system of Example (65). This system is of the form (1) – (3) with

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}, \; \mathbf{g}(\mathbf{x}) = \mathbf{B}\mathbf{x}, \; h(\mathbf{x}) = \mathbf{c}\mathbf{x}.$$

Hence, to apply Theorem (73), it is necessary to examine the row vectors

$$\mathbf{d}h(\mathbf{x}_0), (\mathbf{d}L_{\mathbf{f}}h)(\mathbf{x}_0), (\mathbf{d}L_{\mathbf{g}}h)(\mathbf{x}_0), (\mathbf{d}L_{\mathbf{f}}L_{\mathbf{g}}h)(\mathbf{x}_0), (\mathbf{d}L_{\mathbf{g}}L_{\mathbf{f}}h)(\mathbf{x}_0), \cdots$$

Routine calculations show that these vectors are independent of $\mathbf{x}_0$, and are respectively equal to

$$\mathbf{c}, \mathbf{cA}, \mathbf{cB}, \mathbf{cBA}, \mathbf{cAB}, \cdots$$

However,

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{cA} \\ \mathbf{cB} \end{bmatrix} = \begin{bmatrix} 0\ 1\ 0 \\ 0\ 0\ 1 \\ 1\ 0\ 0 \end{bmatrix},$$

which has rank 3. Hence the system is observable. ■

Theorem (73) gives only a *sufficient* condition for observability. Is it also necessary? Theorem (97) below gives a decomposition result which shows that the condition of Theorem (73) is "almost" necessary; see Corollary (112).

97 Theorem (Decomposition of Unobservable Systems) *Consider the system described by (1) and (3), and let* $\mathbf{x}_0 \in X$ *be given. Let* $\mathbf{O}$ *be as in Definition (82). For each* $\mathbf{x} \in X$, *let* $\mathbf{dO}(\mathbf{x})$ *denote the subspace of* $(\mathbf{R}^n)^*$ *consisting of all row vectors* $\mathbf{d}\alpha(\mathbf{x})$, $\alpha \in \mathbf{O}$. *Suppose there exists a neighborhood* N *of* $\mathbf{x}_0$ *such that*

98 $\dim \mathbf{dO}(\mathbf{x}) = k < n, \ \forall \mathbf{x} \in N.$

Then there exists a diffeomorphism T *on* N *such that, if we make the state variable transformation* $\mathbf{z} = T(\mathbf{x})$ *and partition* $\mathbf{z}$ *as*

$$\mathbf{99} \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_a \\ \mathbf{z}_b \end{bmatrix}, \ \mathbf{z}_a \in \mathbf{R}^k, \ \mathbf{z}_b \in \mathbf{R}^{n-k},$$

then the transformed vector fields $\mathbf{f}_T$ *and* $\mathbf{g}_{iT}$ *have the form*

$$\mathbf{100} \quad \mathbf{f}_T(\mathbf{z}) = \begin{bmatrix} \mathbf{f}_a(\mathbf{z}_a) \\ \mathbf{f}_b(\mathbf{z}_a, \mathbf{z}_b) \end{bmatrix}, \ \mathbf{g}_{iT}(\mathbf{z}) = \begin{bmatrix} \mathbf{g}_{ia}(\mathbf{z}_a) \\ \mathbf{g}_{ib}(\mathbf{z}_a, \mathbf{z}_b) \end{bmatrix},$$

where all vectors with the subscript a belong to $\mathbf{R}^k$, *all vectors with the subscript b belong to* $\mathbf{R}^{n-k}$, *and the function* $\mathbf{h}_T$ *defined by*

$$\mathbf{101} \quad \mathbf{h}_T(\mathbf{z}) := \mathbf{h}[T^{-1}(\mathbf{z})]$$

depends only on $\mathbf{z}_a$.

Remarks Equations (100) and (101) imply that, after the change of coordinates, the system equations assume the form

$$\mathbf{102} \quad \begin{bmatrix} \dot{\mathbf{z}}_a \\ \dot{\mathbf{z}}_b \end{bmatrix} = \begin{bmatrix} \mathbf{f}_a(\mathbf{z}_a) \\ \mathbf{f}_b(\mathbf{z}_a, \mathbf{z}_b) \end{bmatrix} + \sum_{i=1}^{m} u_i \begin{bmatrix} \mathbf{g}_{ia}(\mathbf{z}_a) \\ \mathbf{g}_{ib}(\mathbf{z}_a, \mathbf{z}_b) \end{bmatrix},$$

$$\mathbf{y} = \mathbf{h}_T(\mathbf{z}_a).$$

These equations make clear the fact that the vector $\mathbf{z}_b$ is "unobservable," since it does not influence $\mathbf{z}_a$ in any way, nor is it reflected in the output measurement. Actually, the

statement of the theorem can be strengthened by pointing out that the vector $\mathbf{z}_a$ *is* observable. The precise mathematical statement is as follows: Define $M = T(N)$; then M is the state space in the $\mathbf{z}$ coordinates. For a given $\mathbf{z}_0 \in M$, define $I(\mathbf{z}_0)$ to be the set of states in M that are *indistinguishable* from $\mathbf{z}_0$. Then the claim is that

103 $I(\mathbf{z}_0) = \{\mathbf{z} \in M : \mathbf{z}_a = \mathbf{z}_{0a}\}$.

In other words, the first k components of $\mathbf{z}$ are completely distinguishable, while the last $n-k$ components of $\mathbf{z}$ are completely indistinguishable. This stronger form of Theorem (97) is not difficult to prove, once Theorem (97) itself is established. The proof of this extension is left as an exercise.

Proof Choose smooth functions $\alpha_1, \cdots, \alpha_k \in S(X)$ such that

104 $\mathbf{dO}(\mathbf{x}_0) = \text{span}\{\mathbf{d}\alpha_1(\mathbf{x}_0), \cdots, \mathbf{d}\alpha_k(\mathbf{x}_0)\}$,

and choose a sufficiently small neighborhood N of $\mathbf{x}_0$ such that

105 $\mathbf{dO}(\mathbf{x}) = \text{span}\{\mathbf{d}\alpha_1(\mathbf{x}), \cdots, \mathbf{d}\alpha_k(\mathbf{x})\}, \forall \mathbf{x} \in N$.

Now choose some other $n-k$ functions $\beta_{k+1}, \cdots, \beta_n \in S(X)$ such that

106 $$T(\mathbf{x}) := \begin{bmatrix} \alpha_1(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \alpha_k(\mathbf{x}) \\ \beta_{k+1}(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \beta_n(\mathbf{x}) \end{bmatrix}$$

is a diffeomorphism on N, and define $\mathbf{z} = T(\mathbf{x})$. The question is: What do $\mathbf{f}_T$, $\mathbf{g}_{iT}$ and $\mathbf{h}_T$ look like?

Suppose α is a function belonging to the observation space $\mathbf{O}$. Then of course $\mathbf{d}\alpha(\mathbf{x}) \in \mathbf{dO}(\mathbf{x})\ \forall \mathbf{x} \in N$. Hence (105) implies that, after the coordinate transformation, every function in $\mathbf{O}$ depends only on $\mathbf{z}_a$. In other words, the gradient of every transformed function α_T has the form

107 $\mathbf{d}\alpha_T(\mathbf{z}) = [\mathbf{w}(\mathbf{z}_a)\ \ \mathbf{0}_{n-k}], \forall \mathbf{z} \in M = T(N)$.

Now, note first that each component h_j of the output function belongs to the observation

space $\mathbf{O}$. Hence $\mathbf{h}_T$ has the form (102), i.e., $\mathbf{h}_T$ depends only on $\mathbf{z}_a$. Next, let $\mathbf{v}$ be any one of the vector fields $\mathbf{f}_T, \mathbf{g}_{iT}$, $i = 1, \cdots, m$. Partition $\mathbf{v}$ as

$$\mathbf{108} \quad \mathbf{v}(\mathbf{z}) = \begin{bmatrix} \mathbf{v}_a(\mathbf{z}) \\ \mathbf{v}_b(\mathbf{z}) \end{bmatrix},$$

where, as usual, $\mathbf{v}_a \in \mathbf{R}^k$, and $\mathbf{v}_b \in \mathbf{R}^{n-k}$. Then, for $\alpha_T \in \mathbf{O}$, we have from (107) that

$$\mathbf{109} \quad (L_{\mathbf{v}}\alpha_T)(\mathbf{z}) = \mathbf{w}(\mathbf{z}_a)\,\mathbf{v}_a(\mathbf{z}).$$

The observation space is closed under Lie differentiation with respect to each of the vector fields $\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m$; in other words, if $\alpha \in \mathbf{O}$, then $L_{\mathbf{f}}\alpha \in \mathbf{O}$ and $L_{\mathbf{g}_i}\alpha \in \mathbf{O}\ \forall i$. This is clear from Definition (82). So it follows that, in (109), $L_{\mathbf{v}}\alpha_T$ is *also* independent of $\mathbf{z}_b$. Now let us choose $\alpha = \alpha_i$, the i-th "basis" function as in (105). Then it follows from (106) that

$$\mathbf{110} \quad \alpha_{iT} = z_i.$$

In other words, with the change of coordinates, α_i is just the i-th component of $\mathbf{z}$. Therefore,

$$\mathbf{111} \quad \mathbf{d}\alpha_{iT} = [0\ 0 \cdots 1 \cdots 0\ 0],$$

where the "1" occurs in the i-th position. Substituting this into (109) shows that the i-th component of $\mathbf{v}_a(\mathbf{z})$ is actually independent of $\mathbf{z}_b$. Since $\mathbf{v}$ is any one of the vector fields $\mathbf{f}$, $\mathbf{g}_1, \cdots, \mathbf{g}_m$, this is just putting (100) into words. ■

112 Corollary *Suppose the system described by (1) – (3) is observable. Then* $\mathbf{dO}(\mathbf{x})$ *has dimension n for all* $\mathbf{x}$ *belonging to an open dense subset of X.*

The proof is based on the fact that if $\dim \mathbf{dO}(\mathbf{x}) < n$ over some set which contains an interior point, then it is possible to decompose the system as in Theorem (97), and as a result the system is not locally observable. There are however some technicalities, so the reader is referred to Nijmeijer and van der schaft (1990), p. 97.

Thus far we have seen a decomposition of a nonlinear system into its reachable and its unreachable parts [cf. Lemma (17)], and another decomposition into its observable and its unobservable parts [cf. Theorem (97)]. Is it possible to combine the two decompositions? Recall that a similar decomposition is possible in the case of the linear system (2) – (4). Specifically, given the system (2) – (4), it is possible to find a nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{113} \quad \mathbf{TAT}^{-1} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \mathbf{A}_{13} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \mathbf{A}_{24} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{33} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{bmatrix}, \mathbf{TB} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

$$\mathbf{CT}^{-1} = [\mathbf{C}_1 \ \mathbf{0} \ \mathbf{C}_3 \ \mathbf{0}].$$

Hence, if we define a new state vector $\mathbf{z} = \mathbf{Tx}$ and partition $\mathbf{z}$ as

$$\mathbf{114} \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \mathbf{z}_4 \end{bmatrix},$$

then the various components of $\mathbf{z}$ can be identified as follows: (The state components comprising) $\mathbf{z}_1$ (are): controllable and observable; $\mathbf{z}_2$: controllable but unobservable; $\mathbf{z}_3$: observable but uncontrollable; $\mathbf{z}_4$: both uncontrollable and unobservable. Further, the input-output behavior of the system is determined solely by the matrices $\mathbf{C}_1, \mathbf{A}_{11}, \mathbf{B}_1$. Is there an analogous result for nonlinear systems?

The answer is yes. To state the result precisely, two preliminary definitions are introduced.

115 Definition *Let* $\mathbf{O}(\mathbf{x})$ *be the observation space introduced in Definition (82). For each* $\mathbf{x} \in X$, *define* $\ker \mathbf{dO}(\mathbf{x})$ *to be the subspace of* $\mathbf{R}^n$ *given by*

$$\mathbf{116} \quad \ker \mathbf{dO}(\mathbf{x}) = \{\mathbf{f} \in V(X): \langle \mathbf{d}\alpha(\mathbf{x}), \mathbf{f}(\mathbf{x}) \rangle = 0, \ \forall \alpha \in \mathbf{O}\}.$$

Define $\ker \mathbf{dO}$ *to be the distribution which assigns the subspace* $\ker \mathbf{dO}(\mathbf{x})$ *to each* $\mathbf{x} \in X$.

Thus $\ker \mathbf{dO}$ consists of all vector fields that are annihilated by each of the forms $\mathbf{d}\alpha$ as α varies over the observation space $\mathbf{O}$. Suppose $\mathbf{x}_0 \in X$ is given, and that $\dim \mathbf{dO}(\mathbf{x})$ is constant for all $\mathbf{x}$ in some neighborhood N of $\mathbf{x}_0$. To be specific, suppose $\dim \mathbf{dO}(\mathbf{x}) = k < n \ \forall \mathbf{x} \in N$. Then $\ker \mathbf{dO}$ is a distribution of dimension $n - k$; moreover, $\ker \mathbf{dO}$ is *automatically* involutive, since the condition (7.2.25) is satisfied.

117 Definition *Suppose* Δ_1 *and* Δ_2 *are distributions on* X. *Then* $\Delta_1 \dot{+} \Delta_2$ *is the distribution defined by*

$$\mathbf{118} \quad (\Delta_1 \dot{+} \Delta_2)(\mathbf{x}) = \Delta_1(\mathbf{x}) \dot{+} \Delta_2(\mathbf{x}), \ \forall \mathbf{x} \in X.$$

Recall that, for each $\mathbf{x} \in X$, $\Delta_1(\mathbf{x})$ and $\Delta_2(\mathbf{x})$ are subspaces of $\mathbf{R}^n$. Moreover, if M_1 and M_2 are subspaces of $\mathbf{R}^n$, then

$$\mathbf{119} \quad M_1 \dot{+} M_2 = \{\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2: \mathbf{v}_1 \in M_1, \ \mathbf{v}_2 \in M_2\}.$$

Now we come to the main decomposition result; the proof can be found in Nijmeijer and van der Schaft (1990), pp. 110-111.

120 Theorem *Consider the system (1) – (2). Let Δ_c be defined as the outcome of applying Procedure (28), and define* $\ker \mathbf{dO}$ *by (116). Suppose $\mathbf{x}_0 \in X$ is given, and suppose there exists a neighborhood N of $\mathbf{x}_0$ such that $\Delta_c(\mathbf{x})$,* $\ker \mathbf{dO}(\mathbf{x})$, *and $\Delta_c(\mathbf{x}) \dotplus \ker \mathbf{dO}(\mathbf{x})$ all have constant dimension as $\mathbf{x}$ varies over N. Then there exists a local diffeomorphism T on N such that the transformed vector fields $\mathbf{f}_T$, $\mathbf{g}_{iT}$ and the transformed function $\mathbf{h}_T$ have the following special forms: Partition $\mathbf{z} = T(\mathbf{x})$ as in (114). Then*

$$\textbf{121} \quad \mathbf{f}_T(\mathbf{x}) = \begin{bmatrix} \mathbf{f}_1(\mathbf{z}_1, \mathbf{z}_3) \\ \mathbf{f}_2(\mathbf{z}) \\ \mathbf{f}_3(\mathbf{z}_3) \\ \mathbf{f}_4(\mathbf{z}_3, \mathbf{z}_4) \end{bmatrix}, \mathbf{g}_{iT}(\mathbf{z}) = \begin{bmatrix} \mathbf{g}_1(\mathbf{z}_1, \mathbf{z}_3) \\ \mathbf{g}_2(\mathbf{z}) \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

$$\mathbf{h}_T(\mathbf{z}) = \mathbf{h}_T(\mathbf{z}_1, \mathbf{z}_3).$$

Moreover,

$$\textbf{122} \quad (\Delta_c)_T = \text{span}\,\{[\times \ \times \ \mathbf{0} \ \mathbf{0}]\},$$

$$\textbf{123} \quad (\ker \mathbf{dO})_T = \text{span}\,\{[\mathbf{0} \ \times \ \mathbf{0} \ \times]\}.$$

In (122) and (123), the symbol $\times$ is used as a shorthand for an arbitrary vector belonging to the appropriate subspace of $(\mathbf{R}^n)^*$. Thus (122) means: $(\Delta_c)_T$ consists of *all* vectors having that particular form; (123) should be interpreted similarly.

Equation (122) states that the state components $\mathbf{z}_1$ and $\mathbf{z}_2$ are controllable, while (123) states that the state components $\mathbf{z}_2$ and $\mathbf{z}_4$ are unobservable.

Problem 7.8 Consider the problem of controlling a satellite in space using gas jet actuators. The angular velocity of such a satellite is governed by

$$\mathbf{I}\dot{\boldsymbol{\omega}} = \boldsymbol{\omega} \times (\mathbf{I}\boldsymbol{\omega}) + \boldsymbol{\tau},$$

where $\mathbf{I}$ is the 3×3 inertia matrix of the satellite in a body-attached coordinate frame, $\boldsymbol{\omega}$ is the angular velocity vector, and $\boldsymbol{\tau}$ is the vector of externally applied torque. Suppose the coordinate frame is chosen to correspond to the principal axes of the satellite. Then [cf. Example (5.3.19)]

$$\mathbf{I} = \text{Diag}\,\{I_x, I_y, I_z\},$$

and the equations governing the motion become

$$\dot{x} = ayz + u_1,$$

$$\dot{y} = -bxz + u_2,$$

$$\dot{z} = cxy + u_3,$$

where

$$a = \frac{I_y - I_z}{I_x}, b = \frac{I_x - I_z}{I_y}, c = \frac{I_x - I_y}{I_z},$$

$$u_1 = \frac{\tau_1}{I_x}, u_2 = \frac{\tau_2}{I_y}, u_3 = \frac{\tau_3}{I_z},$$

and x, y, z are respectively short-forms of ω_x, ω_y, ω_z.

(a) Show that the system is *not* reachable if only one actuator is used. Show that, for example, if only u_1 is used, then the quantity $\phi_1 = cy^2 + bz^2$ is constant, whatever be $u_1(\cdot)$. Derive similar results for the case where only u_2 is used, and where only u_3 is used.

(b) Suppose $I_x > I_y > I_z > 0$. Show that the system is reachable if *any two* out of the three actuators are used.

(c) Suppose $I_x = I_y > I_z$, so that $c = 0$. Show that the system is reachable if u_1 and u_3 are used, and if u_2 and u_3 are used. Show that the system is *not* reachable if only u_1 and u_2 are used.

(d) Suppose the satellite is prefectly spherical, so that $I_x = I_y = I_z$. Show that the system is *not* reachable unless *all three* actuators are used.

Note: In the above simplified version of the problem, we are considering the reachability of the angular *velocity* vector alone. For a more thorough treatment which considers the reachability of both the angular *position* as well as the angular velocity, see Crouch (1984).

Problem 7.9 A simplified model for steering an automobile in a plane is given by [see Nijmeijer and van der Schaft (1990), p. 52]

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} \sin\theta \\ \cos\theta \\ 0 \end{bmatrix} u_1 + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u_2,$$

where (x, y) is the position of the centroid of the automobile, and θ is its orientation. Show that the system is reachable.

7.4 FEEDBACK LINEARIZATION: SINGLE-INPUT CASE

In this section and the next, we study an important application of differential-geometric methods, namely the possibility of transforming a given nonlinear system to a linear system via feedback control and a transformation of the state vector. In this section the case of single-input systems is considered, while multi-input systems are studied in the next section.

The problem studied in this section can be described as follows: Consider a system described by

$$\mathbf{1} \qquad \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + u\,\mathbf{g}(\mathbf{x}),$$

where $\mathbf{f}$ and $\mathbf{g}$ are smooth vector fields on some open set $X \subseteq \mathbf{R}^n$ containing $\mathbf{0}$, and $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. It is desired to know when there exist smooth functions $q, s \in S(X)$ with $s(\mathbf{x}) \neq 0$ for all $\mathbf{x}$ in some neighborhood of the origin, and a local diffeomorphism T on $\mathbf{R}^n$ with $T(\mathbf{0}) = \mathbf{0}$, such that if we define

$$\mathbf{2} \qquad v = q(\mathbf{x}) + s(\mathbf{x})\,u,$$

$$\mathbf{3} \qquad \mathbf{z} = T(\mathbf{x}),$$

then the resulting variables $\mathbf{z}$ and v satisfy a *linear* differential equation of the form

$$\mathbf{4} \qquad \dot{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{b}v,$$

where the pair $(\mathbf{A}, \mathbf{b})$ is controllable. If this is the case, then the system (1) is said to be **feedback linearizable.** Note that since $s(\mathbf{x}) \neq \mathbf{0}$ in some neighborhood of $\mathbf{0}$, (2) can be rewritten as

$$\mathbf{5} \qquad u = -\frac{q(\mathbf{x})}{s(\mathbf{x})} + \frac{1}{s(\mathbf{x})}v,$$

where $-q(\mathbf{x})/s(\mathbf{x})$ and $1/s(\mathbf{x})$ are also smooth functions. Hence, if we think of v as the external input applied to the system, then (2) [or equivalently (5)] represents nonlinear state feedback, and a nonlinear state-dependent pre-filter, applied to the system (1). Similarly, (3) represents a nonlinear state-variable transformation. Hence the overall effect of (2) and (3) can be depicted as shown in Figure 7.2.

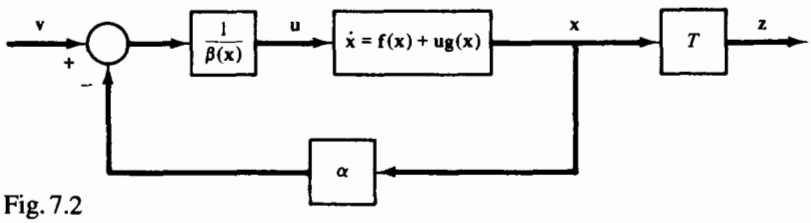


Fig. 7.2

The problem statement can be simplified further. Suppose suitable functions q, s, T can be found such that the resulting state vector $\mathbf{z}$ and input v satisfy (4), and the pair $(\mathbf{A}, \mathbf{b})$ is controllable. Then it is possible to apply a further state transformation

$$\mathbf{6} \qquad \bar{\mathbf{z}} = \mathbf{M}^{-1}\mathbf{z}$$

such that the resulting system is in **controllable canonical form** [see Chen (1984)]. Thus

$$\mathbf{7}\qquad \dot{\bar{\mathbf{z}}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}\bar{\mathbf{z}} + \mathbf{M}^{-1}\mathbf{b}v,$$

where

$$\mathbf{8}\qquad \mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \mathbf{M}^{-1}\mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix},$$

and the a_i's are the coefficients of the characteristic polynomial

$$\mathbf{9}'\qquad |sI - \mathbf{A}| = s^n + \sum_{i=0}^{n-1} a_i s^i.$$

Now a further state feedback of the form

$$\mathbf{10}\qquad v = \bar{v} + [a_0\ a_1\ \cdots\ a_{n-1}]\bar{\mathbf{z}}$$

results in the closed-loop system

$$\mathbf{11}\qquad \dot{\bar{\mathbf{z}}} = \bar{\mathbf{A}}\bar{\mathbf{z}} + \bar{\mathbf{b}}\bar{v},$$

where

$$\mathbf{12}\qquad \bar{\mathbf{A}} = \begin{bmatrix} 0\ 1\ 0 & \cdots & 0 \\ 0\ 0\ 1 & \cdots & 0 \\ \cdot\ \cdot\ \cdot & \cdot & \cdot \\ \cdot\ \cdot\ \cdot & \cdot & \cdot \\ \cdot\ \cdot\ \cdot & \cdot & \cdot \\ 0\ 0\ 0 & \cdots & 1 \\ 0\ 0\ 0 & \cdots & 0 \end{bmatrix}, \bar{\mathbf{b}} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix}.$$

It is easy to see that

$$\mathbf{13}\qquad \dot{\bar{\mathbf{z}}} = \mathbf{M}^{-1}T(\mathbf{x}),\ \bar{v} = v - \mathbf{a}\bar{\mathbf{z}} = q(\mathbf{x}) - \mathbf{a}'\mathbf{M}^{-1}T(\mathbf{x}) + s(\mathbf{x})u,$$

where

14 $\mathbf{a}' = [a_0\ a_1\ \cdots\ a_{n-1}]$.

Note that the transformation (13) is of the same *form* as (2) – (3). Hence if the system (1) is feedback linearizable at all, then it can be transformed to the simpler system (11).

With these observations, the problem under study in this section can be precisely stated as follows:

Feedback Linearization Problem (Single-Input Case) *Given the system (1), do there exist (i) a smooth function* $q \in S(X)$, *(ii) a smooth function* $s \in S(X)$ *such that* $s(\mathbf{x}) \neq \mathbf{0}$ *for all* $\mathbf{x}$ *in some neighborhood of* $\mathbf{0}$, *and (iii) a local diffeomorphism* $T: \mathbf{R}^n \to \mathbf{R}^n$ *such that* $T(\mathbf{0}) = \mathbf{0}$, *satisfying the following conditions: If new variables* v *and* $\mathbf{z}$ *are defined in accordance with (2) and (3) respectively, then*

15 $\dot{z}_1 = z_2, \dot{z}_2 = z_3, \cdots, \dot{z}_{n-1} = z_n, \dot{z}_n = v$?

Note that (15) is just (11) written differently.

Now the main result of this section is presented.

16 Theorem *The feedback linearization problem for the single-input case has a solution if and only if the following two conditions are satisfied in some neighborhood of the origin:*

(i) *The set of vector fields* $\{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g},\ 0 \leq i \leq n-1\}$ *is linearly independent.*

(ii) *The set of vector fields* $\{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g},\ 0 \leq i \leq n-2\}$ *is involutive.*

Proof "Only if" Suppose the problem has a solution; i.e., suppose there exist a neighborhood U of the origin, a diffeomorphism $T: U \to \mathbf{R}^n$, and suitable smooth functions $q, s \in S(X)$. It is to be shown that (i) and (ii) above are satisfied.

By assumption, the transformed variables $\mathbf{z}$ and v satisfy (15). Let T_i denote the i-th component of the vector-valued function $T(\mathbf{x})$. Thus

17 $z_i = T_i(\mathbf{x})$.

Now

18 $\dot{z}_1 = \nabla T_1\, \dot{\mathbf{x}} = \langle \mathbf{d}T_1, \mathbf{f} \rangle + u \langle \mathbf{d}T_1, \mathbf{g} \rangle$ by (1)

$= z_2 = T_2(\mathbf{x})$ by (15).

