Lecture Notes
in Control and Information Sciences 400

María Tomás-Rodríguez and Stephen P. Banks

Linear, Time-varying Approximations to Nonlinear Dynamical Systems

with Applications in Control and Optimization

Springer

Authors

Dr. María Tomás-Rodríguez
City University London
School of Engineering &
Mathematical Sciences
Northampton Square
London
United Kingdom
E-mail: Maria.Tomas-Rodriguez.1@city.ac.uk

Prof. Stephen P. Banks
University of Sheffield
Dept. Automatic Control &
Systems Engineering
Mappin Street
Sheffield
United Kingdom
E-mail: s.banks@sheffield.ac.uk

ISBN 978-1-84996-100-4 e-ISBN 978-1-84996-101-1

DOI 10.1007/978-1-84996-101-1

Lecture Notes in Control and Information Sciences ISSN 0170-8643

Library of Congress Control Number: 2009942768

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

5 4 3 2 1 0

springer.com

To my parents and brother, (M T-R).
To David and Xu, (S P B).

Preface

This book is the culmination of several years' research on nonlinear systems. In contrast to the case of linear systems, where a coherent and well-defined theory has existed for many years (indeed, in many respects, we may regard linear systems theory as 'complete'), nonlinear systems theory has tended to be a set of disparate results on fairly specific kinds of systems. Of course, there are coherent theories of nonlinear systems using differential and/or algebraic geometric methods, but, in many cases, these have very strong conditions attached which are not satisfied in general. In an attempt to build a theory which has great generality we have been led to consider systems with the structure

$$\dot{x} = A(x;u)x+B(x;u)u$$

(possibly also with a measurement equation).

This appears to be quite restrictive, but, as we shall see, almost every system can be put in this form, so that the theory is, in fact, almost completely general. We shall show that systems of this form can be approximated arbitrarily closely (on any finite time interval - no matter how large) by a sequence of linear, time-varying systems. This opens up the prospect of using existing linear theory in the (global) solution of nonlinear problems and it is this with which the book is concerned. It is a research monograph, but it could be used as a graduate-level text; we have tried to keep the notation standard, so that, for the most part, the mathematical language is well-known. In some parts of the book, some previous knowledge of Lie algebras, differential geometry and functional analysis is necessary. Since there are, of course, many excellent (classical) texts on these subjects, we have merely given references to the requisite mathematical ideas.

Now we outline the detailed contents of the book. In Chapter 1 we introduce the systems with which we shall be interested and show how they are related to the most general nonlinear systems. Chapter 2 begins the detailed analysis of these systems, and in particular, we discuss the 'iteration scheme' which is the main technical tool of the approach. Since the method gives rise to sequences of linear, time-varying systems, Chapter 3 is a detailed analysis of such systems; in particular, we determine some explicit solutions and study the stability and spectral theory of these systems. In Chapter 4 we show that much of the linear spectral theory of systems can be generalised to nonlinear systems and in Chapter 5 we give a main application of the ideas to the spectral assignment problem in nonlinear systems. Optimal control of linear systems is a major part of 'classical' control theory and in Chapter 6 we show how to use the iteration method to extend the theory to nonlinear systems. We also discuss the optimality via the Hamilton-Jacobi-Bellman equation. The need for more robust controllers led to the discovery of sliding controllers, which again are generalised to nonlinear systems (and nonlinear sliding surfaces) in Chapter 7. In Chapter 8 we show how the method relates to fixed-point theory and how it can be used inductively to derive certain conditions on nonlinear systems. The generalisation of the technique to partial differential equations and systems is given in Chapter 9, together with examples from moving boundary problems and solitons (nonlinear waves). Lie algebraic methods have significant impact on linear systems theory and in Chapter 10 we see that it can also give a powerful structure theory for nonlinear systems. The global theory of nonlinear systems on manifolds is outlined in Chapter 11 where we show how to piece together a number of local systems into a global one by use of the theory of connections. Low-dimensional systems on manifolds are considered in the cases of 2, 3 and 4 dimensions. Finally, in Chapter 12, we speculate on the future possibilities of the iteration method and show that it is likely to be applicable in many other circumstances in nonlinear systems theory. The appendices give some background on linear algebra, Lie algebras, manifold theory and functional analysis.

Finally, we should acknowledge the influence of many of our students and colleagues who have been associated with this work over the years and, in particular, Metin Salamci, Tayfun Cimen, David McCaffrey, Claudia Navarro-Hernandez, Oscar Hugues-Salas, Zahra Sangelaji, Sherif Fahmy, Yi Song, Evren Gurkan Covasoglu, Xianhua Zheng, Chunyan Du, Wei Chen, Xu Xu, Salman Khalid, Serdar Tombul and Mehmet Itik. They have all contributed in various ways to the evolution of this technique.