Hence (18) implies that

19 $T_2 = \langle \mathbf{d}T_1, \mathbf{f} \rangle = L_{\mathbf{f}} T_1$, and $0 = \langle \mathbf{d}T_1, \mathbf{g} \rangle = L_{\mathbf{g}} T_1$.

Repeating the reasoning for $z_2, \cdots, z_{n-1}$ shows that

20 $L_{\mathbf{f}}T_i = T_{i+1}$, $L_{\mathbf{g}}T_i = 0$, for $i = 1, \cdots, n-1$.

Finally, for $i = n$ we have

21 $\dot{z}_n = L_{\mathbf{f}}T_n + u\, L_{\mathbf{g}}T_n$ by (1)

$= v = q + us$ by (15) and (2).

Hence

22 $L_{\mathbf{f}}T_n = q$, $L_{\mathbf{g}}T_n = s$.

All of these equations hold for all $\mathbf{x} \in U$.

Next, it is claimed that

23′ $L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}}T_j = 0$ for $0 \le i \le n-j-1$, $1 \le j \le n-1$.

In other words, it is claimed that

24 $L_{\mathbf{g}}T_{n-1} = 0$,

$L_{\mathbf{g}}T_{n-2} = 0$, $L_{[\mathbf{f},\mathbf{g}]}T_{n-2} = 0$, $\cdots$

$L_{\mathbf{g}}T_1 = 0$, $\cdots$, $L_{\mathrm{ad}_{\mathbf{f}}^{n-2}\mathbf{g}}T_1 = 0$.

The claim (23) is proved by induction on j, starting with the largest value of j, namely $j = n-1$. The first line in (24) follows from (20), so (23) is true when $j = n-1$. Suppose (23) is true for $j+1, \cdots, n-1$, and let us establish (23) for j. For this fixed value of j, the proof of (23) is again by induction, this time on i, starting with $i = 0$. Clearly $L_{\mathbf{g}}T_j = 0$ by (20), so the statement is true when $i = 0$. Suppose it is true for $0, \cdots, i-1$. Then

25 $L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}}T_j = L_{[\mathbf{f}, \mathrm{ad}_{\mathbf{f}}^{i-1}\mathbf{g}]}T_j$.

By Lemma (7.1.59),

26 $L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}}T_j = L_{\mathbf{f}}L_{\mathrm{ad}_{\mathbf{f}}^{i-1}\mathbf{g}}T_j - L_{\mathrm{ad}_{\mathbf{f}}^{i-1}\mathbf{g}}L_{\mathbf{f}}T_j$.

Now the first term is zero by the inductive hypothesis on $i-1$. So

27 $L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}}T_j = -L_{\mathrm{ad}_{\mathbf{f}}^{i-1}\mathbf{g}}L_{\mathbf{f}}T_j = -L_{\mathrm{ad}_{\mathbf{f}}^{i-1}\mathbf{g}}T_{j+1}$ by (20) $= 0$

by the inductive hypothesis on $j+1$. This establishes (23).

To complete the picture, let us compute $L_{\mathrm{ad}_{\mathbf{f}}^{n-j}\mathbf{g}}T_j$. We have

$$\mathbf{28}\quad L_{\mathrm{ad}_{\mathbf{f}}^{n-j}\mathbf{g}}T_j = L_{[\mathbf{f},\,\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}]}T_j$$

$$= L_{\mathbf{f}}L_{\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}}T_j - L_{\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}}L_{\mathbf{f}}T_j$$

$$= -L_{\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}}T_{j+1},\ 1 \le j \le n-1,$$

where (23) is used to eliminate the first term in the second line, and (20) is used to replace $L_{\mathbf{f}}T_j$ by T_{j+1} in the last line. Now it is claimed that

$$\mathbf{29}\quad L_{\mathrm{ad}_{\mathbf{f}}^{n-j}\mathbf{g}}T_j = (-1)^{n-j}s,\ 1 \le j \le n-1.$$

The proof is by (what else?) induction on j, starting with the highest value. When $j = n-1$, (29) follows from (28) and (22). Now suppose (29) is true for $j+1, \cdots, n-1$. Then

$$\mathbf{30}\quad L_{\mathrm{ad}_{\mathbf{f}}^{n-j}\mathbf{g}}T_j = L_{[\mathbf{f},\,\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}]}T_j$$

$$= L_{\mathbf{f}}L_{\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}}T_j - L_{\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}}L_{\mathbf{f}}T_j$$

$$= -L_{\mathrm{ad}_{\mathbf{f}}^{n-j-1}\mathbf{g}}T_{j+1},$$

where (23) is used to eliminate the first term in the second line, and (20) is used to replace $L_{\mathbf{f}}T_j$ by T_{j+1} in the last line. Now by the inductive hypothesis, the last line equals

$$\mathbf{31}\quad -(-1)^{n-j-1}s = (-1)^{n-j}s.$$

This establishes (29).

Now we are ready to complete the proof. Applying (23) and (29) with $j = 1$ gives

$$\mathbf{32}\quad \langle \mathbf{d}T_1,\ \mathrm{ad}_{\mathbf{f}}^{i}\mathbf{g} \rangle = 0,\ 0 \le i \le n-2,$$

$$\mathbf{33}\quad \langle \mathbf{d}T_1,\ \mathrm{ad}_{\mathbf{f}}^{n-1}\mathbf{g} \rangle = (-1)^{n-1}s.$$

First it is established that Condition (i) is true, i.e., that the set $\{\mathrm{ad}_{\mathbf{f}}^{i}\mathbf{g},\ 0 \le i \le n-1\}$ is linearly independent. For this purpose, suppose $c_0, \cdots, c_{n-1}$ are real numbers, and $\mathbf{x}_0 \in U$ is a point such that

$$\mathbf{34}\quad \sum_{i=0}^{n-1} c_i\,[\mathrm{ad}_{\mathbf{f}}^{i}\mathbf{g}](\mathbf{x}_0) = \mathbf{0}.$$

Take the inner product of this vector with $\mathbf{d}T_1(\mathbf{x}_0)$, and use (32) and (33). This gives

$$\mathbf{35}\quad (-1)^{n-1}c_{n-1}\,s(\mathbf{x}_0) = 0.$$

But since $s(\mathbf{x}_0) \neq 0$, we conclude that $c_{n-1} = 0$. So we can drop the last term from the summation in (34) and take the inner product of the resulting vector with $\mathbf{d}T_{n-2}(\mathbf{x}_0)$. Again using

(23) and (29) gives

36 $(-1)^{n-2}c_{n-2}\,s(\mathbf{x}_0)=0,$

which in turn implies that $c_{n-2}=0$. Repeating this procedure shows that all c_i's equal zero. This shows that Condition (i) of the theorem is true. To prove Condition (ii), observe that the set $\{\mathrm{ad}_{\mathbf{f}}^i\mathbf{g},\ 0\le i\le n-2\}$ contains $n-1$ linearly independent vector fields, since it is a subset of the linearly independent set $\{\mathrm{ad}_{\mathbf{f}}^i\mathbf{g},\ 0\le i\le n-1\}$. Let Δ be the distribution spanned by these $n-1$ vector fields. Then (32) shows that the exact differential form $\mathbf{d}T_1$ annihilates Δ. Moreover, since T_1 is the first component of a local diffeomorphism, $\mathbf{d}T_1(\mathbf{x})\neq\mathbf{0}$ for all $\mathbf{x}\in U$. By the Frobenius theorem [Theorem (7.2.24)], it follows that Δ is involutive, which is precisely Condition (ii).

"If" This part of the proof consists essentially of reversing the above steps. Suppose Conditions (i) and (ii) of the theorem hold over some neighborhood U of $\mathbf{0}$. Then from the involutivity of the set $\{\mathrm{ad}_{\mathbf{f}}^i\mathbf{g},\ 0\le i\le n-2\}$ and the Frobenius theorem, it follows that there exists a nonconstant smooth function T_1 such that

37 $\langle \mathbf{d}T_1, \mathrm{ad}_{\mathbf{f}}^i\mathbf{g}\rangle = 0,\ 0\le i\le n-2,\ \forall \mathbf{x}\in U.$

Without loss of generality, it can be assumed that $T_1(\mathbf{0})=0$; if $T_1(\mathbf{0})\neq 0$, then the constant $T_1(\mathbf{0})$ can be subtracted from T_1 without affecting $\mathbf{d}T_1$ and the validity of (37). Now define $T_2, \cdots, T_n$ recursively by

38 $T_{i+1}=\langle \mathbf{d}T_i, \mathbf{f}\rangle = L_{\mathbf{f}}T_i,\ i=1, \cdots, n-1.$

Then $T_i(\mathbf{0})=0$ for all i, since $\mathbf{f}(\mathbf{0})=\mathbf{0}$. Finally, define

39 $q=\langle \mathbf{d}T_n, \mathbf{f}\rangle,\ s=\langle \mathbf{d}T_n, \mathbf{g}\rangle,$

40 $v=q+su,$

41 $\mathbf{z}=[T_1(\mathbf{x})\ \cdots\ T_n(\mathbf{x})]'.$

Let us see what equations these new variables satisfy. For this purpose, let us first show that

42 $\langle \mathbf{d}T_j, \mathrm{ad}_{\mathbf{f}}^i\mathbf{g}\rangle = 0,\ \text{for}\ 0\le i\le n-j-1,\ 1\le j\le n-1.$

This, it can be observed, is the same as (23); however, the above statement needs to be proved starting from (37) and (38), whereas (23) was proved starting from (20). Nevertheless, the proof of (42) by double induction on i and j is only a minor variation of the proof of (23), and is left as an exercise to the reader. Similarly, it can be shown that

43 $\langle \mathbf{d}T_j, \mathrm{ad}_{\mathbf{f}}^{n-j}\mathbf{g}\rangle = (-1)^{n-j}s,$

which is the same as (29).

Using (42) and (43), it is a simple matter to find a set of equations for the new variables $\mathbf{z}$ and v. First, for $1 \le i \le n-1$, we have

44 $\dot{z}_i = \nabla T_i\,\dot{\mathbf{x}} = \langle \mathbf{d}T_i, \mathbf{f}\rangle + u\,\langle \mathbf{d}T_i, \mathbf{g}\rangle = T_{i+1} = z_{i+1}.$

Finally,

45 $\dot{z}_n = \nabla T_n\,\dot{\mathbf{x}} = \langle \mathbf{d}T_n, \mathbf{f}\rangle + u\,\langle \mathbf{d}T_n, \mathbf{g}\rangle = q + us = v.$

These are precisely the equations (15).

Hence, to complete the proof, it only remains to show that $s(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in U$, and that T is a diffeomorphism on U. To prove the first statement, suppose there exists a point $\mathbf{x}_0 \in U$ such that $s(\mathbf{x}_0) = 0$. Then, combining (37) with (43) evaluated with $j = 1$ gives

46 $\langle \mathbf{d}T_1, \mathrm{ad}_{\mathbf{f}}^{i}\mathbf{g}\rangle(\mathbf{x}_0) = 0, \text{ for } 0 \le i \le n-1.$

This contradicts the hypothesis that the set $\{\mathrm{ad}_{\mathbf{f}}^{i}\mathbf{g},\ i = 0, \cdots, n-1\}$ is linearly independent at all $\mathbf{x} \in U$. Hence $s(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in U$.

Finally, the proof is completed by showing that T is a local diffeomorphism on some neighborhood of $\mathbf{0}$. Suppose $a_1, \cdots, a_n$ are real numbers such that the *row* vector

47 $\sum_{i=1}^{n} a_i\,\mathbf{d}T_i(\mathbf{0}) = \mathbf{0}.$

First take the inner product of this row vector with the column vector $\mathbf{g}(\mathbf{0})$. Using (42) with $i = 1$ gives

48 $0 = a_n\,\langle \mathbf{d}T_n, \mathbf{g}\rangle(\mathbf{0}) = a_n\,s(\mathbf{0}).$

But since $s(\mathbf{0}) \neq 0$, this implies that $a_n = 0$. Let us therefore drop the term $a_n\mathbf{d}T_n$ from the summation in (47), and take the inner product of the resulting sum with the column vector $\mathrm{ad}_{\mathbf{f}}\mathbf{g}(\mathbf{0})$. Using (42) with $i = 1$ and (43) with $j = n-1$ gives

49 $0 = a_{n-1}\,\langle \mathbf{d}T_{n-1}, \mathrm{ad}_{\mathbf{f}}\mathbf{g}\rangle(\mathbf{0}) = -a_{n-1}\,s(\mathbf{0}),$

which in turn implies that $a_{n-1} = 0$. The process can be continued to show that $a_i = 0$ for all i. Hence the differentials $\mathbf{d}T_i(\mathbf{0})$, $i = 1, \cdots, n$ are linearly independent. By the inverse function, this shows that T is a local diffeomorphism on some neighborhood of $\mathbf{0}$. ■

Application: Robot with Flexible Joint

Consider the problem of positioning a link using a motor, where the coupling between the motor and the link has some flexibility. The development below is taken from Marino and Spong (1986). The system under study can be modeled as shown in Figure 7.3, where the motor shaft is coupled to the link by a linear spring. This system can be modeled by the two equations

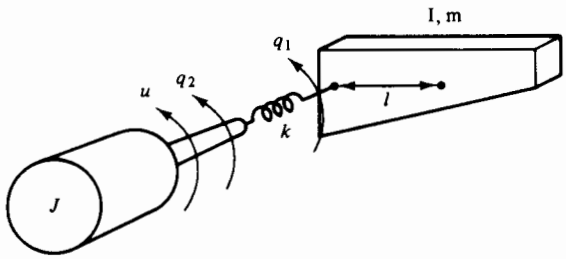


Fig. 7.3

$$I\ddot{q}_1 + mgl \sin q_1 + k(q_1 - q_2) = 0,$$

$$J\ddot{q}_2 - k(q_1 - q_2) = u,$$

where the various physical constants are defined as follows:

J Moment of inertia of the motor about its axis of rotation.

I Moment of inertia of the link about the axis of rotation of the motor shaft.

l Distance from the motor shaft to the center of mass of the link.

m Mass of the link.

g Acceleration due to gravity.

k Torsional spring constant.

q_1 Angle of the link.

q_2 Angle of the motor shaft.

u Torque applied to the motor shaft.

If we choose the natural set of state variables

$$\mathbf{x} = \begin{bmatrix} q_1 \\ \dot{q}_1 \\ q_2 \\ \dot{q}_2 \end{bmatrix},$$

then the system equations assume the form (1) with

$$\mathbf{f} = \begin{bmatrix} x_2 \\ -\dfrac{mgl}{I}\sin x_1 - \dfrac{k}{I}(x_1 - x_3) \\ x_4 \\ \dfrac{k}{J}(x_1 - x_3) \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1/J \end{bmatrix}.$$

From Theorem (16), this system can be transformed to the form (15) if and only if the following two conditions hold over some neighborhood of $\mathbf{0}$: (i) The set $\{\mathbf{g}, \mathrm{ad}_f\mathbf{g}, \mathrm{ad}_f^2\mathbf{g}, \mathrm{ad}_f^3\mathbf{g}\}$ is linearly independent, and (ii) the set $\{\mathbf{g}, \mathrm{ad}_f\mathbf{g}, \mathrm{ad}_f^2\mathbf{g}\}$ is involutive. In the present case all of these vector fields are constant, and

$$[\mathbf{g}\ \mathrm{ad}_f\mathbf{g}\ \mathrm{ad}_f^2\mathbf{g}\ \mathrm{ad}_f^3\mathbf{g}] = \begin{bmatrix} 0 & 0 & 0 & k/IJ \\ 0 & 0 & k/IJ & 0 \\ 0 & 1/J & 0 & -k/J^2 \\ 1/J & 0 & -k/J^2 & 0 \end{bmatrix}.$$

The determinant of this matrix is k^2/I^2J^4 which is obviously nonzero. Hence Condition (i) holds. As for Condition (ii), *any* set of constant vector fields is involutive, since the Lie bracket of two constant vector fields is zero. Hence, by Theorem (16), this system is feedback linearizable.

To construct a linearizing transformation, one can proceed, as in the proof of Theorem (16), to find a nonconstant function T_1 such that $T_1(\mathbf{0}) = 0$ and

$$\langle \mathbf{d}T_1, \mathbf{g} \rangle = 0, \ \langle \mathbf{d}T_1, \mathrm{ad}_f\mathbf{g} \rangle = 0, \ \langle \mathbf{d}T_1, \mathrm{ad}_f^2\mathbf{g} \rangle = 0.$$

In the present instance, since each of the three vector fields is constant, the conditions reduce simply to

$$\frac{\partial T_1}{\partial x_2} = 0, \ \frac{\partial T_1}{\partial x_3} = 0, \ \frac{\partial T_1}{\partial x_4} = 0.$$

So a logical (and simple) choice is

$$T_1(\mathbf{x}) = x_1 =: z_1.$$

Of course this choice is not unique. Now, from (38),

$$z_2 = \langle \mathbf{d}T_1, \mathbf{f} \rangle = x_2,$$

$$z_3 = \langle \mathbf{d}T_2, \mathbf{f} \rangle = -\frac{mgl}{I} \sin x_1 - \frac{k}{I}(x_1 - x_3),$$

$$z_4 = \langle \mathbf{d}T_3, \mathbf{f} \rangle = -\frac{mgl}{I} x_2 \cos x_1 - \frac{k}{I}(x_2 - x_4).$$

This gives the nonlinear state transformation. To obtain the feedback control law (2), we use (39). This gives

$$q = \langle \mathbf{d}T_4, \mathbf{f} \rangle = \frac{mgl}{I} \sin x_1 (x_2^2 + \frac{mgl}{I} \cos x_1 + \frac{k}{I}) + \frac{k}{I}(x_1 - x_3)(\frac{k}{I} + \frac{k}{J} + \frac{mgl}{J} \cos x_1),$$

$$s = \langle \mathbf{d}T_4, \mathbf{g} \rangle = \frac{k}{IJ}.$$

In the new variables, the system equations are

$$\dot{z}_1 = z_2, \dot{z}_2 = z_3, \dot{z}_3 = z_4, \dot{z}_4 = v.$$

Two comments are worth making at this point. First, since $z_1 = x_1$ and z_2, z_3, z_4 are its higher derivatives, the new state variables are physically meaningful: They are respectively the link angle, angular velocity, angular acceleration, and jerk. Second, though Theorem (16) only considers *local* feedback linearization, in the present example the linearization is actually *global*. This can be seen by noting that the transformation T mapping $\mathbf{x}$ into $\mathbf{z}$ is actually globally invertible, with the inverse mapping

$$x_1 = z_1,$$

$$x_2 = z_2,$$

$$x_3 = z_1 + \frac{I}{k}(z_3 + \frac{mgl}{I} \sin z_1),$$

$$x_4 = z_2 + \frac{I}{k}(z_4 + \frac{mgl}{I} z_2 \cos z_1).$$

7.5 FEEDBACK LINEARIZATION: MULTI-INPUT CASE

In this section, the results of the preceding section, and in particular Theorem (7.4.16), are extended to multi-input systems of the form

$$\mathbf{1} \qquad \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \sum_{i=1}^{m} u_i \, \mathbf{g}_i(\mathbf{x}),$$

where $\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m$ are smooth vector fields on some neighborhood X of $\mathbf{0}$, and $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. It is reasonable to assume that the vector fields $\{\mathbf{g}_1, \cdots, \mathbf{g}_m\}$ are linearly independent in some neighborhood of the origin. Otherwise, some of the inputs are redundant, and by redefining the inputs and reducing their number, the linear independence assumption can be satisfied. The objective is to determine whether it is possible to transform the system (1) to a linear system via nonlinear feedback and a state transformation.

The major difference between the single-input case and the multi-input case is that, while there is a single canonical form to which all controllable single-input linear systems can be transformed, there is more than one canonical form for multi-input systems. This necessitates the introduction of some extra concepts, namely controllability indices and the Brunovsky canonical form.

Consider the linear time-invariant system

$$\mathbf{2} \qquad \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u},$$

where $\mathbf{A} \in \mathbf{R}^{n \times n}$, $\mathbf{B} \in \mathbf{R}^{n \times m}$, and $\mathbf{B}$ has rank m. If the system is controllable, then

$$\mathbf{3} \qquad \text{rank } [\mathbf{B} \ \mathbf{AB} \ \mathbf{A}^2\mathbf{B} \cdots \mathbf{A}^{n-1}\mathbf{B}] = n.$$

Define $r_0 = \text{rank } \mathbf{B} = m$, and for $i \geq 1$ define

$$\mathbf{4} \qquad r_i = \text{rank } [\mathbf{B} \ \mathbf{AB} \cdots \mathbf{A}^i\mathbf{B}] - \text{rank } [\mathbf{B} \ \mathbf{AB} \cdots \mathbf{A}^{i-1}\mathbf{B}].$$

Then clearly $0 \leq r_i \leq m$ for all i. It can also be shown that $r_i \geq r_{i+1}$. The proof is quite easy, and in any case the statement follows from the more general version for nonlinear systems given below; see Lemma (16). Thus

$$\mathbf{5} \qquad m = r_0 \geq r_1 \geq \cdots \geq r_{n-1}, \ \sum_{i=0}^{n-1} r_i = n.$$

Now define the **Kronecker indices** $\kappa_1, \cdots, \kappa_m$ of the system (2) as follows: κ_i is the number of the integers r_i that are greater than or equal to i. Clearly

$$\mathbf{6} \qquad \kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_m \geq 0, \text{ and } \sum_{i=1}^{m} \kappa_i = n.$$

For convenience, introduce the integers

$$\mathbf{7} \qquad \sigma_i = \sum_{j=1}^{i} \kappa_j, \; i = 1, \cdots, m.$$

Now the **Brunovsky canonical form** of the system (2) is another linear time-invariant system of the form

$$\mathbf{8} \qquad \dot{\mathbf{z}} = \hat{\mathbf{A}}\mathbf{z} + \hat{\mathbf{B}}\mathbf{v},$$

where $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ have the following special structures:

$$\mathbf{9} \qquad \hat{\mathbf{A}} = \text{Block Diag}\{\hat{\mathbf{A}}_1, \cdots, \hat{\mathbf{A}}_m\},$$

where $\hat{\mathbf{A}}_i$ has dimensions $\kappa_i \times \kappa_i$ and is the companion matrix corresponding to the polynomial s^{κ_i}. In other words,

$$\mathbf{10} \qquad \hat{\mathbf{A}}_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbf{R}^{\kappa_i \times \kappa_i}.$$

The j-th column of the matrix $\hat{\mathbf{B}}$ has all zeros except for a "1" in the σ_j-th row. It can be shown that the system (2) can be put into its Brunovsky canonical form by state feedback and a state-space transformation.

It is easy to see that if the number of inputs m equals 1, then $r_i = 1$ for all i, $\kappa_1 = m$, and the Brunovsky canonical form is just (7.4.12). For multi-input systems, however, several canonical forms are possible. But if two multi-input systems have the same number of states and inputs, then they are "feedback equivalent," in the sense that one system can be transformed into the other via state feedback and a state-space transformation, if and only if they have the same Brunovsky canonical form.

To define a Brunovsky canonical form for nonlinear systems of the form (1), we proceed in a manner entirely analogous to the above, except for a few additional regularity assumptions which are not needed in the linear case. Given the system (1), define

$$\mathbf{11} \qquad C_i = \{\text{ad}_{\mathbf{f}}^k \mathbf{g}_j, \; 1 \le j \le m, \; 0 \le k \le i\}, \; 0 \le i \le n-1,$$

$$\mathbf{12} \qquad \Delta_i = \text{span } C_i.$$

Thus

13 $\Delta_0 = \text{span}\{\mathbf{g}_1, \cdots, \mathbf{g}_m\}$,

14 $\Delta_1 = \text{span}\{\mathbf{g}_1, \cdots, \mathbf{g}_m, [\mathbf{f}, \mathbf{g}_1], \cdots, [\mathbf{f}, \mathbf{g}_m]\}$,

and so on. Now define

15 $r_0 = \dim \Delta_0 = m$, $r_i = \dim \Delta_i - \dim \Delta_{i-1}$, for $i \geq 1$.

It is routine to verify that the above definition reduces to (4) for linear systems of the form (2). But there is a potential complication in the case of nonlinear systems. In the linear case, each distribution Δ_i is generated by a set of *constant* vector fields, and thus has the same dimension at all $\mathbf{x}$. But in the nonlinear case, $\dim \Delta_i(\mathbf{x})$ can vary as $\mathbf{x}$ varies. One way to forestall this difficulty is to assume that $\mathbf{0}$ is a *regular* point of each of the distributions Δ_i, $i = 0, \cdots, n-1$. If such an assumption is made, then $\dim \Delta_i(\mathbf{x})$ is constant for all $\mathbf{x}$ in some neighborhood of $\mathbf{0}$, and as a result the integer r_i is well-defined by (15) for all i. Once the integers r_i are defined, the Kronecker indices $\kappa_1, \cdots, \kappa_m$ and the Brunovsky canonical form are defined exactly as in the linear case.

16 Lemma *We have*

17 $0 \leq r_{i+1} \leq r_i$, $\forall i$.

Proof Clearly $r_{i+1} \geq 0$ since Δ_i is a subset of Δ_{i+1}. To show that $r_{i+1} \leq r_i$, note that, by definition,

18 $\dim \Delta_i = \dim \Delta_{i-1} + r_i$,

or, in other words,

19 $\dim \text{span}\{C_{i-1} \cup \{\text{ad}_{\mathbf{f}}^i \mathbf{g}_1, \cdots, \text{ad}_{\mathbf{f}}^i \mathbf{g}_m\}\} = \dim \text{span}\, C_{i-1} + r_i$.

Thus there are exactly $m - r_i$ vector fields among $\{\text{ad}_{\mathbf{f}}^i \mathbf{g}_1, \cdots, \text{ad}_{\mathbf{f}}^i \mathbf{g}_m\}$ that are linear combinations of those in C_{i-1} and the remaining r_i vector fields among $\{\text{ad}_{\mathbf{f}}^i \mathbf{g}_1, \cdots, \text{ad}_{\mathbf{f}}^i \mathbf{g}_m\}$. For notational convenience only, suppose these are the last $m - r_i$ vector fields, i.e., suppose

20 $\text{ad}_{\mathbf{f}}^i \mathbf{g}_j \in \text{span}\, C_{i-1} \cup \{\text{ad}_{\mathbf{f}}^i \mathbf{g}_1, \cdots, \text{ad}_{\mathbf{f}}^i \mathbf{g}_{r_i}\}$, $j = r_i + 1, \cdots, m$.

Now the linear dependence is maintained when we take the Lie bracket of the two sides with $\mathbf{f}$; thus

21 $\text{ad}_{\mathbf{f}}^{i+1} \mathbf{g}_j \in \text{span}\, C_i \cup \{\text{ad}_{\mathbf{f}}^{i+1} \mathbf{g}_1, \cdots, \text{ad}_{\mathbf{f}}^{i+1} \mathbf{g}_{r_i}\}$, $j = r_i + 1, \cdots, m$.

Hence

22 $\dim \text{span}\, C_{i+1} \leq \dim \text{span}\, C_i + r_i$,

which is the same as saying that $r_{i+1} \leq r_i$. ■

To state the necessary and sufficient conditions for feedback linearizability in the multi-input case, we introduce one last set of integers. Let δ denote the largest value of i such that $r_i \neq 0$. Thus $r_\delta > 0$ but $r_i = 0 \; \forall i > \delta$. Now define

23 $m_\delta = r_\delta, \; m_i = r_i - r_{i+1}$ for $i = 0, \cdots, \delta - 1$.

By Lemma (16), each of the m_i's is nonnegative. Note that, for a multi-input system, both r_i and m_i could be zero for sufficiently large i. However, for a single input system, $\delta = n - 1$, $m_{n-1} = 1$, and $m_i = 0$ for $0 \leq i \leq n-2$. Now it follows readily from (23) that

24 $$r_i = \sum_{j=i}^{\delta} m_j.$$

Since, from (15),

25 $$\dim \Delta_i = \sum_{j=0}^{i} r_j,$$

we get from (24) and (25) that

26 $$\begin{aligned} \dim \Delta_\delta - \dim \Delta_i &= \sum_{j=0}^{\delta} r_j - \sum_{j=0}^{i} r_j = \sum_{j=i+1}^{\delta} r_j \\ &= \sum_{j=i+1}^{\delta} \sum_{k=j}^{\delta} m_k = \sum_{k=i+1}^{\delta} (k-i)\, m_k \\ &= m_{i+1} + 2m_{i+2} + \cdots + (\delta - i)\, m_\delta, \end{aligned}$$

after interchanging the order of summation in the last step.

In order to follow the proof below, one should be able to interpret and understand the integers m_i and κ_i in a variety of ways. The Kronecker indices $\kappa_1, \cdots, \kappa_m$ are just the sizes of the various blocks in (9); there is no loss of generality in assuming that the blocks are arranged in nonincreasing size, which is implied by the first part of (6). The integer δ is equal to the size of the largest block minus 1, i.e., $\delta = \kappa_1 - 1$. Now the integer m_i is just the *number* of blocks of size $i + 1$. Hence

27 $$\sum_{i=0}^{\delta} m_i = m, \; \sum_{i=0}^{\delta} (i+1)\, m_i = n.$$

Now we can state precisely the problem under study.

Feedback Linearization Problem (Multi-Input Case) *Given the system (1) together with a set of integers $\kappa_1, \cdots, \kappa_m$ satisfying (6), do there exist (i) a neighborhood U of $\mathbf{0}$, (ii) a smooth function $\mathbf{q}$: $U \to \mathbf{R}^m$, (iii) a smooth function $\mathbf{S}$: $U \to \mathbf{R}^{m \times m}$ such that* $\det \mathbf{S}(\mathbf{0}) \neq 0$, *and*

(iv) a local diffeomorphism $T:U \to \mathbf{R}^n$ such that $T(\mathbf{0}) = \mathbf{0}$, satisfying the following conditions: If new variables $\mathbf{z}$ and $\mathbf{v}$ are defined according to

28 $\mathbf{z} = T(\mathbf{x})$,

29 $\mathbf{v} = \mathbf{q}(\mathbf{x}) + \mathbf{S}(\mathbf{x})\mathbf{u}$,

where

30 $\mathbf{u} = [u_1 \cdots u_m]'$,

then the new variables $\mathbf{z}$ and $\mathbf{v}$ satisfy the linear differential equation

31 $\dot{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{B}\mathbf{v}$,

with $\mathbf{A}$ and $\mathbf{B}$ in Brunovsky canonical form corresponding to the indices $\kappa_1, \cdots, \kappa_m$?

Now we state the main result of this section.

32 Theorem *Consider the system (1), and assume that the following are true: (a) The vector fields $\mathbf{g}_1, \cdots, \mathbf{g}_m$ are linearly independent at $\mathbf{0}$, so that $\dim \Delta_0 = r_0 = m$ as in (15); and (b) $\mathbf{0}$ is a regular point of the distribution Δ_i for each $i \geq 0$. Under these conditions, the feedback linearization problem in the multi-input case has a solution if and only if the following two conditions are satisfied:*

(i) $\dim \Delta_\delta = n$, *and*

(ii) *The distribution Δ_{i-1} is involutive whenever $m_i \neq 0$.*

Remarks Comparing Theorem (32) with Theorem (7.4.16) for the single-input case, one can spot an extra hypothesis in the present instance, namely the assumption that $\mathbf{0}$ is a regular point of each of the distributions Δ_i. But in the single-input case, this regularity assumption is a consequence of Condition (i). To see this, note that in the single-input case, the distribution Δ_{n-1} is the span of *exactly* $n-1$ vector fields, namely $\mathbf{g}, \mathrm{ad}_{\mathbf{f}}\mathbf{g}, \cdots, \mathrm{ad}_{\mathbf{f}}^{n-1}\mathbf{g}$. Hence, if $\dim \Delta_{n-1} = n$, which is Condition (i) in the single-input case, then this set of vector fields is linearly independent. This implies that Δ_i is the span of the $i+1$ vector fields $\mathbf{g}, \mathrm{ad}_{\mathbf{f}}\mathbf{g}, \cdots, \mathrm{ad}_{\mathbf{f}}^{i}\mathbf{g}$, which in turn implies that $\mathbf{0}$ is automatically a regular point of each distribution Δ_i if Condition (i) holds. Thus, in the single-input case, Theorem (32) reduces *precisely* to Theorem (7.4.16). It should be noted however that Isidori (1989), Theorem 2.2, p. 250, states that even the regularity of $\mathbf{0}$ can be incorporated as part of the necessary and sufficient conditions for feedback linearizability.

Proof "If" Suppose Conditions (i) and (ii) hold. Let δ be, as before, the largest integer i such that $m_i \neq 0$. By (23) this implies that $r_i = 0$ for all $i > \delta$, and hence that $\dim \Delta_\delta = n$. Now

33 $\dim \Delta_{\delta-1} = \dim \Delta_\delta - r_\delta = n - m_\delta$;

that is, the codimension of $\Delta_{\delta-1}$ is m_δ. Now Condition (ii) states that $\Delta_{\delta-1}$ is involutive.

Hence, by the Frobenius theorem, there exist a neighborhood U_0 of $\mathbf{0}$ and smooth functions $\{h_{\delta,i},\ 1 \le i \le m_\delta\}$, such that their differentials are linearly independent at $\mathbf{x} = \mathbf{0}$, and

$$\mathbf{34} \quad <\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{l}\mathbf{g}_j>(\mathbf{x}) = 0,\ \forall \mathbf{x} \in U_0,\ 0 \le l \le \delta - 1,\ 1 \le i \le m_\delta,\ 1 \le j \le m.$$

These functions will form part of the transformation T, as we shall see shortly.

It is claimed that the $m_\delta \times m$ matrix $\mathbf{M}_\delta$ defined by

$$\mathbf{35} \quad (\mathbf{M}_\delta)_{ij}(\mathbf{x}) = <\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta}\mathbf{g}_j>(\mathbf{x})$$

has rank m_δ, i.e., full row rank, at all $\mathbf{x} \in U_0$. To see this, assume that there is a point $\mathbf{x}_0 \in U_0$ and constants $c_1, \cdots, c_{m_\delta}$ such that

$$\mathbf{36} \quad \left[\sum_{i=1}^{m_\delta} c_i <\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta}\mathbf{g}_j>\right](\mathbf{x}_0) = 0, \text{ for } j = 1, \cdots, m.$$

This means, together with (34), that the row vector

$$\mathbf{37} \quad \mathbf{r} = \sum_{i=1}^{m_\delta} c_i\, \mathbf{d}h_{\delta,i}(\mathbf{x}_0)$$

annihilates each of the column vectors $(\mathrm{ad}_{\mathbf{f}}^{l}\mathbf{g}_j)(\mathbf{x}_0)$ for $0 \le l \le \delta$, $1 \le j \le m$. But these vectors are precisely the ones that generate $\Delta_\delta(\mathbf{x}_0)$ and hence span $\mathbf{R}^n$. Thus $\mathbf{r}$ must be the zero (row) vector. Since the row vectors $\mathbf{d}h_{\delta,i}(\mathbf{x}_0)$, $1 \le i \le m_\delta$ are linearly independent, it follows that $c_i = 0\ \forall i$, i.e., that $\mathbf{M}_\delta$ has full row rank.

Next, consider the distribution $\Delta_{\delta-2}$. By (26),

$$\mathbf{38} \quad \dim \Delta_{\delta-2} = \dim \Delta_n - 2m_\delta - m_{\delta-1} = n - 2m_\delta - m_{\delta-1}.$$

Now there are two cases to consider, namely $m_{\delta-1} = 0$ and $m_{\delta-1} \neq 0$. Suppose first that $m_{\delta-1} = 0$. Then it is claimed that the differentials of the $2m_\delta$ functions $\{h_{\delta,i}, L_{\mathbf{f}}h_{\delta,i},\ 1 \le i \le m_\delta\}$ are linearly independent and annihilate each of the vector fields in $\Delta_{\delta-2}$. Note that if $m_{\delta-1} = 0$, then by (38) the codimension of $\Delta_{\delta-2}$ is exactly $2m_\delta$, so what is being claimed is that if $m_{\delta-1} = 0$, then the distribution $\Delta_{\delta-2}$ is *automatically* involutive. Since $\Delta_{\delta-2} \subseteq \Delta_{\delta-1}$, it follows from (34) that $\{\mathbf{d}h_{\delta,i},\ 1 \le i \le m_\delta\}$ annihilate all vector fields in $\Delta_{\delta-2}$. Now, by Lemma (7.1.59), it follows that

$$\mathbf{39} \quad <\mathbf{d}L_{\mathbf{f}}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{l}\mathbf{g}_j> = L_{\mathbf{f}}<\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{l}\mathbf{g}_j> - <\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{l+1}\mathbf{g}_j>$$

$$= 0 \text{ if } 0 \le l \le \delta - 2,$$

since both terms on the right side of (39) are zero by (34). Now it is shown that the $2m_\delta$ differentials $\{\mathbf{d}h_{\delta,i}, \mathbf{d}L_{\mathbf{f}}h_{\delta,i},\ 1 \le i \le m_\delta\}$ are linearly independent. To show this, suppose there

exist constants c_{0i}, c_{1i}, $1 \le i \le m_\delta$ and a point $\mathbf{x}_0 \in U_0$ such that

$$\mathbf{40} \quad \sum_{i=1}^{m_\delta} [c_{0i}\, \mathbf{d}h_{\delta,i} + c_{1i}\, \mathbf{d}L_{\mathbf{f}} h_{\delta,i}](\mathbf{x}_0) = 0.$$

Then

$$\mathbf{41} \quad \sum_{i=0}^{m_\delta} [c_{0i} \langle \mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^l \mathbf{g}_j \rangle + c_{1i} \langle \mathbf{d}L_{\mathbf{f}} h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^l \mathbf{g}_j \rangle (\mathbf{x}_0) = 0 \text{ for } 0 \le l \le \delta - 1,\ 1 \le j \le m.$$

By (34), all terms multiplying the c_{0i}'s are zero. Also, by (39), all coefficients of c_{1i} are zero if $l \le \delta - 2$. So we are left with

$$\mathbf{42} \quad \sum_{i=1}^{m_\delta} c_{1i} \langle \mathbf{d}L_{\mathbf{f}} \mathbf{h}_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1} \mathbf{g}_j \rangle (\mathbf{x}_0) = 0, \text{ for } 1 \le j \le m.$$

Again by Lemma (7.1.59),

$$\mathbf{43} \quad \langle \mathbf{d}L_{\mathbf{f}} h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1} \mathbf{g}_j \rangle = L_{\mathbf{f}} \langle \mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1} \mathbf{g}_j \rangle - \langle \mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta} \mathbf{g}_j \rangle$$

$$= - \langle \mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta} \mathbf{g}_j \rangle,$$

since the first term is zero by (34). Hence (42) reduces to

$$\mathbf{44} \quad -[c_{11} \cdots c_{1m_\delta}]\, \mathbf{M}_\delta = \mathbf{0},$$

where $\mathbf{M}_\delta$ is defined in (35). Since it has already been established that $\mathbf{M}_\delta$ has full row rank, (44) implies that $c_{1i} = 0$ for $1 \le i \le m_\delta$. With this (40) reduces to

$$\mathbf{45} \quad \sum_{i=1}^{m_\delta} c_{0i}\, \mathbf{d}h_{\delta,i}(\mathbf{x}_0) = 0.$$

But since the differentials $\{\mathbf{d}h_{\delta,i},\ 1 \le i \le m_\delta\}$ have been chosen so as to be linearly independent, it follows that $c_{0i} = 0$ for all i as well. This establishes the desired linear independence.

Now suppose $m_{\delta-1} \neq 0$. Then by Condition (ii), the distribution $\Delta_{\delta-2}$ is involutive. It is already known from (38) that the codimension of $\Delta_{\delta-2}$ is $2m_\delta + m_{\delta-1}$. By the Frobenius theorem, there exist a neighborhood U_1 of $\mathbf{0}$ and $2m_\delta + m_{\delta-1}$ linearly independent, exact differential forms which annihilate $\Delta_{\delta-2}(\mathbf{x})$ at all $\mathbf{x} \in U_1$. Without loss of generality, it can be assumed that $U_1 \subseteq U_0$, which is the neighborhood over which the distribution $\Delta_{\delta-1}$ is involutive. Now, of the $2m_\delta + m_{\delta-1}$ exact differential forms which annihilate $\Delta_{\delta-2}$, we have already found $2m_\delta$ of them, namely the set $\{\mathbf{d}h_{\delta,i}, \mathbf{d}L_{\mathbf{f}} h_{\delta,i}\}$. So let us select another $m_{\delta-1}$ smooth functions $\{h_{\delta-1,i},\ 1 \le i \le m_{\delta-1}\}$ such that the set $\{\mathbf{d}h_{\delta,i},\ \mathbf{d}L_{\mathbf{f}} h_{\delta,i},\ 1 \le i \le m_\delta\} \cup \{\mathbf{d}h_{\delta-1,i},\ 1 \le i \le m_{\delta-1}\}$ is linearly independent, and each vector in the set annihilates $\Delta_{\delta-2}(\mathbf{x})$ at all $\mathbf{x} \in U_1$. Define the $m_{\delta-1} \times m$ matrix $\mathbf{M}_{\delta-1}$ by

46 $(\mathbf{M}_{\delta-1})_{ij}(\mathbf{x}) = <\mathbf{d}h_{\delta-1,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1}\mathbf{g}_j>(\mathbf{x}),\ 1 \le i \le m_{\delta-1},\ 1 \le j \le m,$

and consider the $(m_\delta + m_{\delta-1}) \times m$ matrix

47 $$\begin{bmatrix} \mathbf{M}_\delta \\ \mathbf{M}_{\delta-1} \end{bmatrix}.$$

It is claimed that this matrix has full row rank at all $\mathbf{x} \in U_1$. To show this, suppose there exist a point $\mathbf{x}_0 \in U_1$ and a $1 \times (m_\delta + m_{\delta-1})$ row vector that annihilates the above matrix at $\mathbf{x}_0$. In other words, suppose there exist constants c_{0i}, $1 \le i \le m_\delta$, c_{1i}, $1 \le i \le m_{\delta-1}$ such that

48 $$\sum_{i=1}^{m_\delta} c_{0i} <\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta}\mathbf{g}_j>(\mathbf{x}_0) + \sum_{i=1}^{m_{\delta-1}} c_{1i} <\mathbf{d}h_{\delta-1,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1}\mathbf{g}_j>(\mathbf{x}_0) = 0, \text{ for } 1 \le j \le m.$$

Again by Lemma (7.1.59), it follows that

49 $$<\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta}\mathbf{g}_j> = L_{\mathbf{f}} <\mathbf{d}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1}\mathbf{g}_j> - <\mathbf{d}L_{\mathbf{f}}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1}\mathbf{g}_j>$$

$$= - <\mathbf{d}L_{\mathbf{f}}h_{\delta,i}, \mathrm{ad}_{\mathbf{f}}^{\delta-1}\mathbf{g}_j>,$$

since the first term is zero by (34). Hence (48) becomes

50 $$<-\{\sum_{i=1}^{m_\delta} c_{0i}\, \mathbf{d}L_{\mathbf{f}}h_{\delta,i} + \sum_{i=0}^{m_{\delta-1}} c_{1i}\, \mathbf{d}h_{\delta-1,i}\}, \mathrm{ad}_{\mathbf{f}}^{\delta-1}\mathbf{g}_j>(\mathbf{x}_0) = 0, \text{ for } j = 1, \cdots, m.$$

Let us define the vector

51 $$\mathbf{v} = \left[-\sum_{i=1}^{m_\delta} c_{0i}\, \mathbf{d}L_{\mathbf{f}}h_{\delta,i} + \sum_{i=0}^{m_{\delta-1}} c_{1i}\, \mathbf{d}h_{\delta-1,i} \right](\mathbf{x}_0).$$

We know that this vector $\mathbf{v}$ already annihilates $\Delta_{\delta-2}$ at $\mathbf{x}_0$. Now (50) implies that $\mathbf{v}$ also annihilates $\mathrm{ad}_{\mathbf{f}}^{\delta-1}\mathbf{g}_j(\mathbf{x}_0)$ for $1 \le j \le m$. Hence $\mathbf{v}$ annihilates $\Delta_{\delta-1}(\mathbf{x}_0)$, whence it must belong to the span of the exact differentials that annihilate $\Delta_{\delta-1}$, namely $\{\mathbf{d}h_{\delta,i}(\mathbf{x}_0),\ 1 \le i \le m_\delta\}$. But since the set of differentials $\{\mathbf{d}h_{\delta,i},\ 1 \le i \le m_\delta\} \cup \{\mathbf{d}L_{\mathbf{f}}h_{\delta,i},\ 1 \le i \le m_\delta\} \cup \{\mathbf{d}h_{\delta-1,i},\ 1 \le i \le m_{\delta-1}\}$ is linearly independent, this implies that $\mathbf{v}$ must the zero vector and that all constants c_{0i}, c_{1i} must be zero. This shows that the matrix in (47) has full row rank.

This process can be repeated at each stage. When it is finished, we will have functions

52 $$\{h_{\delta,i}, L_{\mathbf{f}}h_{\delta,i}, \cdots, L_{\mathbf{f}}^{\delta-1}h_{\delta,i},\ 1 \le i \le m_\delta\},$$

$$\cdots, \{h_{2,i}, L_{\mathbf{f}}h_{2,i},\ 1 \le i \le m_2\}, \{h_{1,i},\ 1 \le i \le m_1\}.$$

Of course, if $m_i = 0$ for some i, then the corresponding set of functions will be absent. Now the total number of functions is

53 $\delta m_\delta + (\delta - 1)m_{\delta-1} + \cdots + 2m_2 + m_1 = n - m = n - \dim \Delta_0,$

from (26). The differentials of these functions are linearly independent in some neighborhood of $\mathbf{0}$, and they all annihilate every vector field in Δ_0. To complete the process, it is necessary to consider two cases, namely $m_0 = 0$, and $m_0 \neq 0$. In either case we have from (24) that

54 $$\sum_{i=0}^{\delta} m_i = r_0 = m.$$

If $m_0 = 0$, then the set

55 $$\{L_{\mathbf{f}}^l h_{k,i},\ 1 \le i \le m_k,\ 0 \le l \le k,\ 1 \le k \le \delta\}$$

contains exactly n functions. [Observe that there is one more Lie derivative of each $h_{k,i}$ in the set (55) than there is in the set (52).] This is because, if $m_0 = 0$, then from (54),

56 $$\sum_{i=1}^{\delta} m_i = m,$$

so from (53) and (56),

57 $(\delta + 1)m_\delta + \delta m_{\delta-1} + \cdots + 3m_2 + 2m_1 = n.$

Also, by arguments that are now (I hope!) familiar, it follows that the differentials of these functions are linearly independent over some neighborhood U of $\mathbf{0}$. Now suppose $m_0 > 0$. (This cannot happen in the single-input case.) Then there are just $n - m_0$ functions in the set (55). In this case we must choose m_0 other functions, call them $\{h_{0,i},\ 1 \le i \le m_0\}$, such that the differentials of the n functions in the set

58 $$\{L_{\mathbf{f}}^l h_{k,i},\ 1 \le i \le m_k,\ 0 \le l \le k,\ 0 \le k \le \delta\}$$

are linearly independent over some neighborhood of the origin. The selection of these additional m_0 functions is not difficult. Even if $m_0 > 0$, the differentials of the $n - m_0$ functions in (55) are linearly independent. So one can just choose some *constant* row vectors $\alpha_1, \cdots, \alpha_{m_0}$ to complete a row basis, and just let $h_{0,i}(\mathbf{x}) = \alpha_i \mathbf{x}$. In any case, one now has the n functions of (58). The linear independence property means that the map taking the vector $\mathbf{x}$ into the n-vector whose components are the n functions in (58) is a local diffeomorphism. Next, in analogy with (47), form the matrix

$$\mathbf{59}\quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_\delta \\ \mathbf{M}_{\delta-1} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{M}_1 \\ \mathbf{M}_0 \end{bmatrix},$$

where

$$\mathbf{60}\quad (\mathbf{M}_k)_{ij}(\mathbf{x}) = <\mathbf{d}h_{k,i}, \mathrm{ad}_{\mathbf{f}}^k \mathbf{g}_j>(\mathbf{x}),\ 1 \le i \le m_k,\ 1 \le j \le m.$$

The number of rows of this matrix is

$$\mathbf{61}\quad \sum_{i=0}^{\delta} m_i = m.$$

In other words, the matrix $\mathbf{M}$ is square. At each stage, it has been ensured that the matrix continues to have full row rank even as one more block is added at the bottom. So the final matrix $\mathbf{M}$ is nonsingular for all $\mathbf{x}$ in some neighborhood of $\mathbf{0}$.

We are now in the home stretch. Let us first make a couple of observations. 1) Amongst the integers m_i, no more than m can be positive; this is clear from (54). 2) Suppose we define $\kappa_i = i + 1$ for those values of i for which $m_i > 0$; then these are precisely the Kronecker indices. Also, $m_k > 0$ only when $k = \kappa_i$ for some i, so that only the corresponding functions $h_{k,i}$ appear in the list (58). Now define

$$\mathbf{62}\quad \phi_1 = h_{\delta,1},\ \cdots,\ \phi_{m_\delta} = h_{\delta,m_\delta},$$

$$\text{If } m_{\delta-1} > 0,\ \phi_{m_\delta+1} = h_{\delta-1,1},\ \cdots,\ \phi_{r_{\delta-1}} = h_{\delta-1,m_{\delta-1}},\ \cdots,$$

$$\text{If } m_0 > 0,\ \phi_{r_1+1} = h_{0,1},\ \cdots,\ \phi_{r_0} = h_{0,m_0}.$$

Note that, in computing the subscripts of the ϕ's, we have made use of the relationship (24). There are exactly m of these functions ϕ_i. For each index i, note that κ_i is the number of times the function ϕ_i is Lie-differentiated in the list (58). Define a map $\mathbf{z}_i : U \to \mathbf{R}^{\kappa_i}$ by

$$\mathbf{63}\quad \mathbf{z}_i(\mathbf{x}) = \begin{bmatrix} \phi_i(\mathbf{x}) \\ L_{\mathbf{f}}\phi_i(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ L_{\mathbf{f}}^{\kappa_i - 1}\phi_i(\mathbf{x}) \end{bmatrix},$$

and define a map $T: U \to \mathbf{R}^n$ by

$$\mathbf{64}\quad T(\mathbf{x}) = \begin{bmatrix} \mathbf{z}_1(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{z}_m(\mathbf{x}) \end{bmatrix}.$$

Then T is a local diffeomorphism, as discussed earlier. If each ϕ_i is chosen such that $\phi_i(\mathbf{0}) = 0$, which is easy to achieve since a suitable constant can be added to each $h_{k,i}$, then $T(\mathbf{0}) = \mathbf{0}$. We shall see very shortly that this is the desired state transformation.

What are the differential equations governing these new variables $T_i(\mathbf{x})$? By definition,

$$\mathbf{65}\quad \frac{d}{dt} z_{i,1} = \langle \mathbf{d}\phi_i(\mathbf{x}), \dot{\mathbf{x}} \rangle$$

$$= \langle \mathbf{d}\phi_i(\mathbf{x}), [\mathbf{f}(\mathbf{x}) + \sum_{j=1}^{m} u_j \mathbf{g}_j(\mathbf{x})] \rangle$$

$$= L_{\mathbf{f}}\phi_i + \sum_{j=1}^{m} u_j L_{\mathbf{g}_j}\phi_i$$

$$= L_{\mathbf{f}}\phi_i = z_{i,2},$$

since $\langle \mathbf{d}\phi_i, \mathbf{g}_j \rangle = 0$ for all i, j. Similarly

$$\mathbf{66}\quad \frac{d}{dt} z_{i,l} = z_{i,l+1}, \text{ for } 1 \le l \le \kappa_i - 1.$$

All this should look familiar, because this is the same set of manipulations as in the single-input case. Now

$$\mathbf{67} \quad \frac{d}{dt} z_{i,\kappa_i} = L_{\mathbf{f}}^{\kappa_i} \phi_i + \sum_{j=1}^{m} u_j L_{\mathbf{g}_j} L_{\mathbf{f}}^{\kappa_i - 1} \phi_i$$

At this stage, observe that, by using Lemma (7.1.59) repeatedly, we can write

$$\mathbf{68} \quad L_{\mathbf{g}_j} L_{\mathbf{f}}^{\kappa_i - 1} \phi_i = \langle \mathbf{d}\phi_i, \mathrm{ad}_{\mathbf{f}}^{\kappa_i - 1} \mathbf{g}_j \rangle = (-1)^{\kappa_i - 1} M_{ij},$$

where the matrix $\mathbf{M}$ is defined in (59). Thus, collecting the equations (67) as i varies from 1 to m gives

$$\mathbf{69} \quad \frac{d}{dt} \begin{bmatrix} z_{1,\kappa_1} \\ \cdot \\ \cdot \\ \cdot \\ z_{m,\kappa_m} \end{bmatrix} = \mathbf{q}(\mathbf{x}) + \mathbf{S}\mathbf{u},$$

where $q_i(\mathbf{x}) = L_{\mathbf{f}}^{\kappa_i} \phi_i$, $s_{ij} = (-1)^{\kappa_i - 1} M_{ij}$, and $\mathbf{M}$ is the matrix of (59). Let us define

$$\mathbf{70} \quad \mathbf{v} = \mathbf{q}(\mathbf{x}) + \mathbf{S}(\mathbf{x})\mathbf{u}.$$

Then (69) shows that

$$\mathbf{71} \quad \frac{d}{dt} z_{i,\kappa_i} = v_i.$$

Finally, (66) and (71) show that, in terms of the new variables $\mathbf{z}$ and $\mathbf{v}$, the system is in Brunovsky canonical form.

"Only if" Suppose the feedback linearization problem has a solution. It must be shown that Conditions (i) and (ii) of the theorem are satisfied.

It is notationally convenient to renumber the m inputs as follows: To begin with, the m vector fields $\mathbf{g}_1, \cdots, \mathbf{g}_m$ are linearly independent. Now consider the set spanning Δ_1, namely

$$\mathbf{72} \quad C_1 = \{\mathbf{g}_1, \cdots, \mathbf{g}_m, [\mathbf{f}, \mathbf{g}_1], \cdots, [\mathbf{f}, \mathbf{g}_m]\}.$$

Suppose $m_0 \neq 0$. This implies [see (23)] that exactly m_0 vector fields among $\{[\mathbf{f}, \mathbf{g}_1, \cdots, [\mathbf{f}, \mathbf{g}_m]\}$ are linearly dependent on the rest and on $\mathbf{g}_1, \cdots, \mathbf{g}_m$. Renumber the inputs so that these linearly dependent vector fields are the *last* m_0 vector fields; therefore,

$$\mathbf{73} \quad [\mathbf{f}, \mathbf{g}_i] \in \mathrm{span}\,\{\mathbf{g}_1, \cdots, \mathbf{g}_m, [\mathbf{f}, \mathbf{g}_1], \cdots, [\mathbf{f}, \mathbf{g}_{m-m_0}]\}, \text{ for } i = m - m_0 + 1, \cdots, m.$$

Note that the above linear dependence is preserved when we take higher order Lie brackets as well. Thus

74 $\mathrm{ad}_{\mathbf{f}}^2\mathbf{g}_i \in \text{span}\{\mathbf{g}_i, [\mathbf{f}, \mathbf{g}_i], 1 \le i \le m\} \cup \{\mathrm{ad}_{\mathbf{f}}^2\mathbf{g}_1, \cdots, \mathrm{ad}_{\mathbf{f}}^2\mathbf{g}_{m-m_0}\}$, for $i = m - m_0 + 1, \cdots, m$.

But if $m_1 \neq 0$, this means that there is an *additional* linear dependence above and beyond (74). Exactly m_1 vector fields from the left side of (74) are linearly dependent on the rest, plus all the vector fields on the right side of (74). Renumber the inputs again such that these linearly dependent ones are the last ones, namely $\mathrm{ad}_{\mathbf{f}}^2\mathbf{g}_i$ for $i = m - m_0 - m_1 + 1, \cdots, m$. In other words,

75 $\mathrm{ad}_{\mathbf{f}}^2\mathbf{g}_i \in \text{span}\{\mathbf{g}_j, 1 \le j \le m\} \cup \{\mathrm{ad}_{\mathbf{f}}\mathbf{g}_j, 1 \le j \le m - m_0\} \cup \{\mathrm{ad}_{\mathbf{f}}^2\mathbf{g}_j, 1 \le j \le m - m_0 - m_1\}$,

$$\text{for } i = m - m_0 - m_1 + 1, \cdots, m.$$

Observe the simplification incorporated in (75): In the term $\mathrm{ad}_{\mathbf{f}}\mathbf{g}_j$, the subscript j only ranges from 1 to $m - m_0$; this is made possible by the linear dependence (73).

The process can be repeated for each l from 1 to δ. If $m_l = 0$, no renumbering of inputs is needed. If $m_l \neq 0$, the inputs from 1 to $m - \sum_{j=0}^{l-1} m_j$ are renumbered such that

76 $\mathrm{ad}_{\mathbf{f}}^l\mathbf{g}_i \in \text{span} \bigcup_{k=0}^{l} \{\mathrm{ad}_{\mathbf{f}}^k\mathbf{g}_j, 1 \le j \le m - \sum_{s=0}^{k-1} m_s\}$,

where the empty sum is taken as zero. Note that (73) and (75) are special cases of (76), corresponding to the values $l = 1$ and $l = 2$, respectively.

By assumption, the feedback linearization problem has a solution. Accordingly, choose T, $\mathbf{q}$ and $\mathbf{S}$ as in (28) and (29). By expanding (31), the Brunovsky canonical form can be written as follows:

77 $\dot{z}_{i,l} = z_{i,l+1}$, for $1 \le l \le \kappa_i - 1$, $\dot{z}_{i,\kappa_i} = v_i$, for $1 \le i \le m$.

This is just what (66) and (71) say. Define the m functions $\phi_1, \cdots, \phi_m$ by

78 $\phi_i(\mathbf{x}) = z_{i,1}$, $i = 1, \cdots, m$.

Now the Brunovsky canonical form consists essentially of m decoupled single-input systems, with the number of states of the i-th system equaling κ_i. Hence, proceeding exactly as in the proof of Theorem (7.4.16), one can show [cf. (7.4.32)] that

79 $\langle \mathbf{d}\phi_i, \mathrm{ad}_{\mathbf{f}}^l\mathbf{g}_j \rangle = 0$, for $0 \le l \le \kappa_i - 2$, $1 \le j, i \le m$,

and that [cf. (7.4.33)]

80 $\langle \mathbf{d}\phi_i, \mathrm{ad}_{\mathbf{f}}^{\kappa_i - 1}\mathbf{g}_j \rangle = (-1)^{\kappa_i - 1} s_{ij}$, $1 \le i, j \le m$,

where s_{ij} is the ij-th element of the matrix $\mathbf{S}$ of (29).

By assumption, the $m \times m$ matrix $\mathbf{S}$ of (29) is nonsingular for all $\mathbf{x}$ sufficiently near $\mathbf{0}$. For each $k = 1, \cdots, \delta$, define

$$\mathbf{81} \quad \eta_k = \sum_{s=k}^{\delta} m_s.$$

Now it is shown that, for each $k = \delta, \cdots, 1$, the $\eta_k \times \eta_k$ submatrix of $\mathbf{S}$ consisting of the first η_k rows and the first η_k columns is nonsingular. The significance of this claim is as follows: Fix k, and consider the $\eta_k \times m$ submatrix $\mathbf{S}_k$ of $\mathbf{S}$ consisting of the first η_k rows of $\mathbf{S}$. Since $\mathbf{S}$ itself is nonsingular and hence of full row rank, this submatrix $\mathbf{S}_k$ has rank η_k and thus contains η_k linearly independent columns. The point of the claim is that actually the *first* η_k columns of $\mathbf{S}_k$ are linearly independent.

To prove the claim, suppose first that $k = \delta$, so that $\eta_k = \eta_\delta = m_\delta$. Note that $\delta + 1$ is the size of the largest block in the Brunovsky canonical form, and that m_δ is the number of such blocks. Thus $\kappa_1 = \kappa_2 = \cdots = \kappa_{m_\delta} = \delta + 1$. Now consider the $m_\delta \times m$ matrix $\mathbf{S}_\delta$ consisting of the first m_δ rows of $\mathbf{S}$. Letting i in (80) range over $1, \cdots, m_\delta$ gives

$$\mathbf{82} \quad \langle \mathbf{d}\phi_i, \mathrm{ad}_{\mathbf{f}}^{\delta} \mathbf{g}_j \rangle = (-1)^{\delta} s_{ij}, \; 1 \le i \le m_\delta, \; 1 \le j \le m.$$

But (76) states that, if $j > m_\delta$, the vector field $\mathrm{ad}_{\mathbf{f}}^{\delta} \mathbf{g}_j$ is a linear combination [with coefficients from $S(X)$] of $\mathrm{ad}_{\mathbf{f}}^{\delta} \mathbf{g}_1, \cdots, \mathrm{ad}_{\mathbf{f}}^{\delta} \mathbf{g}_{m_\delta}$, and some other vector fields $\mathrm{ad}_{\mathbf{f}}^{l} \mathbf{g}_s$ where $l < \delta$. Now using (79) shows that, if $j > m_\delta$, the function $\langle \mathbf{d}\phi_i, \mathrm{ad}_{\mathbf{f}}^{\delta} \mathbf{g}_j \rangle$ is a linear combination of $s_{i1}, \cdots, s_{im_\delta}$. Repeating this argument as i varies over $1, \cdots, m_\delta$ shows that the last $n - m_\delta$ columns of $\mathbf{S}_\delta$ are linear combinations of the first m_δ columns. Since $\mathbf{S}_\delta$ has rank m_δ, it must be that the first m_δ columns of $\mathbf{S}_\delta$ are linearly independent.

Next, let $k = \delta - 1$, so that $\eta_{\delta-1} = m_\delta + m_{\delta-1}$. If $m_{\delta-1} = 0$, then $\eta_{\delta-1} = m_\delta$ and no further argument is needed, so suppose $m_{\delta-1} \neq 0$. Then $\kappa_{m_\delta+1} = \cdots = \kappa_{m_\delta + m_{\delta-1}} = \delta$. In this case, in addition to (82), we have the relationships [cf. (7.4.29)]

$$\mathbf{83} \quad \langle \mathbf{d}L_{\mathbf{f}}\phi_i, \mathrm{ad}_{\mathbf{f}}^{\delta-1} \mathbf{g}_j \rangle = (-1)^{\delta-1} s_{ij}, \; 1 \le i \le m_\delta, \; 1 \le j \le m,$$

$$\mathbf{84} \quad \langle \mathbf{d}\phi_i, \mathrm{ad}_{\mathbf{f}}^{\delta-1} \mathbf{g}_j \rangle = (-1)^{\delta-1} s_{ij}, \; m_\delta + 1 \le i \le m_\delta + m_{\delta-1}, \; 1 \le j \le m.$$

Now consider the $(m_\delta + m_{\delta-1}) \times m$ submatrix $\mathbf{S}_{\delta-1}$. In this case (76) shows that if $j > m_\delta + m_{\delta-1}$, then $\mathrm{ad}_{\mathbf{f}}^{\delta-1} \mathbf{g}_j$ is a linear combination of $\mathrm{ad}_{\mathbf{f}}^{\delta-1} \mathbf{g}_k$ for $k = 1, \cdots, m_\delta + m_{\delta-1}$, as well as of $\mathrm{ad}_{\mathbf{f}}^{l} \mathbf{g}_k$ for $l < \delta - 1$. Now by (79) and (7.4.23), each of the forms $\mathbf{d}L_{\mathbf{f}}\phi_i$ for $1 \le i \le m_\delta$ and $\mathbf{d}\phi_i$ for $m_\delta + 1 \le i \le m_\delta + m_{\delta-1}$ annihilate each vector field $\mathrm{ad}_{\mathbf{f}}^{l} \mathbf{g}_k$ whenever $l < \delta - 1$. This, together with (83) and (84), shows that the last $m - m_\delta - m_{\delta-1}$ columns of $\mathbf{S}_{\delta-1}$ are linear combinations of the first $m_\delta + m_{\delta-1}$ columns. Since the rank of $\mathbf{S}_{\delta-1}$ is $m_\delta + m_{\delta-1}$, this implies that these first columns of $\mathbf{S}_{\delta-1}$ must be linearly independent. The same reasoning can be repeated for each k from $\delta - 1$ down to 1, thus establishing the claim.

Now it is shown that Condition (i) of the theorem is satisfied, i.e., that $\dim \Delta_\delta = n$. This is the same as showing that the set $\{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}_j, 0 \le i \le \delta, 1 \le j \le m\}$ contains n linearly independent vector fields. For this purpose, suppose there exist constants $c_{ij}, 0 \le i \le \delta, 1 \le j \le m$, such that

$$\mathbf{85} \quad \mathbf{w}(\mathbf{0}) := \sum_{i=0}^{\delta} \sum_{j=1}^{m} c_{ij} (\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}_j)(\mathbf{0}) = \mathbf{0},$$

and define $\mathbf{w} \in V(X)$ as the summation above. In view of the linear dependencies in (76), it can be assumed without loss of generality that

$$\mathbf{86} \quad c_{ij} = 0 \text{ if } j > m - \sum_{s=0}^{i-1} m_s = \eta_i,$$

where η_i is defined in (81) and η_0 is taken as 0. Thus (85) can be rewritten as

$$\mathbf{87} \quad \mathbf{w}(\mathbf{0}) = \sum_{i=0}^{\delta} \sum_{j=1}^{\eta_i} c_{ij} (\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}_j)(\mathbf{0}) = \mathbf{0}.$$

First apply the forms $\mathbf{d}\phi_1, \cdots, \mathbf{d}\phi_{m_\delta}$ to $\mathbf{w}$. Using (79) and (82), this gives

$$\mathbf{88} \quad 0 = \langle \mathbf{d}\phi_i, \mathbf{w} \rangle (\mathbf{0}) = \sum_{j=1}^{m_\delta} s_{ij}(\mathbf{0})\, c_{\delta j}, \text{ for } 1 \le i \le m_\delta,$$

or

$$\mathbf{89} \quad \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = \begin{bmatrix} s_{11} & \cdots & s_{1,m_\delta} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ s_{m_\delta,1} & \cdots & s_{m_\delta,m_\delta} \end{bmatrix} (\mathbf{0}) \begin{bmatrix} c_{\delta,1} \\ \cdot \\ \cdot \\ \cdot \\ c_{\delta,m_\delta} \end{bmatrix}.$$

But the coefficient matrix on the right side of (89) is already known to be nonsingular. Hence it can be concluded that

$$\mathbf{90} \quad c_{\delta j} = 0 \text{ for } j = 1, \cdots, m_\delta.$$

As a result, the expression for $\mathbf{w}$ can be simplified to

$$\mathbf{91} \quad \mathbf{w} = \sum_{i=0}^{\delta-1} \sum_{j=1}^{\eta_i} c_{ij} \mathrm{ad}_{\mathbf{f}}^i \mathbf{g}_j.$$

Now apply the m_δ forms $\mathbf{d}L_{\mathbf{f}}\phi_j, j = 1, \cdots, m_\delta$, and if $m_{\delta-1} \neq 0$ the $m_{\delta-1}$ forms $\mathbf{d}\phi_j, j = m_\delta + 1, \cdots, m_\delta + m_{\delta-1}$ to $\mathbf{q}$. This gives the equation

$$\mathbf{92}\qquad \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = \begin{bmatrix} s_{11} & \cdots & s_{1,m_\delta + m_{\delta-1}} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ s_{m_\delta + m_{\delta-1},1} & \cdots & s_{m_\delta + m_{\delta-1}, m_\delta + m_{\delta-1}} \end{bmatrix} (\mathbf{0}) \begin{bmatrix} c_{\delta-1,1} \\ \cdot \\ \cdot \\ \cdot \\ c_{\delta-1, m_\delta + m_{\delta-1}} \end{bmatrix}.$$

Again, since the coefficient matrix in (92) is known to be nonsingular, it follows that $c_{\delta-1,j} = 0$ for all j. The process can be repeated and shows that all c_{ij} are zero. This shows that the set of vector fields $\{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}_j,\ 0 \le i \le \delta,\ 1 \le j \le m\}$ contains n linearly independent vector fields, and establishes Condition (i) of the theorem.

Finally, it is shown that Condition (ii) of the theorem holds. Since $\dim \Delta_\delta = n$, (26) shows that

$$\mathbf{93}\qquad \dim \Delta_{\delta-1} = n - m_\delta.$$

But (79) shows that the m_δ exact differentials $\{\mathbf{d}\phi_i,\ 1 \le i \le m_\delta\}$ annihilate $\Delta_{\delta-1}$. Moreover, these differentials are linearly independent, since they are rows of the Jacobian of the local diffeomorphism T. By the Frobenius theorem, this implies that $\Delta_{\delta-1}$ is involutive.

Now let us go to $\Delta_{\delta-2}$. There are two cases to consider, namely: $m_{\delta-1} = 0$, and $m_{\delta-1} \neq 0$. Suppose first that $m_{\delta-1} = 0$. In this case [cf. (7.4.23)], in addition to (79), we also have

$$\mathbf{94}\qquad \langle \mathbf{d}L_{\mathbf{f}}\phi_i,\ \mathrm{ad}_{\mathbf{f}}^l \mathbf{g}_j \rangle = 0,\ \text{for } 0 \le l \le \delta - 2,\ 1 \le j \le m,\ 1 \le i \le m_\delta.$$

Since $m_{\delta-1} = 0$, it follows from (26) that

$$\mathbf{95}\qquad \dim \Delta_{\delta-2} = n - 2m_\delta.$$

But (79) and (94) show that the $2m_\delta$ exact and linearly independent differentials $\{\mathbf{d}\phi_i, \mathbf{d}L_{\mathbf{f}}\phi_i,\ 1 \le i \le m_\delta\}$ annihilate $\Delta_{\delta-2}$. Hence $\Delta_{\delta-2}$ is involutive, by the Frobenius theorem. Now suppose $m_{\delta-1} \neq 0$. Then (79) and (94) are still true. But now, from (26),

$$\mathbf{96}\qquad \dim \Delta_{\delta-2} = n - 2m_\delta - m_{\delta-1}.$$

Since $m_{\delta-1} \neq 0$, the Brunovsky canonical form contains $m_{\delta-1}$ blocks of size $\delta \times \delta$; i.e., $\kappa_{m_\delta+1} = \cdots = \kappa_{m_\delta + m_{\delta-1}} = \delta$. So from (79) it follows that

$$\mathbf{97}\qquad \langle \mathbf{d}\phi_i,\ \mathrm{ad}_{\mathbf{f}}^l \mathbf{g}_j \rangle = 0,\ \text{for } 0 \le l \le \delta - 2,\ 1 \le j \le m,\ m_\delta + 1 \le i \le m_\delta + m_{\delta-1}.$$

Thus (79), (94), and (97) demonstrate $2m_\delta + m_{\delta-1}$ exact, linearly independent differentials which annihilate $\Delta_{\delta-2}$. Hence $\Delta_{\delta-2}$ is involutive by the Frobenius theorem. By repeating the process, one can conclude that Δ_i is involutive for all i from 1 to $\delta - 1$. This shows that Condition (ii) of the theorem is also necessary. ■

Application: Robot with Flexible Joints

As an illustration of the feedback linearization of multi-input systems, we study an m-link robot where each joint is modelled as a torsional spring. The system under study is a generalization of the single-link system of Section 7.4.

The development below follows Spong (1987). In this paper, it is shown that, subject to a few simplifying assumptions, the dynamics of the robot can be modelled by the following Euler-Lagrange equations:

$$\mathbf{D}(\mathbf{q}_1)\ddot{\mathbf{q}}_1 + \mathbf{c}(\mathbf{q}_1, \dot{\mathbf{q}}_1) + \mathbf{K}(\mathbf{q}_1 - \mathbf{q}_2) = \mathbf{0},$$

$$\mathbf{J}(\mathbf{q}_2)\ddot{\mathbf{q}}_2 - \mathbf{K}(\mathbf{q}_1 - \mathbf{q}_2) = \mathbf{u},$$

where the various symbols are defined as follows:

$\mathbf{q}_1$ m-vector of link angles.

$\mathbf{q}_2$ m-vector of motor angles.

$\mathbf{D}$ $m \times m$ mechanical inertia matrix.

$\mathbf{J}$ $m \times m$ electrical inertia matrix.

$\mathbf{c}$ m-vector of Coriolis, centripetal, and gravity terms.

$\mathbf{K}$ $m \times m$ diagonal matrix of spring constants.

$\mathbf{u}$ m-vector of externally applied torques.

If each component of $\mathbf{K}$ approaches infinity, the joint springs become infinitely stiff, i.e., rigid. In this case, there is no "play" in the spring, and as a consequence, $\mathbf{q}_1 = \mathbf{q}_2$. Hence the system equations simplify to

$$[\mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})]\ddot{\mathbf{q}} + \mathbf{c}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{u},$$

where $\mathbf{q} = \mathbf{q}_1 = \mathbf{q}_2$. This is a system of m second-order equations. But if each spring constant is finite, then the system is represented by $2m$ second-order equations.

By selecting the state variables

$$\mathbf{x}_1 = \mathbf{q}_1, \mathbf{x}_2 = \dot{\mathbf{q}}_1, \mathbf{x}_3 = \mathbf{q}_2, \mathbf{x}_4 = \dot{\mathbf{q}}_2,$$

the system equations are of the form (1) with $n = 4m$,

$$\mathbf{f} = \begin{bmatrix} \mathbf{x}_2 \\ -\mathbf{D}^{-1}(\mathbf{c}+\mathbf{s}) \\ \mathbf{x}_4 \\ \mathbf{J}^{-1}\mathbf{s} \end{bmatrix}, [\mathbf{g}_1 \cdots \mathbf{g}_m] = \begin{bmatrix} \mathbf{0}_{3m\times m} \\ \mathbf{J}^{-1} \end{bmatrix},$$

and the following shorthand notation is used:

$$\mathbf{D} = \mathbf{D}(\mathbf{x}_1),\ \mathbf{c} = \mathbf{c}(\mathbf{x}_1, \mathbf{x}_2),\ \mathbf{s} = \mathbf{K}(\mathbf{x}_1 - \mathbf{x}_3),\ \mathbf{J} = \mathbf{J}(\mathbf{x}_3).$$

The feedback linearizability of the above system can be tested using Theorem (32), but this is quite messy. Instead, we draw inspiration from the scalar case studied in Section 7.4, and select the new coordinates

$$\mathbf{z}_1 = \mathbf{x}_1,\ \mathbf{z}_2 = \dot{\mathbf{x}}_1,\ \mathbf{z}_3 = \ddot{\mathbf{x}}_1,\ \mathbf{z}_4 = \frac{d^3\mathbf{x}_1}{dt^3}.$$

The first step is to verify that the above transformation is indeed a diffeomorphism. For this purpose, observe that

$$\mathbf{z}_1 = \mathbf{x}_1,\ \mathbf{z}_2 = \mathbf{x}_2,$$

$$\mathbf{z}_3 = -\mathbf{D}^{-1}(\mathbf{x}_1)\,[\mathbf{c}(\mathbf{x}_1, \mathbf{x}_2) + \mathbf{K}(\mathbf{x}_1 - \mathbf{x}_3)],$$

$$\mathbf{z}_4 = \dot{\mathbf{x}}_3 = \mathbf{h}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) + \mathbf{D}^{-1}(\mathbf{x}_1)\,\mathbf{K}\mathbf{x}_4,$$

where $\mathbf{h}$ denotes some (rather messy) function. In deriving the expression for $\mathbf{z}_4$, we have used the fact that $\mathbf{x}_4$ appears only in the time derivative of $\mathbf{D}^{-1}(\mathbf{x}_1)\,\mathbf{K}\mathbf{x}_3$. Thus the map T taking $\mathbf{x}$ into $\mathbf{z}$ is obviously smooth. The map T also has a smooth inverse given by

$$\mathbf{x}_1 = \mathbf{z}_1,\ \mathbf{x}_2 = \mathbf{z}_2,$$

$$\mathbf{x}_3 = \mathbf{z}_1 + \mathbf{K}^{-1}\,[\mathbf{D}(\mathbf{z}_1)\,\mathbf{z}_3 + \mathbf{c}(\mathbf{z}_1, \mathbf{z}_2)],$$

$$\mathbf{x}_4 = \mathbf{K}^{-1}\,\mathbf{D}(\mathbf{z}_1)\,[\mathbf{z}_4 - \mathbf{h}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)]$$

$$= \mathbf{K}^{-1}\,\mathbf{D}(\mathbf{z}_1)\,[\mathbf{z}_4 - \bar{\mathbf{h}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)],$$

after substituting for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ in terms of $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. Thus T is actually a *global* diffeomorphism.

In terms of the new coordinates, the system equations become

$$\dot{\mathbf{z}}_1 = \mathbf{z}_2,\ \dot{\mathbf{z}}_2 = \mathbf{z}_3,\ \dot{\mathbf{z}}_3 = \mathbf{z}_4,$$

$$\dot{\mathbf{z}}_4 = \frac{d}{dt}\mathbf{h}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) + \frac{d}{dt}[\mathbf{D}^{-1}(\mathbf{x}_1)\,\mathbf{K}]\,\mathbf{x}_4 + \mathbf{D}^{-1}(\mathbf{x}_1)\,\mathbf{K}\dot{\mathbf{x}}_4$$

$$= \mathbf{a}(\mathbf{x}) + \mathbf{D}^{-1}(\mathbf{x}_1)\mathbf{K}\mathbf{J}^{-1}(\mathbf{x}_3)\,[\mathbf{u} + \mathbf{s}(\mathbf{x}_1, \mathbf{x}_3)]$$

$$= \mathbf{q}(\mathbf{x}) + \mathbf{S}(\mathbf{x})\,\mathbf{u},$$

where $\mathbf{a}$ is a smooth m-vector-valued function denoting the sum of the first two terms on the right side of the first equation,

$$\mathbf{q} = \mathbf{a} + \mathbf{D}^{-1}\,\mathbf{K}\,\mathbf{J}^{-1}\mathbf{s}$$

is a smooth ***m*-vector-valued** function, and

$$\mathbf{S} = \mathbf{D}^{-1}\mathbf{K}\mathbf{J}^{-1}$$

is everywhere nonsingular. (Note that the usage of the symbol $\mathbf{q}$ differs from earlier usage to denote various angles.) Thus, if the new input vector $\mathbf{v}$ is defined in accordance with (29), then

$$\dot{\mathbf{z}}_4 = \mathbf{v}.$$

Thus the system equations look like

$$\begin{bmatrix} \dot{\mathbf{z}}_1 \\ \dot{\mathbf{z}}_2 \\ \dot{\mathbf{z}}_3 \\ \dot{\mathbf{z}}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \mathbf{z}_4 \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ I \end{bmatrix} \mathbf{v},$$

where all identity matrices have dimensions $m \times m$. Hence the system is in Brunovsky canonical form, with $\kappa_i = 4$ for all i.

7.6 INPUT-OUTPUT LINEARIZATION

The problem studied in this section is the following: Consider a system with m inputs, m outputs, and n states, described by

$$\mathbf{1} \qquad \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \sum_{j=1}^{m} u_j\,\mathbf{g}_j(\mathbf{x}),$$

$$\mathbf{2} \qquad y_i = h_i(\mathbf{x}),\ i = 1, \cdots, m,$$

where $\mathbf{f}, \mathbf{g}_1, \cdots, \mathbf{g}_m$ are vector fields on some open subset $X \subseteq \mathbf{R}^n$, and $h_i \in S(X)$ for all i. It is desired to apply a state feedback control law of the form

3 $\mathbf{v} = \mathbf{q}(\mathbf{x}) + \mathbf{S}(\mathbf{x})\,\mathbf{u},$

where $\mathbf{q}(\mathbf{x}) \in \mathbf{R}^m$ and $\mathbf{S}(\mathbf{x}) \in \mathbf{R}^{n \times n}$ are smooth functions of $\mathbf{x}$, and in addition $\mathbf{S}(\mathbf{0})$ is nonsingular, such that the following is true: There exist integers $r_i \geq 1$ for $i = 1, \cdots, m$, such that, with the control law (3), the input-output relationship of the system is described by

4 $$\frac{d^{r_i} y_i}{dt^i} = v_i,\ i = 1, \cdots, m.$$

In other words, the feedback law (2) is expected to achieve two things: First, the input-output relationship is **decoupled**, in that y_i depends only on v_i. Second, in the i-th "channel" this input-output relationship is that of an integrator of order r_i. The scheme can be depicted as shown in Figure 7.4.

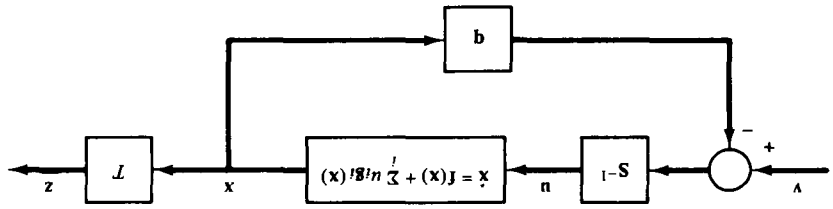


Fig. 7.4

Input-output linearization is in some sense a much less ambitious goal than input-state linearization, which is the subject of Sections 7.4 and 7.5. To illustrate the basic idea, consider a single-input, single-output system described by

5 $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + u\,\mathbf{g}(\mathbf{x}),\ y = h(\mathbf{x}).$

Then

6 $\dot{y} = \nabla h\,\dot{\mathbf{x}} = \langle \mathbf{d}h, \mathbf{f} \rangle + u \langle \mathbf{d}h, \mathbf{g} \rangle.$

Suppose $\langle \mathbf{d}h, \mathbf{g} \rangle(\mathbf{0}) \neq 0$. Then continuity implies that $\langle \mathbf{d}h, \mathbf{g} \rangle(\mathbf{x}) \neq 0$ for all $\mathbf{x}$ sufficiently close to $\mathbf{0}$. So if we define

7 $q(\mathbf{x}) = \langle \mathbf{d}h, \mathbf{f} \rangle,\ s(\mathbf{x}) = \langle \mathbf{d}h, \mathbf{g} \rangle,$

8 $v = q(\mathbf{x}) + s(\mathbf{x})\,u,$

then (6) reduces to

9 $\dot{y} = v.$

Suppose $\langle \mathbf{d}h, \mathbf{g} \rangle(\mathbf{0}) = 0$. Then the above scheme does not work. But suppose

$\langle \mathbf{d}h, \mathbf{g} \rangle(\mathbf{x}) \equiv 0$ for all $\mathbf{x}$ in some neighborhood of $\mathbf{0}$. [This is not always true, even if $\langle \mathbf{d}h, \mathbf{g} \rangle(\mathbf{0}) = 0$.] Then (6) becomes

10 $\dot{y} = \langle \mathbf{d}h, \mathbf{f} \rangle = L_{\mathbf{f}} h.$

Hence

11 $\ddot{y} = \langle \mathbf{d}L_{\mathbf{f}} h, \mathbf{f} \rangle + u \langle \mathbf{d}L_{\mathbf{f}} h, \mathbf{g} \rangle = L_{\mathbf{f}}^2 h + u L_{\mathbf{g}} L_{\mathbf{f}} h,$

where $L_{\mathbf{f}}^2$ is a shorthand for $L_{\mathbf{f}} L_{\mathbf{f}}$. Again, if $(L_{\mathbf{g}} L_{\mathbf{f}} h)(\mathbf{0}) \neq 0$, then we can define

12 $q = L_{\mathbf{f}}^2 h, \; s = L_{\mathbf{g}} L_{\mathbf{f}} h,$

and v by (8). Then (11) implies that

13 $\ddot{y} = v.$

In general, suppose there is a neighborhood U of $\mathbf{0}$ and an integer $r \geq 2$ such that

14 $(L_{\mathbf{g}} L_{\mathbf{f}}^k h)(\mathbf{x}) \equiv 0 \; \forall \mathbf{x} \in U, \text{ for } k = 0, \cdots, r-2,$

15 $(L_{\mathbf{g}} L_{\mathbf{f}}^{r-1} h)(\mathbf{0}) \neq 0.$

Then we can define

16 $q = L_{\mathbf{f}}^r h, \; s = L_{\mathbf{g}} L_{\mathbf{f}}^{r-1} h,$

and v by (8). This results in

17 $y^{(r)} = v.$

In such a case the system (5) is said to **have relative degree** r. This terminology is quite consistent with the linear case. Consider a single-input, single-output system of the form

18 $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u, \; y = \mathbf{c}\mathbf{x}.$

Then the relative degree of this system is the smallest integer r such that

19 $\mathbf{c}\mathbf{A}^k \mathbf{b} = 0 \text{ for } k = 0, \cdots, r-2, \text{ and } \mathbf{c}\mathbf{A}^{r-1}\mathbf{b} \neq 0.$

For a linear system of the form (18), we know that the relative degree r is less than or equal to the system order n. This follows readily from (19). If

20 $\mathbf{c}\mathbf{A}^k \mathbf{b} = 0 \text{ for } k = 0, \cdots, n-1,$

then the Cayley-Hamilton theorem implies that

21 $\mathbf{cA}^k\mathbf{b} = 0 \; \forall k \geq 0,$

or in other words, the transfer matrix is identically zero, which contradicts the fact that $\mathbf{cA}^{r-1}\mathbf{b} \neq 0$. But for nonlinear systems there is no analog of the Cayley-Hamilton theorem. So one is forced to adopt a different strategy.

22 Lemma *Consider the system (5), and suppose there exists an integer r satisfying (14) – (15). Then* $r \leq n$.

This follows immediately from another result, which is useful in its own right.

23 Lemma *Consider the system (5), and suppose (14) – (15) hold for some integer r. Then the row vectors* $\{\mathbf{d}h, \mathbf{d}L_{\mathbf{f}}h, \cdots, \mathbf{d}L_{\mathbf{f}}^{r-1}h\}$ *are linearly independent in some neighborhood of* $\mathbf{0}$.

Proof Choose real constants $\alpha_0, \cdots, \alpha_{r-1}$ such that

24 $$\sum_{i=0}^{r-1} \alpha_i \, \mathbf{d}L_{\mathbf{f}}^i h(\mathbf{0}) = \mathbf{0}.$$

It is shown that $\alpha_i = 0$ for all i. For this purpose, it is claimed that

25 $$L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}} L_{\mathbf{f}}^k h = \begin{cases} 0, & \text{if } i + k \leq r-1, \\ (-1)^i L_{\mathbf{g}} L_{\mathbf{f}}^{r-1} h, & \text{if } i + k = r-1 \end{cases}$$

The proof of (25) is by induction on i for fixed k. If $i = 0$, then (25) reduces to (14). Now suppose (25) is true for $0, \cdots, i-1$. Note that $\mathrm{ad}_{\mathbf{f}}^i \mathbf{g} = [\mathbf{f}, \mathrm{ad}_{\mathbf{f}}^{i-1} \mathbf{g}]$. Hence, by Lemma (7.1.59), we have

26 $$L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}} L_{\mathbf{f}}^k h = L_{\mathbf{f}} L_{\mathrm{ad}_{\mathbf{f}}^{i-1} \mathbf{g}} L_{\mathbf{f}}^k h - L_{\mathrm{ad}_{\mathbf{f}}^{i-1} \mathbf{g}} L_{\mathbf{f}} L_{\mathbf{f}}^k h = -L_{\mathrm{ad}_{\mathbf{f}}^{i-1} \mathbf{g}} L_{\mathbf{f}}^{k+1} h,$$

since the first term is zero by the inductive assumption. Now, if $i + k \leq r - 1$, so of course is $(i-1) + (k+1)$. Hence, once again by the inductive assumption, it follows that

27 $L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}} L_{\mathbf{f}}^k h = 0$ if $i + k \leq r - 1$.

On the other hand, if $i + k = r - 1$, then the inductive assumption implies that

28 $$L_{\mathrm{ad}_{\mathbf{f}}^i \mathbf{g}} L_{\mathbf{f}}^k h = -L_{\mathrm{ad}_{\mathbf{f}}^{i-1} \mathbf{g}} L_{\mathbf{f}}^{k+1} h = \cdots = (-1)^i L_{\mathbf{g}} L_{\mathbf{f}}^{r-1} h.$$

To prove that $\alpha_i = 0 \; \forall i$ if (24) holds, define

29 $$a = \sum_{i=0}^{r-1} \alpha_i L_{\mathbf{f}}^i h \in S(X).$$

Suppose (24) is true, i.e., $\mathbf{d}a(\mathbf{0}) = \mathbf{0}$. Then

30 $0 = \langle \mathbf{d}a, \mathbf{g} \rangle(\mathbf{0}) = (L_{\mathbf{g}} a)(\mathbf{0})$.

However,

31 $L_{\mathbf{g}} a = \sum_{i=0}^{r-1} \alpha_i L_{\mathbf{g}} L_{\mathbf{f}}^i h = \alpha_{r-1} L_{\mathbf{g}} L_{\mathbf{f}}^{r-1} h,$

where the last step follows from (14). Combining (30) and (31) shows that

32 $0 = \alpha_{r-1} (L_{\mathbf{g}} L_{\mathbf{f}}^{r-1} h)(\mathbf{0})$.

Now (15) implies that $\alpha_{r-1} = 0$, i.e., that

33 $a = \sum_{i=0}^{r-2} \alpha_i L_{\mathbf{f}}^i h.$

Next, we have

34 $\langle \mathbf{d}a, \mathrm{ad}_{\mathbf{f}} \mathbf{g} \rangle(\mathbf{0}) = (L_{\mathrm{ad}_{\mathbf{f}} \mathbf{g}} a)(\mathbf{0}) = 0,$

since $\mathbf{d}a(\mathbf{0}) = \mathbf{0}$ by assumption. Substituting for a from (33) and using (25) leads to

35 $0 = -\alpha_{r-2} (L_{\mathbf{g}} L_{\mathbf{f}}^{r-1} h)(\mathbf{0})$.

In turn, this combined with (15) shows that $\alpha_{r-2} = 0$. The argument can be repeated all the way down to show that $\alpha_i = 0 \; \forall i$. ■

Proof of Lemma (22) Since the set of row vectors $\{\mathbf{d}L_{\mathbf{f}}^i h, \; 0 \leq i \leq r-1\}$ is linearly independent, it is obvious that $r \leq n$. ■

In the case of multi-input, multi-output systems, the idea is essentially the same, except that the relative degree is now a vector and not a scalar. The system (1) – (2) is said to **have the relative degree vector** $\mathbf{r} = [r_1 \cdots r_m]'$ if (i) there exists a neighborhood U of $\mathbf{0}$ such that

36 $(L_{\mathbf{g}_j} L_{\mathbf{f}}^k h_i)(\mathbf{x}) \equiv 0$, for $k = 0, \cdots, r_i - 2$,

and (ii) the $m \times m$ matrix $\mathbf{S}$ defined by

37 $s_{ij} = L_{\mathbf{g}_j} L_{\mathbf{f}}^{r_i - 1} h_i$

is nonsingular at $\mathbf{x} = \mathbf{0}$. In this case, if we define the smooth m-vector valued function $\mathbf{q}$ by

38 $q_i = L_{\mathbf{f}}^{r_i} h_i,$

define the smooth $m \times m$ matrix $\mathbf{S}$ as in (36), and the new control vector $\mathbf{v}$ by (3), then the resulting input-output relationship is of the form (4).

The next result is the MIMO analog of Lemma (22).

39 Lemma *Suppose the system (1) – (2) has the relative degree vector* $\mathbf{r} = [r_1 \cdots r_m]'$. *Then*

$$\mathbf{40} \quad \sum_{i=1}^{m} r_i \leq n.$$

As in the scalar case, the proof of this lemma is based on another result which is of independent interest.

41 Lemma *Consider the system (1) – (2), and suppose it has the relative degree vector* $\mathbf{r} = [r_1 \cdots r_n]'$. *Then the* $r = \sum_{i=1}^{m} r_i$ *row vectors*

$$\mathbf{42} \quad \{\mathbf{d}L_{\mathbf{f}}^{k} h_i,\ 0 \leq k \leq r_i - 1,\ 1 \leq i \leq m\}$$

are linearly independent in some neighborhood of $\mathbf{0}$.

Proof Select real constants α_{ik}, $0 \leq k \leq r_i - 1$, $1 \leq i \leq m$, such that

$$\mathbf{43} \quad \sum_{i=1}^{m} \sum_{k=1}^{r_i - 1} \alpha_{ik} (\mathbf{d}L_{\mathbf{f}}^{k} h_i)(\mathbf{0}) = \mathbf{0}.$$

Define

$$\mathbf{44} \quad a = \sum_{i=1}^{m} \sum_{k=1}^{r_i - 1} \alpha_{ik} L_{\mathbf{f}}^{k} h_i \in S(X).$$

Then, using (36), it follows that

$$\mathbf{45} \quad L_{\mathbf{g}_j} a = \sum_{i=1}^{m} \alpha_{i,r_i - 1} L_{\mathbf{g}_j} L_{\mathbf{f}}^{r_i - 1} h_i = \sum_{i=1}^{m} \alpha_{i,r_i - 1} s_{ij},$$

where s_{ij} is defined in (37). Now (43) shows that

$$\mathbf{46} \quad 0 = \langle \mathbf{d}a, \mathbf{g}_j \rangle (\mathbf{0}) = (L_{\mathbf{g}_j} a)(\mathbf{0}), \text{ for } 1 \leq j \leq m.$$

Equivalently,

$$\mathbf{47} \quad \mathbf{0} = [\alpha_{1,r_i - 1} \cdots \alpha_{m,r_m - 1}]\, \mathbf{S}(\mathbf{0}).$$

But since $\mathbf{S}(\mathbf{0})$ is nonsingular, this shows that

$$\mathbf{48} \quad \alpha_{i,r_i - 1} = 0, \text{ for } i = 1, \cdots, m.$$

Hence

$$\textbf{49} \quad a = \sum_{i=1}^{m} \sum_{k=1}^{r_i - 2} \alpha_{ik} L_{\mathbf{f}}^k h_i \in S(X).$$

Now the Lie derivative $L_{\mathrm{ad}_{\mathbf{f}}\mathbf{g}_j} a$ is computed for each j. The details follow along the same lines as (47), and lead to

$$\textbf{50} \quad \mathbf{0} = -[\alpha_{1,r_1-2} \cdots \alpha_{m,r_m-2}]\, \mathbf{S}(\mathbf{0}),$$

which in turn shows that

$$\textbf{51} \quad \alpha_{i,r_i-2} = 0, \text{ for } i = 1, \cdots, m.$$

The argument can be repeated all the way down to show that $\alpha_{i,k} = 0 \; \forall i, \; k$. ■

Proof of Lemma (39) Since the set of row vectors $\{\mathbf{d}L_{\mathbf{f}}^k h_i, \, 0 \le k \le r_i - 1, \, 1 \le i \le m\}$ is linearly independent, the number of vectors cannot exceed n. It follows that (40) is true. ■

Thus far the process of input-output linearization has been purely mechanical with very little geometric insight. The next result shows that input-output decoupling is achieved, in effect, by making some system states unobservable via state feedback.

52 Theorem *Consider the system (1) – (2), and suppose it has the relative degree vector* $\mathbf{r} = [r_1 \cdots r_m]'$. *Define* $r = \sum_{i=1}^{m} r_i$. *Then there exists a local diffeomorphism T around* $\mathbf{0}$ *such that, in terms of the transformed state vector* $\mathbf{z} = T(\mathbf{x})$, *the system equations assume the following form: Partition* $\mathbf{z} \in \mathbf{R}^n$ *as*

$$\textbf{53} \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ . \\ . \\ . \\ \mathbf{z}_m \\ \mathbf{z}_u \end{bmatrix},$$

where

$$\textbf{54} \quad \mathbf{z}_i \in \mathbf{R}^{r_i}, \text{ for } 1 \le i \le m, \; \mathbf{z}_i \in \mathbf{R}^{n-r}.$$

Then, with the control law $\mathbf{v}$ *of (3) with* $\mathbf{q}$ *and* $\mathbf{S}$ *given by (38) and (37) respectively, we have*

$$\textbf{55} \quad \dot{z}_{i,1} = z_{i,2}, \; \cdots, \; \dot{z}_{i,r_i-2} = z_{i,r_i-1}, \; \dot{z}_{i,r_i} = v_i, \; 1 \le i \le m,$$

56 $\dot{\mathbf{z}}_u = \mathbf{f}_u(\mathbf{z}) + \sum_{j=1}^{m} \mathbf{g}_{uj}(\mathbf{z})\, v_j,$

57 $y_i = z_{i,1},\ 1 \le i \le m.$

Remarks Let $\mathbf{A}_i$ be an $r_i \times r_i$ matrix in companion form, with characteristic polynomial s^{r_i}; let $\mathbf{b}_i$ be an $r_i \times 1$ column vector with a "1" in the last row and zeros elsewhere. Finally, let $\mathbf{c}_i$ be a $1 \times r_i$ row vector with a "1" in the first column and zeros elsewhere. Then (55) and (57) can be expressed as

58 $\dot{\mathbf{z}}_i = \mathbf{A}_i \mathbf{z}_i + \mathbf{b}_i v_i,\ y_i = \mathbf{c}_i \mathbf{z}_i,$

while (56) remains as is. Thus, after applying the feedback control law (3), (37) – (38) and changing coordinates, the system looks like m decoupled systems, plus (56). Alternatively, define

59 $$\mathbf{z}_o = \begin{bmatrix} \mathbf{z}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{z}_m \end{bmatrix} \in \mathbf{R}^r,$$

and let $(\mathbf{A}_o, \mathbf{B}_o)$ be a Brunovsky canonical form corresponding to the block sizes $r_1, \cdots, r_m$ [cf. (7.5.10) *et seq.*]. Then (55) and (57) can be expressed as

60 $\dot{\mathbf{z}}_o = \mathbf{A}_o \mathbf{z}_o + \mathbf{B}_o \mathbf{v},\ \mathbf{y} = \mathbf{C}_o \mathbf{z}_o,$

where the definition of $\mathbf{C}$ is self-evident. Note that $\mathbf{z}_u$ does not affect any of the $\mathbf{z}_i$ or y_i, and is thus "unobservable."

Proof By Lemma (41), the r vectors of (42) are linearly independent. Accordingly, we can define

61 $z_{i,k+1} = L_{\mathbf{f}}^k h_j,\ 0 \le k \le r_i - 1,\ 1 \le i \le m$

as the first r components of a local diffeomorphism T around $\mathbf{0}$. The last $n - r$ components of T can be chosen arbitrarily, so long as the row vectors $\mathbf{d}T_{r+1}(\mathbf{0}), \cdots, \mathbf{d}T_n(\mathbf{0})$, together with the r row vectors of (41), form a row basis. Then

62 $\dot{z}_{i,k+1} = L_{\mathbf{f}}^{k+1} h_i + \sum_{j=1}^{m} u_j\, L_{\mathbf{g}_j} L_{\mathbf{f}}^k h_i.$

If $k \le r_i - 2$, then from (36) and (55), we have

63 $\dot{z}_{i,k+1} = z_{i,k+2}.$

If $k = r_i - 1$, then from (37) and (38), we get

64 $\dot{z}_{i,r_i} = L_{\mathbf{f}}^{r_i} h_i + \sum_{j=1}^{m} u_j L_{\mathbf{g}_j} L_{\mathbf{f}}^{r_i - 1} h_i = q_i + \sum_{j=1}^{m} s_{ij} u_j = v_i.$

This proves (55). If $\mathbf{z}_u$ denotes the vector consisting of the last $n - r$ components of $\mathbf{z}$, then $\dot{\mathbf{z}}_u$ has no special form, and just looks like (56). Finally, the relation (57) follows from (61) by setting $k = 0$. ■

7.7 STABILIZATION OF LINEARIZABLE SYSTEMS

In this brief section, the various results presented thus far are combined into procedures for stabilizing nonlinear systems that are linearizable, either in the input-state sense or in the input-output sense.

Throughout the section, attention is focused on the familiar system model

1 $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \sum_{i=1}^{m} u_i \mathbf{g}_i(\mathbf{x}),$

2 $\mathbf{y} = \mathbf{h}(\mathbf{x}),$

where $\mathbf{f}, \mathbf{g}_i \in V(X)$, and $\mathbf{h}: X \to \mathbf{R}^m$ is smooth. Notice the assumption that the system is *square*, i.e., has the same number of inputs and outputs. In addition, it is also assumed that

3 $\mathbf{f}(\mathbf{0}) = \mathbf{0}, \mathbf{h}(\mathbf{0}) = \mathbf{0}.$

Hence the state $\mathbf{x}_0 = \mathbf{0}$ is an equilibrium of the system in the sense that if the system is started in this initial state and no input is applied, then the system remains in that state thereafter. This does not mean, however, that the equilibrium $\mathbf{0}$ is asymptotically stable in the sense of Lyapunov (cf. Chapter 5).

The problem studied in this section is the stabilization of the system (1) – (2) via a feedback control law of the form

4 $\mathbf{v} = \mathbf{q}(\mathbf{x}) + \mathbf{S}(\mathbf{x})\,\mathbf{u},$

where $\mathbf{q}: X \to \mathbf{R}^m$ and $\mathbf{S}: X \to \mathbf{R}^{m \times m}$ are smooth functions, and in addition, $\mathbf{q}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{S}(\mathbf{0})$ is nonsingular. Two distinct cases are considered, namely: (i) the system (1) – (2) is input-state linearizable in the sense of Section 7.5, and (ii) the system (1) – (2) is input-output linearizable in the sense of Section 7.6.

Accordingly, suppose first that the system (1) – (2) satisfies the conditions of Theorem (7.5.32). Choose the feedback functions $\mathbf{q}$ and $\mathbf{S}$ in (4) as well as a local diffeomorphism T around $\mathbf{0}$, such that the transformed state vector $\mathbf{z} = T(\mathbf{x})$ is related to $\mathbf{v}$ by

5 $\dot{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{B}\mathbf{v}$,

where $(\mathbf{A}, \mathbf{B})$ is in Brunovsky canonical form. Next, note from the line below (7.5.69) that $q_i = L_{\mathbf{f}}^{\kappa_i} \phi_i$ for some integer κ_i and smooth function ϕ_i, for each i. Since $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, it follows that $\mathbf{q}(\mathbf{0}) = \mathbf{0}$. Also, it can be assumed without loss of generality that $T(\mathbf{0}) = \mathbf{0}$. Now the eigenvalues of the matrix $\mathbf{A}$ are all at the origin; hence the system (5) is *not* asymptotically stable. However, since the pair $(\mathbf{A}, \mathbf{B})$ is controllable, it *is* possible to choose an $m \times n$ matrix $\mathbf{K}$ such that the matrix $\mathbf{A} - \mathbf{B}\mathbf{K}$ is Hurwitz, i.e., such that all eigenvalues of $\mathbf{A} - \mathbf{B}\mathbf{K}$ have negative real parts. Hence, if we apply the further feedback control

6 $\mathbf{v} = -\mathbf{K}\mathbf{z}$,

then $\mathbf{z} = \mathbf{0}$ (or equivalently, $\mathbf{x} = \mathbf{0}$) is an exponentially stable equilibrium of the resulting system. The scheme can be depicted as in Figure 7.5. For obvious reasons, the control law (4) is referred to as the **inner-loop control**, while (6) is referred to as the **outer-loop control**.

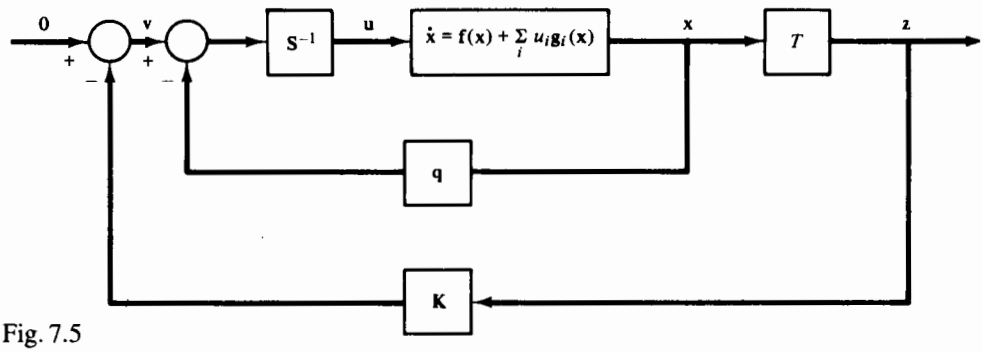


Fig. 7.5

7 **Example** A well-studied application of the above strategy is from the world of robotics. Consider a *rigid* m-link robot operating in a gravity-free environment. Thus the robot can be operating in a "horizontal" plane, perpendicular to gravity, or else it can be operating in outer space. In either case, it is known [see Spong and Vidyasagar (1989), Chapters 6 and 7] that the dynamics of such a robot are described by the Euler-Lagrange equations

$$\mathbf{M}(\mathbf{y})\ddot{\mathbf{y}} + \mathbf{h}(\mathbf{y}, \dot{\mathbf{y}}) = \mathbf{u},$$

where $\mathbf{y}$ is the m-vector of link angles, and $\mathbf{u}$ is the m-vector of the torques applied at the various joints. The matrix $\mathbf{M}$, known as the **inertia matrix**, is symmetric and positive definite for each $\mathbf{y}$. The vector $\mathbf{h}$ incorporates all the Coriolis and centripetal terms. *In the absence of gravity*, $\mathbf{h}$ satisfies

$$\mathbf{h}(\mathbf{0}, \mathbf{0}) = \mathbf{0}.$$

For the system above, let us choose the natural state vector

$$\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \dot{\mathbf{y}} \end{bmatrix},$$

and choose the inner-loop control

$$\mathbf{u} = \mathbf{M}(\mathbf{y})\,\mathbf{v} + \mathbf{h}(\mathbf{y}, \dot{\mathbf{y}}),$$

which is clearly of the form (4). Then the resulting closed-loop system is described by

$$\mathbf{M}(\mathbf{y})\,\ddot{\mathbf{y}} = \mathbf{M}(\mathbf{y})\,\mathbf{v}.$$

Since $\mathbf{M}$ is nonsingular (by virtue of being positive definite), both sides of the above equation can be multiplied by $\mathbf{M}^{-1}$, giving

$$\ddot{\mathbf{y}} = \mathbf{v}.$$

Hence the system looks like a set of m double integrators. This shows that all the Kronecker indices $\kappa_1, \cdots, \kappa_m$ are all equal to 2 [cf. (7.5.6)]. Now a standard proportional plus derivative feedback of the form

$$\mathbf{v} = -\mathbf{K}_p\mathbf{y} - \mathbf{K}_d\dot{\mathbf{y}},$$

chosen as the outer-loop control, stabilizes the system, provided of course that $\mathbf{K}_p$ and $\mathbf{K}_d$ are chosen suitably. A common choice is

$$\mathbf{K}_p = k_p I,\ \mathbf{K}_d = k_d I,$$

in which case *any* $k_p > 0$, $k_d > 0$ will stabilize the system.

The above strategy is often referred to in the robotics community as the **computed torque method.** ■

Now suppose the system (1) – (2) fails to be input-state linearizable, but does satisfy the conditions (7.6.36-37). Define $\mathbf{S}$ and $\mathbf{q}$ in (4) in accordance with (7.6.37) and (7.6.38) respectively, and assume that $\mathbf{S}(\mathbf{0})$ is nonsingular. This means that the system at hand is input-output linearizable. Let $\mathbf{r} = [r_1 \cdots r_m]'$ denote the relative degree vector of the system. By Theorem (7.6.52) and the remarks thereafter, it follows that after applying the control (4), the closed-loop system is described by [cf. (7.6.56) and (7.6.60)]

$$\mathbf{8} \qquad \dot{\mathbf{z}}_o = \mathbf{A}_o\mathbf{z}_o + \mathbf{B}_o\mathbf{v},\ \mathbf{y} = \mathbf{C}_o\mathbf{z}_o,$$

$$\mathbf{9} \qquad \dot{\mathbf{z}}_u = \mathbf{f}_u(\mathbf{z}) + \sum_{i=1}^{m} \mathbf{g}_{ui}(\mathbf{z}_o, \mathbf{z}_u)\,v_j,$$

where $\mathbf{A}_o \in \mathbf{R}^{r\times r}$, $\mathbf{B}_o \in \mathbf{R}^{r\times m}$, $\mathbf{C}_o \in \mathbf{R}^{m\times r}$, and $(\mathbf{A}_o, \mathbf{B}_o)$ is in the Brunovsky canonical form.

Hence, by applying the outer-loop control law

10 $\mathbf{v} = -\mathbf{K}_o \mathbf{z}_o,$

and choosing $\mathbf{K}_o$ suitably, one can ensure that the resulting closed-loop system

11 $\dot{\mathbf{z}}_o = (\mathbf{A}_o - \mathbf{B}_o \mathbf{K}_o)\mathbf{z}_o$

is exponentially stable. But what happens to the "unobservable" part $\mathbf{z}_u$?

12 Definition *The system*

13 $\dot{\mathbf{z}}_u = \mathbf{f}_u(\mathbf{0}, \mathbf{z}_u)$

evolving on $\mathbf{R}^{n-r}$ *is called the* **zero dynamics** *of the system (1) – (2).*

The zero dynamics represent the dynamics of the unobservable part $\mathbf{z}_u$ when (i) the input is set equal to zero, and (ii) the output is constrained to be identically zero (which in turn implies that $\mathbf{z}_o \equiv \mathbf{0}$). It is possible to define the zero dynamics directly, without going through the state feedback control law and the state variable transformation; the interested reader is referred to Nijmeijer and van der Schaft (1990), Chapter 12.

The next two results show that, if the zero dynamics are stable, then the outer-loop control (11) *does* indeed stabilize the system.

14 Theorem *Suppose* $\mathbf{z}_u = \mathbf{0}$ *is an asymptotically stable equilibrium of (13), and that the matrix* $\mathbf{A}_o - \mathbf{B}_o \mathbf{K}_o$ *is Hurwitz. Then* $(\mathbf{z}_o, \mathbf{z}_u) = (\mathbf{0}, \mathbf{0})$ *is an asymptotically stable equilibrium of the system*

15 $\dot{\mathbf{z}}_o = \mathbf{A}_o \mathbf{z}_o + \mathbf{B}_o \mathbf{v},$

16 $\dot{\mathbf{z}}_u = \mathbf{f}_u(\mathbf{z}_o, \mathbf{z}_u) + \sum_{i=1}^{m} \mathbf{g}_{ui}(\mathbf{z}_o, \mathbf{z}_u)\, v_i,$

17 $\mathbf{v} = -\mathbf{K}_o \mathbf{z}_o.$

Proof The result follows readily from Theorem (5.8.84). First, the system (15) – (17) has the "triangular" form needed to apply the theorem. Second, if we substitute $\mathbf{z}_o = \mathbf{0}$ (and hence $\mathbf{v} = \mathbf{0}$) in (16), then $\dot{\mathbf{z}}_u$ is given by (13). ■

18 Theorem *Suppose* $\mathbf{z}_u = \mathbf{0}$ *is an exponentially stable equilibrium of (15), and that the matrix* $\mathbf{A}_o - \mathbf{B}_o \mathbf{K}_o$ *is Hurwitz. Then* $(\mathbf{z}_o, \mathbf{z}_u) = (\mathbf{0}, \mathbf{0})$ *is an exponentially stable equilibrium of the system (15) – (17).*

Proof The result follows readily from Theorem (5.8.103). ■

Notes and References

What follows is a very brief description of the evolution of differential-geometric control theory. Thorough treatments of the theory as well as detailed attributions of individual contributions can be found in the recent excellent texts by Isidori (1985) and Nijmeijer and van der Schaft (1990).

While the treatment of the theory in the present text is heavily slanted towards explicit computation in a given coordinate system, the reader would find it useful to have at least an introduction to the modern "coordinate-free" approach to differential geometry. The book by Guillemin and Pollack (1974) is recommended. In addition, the afore-mentioned books by Isidori and by Nijmeijer and van der Schaft contain a lot of background material. The treatment of distributions is standard, and the proof of the Frobenius theorem (given in Appendix C) follows Warner (1971).

To many people, the influential paper by Brockett (1972) provided the first glimpse of the potential of differential-geometric methods in solving nonlinear control problems. The reachability problem was treated by Lobry (1970) and Krener (1974). The paper by Krener contains Theorem (7.3.41). Theorem (7.3.46) regarding reachability around an equilibrium can be found in Lee and Markus (1967). The idea of observation space was introduced in the papers by Hermann and Krener (1977) and Isidori et al. (1981).

The feedback linearization problem, of the input-state relationship but without a state-space transformation, was posed by Brockett (1978). The problem was solved, after the class of control laws was enlarged to include a state-space transformation, in the single-input case by Su (1982), and in the multi-input case by Hunt and Su (1981) and Hunt et al. (1983a, 1983b); see also Isidori and Ruberti (1984). The proof of Theorem (7.5.32) given here follows Isidori (1989).

The input-output linearization problem can be traced to the work of Singh and Rugh (1972); the decoupling problem was also solved in Isidori et al. (1981). The importance of zero dynamics is brought out in Byrnes and Isidori (1984, 1988).

Finally, the theory has been enriched by several examples and applications. While it is not always possible to assign proper credit for examples, it can be mentioned that the study in Problem 7.8 is a simplified version of Crouch (1984). The application in Section 7.4 to a single link with a flexible joint is due to Marino and Spong (1986), while its extension to multiple flexible joints is due to Spong (1987). In Wang (1989), the theory developed in Sections 7.5 and 7.6 is applied to the problem of controlling a multi-link robot with a single flexible *link*. It is shown that the system is *not* input-state feedback linearizable, but *is* input-output feedback linearizable; see also Wang and Vidyasagar (1991).

A. PREVALENCE OF DIFFERENTIAL EQUATIONS WITH UNIQUE SOLUTIONS

In Section 2.4, it has been shown that the differential equation

$$\mathbf{1} \qquad \dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t)]$$

has a unique solution trajectory corresponding to each initial condition

$$\mathbf{2} \qquad \mathbf{x}(t_0) = \mathbf{x}_0$$

provided the function **f** satisfies a local Lipschitz condition. On the surface, this condition would appear to be hopelessly restrictive, since the set of all Lipschitz-continuous functions is a "meager" subset of the set of all continuous functions. (This term is defined shortly.) This might lead one to conjecture that the set of functions **f** for which (1) has a unique solution trajectory passing through each pair $(t_0, \mathbf{x}_0)$ is itself meager. But in 1932, the Polish mathematician Witold Orlicz proved that the set of all such **f** is in fact the *complement* of a meager set. This leads to a paradoxical situation: Orlicz' theorem states that "practically all" differential equations have unique solutions. On the other hand, the only such differential equations that can be characterized *explicitly* are "practically nothing" within this set. This suggests that there is a huge set of differential equations with unique solutions, waiting for an explicit description.

The purpose of this appendix is to make these ideas precise. In the interests of conciseness, it is assumed that the reader knows something about topological spaces and about Cantor's classical theorem on the meagerness of the set of differentiable functions within the set of continuous functions. Though the relevant results are summarized below, a reader who is being exposed to these ideas for the first time is advised first to read the appropriate background material from a standard text on topology, e.g., Kelley (1955).

First some concepts from topology. Suppose X is a topological space. This means that a collection **T** of subsets of X has been identified with the following properties:

1. Both X and the empty set $\varnothing$ belong to **T**.
2. Arbitrary unions and finite intersections of sets belonging to **T** once again belong to **T**.

The sets belonging to **T** are said to be **open.** A set is **closed** if and only if its complement is open. From Conditions 1 and 2 above, it follows that both X and $\varnothing$ are closed, and finite unions and arbitrary intersections of closed sets are again closed. Let $A \subseteq X$. Then the

closure of A, denoted by $\bar{A}$, is the intersection of all closed sets which contain A; clearly $\bar{A}$ is itself closed. The **interior** of A, denoted by A^o, is the union of all open sets contained in A; clearly A^o is itself open. A set $A \subseteq X$ is **dense** if $\bar{A} = X$. It is **nowhere dense** if $(\bar{A})^o = \emptyset$, or equivalently, if the complement of $\bar{A}$ is dense. It is fairly easy to show that a finite intersection of dense sets is itself dense, and that a finite union of nowhere dense sets is itself nowhere dense. A set $A \subseteq X$ is **meager** or **of first category** if it is a *countable* union of nowhere dense sets; it is **nonmeager** or **of second category** otherwise, i.e., if it is not meager. Thus every $A \subseteq X$ is either meager or nonmeager. Contrast this with the fact that a set may be neither open nor closed. If X is a Banach space and $\mathbf{T}$ is the topology derived from the norm on X, then the well-known **Baire category theorem** states that the complement of a meager set is dense.

Now back to differential equations. Let D be a given subset of $\mathbf{R}^{n+1}$ which is open and bounded, where n is the order of the differential equations we wish to study. Let $C(D)$ denote the set of all continuous, bounded functions mapping D into $\mathbf{R}^n$. If we define

$$\mathbf{3} \qquad \|\mathbf{f}\|_D = \sup_{(t,\,\mathbf{x})\in D} \|\mathbf{f}(t, \mathbf{x})\|,$$

then $\|\cdot\|_D$ is a norm on $C(D)$, and $C(D)$ is a Banach space with this norm. Let us equip $C(D)$ with the topology derived from this norm. Now there are two interesting properties of $C(D)$.

4 Fact *Let $C_1(D)$ denote the set of functions in $C(D)$ that are everywhere right-differentiable as well as left-differentiable. Then $C_1(D)$ is a meager subset of $C(D)$.*

This fact was discovered by Georg Cantor. A popular way to express it is: Almost all continuous functions are nondifferentiable.

5 Fact *$C_1(D)$ is a dense subset of $C(D)$. In fact, the set $C^{(1)}(D)$ of functions that are continuously differentiable everywhere is dense in $C(D)$.*

Facts (4) and (5) taken together mean that, even though almost all continuous functions are nondifferentiable, there are enough differentiable functions in the sense that every continuous function can be approximated arbitrarily closely by a differentiable function.

Now consider the differential equation (1). The only known sufficient condition to guarantee that (1) has a unique solution passing through every point is that the function $\mathbf{f}$ be Lipschitz continuous. (See Section 2.4.) Since Lipschitz continuity implies both right- and left-differentiability, it follows from Fact (4) that the set of all Lipschitz continuous functions is a meager set, or a set of first category. Hence Theorem (2.4.3) is applicable only to a meager set of functions $\mathbf{f}$. This leads naturally to the question: Is the set of functions $\mathbf{f}$ for which (1) has a unique solution passing through each point $(t_0, \mathbf{x}_0) \in D$ a meager set?

Theorem (10) below shows that the answer is an emphatic "No." In fact, almost exactly the *opposite* is true. The set of functions $\mathbf{f}$ for which (1) has a unique solution through each point $(t_0, \mathbf{x}_0) \in D$ is the *complement* of a meager set, and by the Baire category theorem, is therefore a *dense* subset of $C(D)$. In order to prove this theorem, a preliminary

result is presented. This preliminary result is not proved, since the proof involves concepts such as equicontinuity, the discussion of which would take us too far afield.

6 Fact *Consider the differential equation (1), and suppose* $\mathbf{f} \in C(D)$. *Then, through each* $(t_0, \mathbf{x}_0) \in D$, *there exists a (not necessarily unique) solution trajectory of (1). In particular, through each* $(t_0, \mathbf{x}_0) \in D$, *there exist an interval* (t_1, t_2) *containing* t_0, *a maximal solution* $\mathbf{x}_{\max}(\cdot)$, *and a minimal solution* $\mathbf{x}_{\min}(\cdot)$, *such that (i) both* $\mathbf{x}_{\max}$ *and* $\mathbf{x}_{\min}$ *satisfy (1), and (ii)*

$$\mathbf{x}_{\min}(t) \leq \mathbf{x}(t) \leq \mathbf{x}_{\max}(t), \quad \forall t \in (t_1, t_2), \tag{7}$$

whenever $\mathbf{x}(\cdot)$ *satisfies (1). Select numbers* $a, b > 0$ *such that*

$$\{(t, \mathbf{x}): |t - t_0| < a, \ \|\mathbf{x} - \mathbf{x}_0\| < b\} \subseteq D. \tag{8}$$

Then $\mathbf{x}_{\max}$ *and* $\mathbf{x}_{\min}$ *are defined at least over the interval* $[t_0 - \alpha, t_0 + \alpha]$, *where*

$$\alpha = \min\{a, \frac{b}{\|\mathbf{f}\|_D}\}. \tag{9}$$

Let $\mathbf{x}_{\max}(\mathbf{f}; t, t_0, \mathbf{x}_0)$ denote the maximal solution of (1) passing through $(t_0, \mathbf{x}_0)$ evaluated at time t, and define $\mathbf{x}_{\min}(\mathbf{f}; t, t_0, \mathbf{x}_0)$ analogously.

10 Theorem *The set U of all functions* $\mathbf{f} \in C(D)$ *for which (1) has a unique solution passing through each* $(t_0, \mathbf{x}_0) \in D$ *is the complement of a meager set.*

Proof Let ∂D denote the boundary of D, and for each integer $s \geq 1$, let $\bar{D}_s$ denote the set of all points $(t, \mathbf{x}) \in D$ which are at a distance of at least $1/s$ from ∂D. Then $\bar{D}_s \subseteq D$, and $\bar{D}_s$ is closed for each s, but

$$D = \bigcup_{s \geq 1} \bar{D}_s. \tag{11}$$

Let m, p, r, s range over the natural numbers. For each s, r, let $\{\tau_1, \tau_2, \cdots\}$ denote the rational numbers in the interval $(-1/sr, 1/sr)$. Suppose $(t_0, \mathbf{x}_0) \in \bar{D}_s$. Then (8) is satisfied with $a = b = 1/s$. Hence, if $\|\mathbf{f}\|_D \leq r$, then by Fact (6), the maximal and minimal solutions of (1) – (2) are defined at least over the interval $[t_0 - 1/sr, t_0 + 1/sr]$. Now define the set F_{mprs} as the set of all $f \in C(D)$ satisfying two conditions: (i)

$$\|\mathbf{f}\|_D \leq r, \tag{12}$$

and (ii) there exists a point $(t, \mathbf{x}) \in \bar{D}_s$ such that

$$\mathbf{x}_{\max}(\mathbf{f}; t + \tau_m, t, \mathbf{x}) - \mathbf{x}_{\min}(\mathbf{f}; t + \tau_m, t, \mathbf{x}) \geq 1/p. \tag{13}$$

Define

14 $F = \bigcup_{m,p,r,s} F_{mprs}$.

It is claimed that U is the complement of F in $C(D)$. It is easy to see that U and F are disjoint: If $\mathbf{f} \in U$, then by the uniqueness of solutions,

15 $\mathbf{x}_{\max}(\mathbf{f}; t+\tau, t, \mathbf{x}) = \mathbf{x}_{\min}(\mathbf{f}; t+\tau, t, \mathbf{x}), \forall t, \tau, \mathbf{x}.$

So (13) can never be satisfied, and so $\mathbf{f}$ does not belong to F. To show that every $\mathbf{f}$ not in U belongs to F, we show that if $\mathbf{f}U$, then $\mathbf{f} \in F_{mprs}$ for sufficiently large m, p, r, s. Since $\mathbf{f}U$, there is a point $(t, \mathbf{x}) \in D$ and a $\tau > 0$ such that

16 $\mathbf{x}_{\max}(\mathbf{f}; t+\tau, t, \mathbf{x}) > \mathbf{x}_{\min}(\mathbf{f}; t+\tau, t, \mathbf{x}).$

Since both $\mathbf{x}_{\max}$ and $\mathbf{x}_{\min}$ are continuous, (16) holds if τ is replaced by a rational number τ_m sufficiently close to τ. Now (16) implies (13) for sufficiently large p. Since $(t, \mathbf{x}) \in D$, it also belongs to $\bar{D}_s$ for sufficiently large s. Finally, (12) is satisfied for sufficiently large r. Hence $\mathbf{f} \in F$. This shows that U is the complement of F.

Thus F is the set of those $\mathbf{f} \in C(D)$ for which (1) – (2) does not *always* have a unique solution. Since U is the complement of F, the theorem statement is that F is meager. Note that F is a countable union of the sets F_{mprs}. The proof of the theorem consists of showing that each F_{mprs} is closed and nowhere dense. Assume for a moment that F_{mprs} is closed (this is shown next). Since $C^{(1)}(D)$ is dense in $C(D)$, every open ball in $C(D)$ contains an element of $C^{(1)}(D)$. Note that $C^{(1)}(D) \subseteq U$ by Corollary (2.4.23). Hence every open ball in $C(D)$ contains an element of U, and as a result, no open ball can be a subset of F_{mprs}. In other words, if F_{mprs} is closed, it is nowhere dense.

Thus the proof is completed by showing that F_{mprs} is closed for each m, p, r, s. To show this, suppose $\{\mathbf{f}_i\}$ is a sequence in F_{mprs} converging to $\mathbf{f} \in C(D)$; it is desired to show that $\mathbf{f} \in F_{mprs}$. Since $\mathbf{f}_i \in F_{mprs}$, for each i there exists a $(t_i, \mathbf{x}_i) \in \bar{D}_s$ such that

17 $\mathbf{x}_{\max}(\mathbf{f}_i; t_i+\tau_m, t_i, \mathbf{x}_i) - \mathbf{x}_{\min}(\mathbf{f}_i; t_i+\tau_m, t_i, \mathbf{x}_i) \geq 1/p.$

Since $\bar{D}_s$ is compact, the sequence $\{(t_i, \mathbf{x}_i)\}$ contains a convergent subsequence. Renumber this subsequence once again as $\{(t_i, \mathbf{x}_i)\}$, and let $(t_0, \mathbf{x}_0) \in \bar{D}_s$ denote its limit. Now $\mathbf{x}_{\max}$ satisfies the following integral equation which is equivalent to (1) – (2):

18 $\mathbf{x}_{\max}(\mathbf{f}_i; t_i+\tau, t_i, \mathbf{x}_i) = \mathbf{x}_i + \int_0^\tau \mathbf{f}_i[t_i+\lambda, \mathbf{x}_{\max}(\mathbf{f}_i; t_i+\lambda, t_i, \mathbf{x}_i)]\, d\lambda.$

$\mathbf{x}_{\min}$ satisfies an analogous equation. Now, for each fixed $\tau \in (-1/sr, 1/sr)$, the sequence $\{\mathbf{x}_{\max}(\mathbf{f}_i; t_i+\tau, t_i, \mathbf{x}_i)\}$ is bounded, and therefore contains a convergent subsequence. Renumber this subsequence again with the same index i, and define

19 $\mathbf{y}(t_0+\tau)=\lim_{i\to\infty}\mathbf{x}_{\max}(\mathbf{f}_i;t_i+\tau,t_i,\mathbf{x}_i).$

Taking limits in (18) shows that

20 $\mathbf{y}(t_0+\tau)=\mathbf{x}_0+\int_0^\tau \mathbf{f}[t_0+\lambda,\mathbf{y}(t_0+\lambda)]\,d\lambda.$

In other words, $\mathbf{y}(\cdot)$ is a solution of (1) – (2). Similarly, define

21 $\mathbf{z}(t_0+\tau)=\lim_{i\to\infty}\mathbf{x}_{\min}(\mathbf{f}_i;t_i+\tau,t_i,\mathbf{x}_i),$

after finding a convergent subsequence if necessary. Then $\mathbf{z}(\cdot)$ is also a solution of (1) – (2). Moreover, taking limits in (17) shows that

22 $\mathbf{y}(t_0+\tau_m)-\mathbf{z}(t_0+\tau_m)\geq 1/p.$

Since $\mathbf{y}(\cdot)$ and $\mathbf{z}(\cdot)$ are both solutions of (1) – (2), it follows from (7) that

23 $\mathbf{x}_{\min}(\mathbf{f};t_0+\tau_m,t_0,\mathbf{x}_0)\leq\mathbf{z}(t_0+\tau_m)$, and

$\mathbf{x}_{\min}(\mathbf{f};t_0+\tau_m,t_0,\mathbf{x}_0)\geq\mathbf{y}(t_0+\tau_m).$

Combining (22) and (23) gives

24 $\mathbf{x}_{\max}(\mathbf{f};t_0+\tau_m,t_0,\mathbf{x}_0)-\mathbf{x}_{\min}(\mathbf{f};t_0+\tau_m,t_0,\mathbf{x}_0)\geq 1/p.$

Hence (13) is satisfied with $(t,\mathbf{x})=(t_0,\mathbf{x}_0)$. This shows that $\mathbf{f}\in F_{mprs}$. Thus F_{mprs} is closed. ■

Finally, note that throughout Chapter 5, it is assumed that the differential equation under study has a unique solution corresponding to each initial condition. Theorem (10) shows that this is quite a reasonable assumption, since it is "almost always" satisfied.

B. PROOF OF THE KALMAN-YACUBOVITCH LEMMA

The objective of this appendix is to provide a proof of the Kalman-Yacubovitch lemma [Theorem (5.6.13)] in full generality.

1 **Theorem** *Consider the system*

2 $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$,

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

where $\mathbf{x}(t) \in \mathbf{R}^n$, *and* $\mathbf{y}(t)$, $\mathbf{u}(t) \in \mathbf{R}^m$ *with* $m < n$. *Suppose (i) the matrix* $\mathbf{A}$ *is Hurwitz; (ii) the pair* $(\mathbf{A}, \mathbf{B})$ *is controllable; and (ii) the pair* $(\mathbf{C}, \mathbf{A})$ *is observable. Define*

3 $\mathbf{H}(s) = \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$.

Then the following two statements are equivalent:

(A) There exist matrices $\mathbf{P} \in \mathbf{R}^{n \times n}$, $\mathbf{Q} \in \mathbf{R}^{m \times n}$, *and* $\mathbf{W} \in \mathbf{R}^{m \times m}$, *such that*

4 $\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}'\mathbf{Q}$,

5 $\mathbf{B}'\mathbf{P} + \mathbf{W}'\mathbf{Q} = \mathbf{C}$,

6 $\mathbf{W}'\mathbf{W} = \mathbf{D} + \mathbf{D}'$,

and in addition, the following conditions are satisfied: (i) $\mathbf{P}$ *is symmetric and positive definite; (ii) the pair* $(\mathbf{Q}, \mathbf{A})$ *is observable; and (iii) if we define*

7 $\mathbf{T}(s) = \mathbf{Q}(sI - \mathbf{A})^{-1}\mathbf{B} + \mathbf{W}$,

then

8 $\operatorname{rank} \mathbf{T}(j\omega) = m$, $\forall \omega$.

(B) The transfer matrix $\mathbf{H}(\cdot)$ *satisfies*

9 $\mathbf{H}(j\omega) + \mathbf{H}^*(j\omega) > 0$, $\forall \omega$,

where * *denotes the conjugate transpose, and* "> 0" *means that the matrix is positive definite.*

Remarks Actually, the theorem is almost always used in the direction "B implies A." The statement "A implies B" is more of academic interest, to show that the hypotheses in (B) are in some sense the minimum required.

The proof of the theorem requires a couple of standard results from linear system theory. Recall that a quadruplet $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ is said to be a **realization** of a proper rational matrix $\mathbf{H}(\cdot)$ if (3) holds; the realization is said to be **minimal** if, in addition, the pair $(\mathbf{A}, \mathbf{B})$ is controllable and the pair $(\mathbf{C}, \mathbf{A})$ is observable.

10 Lemma *Suppose* $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ *and* $(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \bar{\mathbf{C}}, \bar{\mathbf{D}})$ *are both minimal realizations of a proper rational matrix* $\mathbf{H}(\cdot)$. *Then there exists a nonsingular matrix* $\mathbf{M} \in \mathbf{R}^{n \times n}$ *such that*

11 $\mathbf{A} = \mathbf{M}^{-1}\bar{\mathbf{A}}\mathbf{M}, \mathbf{B} = \mathbf{M}^{-1}\bar{\mathbf{B}}, \mathbf{C} = \bar{\mathbf{C}}\mathbf{M}.$

For a proof, see Kailath (1980), p. 364.

12 Lemma *Suppose* $\mathbf{V}(\cdot)$ *is a proper rational matrix of dimensions* $m \times m$, *satisfying two additional conditions:*

13 $\mathbf{V}(s) = \mathbf{V}'(-s)$, and

14 $\mathbf{V}(j\omega) > 0 \ \forall \omega.$

Then there exists a proper stable rational matrix $\mathbf{T}(\cdot)$ *of dimensions* $m \times m$ *such that*

15 $\mathbf{V}(s) = \mathbf{T}'(-s)\,\mathbf{T}(s),$

and in addition,

16 $\operatorname{rank} \mathbf{T}(j\omega) = m, \ \forall \omega.$

For a proof, see Anderson and Moore (1979), p. 240.

Proof of Theorem (1) "(B) $\Rightarrow$ (A)" Define

17 $\mathbf{V}(s) = \mathbf{H}(s) + \mathbf{H}'(-s).$

Then $\mathbf{V}(\cdot)$ satisfies (13). Moreover, since $\mathbf{H}(\cdot)$ satisfies (9), it follows that $\mathbf{V}(\cdot)$ satisfies (14). Thus, by Lemma (12), there exists a proper *stable* rational matrix $T(\cdot)$ such that (15) and (16) hold. Let $(\mathbf{F}, \mathbf{G}, \mathbf{K}, \mathbf{L})$ be a minimal realization of $\mathbf{T}(\cdot)$. Thus

18 $\mathbf{T}(s) = \mathbf{K}(sI - \mathbf{F})^{-1}\mathbf{G} + \mathbf{L},$

and in addition, the pairs $(\mathbf{F}, \mathbf{G})$ and $(\mathbf{K}, \mathbf{L})$ are respectively controllable and observable. Now

$$\mathbf{19}\quad \mathbf{T}'(-s)\mathbf{T}(s) = [\mathbf{L}' - \mathbf{G}'(sI + \mathbf{F}')^{-1}\mathbf{K}'][\mathbf{L} + \mathbf{K}(sI - \mathbf{F})^{-1}\mathbf{G}]$$

$$= \mathbf{L}'\mathbf{L} - \mathbf{G}'(sI + \mathbf{F}')^{-1}\mathbf{K}'\mathbf{L} + \mathbf{L}'\mathbf{K}(sI - \mathbf{F})^{-1}\mathbf{G}$$

$$- \mathbf{G}'(sI + \mathbf{F}')^{-1}\mathbf{K}'\mathbf{K}(sI - \mathbf{F})^{-1}\mathbf{G}.$$

The last term can be simplified further. Since the pair $(\mathbf{K}, \mathbf{F})$ is observable, there exists a symmetric $n \times n$ matrix $\mathbf{R} > 0$ such that

$$\mathbf{20}\quad \mathbf{F}'\mathbf{R} + \mathbf{R}\mathbf{F} = -\mathbf{K}'\mathbf{K}.$$

Now observe that

$$\mathbf{21}\quad -\mathbf{K}'\mathbf{K} = \mathbf{F}'\mathbf{R} + \mathbf{R}\mathbf{F} = (sI + \mathbf{F}')\mathbf{R} - \mathbf{R}(sI - \mathbf{F}).$$

Substituting from (21) into (19) shows that the last term equals

$$\mathbf{22}\quad \mathbf{G}'(sI + \mathbf{F}')^{-1}[(sI + \mathbf{F}')\mathbf{R} - \mathbf{R}(sI - \mathbf{F})](sI - \mathbf{F})^{-1}\mathbf{G}$$

$$= \mathbf{G}'\mathbf{R}(sI - \mathbf{F})^{-1}\mathbf{G} - \mathbf{G}'(sI + \mathbf{F}')^{-1}\mathbf{R}\mathbf{G}.$$

Substituting from (22) into (19) gives

$$\mathbf{23}\quad \mathbf{T}'(-s)\mathbf{T}(s) = \mathbf{L}'\mathbf{L} + (\mathbf{G}'\mathbf{R} + \mathbf{L}'\mathbf{K})(sI - \mathbf{F})^{-1}\mathbf{G} - \mathbf{G}'(sI + \mathbf{F}')^{-1}(\mathbf{R}\mathbf{G} + \mathbf{K}'\mathbf{L}).$$

By (15), we have

$$\mathbf{24}\quad \mathbf{T}'(-s)\mathbf{T}(s) = \mathbf{H}(s) + \mathbf{H}'(-s)$$

$$= \mathbf{D} + \mathbf{D}' + \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{B} - \mathbf{B}'(sI + \mathbf{A}')^{-1}\mathbf{C}.$$

Since $\mathbf{A}$ and $\mathbf{F}$ are both Hurwitz matrices, it is possible to equate the stable and the completely unstable parts in Equations (23) and (24). This gives

$$\mathbf{25}\quad (\mathbf{G}'\mathbf{R} + \mathbf{L}'\mathbf{K})(sI - \mathbf{F})^{-1}\mathbf{G} = \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{B}.$$

By the minimality of the two realizations and Lemma (10), it follows that there exists a non-singular matrix $\mathbf{M} \in \mathbf{R}^{n \times n}$ such that

$$\mathbf{26}\quad \mathbf{C} = (\mathbf{G}'\mathbf{R} + \mathbf{L}'\mathbf{K})\mathbf{M},\ \mathbf{B} = \mathbf{M}^{-1}\mathbf{G},\ \mathbf{A} = \mathbf{M}^{-1}\mathbf{F}\mathbf{M}.$$

Define

$$\mathbf{27}\quad \mathbf{P} = \mathbf{M}'\mathbf{R}\mathbf{M},\ \mathbf{W} = \mathbf{L},\ \mathbf{Q} = \mathbf{K}\mathbf{M}.$$

It is claimed that (4) to (6) are satisfied. To prove (4), pre-multiply both sides of (20) by $\mathbf{M}'$ and post-multiply by $\mathbf{M}$. This gives, after a little manipulation,

28 $\mathbf{M}'\mathbf{F}'(\mathbf{M}^{-1})'\mathbf{M}'\mathbf{R}\mathbf{M}+\mathbf{M}'\mathbf{R}\mathbf{M}\mathbf{M}^{-1}\mathbf{F}\mathbf{M}=-\mathbf{M}'\mathbf{K}'\mathbf{K}\mathbf{M},$

which is (4). Next, from (26) we get

29 $\mathbf{C}=(\mathbf{G}'\mathbf{R}+\mathbf{L}'\mathbf{K})\,\mathbf{M}=\mathbf{B}'\mathbf{M}'\mathbf{R}\mathbf{M}+\mathbf{L}'\mathbf{K}\mathbf{M}=\mathbf{B}'\mathbf{P}+\mathbf{W}'\mathbf{Q},$

which is (5). Finally, equating the constant terms in (23) and (24) shows that

30 $\mathbf{L}'\mathbf{L}=\mathbf{D}+\mathbf{D}',$

which is (6) since $\mathbf{W}=\mathbf{L}$.

"(A) $\Rightarrow$ (B)" The idea is to show that if (4) to (6) hold and $\mathbf{T}(\cdot)$ is defined by (7), then

31 $\mathbf{H}(s)+\mathbf{H}'(-s)=\mathbf{T}'(-s)\,\mathbf{T}(s).$

If (31) is true, then (8) implies (9). Thus the proof is complete once (31) is established. From (3), we have

32 $$\begin{aligned}\mathbf{H}(s)+\mathbf{H}'(-s)&=\mathbf{D}+\mathbf{D}'+\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}-\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\mathbf{C}'\\&=\mathbf{W}'\mathbf{W}+(\mathbf{B}'\mathbf{P}+\mathbf{W}'\mathbf{Q})\,(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}-\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\,(\mathbf{P}\mathbf{B}+\mathbf{Q}'\mathbf{W}),\end{aligned}$$

after using (5) and (6). On the other hand, from (7),

33 $\mathbf{T}'(-s)\mathbf{T}(s)=\mathbf{W}'\mathbf{W}+\mathbf{W}'\mathbf{Q}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}-\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\mathbf{Q}'\mathbf{W}-\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\mathbf{Q}\mathbf{Q}'(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}.$

Comparing (32) and (33) shows that (31) is proved if it can be established that

34 $-\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\,\mathbf{Q}'\mathbf{Q}\,(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}=\mathbf{B}'\mathbf{P}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}-\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\mathbf{P}\mathbf{B}.$

For this purpose, rewrite the right side of (34) as

35 $$\begin{aligned}&\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\,[(s\mathbf{I}+\mathbf{A}')\,\mathbf{P}-\mathbf{P}\,(s\mathbf{I}-\mathbf{A})]\,(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\\&\quad=\mathbf{B}'(s\mathbf{I}+\mathbf{A}')^{-1}\,(\mathbf{A}'\mathbf{P}+\mathbf{P}\mathbf{A})\,(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}.\end{aligned}$$

Finally, (4) implies (34). ■

36 **Corollary** *Let* $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{H}(\cdot)$ *be as in Theorem (1). Suppose*

37 $\inf_{\omega\in\mathbb{R}} \lambda_{\min}[\mathbf{H}(j\omega)+\mathbf{H}^*(j\omega)]>0.$

Under these conditions, there exist a symmetric positive definite matrix $\mathbf{P}\in\mathbf{R}^{n\times n}$, *matrices* $\mathbf{Q}\in\mathbf{R}^{m\times n}$, $\mathbf{W}\in\mathbf{R}^{m}$, *and an* $\varepsilon>0$ *such that*

38 $\mathbf{A}'\mathbf{P}+\mathbf{P}\mathbf{A}=-\varepsilon\mathbf{P}-\mathbf{Q}'\mathbf{Q}$,

39 $\mathbf{B}'\mathbf{P}+\mathbf{W}'\mathbf{Q}=\mathbf{C}$,

40 $\mathbf{W}'\mathbf{W}=\mathbf{D}+\mathbf{D}'$.

Remarks The hypothesis (37) is stronger than (9). For example, the function

41 $h(s)=\dfrac{1}{s+1}$

satisfies (9) but not (37). Correspondingly, the conclusion of Corollary (36) is also stronger, as can be seen by comparing (38) and (4). The right side of (38) is positive *definite*, whereas the right side of (4) is only positive *semidefinite,* since $m < n$. Also, whereas Theorem (1) provides a necessary as well as sufficient condition, the converse of Corollary (36) is false in general.

Proof Define

42 $f(\sigma)=\inf_{\omega\in\mathbb{R}} \lambda_{\min}[\mathbf{H}(\sigma+j\omega)+\mathbf{H}^*(\sigma+j\omega)]$.

Since $\mathbf{A}$ is a Hurwitz matrix, $\mathbf{f}(\sigma)$ is well-defined for all $\sigma \geq 0$, and all sufficiently small $\sigma < 0$. Moreover $f(\cdot)$ is continuous, and (37) states that $f(0) > 0$. Hence for all sufficiently small $\varepsilon > 0$, we have

43 $\inf_{\omega\in\mathbb{R}} \lambda_{\min}[\mathbf{H}(-\varepsilon/2+j\omega)\mathbf{H}^*(-\varepsilon/2+j\omega)] > 0$.

Define

44 $\mathbf{H}_\varepsilon(s)=\mathbf{H}(s-\varepsilon/2)$,

and note that the quadruple $\{\mathbf{A}+(\varepsilon/2)I, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ is a minimal realization of $\mathbf{H}_\varepsilon$. Moreover, $\varepsilon > 0$ can be chosen sufficiently small that $\mathbf{A}+(\varepsilon/2)I$ is also a Hurwitz matrix. Now apply Theorem (1) to the transfer matrix $\mathbf{H}_\varepsilon(\cdot)$, replacing $\mathbf{A}$ by $\mathbf{A}+(\varepsilon/2)I$ throughout. Then (5) and (6) remain unaffected since they do not involve $\mathbf{A}$, and lead to (39) and (40) respectively. Equation (4) is replaced by

45 $[\mathbf{A}+(\varepsilon/2)I]'\mathbf{P}+\mathbf{P}[\mathbf{A}+(\varepsilon/2)I]=-\mathbf{Q}'\mathbf{Q}$,

or equivalently,

46 $\mathbf{A}'\mathbf{P}+\mathbf{P}\mathbf{A}=-\varepsilon\mathbf{P}-\mathbf{Q}'\mathbf{Q}$,

which is (38). ■

C. PROOF OF THE FROBENIUS THEOREM

In this appendix, a proof is given of the Frobenius theorem, which is restated here for convenience. All symbols are as in Chapter 7.

1 **Theorem** *Suppose X is an open connected subset of $\mathbf{R}^n$ containing $\mathbf{0}$, and suppose $\mathbf{f}_1, \cdots, \mathbf{f}_m \in V(X)$, where $m < n$, are linearly independent at all $\mathbf{x} \in X$. Suppose there exist smooth functions $\alpha_{ijk} \in S(X)$, $1 \leq i, j, k \leq m$, such that*

$$2 \qquad [\mathbf{f}_i, \mathbf{f}_j](\mathbf{x}) = \sum_{k=1}^{m} \alpha_{ijk}(\mathbf{x})\, \mathbf{f}_k(\mathbf{x}), \ \forall \mathbf{x} \in X.$$

Then, for each $\mathbf{x}_0 \in X$, there exist an open connected neighborhood $U \subseteq X$ of $\mathbf{x}_0$ and smooth functions $\phi_{m+1}, \cdots, \phi_n \in S(X)$ such that $\mathbf{d}\phi_{m+1}(\mathbf{x}), \cdots, \mathbf{d}\phi_n(\mathbf{x})$ are linearly independent for each $\mathbf{x} \in X$, and

$$3 \qquad <\mathbf{d}\phi_i, \mathbf{f}_j>(\mathbf{x}) = 0 \ \forall \mathbf{x} \in U, \text{ for } j = 1, \cdots, m; i = m+1, \cdots, n.$$

The proof makes use of the following lemma, which is of independent interest.

4 **Lemma** *Suppose $\mathbf{f} \in V(X)$, $\mathbf{x}_0 \in X$, and $\mathbf{f}(\mathbf{x}_0) \neq \mathbf{0}$. Then there exist a neighborhood $U \subseteq X$ of $\mathbf{x}_0$ and a diffeomorphism $T: U \to X$ such that*

$$5 \qquad \mathbf{f}_T(\mathbf{y}) = [1 \ 0 \ \cdots 0]', \ \forall \mathbf{y} \in T(U).$$

Remarks The lemma states that it is always possible to make a local change of coordinates such that *any one* given nonvanishing vector field looks like $[1 \ 0 \cdots 0]'$ in the new coordinates. See (7.1.8) for the definition of the transformed vector field $\mathbf{f}_T$.

Proof of the Lemma For notational convenience only, suppose $\mathbf{x}_0 = \mathbf{0}$; the case where $\mathbf{x}_0 \neq \mathbf{0}$ requires only more elaborate notation, and the required changes are easy to make.

Let $\mathbf{s}_{\mathbf{f},t}$ denote the integral curve of the vector field $\mathbf{f}$, as defined in (7.1.6) *et seq.* Thus $\mathbf{s}_{\mathbf{f},t}(\mathbf{x}_0)$ denotes the solution of the differential equation

$$6 \qquad \dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t)], \ \mathbf{x}(0) = \mathbf{x}_0$$

evaluated at time t.

Select a nonsingular $n \times n$ matrix $\mathbf{M}$ such that its first column equals $\mathbf{f}(\mathbf{0})$; this is possible since $\mathbf{f}(\mathbf{0}) \neq \mathbf{0}$. Given $\mathbf{x} \in X$, define $\bar{\mathbf{x}} \in \mathbf{R}^{n-1}$ by

$$\mathbf{7} \qquad \bar{\mathbf{x}} = [x_2 \cdots x_n]' \in \mathbf{R}^{n-1}.$$

Next, given a vector $\mathbf{x} \in X$, define the n-vector $\mathbf{q}(\mathbf{x})$ by

$$\mathbf{8} \qquad \mathbf{q}(\mathbf{x}) = \mathbf{s}_{\mathbf{f}, x_1}(\mathbf{M}[0 \ \ \bar{\mathbf{x}}']').$$

Equation (8) says the following: Given $\mathbf{x} \in X$, form the vector $[0 \ \ \bar{\mathbf{x}}']' \in \mathbf{R}^{n-1}$. Evaluate the integral curve of the vector field $\mathbf{f}$, passing through the n-vector $\mathbf{M}[0 \ \ \bar{\mathbf{x}}']'$ and at "time" x_1. Call the resulting vector $\mathbf{q}(\mathbf{x})$. It is easy to see that $\mathbf{q}(\mathbf{x})$ is well-defined whenever $\|\mathbf{x}\|$ is sufficiently small, since $\mathbf{M}[0 \ \ \bar{\mathbf{x}}']' \in X$, and (6) has a unique solution up to "time" x_1. Moreover, $\mathbf{q}(\mathbf{x}) \in X$.

Next, it is shown that $\mathbf{q}$ is a local diffeomorphism at $\mathbf{0}$. Towards this end, let us compute the Jacobian matrix of $\mathbf{q}$ at $\mathbf{0}$. For this purpose, two observations are made:

$$\mathbf{9} \qquad \left[\frac{\partial}{\partial t} \mathbf{s}_{\mathbf{f},t}(\mathbf{x}) \right]_{(t,\mathbf{x})=(0,\mathbf{0})} = \mathbf{f}(\mathbf{0}),$$

$$\mathbf{10} \qquad \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},t}(\mathbf{x}) \right]_{(t,\mathbf{x})=(0,\mathbf{0})} = I.$$

The relationship (9) follows readily from the fact that $\mathbf{s}_{\mathbf{f},t}$ satisfies the integral curve relationship

$$\mathbf{11} \qquad \frac{\partial}{\partial t} \mathbf{s}_{\mathbf{f},t}(\mathbf{x}) = \mathbf{f}[\mathbf{s}_{\mathbf{f},t}(\mathbf{x})], \ \mathbf{s}_{\mathbf{f},0}(\mathbf{x}) = \mathbf{x}.$$

To establish (10), note that

$$\mathbf{12} \qquad \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},t}(\mathbf{x}) \right]_{(t,\mathbf{x})=(0,\mathbf{0})} = \left[\frac{\partial}{\partial \mathbf{x}} \mathbf{s}_{\mathbf{f},0}(\mathbf{x}) \right]_{\mathbf{x}=\mathbf{0}} = I.$$

Now it is possible to evaluate the Jacobian of the map $\mathbf{q}$. First, by (9),

$$\mathbf{13} \qquad \left[\frac{\partial \mathbf{q}}{\partial x_1} \right]_{\mathbf{x}=\mathbf{0}} = \mathbf{f}(\mathbf{M0}) = \mathbf{f}(\mathbf{0}).$$

Next, combining (10) with the chain rule gives

$$\mathbf{14}\quad \left[\frac{\partial \mathbf{q}}{\partial \bar{\mathbf{x}}}\right]_{\mathbf{x}=\mathbf{0}} = I \cdot \mathbf{M}_2 = \mathbf{M}_2,$$

where $\mathbf{M}_2$ is the $n \times (n-1)$ matrix consisting of the last $(n-1)$ columns of $\mathbf{M}$. Thus

$$\mathbf{15}\quad \left[\frac{\partial \mathbf{q}}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{0}} = \left[\frac{\partial \mathbf{q}}{\partial x_1} \quad \frac{\partial \mathbf{q}}{\partial \bar{\mathbf{x}}}\right]_{\mathbf{x}=\mathbf{0}} = [\mathbf{f}(\mathbf{0}) \ \mathbf{M}_2] = \mathbf{M},$$

which is nonsingular by construction. Hence it is established that $\mathbf{q}$ is a local diffeomorphism at $\mathbf{0}$. This means that, given any $\mathbf{y} \in X$ sufficiently close to $\mathbf{0}$, there exists a unique vector $\bar{\mathbf{z}} \in \mathbf{R}^{n-1}$ and a "time" τ such that

$$\mathbf{16}\quad \mathbf{y} = \mathbf{s}_{\mathbf{f},\tau}(\mathbf{M}[0 \ \bar{\mathbf{z}}']').$$

Next, define $T = \mathbf{q}^{-1}$; then $T(\mathbf{0}) = \mathbf{0}$, and T is also a local diffeomorphism at $\mathbf{0}$. Now define $\mathbf{g} \in V(X)$ to be the transformed vector field $\mathbf{f}_T$, i.e., the vector field $\mathbf{f}$ transformed by the coordinate change T. There is an easy way to compute $\mathbf{g}$. From (7.1.10), the integral curves of the vector fields $\mathbf{g}$ and $\mathbf{f}$ satisfy the relationship

$$\mathbf{17}\quad \mathbf{s}_{\mathbf{g},t} = T \cdot \mathbf{s}_{\mathbf{f},t} \cdot T^{-1}.$$

Suppose $\mathbf{x} \in X$ is sufficiently close to $\mathbf{0}$. Then

$$\mathbf{18}\quad T^{-1}(\mathbf{x}) = \mathbf{q}(\mathbf{x}) = \mathbf{s}_{\mathbf{f},x_1}(\mathbf{M}[0 \ \bar{\mathbf{x}}']'),$$

$$\mathbf{19}\quad \mathbf{s}_{\mathbf{f},t}[T^{-1}(\mathbf{x})] = \mathbf{s}_{\mathbf{f},t}[\mathbf{s}_{\mathbf{f},x_1}(\mathbf{M}[0 \ \bar{\mathbf{x}}']')] = \mathbf{s}_{\mathbf{f},t+x_1}(\mathbf{M}[0 \ \bar{\mathbf{x}}']') = \mathbf{y}, \text{ say.}$$

Now $T(\mathbf{y}) = \mathbf{q}^{-1}(\mathbf{y})$ consists of the unique "time" τ and vector $\bar{\mathbf{z}} \in \mathbf{R}^{n-1}$ such that (16) holds. Comparing (19) and (16) shows that

$$\mathbf{20}\quad T(\mathbf{y}) = \begin{bmatrix} x_1 + t \\ \bar{\mathbf{x}} \end{bmatrix}.$$

In other words,

$$\mathbf{21}\quad \mathbf{s}_{\mathbf{g},t}(\mathbf{x}) = \begin{bmatrix} x_1 + t \\ \bar{\mathbf{x}} \end{bmatrix}.$$

Now simple differentiation shows that

$$\mathbf{22}\quad \mathbf{g}(\mathbf{x}) = [1 \ 0 \cdots 0]'.$$

Hence T is the required diffeomorphism. ■

Proof of the Theorem The proof is by induction on the integer m. If $m = 1$, then there is a solitary vector field $\mathbf{f}$, and "linear independence" means that $\mathbf{f}(\mathbf{x}) \neq \mathbf{0}\ \forall \mathbf{x} \in X$. Now Lemma (4) shows that there exists a diffeomorphism T such that (5) holds. Thus one can choose the functions

23 $\phi_i(\mathbf{x}) = T_i(\mathbf{x}),\ i = 2, \cdots, n.$

In the transformed coordinates $\mathbf{y} = T(\mathbf{x})$, we have

24 $\mathbf{f}_T(\mathbf{y}) = [1\ 0 \cdots 0]'\ \forall \mathbf{y},$

25 $\phi_{iT}(\mathbf{y}) = y_i,\ i = 2, \cdots, n.$

It is clear that (3) holds. Hence the theorem is true if $m = 1$.

Now suppose by way of induction that the theorem is true up to $m-1$ vector fields. Once again, for notational simplicity only, suppose $\mathbf{x}_0 = \mathbf{0}$. Given the m vector fields $\mathbf{f}_1, \cdots, \mathbf{f}_m$ satisfying the hypotheses, as a first step select a local diffeomorphism T such that $\mathbf{f}_{1T}$ has the form

26 $\mathbf{f}_{1T}(\mathbf{y}) = [1\ 0 \cdots 0]',\ \forall \mathbf{y},$

where "$\forall \mathbf{y}$" really means "for all $\mathbf{y}$ sufficiently close to $\mathbf{0}$." For simplicity, define

27 $\mathbf{g}_i(\mathbf{y}) = \mathbf{f}_{iT}(\mathbf{y}),\ i = 1, \cdots, m,$

and note that

28 $\mathbf{g}_1(\mathbf{y}) = [1\ 0 \cdots 0]',\ \forall \mathbf{y}.$

Since Lie brackets are preserved under coordinate transformations [see (7.1.58)], it follows that the set of vector fields $\{\mathbf{g}_1, \cdots, \mathbf{g}_m\}$ is also involutive. Thus there exist smooth functions $\beta_{ijk} \in S(X)$ such that

29 $[\mathbf{g}_i, \mathbf{g}_j](\mathbf{y}) = \sum_{k=1}^{m} \beta_{ijk}(\mathbf{y})\, \mathbf{g}_k(\mathbf{y}),\ \forall \mathbf{y}.$

Next, define the vector fields

30 $\mathbf{h}_i(\mathbf{y}) = \mathbf{g}_i(\mathbf{y}) - g_{i1}(\mathbf{y})\, \mathbf{g}_1(\mathbf{y}),$ for $i = 2, \cdots, m,$

where $g_{i1}(\mathbf{y})$ is the first component of $\mathbf{g}_i(\mathbf{y})$. Then $h_{i1}(\mathbf{y}) = 0$ for all $\mathbf{y}$. Since the set $\{\mathbf{g}_1, \cdots, \mathbf{g}_m\}$ is linearly independent, an easy calculation shows that so is the set $\{\mathbf{g}_1, \mathbf{h}_2, \cdots, \mathbf{h}_m\}$. Now let $\bar{\mathbf{h}}_i(\mathbf{y})$ denote the vector consisting of the *last* $(n-1)$ components of the vector $\mathbf{h}_i(\mathbf{y})$, for $i = 2, \cdots, m$. Since $h_{i1}(\mathbf{y}) = 0$ for all i, the linear independence of the set $\{\mathbf{h}_2, \cdots, \mathbf{h}_m\}$ implies the linear independence of the set $\{\bar{\mathbf{h}}_2, \cdots, \bar{\mathbf{h}}_m\}$ (over $\mathbf{R}^{n-1}$).

The next step is to show that the set $\{\bar{\mathbf{h}}_2, \cdots, \bar{\mathbf{h}}_m\}$ is involutive over some neighborhood of $\mathbf{0}$ (in $\mathbf{R}^{n-1}$). For this purpose, observe first that, since $\mathbf{h}_2, \cdots, \mathbf{h}_m$ are just "linear" combinations of $\mathbf{g}_1, \cdots, \mathbf{g}_m$, it follows from (29) that there exist smooth functions $a_{ijk} \in S(X)$ such that

$$\mathbf{31} \quad [\mathbf{h}_i, \mathbf{h}_j](\mathbf{y}) = a_{ij1}(\mathbf{y})\,\mathbf{g}_1(\mathbf{y}) + \sum_{k=2}^{m} a_{ijk}(\mathbf{y})\,\mathbf{h}_k(\mathbf{y}), \ \forall \mathbf{y}.$$

Since $h_{i1}(\mathbf{y}) \equiv 0$ for all i, a simple calculation shows that the first component of $[\mathbf{h}_i, \mathbf{h}_j](\mathbf{y})$ is also identically zero, for all i, j. Now $\mathbf{g}_1(\mathbf{y})$ has the form (28), while $h_{k1}(\mathbf{y}) = 0$ for all k. Hence it follows that $a_{ij1}(\mathbf{y}) \equiv 0$, i.e.,

$$\mathbf{32} \quad [\mathbf{h}_i, \mathbf{h}_j](\mathbf{y}) = \sum_{k=2}^{m} a_{ijk}(\mathbf{y})\,\mathbf{h}_k(\mathbf{y}), \ \forall \mathbf{y}.$$

Now look at the "slice" of X defined by $y_1 = 0$, and denote it by $\bar{X}$. This is an open connected subset of $\mathbf{R}^{n-1}$, and it contains $\mathbf{0}_{n-1}$, the origin in $\mathbf{R}^{n-1}$. Define

$$\mathbf{33} \quad \bar{\mathbf{y}} = [y_2 \cdots y_n]' \in \mathbf{R}^{n-1},$$

and substitute $y_1 = 0$ in (32). Define

$$\mathbf{34} \quad \bar{a}_{ijk}(\bar{\mathbf{y}}) = a_{ijk}[(0, \bar{\mathbf{y}})],$$

as an element of $S(\bar{X})$. {Note that the symbol $(0, \bar{\mathbf{y}})$ is used instead of the more correct $[0 \ \bar{\mathbf{y}}']'$ in the interests of simplicity.} Observe (after an easy computation) that the last $(n-1)$ rows of $[\mathbf{h}_i, \mathbf{h}_j]$, evaluated at $y = (0, \bar{\mathbf{y}})$, equal $[\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j](0, \bar{\mathbf{y}})$. Thus (32) leads one to conclude that

$$\mathbf{35} \quad [\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j](0, \bar{\mathbf{y}}) = \sum_{k=2}^{m} \bar{a}_{ijk}(\bar{\mathbf{y}})\,\bar{\mathbf{h}}_k(0, \bar{\mathbf{y}}), \ \forall \bar{\mathbf{y}}.$$

Equation (35) shows that the set $\{\bar{\mathbf{h}}_2, \cdots, \bar{\mathbf{h}}_m\}$ is involutive over some neighborhood of $\mathbf{0}_{n-1}$. Hence, by the inductive hypothesis, there exist smooth functions $\bar{\phi}_{m+1}(\bar{\mathbf{y}}), \cdots, \bar{\phi}_n(\bar{\mathbf{y}}) \in S(\bar{X})$ and a neighborhood $\bar{N} \subseteq \mathbf{R}^{n-1}$ of $\mathbf{0}_{n-1}$ such that $\mathbf{d}\bar{\phi}_{m+1}(\bar{\mathbf{y}}), \cdots, \mathbf{d}\bar{\phi}_n(\bar{\mathbf{y}})$ are linearly independent at each $\bar{\mathbf{y}} \in \bar{N}$, and

$$\mathbf{36} \quad \frac{\partial \bar{\phi}_i(\bar{\mathbf{y}})}{\partial \bar{\mathbf{y}}}\,\bar{\mathbf{h}}_i(0, \bar{\mathbf{y}}) = 0, \ \forall \bar{\mathbf{y}} \in \bar{N}, \text{ for } j = 2, \cdots, m; i = m+1, \cdots, n.$$

Now define functions $\phi_{m+1}, \cdots, \phi_n \in S(X)$ by

$$\mathbf{37} \quad \phi_i(\mathbf{y}) = \bar{\phi}_i(\bar{\mathbf{y}}), \text{ for } i = m+1, \cdots, n.$$

In other words, $\phi_i(\mathbf{y})$ is actually independent of y_1. Define $N \subseteq X$ by

38 $N = \{\mathbf{y} \in X : \bar{\mathbf{y}} \in \bar{N}\}$.

It is claimed that

39 $<\mathbf{d}\phi_i, \mathbf{g}_j>(\mathbf{y}) = 0 \; \forall \mathbf{y} \in \bar{N}$, for $j = 1, \cdots, m$; $i = m+1, \cdots, n$.

To prove (39), it is enough to establish that

40 $<\mathbf{d}\phi_i, \mathbf{g}_1>(\mathbf{y}) = 0 \; \forall \mathbf{y} \in N$, for $i = m+1, \cdots, n$,

$<\mathbf{d}\phi_i, \mathbf{h}_j>(\mathbf{y}) = 0 \; \forall \mathbf{y} \in N$, for $j = 2, \cdots, m$; $i = m+1, \cdots, n$.

The equivalence of (39) and (40) is a ready consequence of the fact that, for each fixed $\mathbf{y}$, the sets of vectors $\{\mathbf{g}_1(\mathbf{y}), \cdots, \mathbf{g}_m(\mathbf{y})\}$ and $\{\mathbf{g}_1(\mathbf{y}), \mathbf{h}_2(\mathbf{y}), \cdots, \mathbf{h}_m(\mathbf{y})\}$ span exactly the same subspace of $\mathbf{R}^n$.

Thus the theorem is proved once (40) is established. The first equation in (40) is immediate. Since $\phi_i(\mathbf{y})$ is independent of y_1, $\mathbf{d}\phi_i(\mathbf{y})$ has the form

41 $\mathbf{d}\phi_i(\mathbf{y}) = [0 \;\; \partial\bar{\phi}_i/\partial\bar{\mathbf{y}}]$,

while $\mathbf{g}_1(\mathbf{y})$ has the form (28). So the inner product of these vectors is zero. To prove the second equation in (40), recall from Lemma (7.1.59) that

42 $L_{[\mathbf{g}_1, \mathbf{h}_j]}\phi_i = L_{\mathbf{g}_1} L_{\mathbf{h}_j} \phi_i - L_{\mathbf{h}_j} L_{\mathbf{g}_1} \phi_i = L_{\mathbf{g}_1} L_{\mathbf{h}_j} \phi_i$,

where we use the fact that $L_{\mathbf{g}_1}\phi_i = <\mathbf{d}\phi_i, \mathbf{g}_1> \equiv 0$. However, by involutivity, there exist smooth functions $b_{ijk} \in S(X)$ such that

43 $[\mathbf{g}_1, \mathbf{h}_j] = b_{j1}\mathbf{g}_1 + \sum_{k=2}^{m} b_{jk}\mathbf{h}_k$,

where the dependence on $\mathbf{y}$ is suppressed in the interests of clarity. Substituting from (43) into (42), using the distributivity of the Lie derivative, and using the fact that $L_{\mathbf{g}_1}\phi_i \equiv 0$, gives

44 $L_{\mathbf{g}_1} L_{\mathbf{h}_j} \phi_i = L_{[\mathbf{g}_1, \mathbf{h}_j]}\phi_i = \sum_{k=2}^{m} b_{jk} L_{\mathbf{h}_k} \phi_i$.

Define

45 $\psi_{ij} = L_{\mathbf{h}_j}\phi_i = <\mathbf{d}\phi_i, \mathbf{h}_j> \in S(X)$.

Observe that, since $\mathbf{g}_1$ has the form (28),

46 $L_{\mathbf{g}_1}\psi_{ij} = \langle \mathbf{d}\psi_{ij}, \mathbf{g}_1 \rangle = \dfrac{\partial}{\partial y_1}\psi_{ij}.$

Thus (44) becomes

47 $$\frac{\partial}{\partial y_1}\psi(y_1, \bar{\mathbf{y}}) = \sum_{k=2}^{m} b_{jk}(y_1, \bar{\mathbf{y}})\,\psi_{jk}(y_1, \bar{\mathbf{y}}).$$

Now fix the index i and the vector $\bar{\mathbf{y}} \in \bar{X}$. Then (47) represents a linear vector differential equation in the $(n-1)$-dimensional unknown vector $[\psi_{i2} \cdots \psi_{im}]'$, and with y_1 as the independent variable. The "initial condition" is provided by evaluating $\psi_{ij}(y_1, \bar{\mathbf{y}})$ at $y_1 = 0$. But, from (36) and (41), it follows that

48 $$\psi_{ij}(0, \bar{\mathbf{y}}) = \frac{\partial \bar{\phi}_i(\bar{\mathbf{y}})}{\partial \bar{\mathbf{y}}}\,\bar{\mathbf{h}}_j(0, \bar{\mathbf{y}}) = 0.$$

Since the vector differential equation (47) is homogeneous (i.e., there is no forcing function) and has zero initial condition, it follows that

49 $\psi_{ij}(y_1, \bar{\mathbf{y}}) = 0, \ \forall (y_1, \bar{\mathbf{y}}) \in N, \ \forall i, j.$

This completes the proof. ■

References

Aizerman, M. A. (1949), "On a problem concerning stability in the large of dynamical systems," *Uspehi Mat. Nauk.*, **4**, 187-188.

Anderson, B. D. O. (1982), "Internal and external stability of linear time-varying systems," *SIAM J. Control*, **20**, 408-413.

Anderson, B. D. O. and Moore, J. B. (1979), *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N.J.

Araki, M. and Saeki, M. (1983), "A quantitative condition for the well-posedness of interconnected dynamical systems," *IEEE Trans. Auto. Control*, **AC-28**, 569-577.

Arnold, V. I. (1973), *Ordinary Differential Equations*, M.I.T. Press, Cambridge, MA.

Barman, J. F. (1973), *Well-posedness of Feedback Systems and Singular Perturbations*, Ph.D. thesis, Dept. of EECS, Univ. of California, Berkeley.

Barbashin, E. A. and Krasovskii, N. N. (1952), "On the stability of motion in the large," (Russian) *Dokl. Akad. Nauk.*, **86**, 453-456.

Bellman, R. E. (1953), *Stability Theory of Differential Equations*, McGraw-Hill, New York.

Bellman, R. E. (1970), *Introduction to Matrix Analysis*, McGraw-Hill.

Boguliuboff, N. N., and Mitropolsky, Y. A. (1961), *Asymptotic Methods in the Theory of Nonlinear Oscillations*, Gordon and Breach, New York.

Boothby, W. M. (1975), *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York.

Bose, N. K. (1982), *Applied Multidimensional System Theory*, Van Nostrand Reinhold, New York.

Brockett, R. W. (1972), "System theory on group manifolds and coset spaces," *SIAM J. Control*, **10**, 265-284.

Brockett, R. W. (1978), "Feedback invariants for nonlinear systems," *Proc. IFAC World Congress*, Helsinki, 1115-1120.

Byrnes, C. I. and Isidori, A. (1984), "A frequency domain philosophy for nonlinear systems," *Proc. IEEE Conf. on Decision and Control*, Las Vegas, NV, 1569-1573.

Byrnes, C. I. and Isidori, A. (1988) "Local stabilization of minimum-phase nonlinear systems," *Systems and Control Letters*, **11**, 9-17.

Callier, F. M. and Desoer, C. A. (1972), "A graphical test for checking the stability of a linear time-invariant feedback system," *IEEE Trans. Auto. Control,* **AC-17**, 773-780.

Callier, F. M. and Desoer, C. A. (1978), "An algebra of transfer functions of distributed linear time-invariant systems," *IEEE Trans. Circ. and Sys.,* **CAS-25**, 651-662.

Callier, F. M. and Desoer, C. A. (1980) "Simplifications and clarifications on the paper 'An algebra of transfer functions of distributed linear time-invariant systems,' " *IEEE Trans. Circ. and Sys.,* **CAS-27**, 320-323.

Coppel, W. A. (1965), *Stability and Asymptotic Behaviour of Differential Equations,* Heath, Boston, MA.

Chen, C. T. (1986), *Introduction to Linear System Theory,* (Second Edition) Holt, Rinehart and Winston, New York.

Crouch, P. E. (1984), "Spacecraft attitude control and stabilization," *IEEE Trans. Auto. Control,* **AC-29**, 321-331.

Dahlquist, G. (1959), "Stability and error bounds in the numerical integrations of ordinary differential equations," *Trans. Roy. Inst. Tech. (Sweden),* **130**.

Desoer, C. A. (1965), "A generalization of the Popov criterion," *IEEE Trans. Auto. Control,* **AC-10**, 182-184.

Desoer, C. A. (1969), "Slowly varying system $\dot{x} = A(t)x$," *IEEE Trans. Auto. Control,"* **AC-14**, 780-781.

Desoer, C. A. (1970), "Slowly varying system $x_{i+1} + A_i x_i$," *Electronics Letters,* **6**, 339-340.

Desoer, C. A. and Haneda, H. (1972), "The measure of a matrix as a tool to analyze algorithms for circuit analysis," *IEEE Trans. Circ. Thy.,* **CT-19**, 480-486.

Desoer, C. A. and Thomasian, A. J. (1963), "A note on zero-state stability of linear systems," *Proc. Allerton Conf.,* 50-52.

Desoer, C. A. and Vidyasagar, M. (1975), *Feedback Systems: Input-Output Properties,* Academic Press, New York.

Eggleston, H. G. (1966), *Convexity,* Cambridge Univ. Press, Cambridge.

Gear, C. W. (1971), *Numerical Initial Value Problems in Ordinary Differential Equations,* Prentice-Hall, Englewood Cliffs, N.J.

Gelb, A. and Vander Velde, W. E. (1968), *Multiple-Input Describing Functions and Nonlinear System Design,* McGraw-Hill, New York.

Guillemin, V. and Pollack, A. (1974), *Differential Topology,* Prentice-Hall, Englewood Cliffs, N.J.

Hahn, W. (1967), *Stability of Motion*, Springer-Verlag, Berlin.

Hale, J. K. (1975), *Theory of Functional Differential Equations*, Springer-Verlag, New York.

Hermann, R. and Krener, A. J. (1977), "Nonlinear controllability and observability," *IEEE Trans. Auto. Control*, **AC-22**, 728-740.

Hill, D. J. and Moylan, P. J. (1980), "Connections between finite gain and asymptotic stability," *IEEE Trans. Auto. Control*, **AC-25**, 931-936.

Hirsch, M. W. and Smale, S. (1974), *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York.

Hunt, L. R. and Su, R. (1981), "Linear equivalents of nonlinear time-varying systems," *Proc. Int'l. Symp. on Math. Thy. of Netw. and Sys.*, Santa Monica, CA, 119-123.

Hunt, L. R., Su, R. and Meyer, G. (1983a), "Design for multi-input nonlinear systems," in *Differential Geometric Control Theory*, R. W. Brockett, R. S. Millman and H. J. Sussman (Eds.), Birkhauser, Boston, 268-298.

Hunt, L. R., Su, R. and Meyer, G. (1983b), "Global transformations of nonlinear systems," *IEEE Trans. Auto. Control*, **AC-28**, 24-31.

Isidori, A. (1989), *Nonlinear Control Systems* (Second Edition), Springer-Verlag, New York.

Isidori, A., Krener, A. J., Gori-Georgi, C. and Monaco, C. (1981), "Nonlinear decoupling via feedback: A differential-geometric approach," *IEEE Trans. Auto. Control*, **AC-26**, 331-345.

Isidori, A. and Ruberti, A. (1984), "On the synthesis of linear input-output responses for nonlinear systems," *Systems and Control Letters*, **4**, 17-22.

Kailath, T. (1980), *Linear Systems*, Prentice-Hall, Englewood Cliffs, N.J.

Kalman, R. E. (1957), "Physical and mathematical mechanisms of instability in nonlinear automatic control systems," *Trans. ASME*, **79**, 553-566.

Kalman, R. E. (1963a), "Lyapunov functions for the problem of Lur'e in automatic control," *Proc. Nat'l. Acad. Sci.*, **49**, 201-205.

Kalman, R. E. (1963b), "Mathematical description of linear dynamical systems," *SIAM. J. Control*, **1**, 152-192.

Kammler, D. W. (1976), "Approximation with sums of exponentials in $L_p[0, \infty)$," *J. Approx. Thy.*, **16**, 384-408.

Kelley, J. L. (1955), *General Topology*, Van Nostrand, New York.

Krasovskii, N. N. (1959) *Problems of the Theory of Stability of Motion* (Russian); English translation, Stanford Univ. Press, Stanford, CA, 1963.

Krener, A. J. (1974), "A generalization of Chow's theorem and the bang-bang theorem to nonlinear control systems," *SIAM J. Control,* **12**, 43-52.

Lagrange, J. L. (1788) *Mécanique Celeste*, Dunod, Paris.

LaSalle, J. P. (1960), "Some extensions of Liapunov's second method," *IRE. Trans. Circ. Thy.,* **CT-7**, 520-527.

Lee, E. B. and Markus, L. (1967), *Foundations of Optimal Control Theory*, Wiley, New York.

Levin, J. L. (1960) "On the global asymptotic behavior of nonlinear systems of differential equations," *Arch. Rat. Mech. and Anal.,* **6**, 65-74.

Lloyd, N. G. (1978), *Degree Theory,* Cambridge Univ. Press, Cambridge.

Lobry, C. (1970), "Contrôllabilité des systèmes non linéaires," *SIAM J. Control,* **8**, 573-605.

Lozano-Leal, R. and Joshi, S. M. (1990), "Strictly positive real functions revisited," *IEEE Trans. Auto. Control,* **AC-35**, 1243-1245.

Lyapunov, A. M. (1892), *Problème Generale de la Stabilité du Mouvement,* French translation in 1907, photo-reproduced in *Annals of Mathematics*, Princeton Univ. Press, Princeton, N.J., 1949.

Malkin, I. G. (1954), "On a question of reversibility of Liapunov's theorem on asymptotic stability," *Prikl. Math. Mech.,* **18**, 129-138; English translation found in *Nonlinear Systems: Stability Analysis*, J. K. Aggarwal and M. Vidyasagar (Eds.), Dowden, Hutchinson and Ross, New York, 1977, pp. 161-170.

Marino, R. and Spong, M. W., "Nonlinear control techniques for flexible joint manipulators: A single link case study," *Proc. IEEE Int'l. Conf. Robotics and Auto.*, San Francisco, CA, 1026-1030.

Massera, J. L. (1949), "On Liapunoff's conditions of stability," *Ann. Math.*, **50**, 705-721.

Massera, J. L. (1956), "Contributions to stability theory," *Ann. Math.*, **64**, 182-206.

Mees, A. I. (1981), *Dynamics of Feedback Systems*, Wiley, New York.

Mees, A. I., and Bergen, A. R. (1975), "Describing Functions Revisited," *IEEE Trans. Auto. Control,* **AC-20**, 473-478.

Michel, A. N., Miller, R. K. and Tang, W. (1978), "Lyapunov stability of interconnected systems: Decomposition into strongly connected subsystems," *IEEE Trans. Circ. and Sys.,* **CAS-25**, 799-809.

Moylan, P. J. and Hill, D. J. (1978), "Stability criteria for large-scale systems," *IEEE Trans. Auto. Control*, **AC-23**, 143-150.

Narendra, K. S. and Goldwyn, R. M. (1964), "A geometrical criterion for the stability of certain nonlinear nonautonomous systems," *IEEE Trans. Circ. Thy.*, **CT-11**, 406-407.

Narendra, K. S. and Taylor, J. H. (1973), *Frequency Domain Stability for Absolute Stability*, Academic Press, New York.

Nemytskii, V. V. and Stepanov, V. V. (1960), *Theory or Differential Equations*, Princeton Univ. Press, Princeton, N.J.

Nijmeijer, H. and van der Schaft, A. (1990), *Nonlinear Dynamical Control Systems*, Springer-Verlag, Berlin.

Orlicz, W. (1932), "Zur theorie der differentialgleichung $y=f(x,y)$," *Bull. Int'l. de L'Académie Polonaise des Sciences et des Lettres, Ser. A*, 221-228.

Popov, V. M. (1961), "Absolute stability of nonlinear systems of automatic control," *Automation and Remote Control*, **22**, 857-875.

Popov, V. M. (1973), *Hyperstability of Control Systems*, Springer-Verlag, New York.

Protonarios, E. N. and Wing, O. (1967), "Theory of nonuniform RC lines," *IEEE Trans. Circ. Thy.*, **CT-14**, 2-20.

Rosenbrock, H. H. (1970), *Computer-Aided Control System Design*, Academic Press, New York.

Rouche, N., Habets, P. and Laloy, M. (1977), *Stability Theory by Liapunov's Direct Method*, Springer-Verlag, Berlin.

Royden, H. L. (1963), *Real Analysis*, MacMillan, New York.

Sandberg, I. W. (1964a), "On the L_2-boundedness of solutions of nonlinear functional equations," *Bell Sys. Tech. J.*, **43**, 1581-1599.

Sandberg, I. W. (1964b), "A frequency-domain condition for stability of feedback systems containing a single time-varying nonlinear element," *Bell Sys. Tech. J.*, **43**, 1601-1608.

Sandberg, I. W. (1965), "Some results on the theory of physical systems governed by nonlinear functional equations," *Bell Sys. Tech. J.*, **44**, 871-898.

Sanders, J. A., and Verhulst, F (1985), *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, New York.

Silverman, L. M. and Anderson, B. D. O. (1968), "Controllability, observability and stability of linear systems," *SIAM J. Control*, **6**, 121-130.

Singh, S. N. and Rugh, W. J. (1972), "Decoupling in a class of nonlinear systems by state variable feedback," *ASME Trans., J. Dyn. Sys. Meas. and Control*, **94**, 323-329.

Spong, M. W. (1987), "Modeling and control of elastic joint robots," *ASME Trans., J. Dyn. Sys. Meas. and Control,* **109**, 310-319.

Spong, M. W. and Vidyasagar, M. (1989), *Robot Dynamics and Control*, Wiley, New York.

Su, R. (1982), "On the linear equivalents of nonlinear systems," *Systems and Control Letters,* **2**, 48-52.

Vidyasagar, M. (1972), "Input-output stability of a broad class of linear time-invariant multivariable feedback systems," *SIAM J. Control,* **10**, 203-209.

Vidyasagar, M. (1977), "L_2- stability of interconnected systems using a reformulation of the passivity theorem," *IEEE Trans. Circ. and Sys.,* **CAS-24**, 637-645.

Vidyasagar, M. (1978a), "On matrix measures and Lyapunov functions," *J. Math. Anal. and Appl.,* **62**, 90-103.

Vidyasagar, M. (1978b), "On the use of right-coprime factorizations in distributed feedback systems containing unstable subsystems," *IEEE Trans. Circ. and Sys.,* **CAS-25**, 916-925.

Vidyasagar, M. (1980a), "On the well-posedness of large-scale interconnected systems," *IEEE Trans. Auto. Control,* **AC-25**, 413-420.

Vidyasagar, M. (1980b), "On the stabilization of nonlinear systems using state detection," *IEEE Trans. Auto. Control,* **AC-25**, 504-509.

Vidyasagar, M. (1980c), "Decomposition techniques for large-scale systems with nonadditive interconnections: Stability and stabilizability," *IEEE Trans. Auto. Control,* **AC-25**, 773-779.

Vidyasagar, M. (1981), *Input-Output Analysis of Large-Scale Interconnected Systems*, Springer-Verlag, New York.

Vidyasagar, M. (1985), *Control System Synthesis: A Factorization Approach*, M.I.T. Press, Cambridge, MA.

Vidyasagar, M. and Bose, N. K. (1975), "Input-output stability of linear systems defined over measure spaces," *Proc. Midwest Symp. Circ. and Sys.*, Montreal, Canada, 396-399.

Vidyasagar, M., Schneider, H. and Francis, B. A. (1982), "Algebraic and topological aspects of feedback stabilization," *IEEE Trans. Auto. Control,* **AC-25**, 880-894.

Vidyasagar, M. and Vannelli, A. (1982), "New relationships between input-output and Lyapunov stability," *IEEE Trans. Auto. Control,* **AC-27**, 481-483.

Vinograd, V. E. (1957), "The inadequacy of the method of characteristic exponents for the study of nonlinear differential equations," *Math. Sbornik,* **41**, 431-438.

Walter, W. (1970), *Differential and Integral Inequalities,* Springer-Verlag, Berlin.

Waltman, P. (1964) *Ordinary Differential Equations*, Wiley, New York.

Wang, D. (1989), *Modelling and Control of Multi-link Robots Containing a Single Flexible Link*, Ph.D. thesis, Dept. of Elect. Engrg., Univ. of Waterloo, Waterloo, Canada.

Wang, D. and Vidyasagar, M. (1991), "Control of a class of manipulators with a single flexible link — Part I: Feedback linearization," *ASME J. Dynamic Sys., Measurement and Control*, **4**, 655-661.

Warner, F. W. (1983) *Foundations of Differentiable Manifolds and Lie Groups*, Springer-Verlag, New York.

Willems, J. C. (1969a), "Stability, instability, invertibility and causality," *SIAM J. Control*, **7**, 645-671.

Willems, J. C. (1969b), "Some results on the L^p-stability of linear time-varying systems," *IEEE Trans. Auto. Control*, **AC-14**, 660-665.

Willems, J. C. (1971), *The Analysis of Feedback Systems*, M.I.T. Press, Cambridge, MA.

Wonham, W. M. (1979), *Linear Multivariable Control*, (Second Edition) Springer-Verlag, New York.

Yakubovic, V. A. (1964), "Solution of certain matrix inequalities encountered in nonlinear control theory," *Soviet Math. Doklady*, **5**, 652-656.

Zames, G. (1966a), "On the input-output stability of time-varying nonlinear feedback systems, Part I," *IEEE Trans. Auto. Control*, **AC-11**, 228-238.

Zames, G. (1966b), "On the input-output stability of time-varying nonlinear feedback systems, Part II," *IEEE Trans. Auto. Control*, **AC-11**, 465-476.

INDEX

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z