Sheffield, London,
January 2010

María Tomás-Rodríguez
Stephen P. Banks

Contents

Chapter 1
Introduction to Nonlinear Systems

1.1 Overview

In this book we shall present a new way of approaching nonlinear systems by regarding them as limits of linear, time-varying ones. In order to explain the method, consider an unforced nonlinear dynamical system defined by the differential equation

$$\dot{x} = f(x,t), x(0) = x_0. \tag{1.1}$$

Suppose that f is differentiable at $x = 0$, for all t, and

$$f(0,t) = 0, \text{for all } t.$$

Then we can write (1.1) in the form

$$\dot{x} = A(x,t)x, x(0) = x_0 \tag{1.2}$$

for some differentiable, matrix-valued function $A(x,t)$. Note that this representation is not unique; however, if

$$\begin{gathered} \dot{x} = A_1(x,t)x \\ \dot{x} = A_2(x,t)x \end{gathered}$$

represent the same system, then

$$(A_1(x,t) - A_2(x,t))x = 0$$

so that A_1 and A_2 differ by a matrix whose kernel contains x for any $x \in \mathbb{R}$. Thus, for example, for the Van der Pol oscillator given by the equations

$$\begin{aligned} \dot{x}_1 &= x_2 - x_1^3 + x_1 \\ \dot{x}_2 &= -x_1 \end{aligned}$$

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 1–10.
springerlink.com

we have

$$\dot{x} = \begin{pmatrix} 1-x_1^2 & 1 \\ -1 & 0 \end{pmatrix} x$$

and, for example,

$$\dot{x} = \begin{pmatrix} 1-x_1^2+x_2 & 1-x_1 \\ -1 & 0 \end{pmatrix} x.$$

The matrices of these systems differ by the matrix

$$\begin{aligned} B(x) &\doteq \begin{pmatrix} 1-x_1^2 & 1 \\ -1 & 0 \end{pmatrix} - \begin{pmatrix} 1-x_1^2+x_2 & 1-x_1 \\ -1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} -x_2 & x_1 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

so that

$$B(x)x = \begin{pmatrix} -x_2 & x_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We shall assume that Equation 1.1 has global solutions, bounded on all compact time intervals (although much of what is done here is valid for systems with finite blow-up times, such as $\dot{x} = x^2$, provided we restrict attention to a compact subinterval of the interval on which solutions are defined). The basic idea is then to replace the system (1.2) by a sequence of linear, time-varying equations of the form

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t),t)x^{[i]},\ i \geq 1,\ x^{[i]}(0) = x_0 \tag{1.3}$$

where $x^{[1]}(t)$ can be any suitable starting function. In many cases we shall take

$$x^{[1]}(t) = x_0,$$

i.e. a constant function with value equal to the initial point. However, the closer $x^{[1]}(t)$ is to the true solution, the quicker the convergence will be. Of course, if we write $\xi(t,x_0)$ for the solution of (1.2) through x_0 (at time $t = 0$), then the linear, time-varying equation

$$\dot{x} = A(\xi(t,x_0),t)x,\ x(0) = x_0$$

has precisely the same solution $x(t) = \xi(t,x_0)$.

1.2 Existence and Uniqueness

In this section we shall state, without proof, standard results on the existence and uniqueness of nonlinear differential equations. These can be found, for example, in [1] or [2].

Theorem 1.1. *The system of equations*

$$\dot{x} = f(x,t),\, x(\tau) = x_0$$

has a unique solution for

$$(x_0, \tau) \in \{(x,t) : |t-\tau| \leq a, \|x-x_0\| \leq b, a, b > 0\} \doteq R$$

if f is Lipschitz continuous in x and measurable in t. Moreover, if

$$M = max|f(t,x)|,\, (t,x) \in R$$

then the solution exists for times t such that

$$|t-\tau| \leq \alpha,$$

where

$$\alpha = min\left(a, \frac{b}{M}\right).$$

For the system (1.2) we have

$$\begin{aligned} \|A(t,x)x\| &\leq \|A(x,t)\| \cdot \|x\| \\ &\leq \|A(x,t)\| \left(\|x_0\| + b\right) \\ &\leq K\left(\|x_0\| + b\right) \end{aligned}$$

if $A(x,t)$ is bounded by K on R. Hence, in this case the solutions exist at least for times satisfying $|t-\tau| \leq \alpha$, where

$$\alpha = min\left(a, \frac{b}{K(\|x_0\| + b)}\right).$$

1.3 Logistic Systems

Before getting into the details of general nonlinear systems to be discussed in the remainder of the book, we shall illustrate the method with a particularly nice class of systems called *logistic systems*. These are used to model many types of biological problems from predator-prey systems to the growth of cancer cells. The basic structure of a logistic system is one of the form

$$\begin{aligned}\dot{x}_1 &= f_1(x_1,\cdots,x_n)\cdot x_1\\ \dot{x}_2 &= f_2(x_1,\cdots,x_n)\cdot x_2\\ &\cdots\\ \dot{x}_n &= f_n(x_1,\cdots,x_n)\cdot x_n\end{aligned} \tag{1.4}$$

i.e. one in which x_i is a factor of the corresponding component of the vector field, for each $1 \le i \le n$. The reason these are so well adapted to our method is that, when we write the equations in the form (1.2), we obtain a diagonal matrix A:

$$\begin{pmatrix}\dot{x}_1\\ \vdots\\ \dot{x}_n\end{pmatrix} = \begin{pmatrix} f_1(x_1,\cdots,x_n) & & 0\\ & \ddots & \\ 0 & & f_n(x_1,\cdots,x_n)\end{pmatrix}\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix} \tag{1.5}$$

so that, when we iterate, we obtain a separable linear system of time-varying equations:

$$\begin{pmatrix}\dot{x}_1^{[i]}\\ \vdots\\ \dot{x}_n^{[i]}\end{pmatrix} = \begin{pmatrix} f_1(x_1^{[i-1]},\cdots,x_n^{[i-1]}) & & 0\\ & \ddots & \\ 0 & & f_n(x_1^{[i-1]},\cdots,x_n^{[i-1]})\end{pmatrix}\begin{pmatrix}x_1^{[i]}\\ \vdots\\ x_n^{[i]}\end{pmatrix}. \tag{1.6}$$

The solutions of these approximating sequences are given by

$$\begin{aligned}x_1^{[i]}(t) &= e^{\int_0^t f_1(x_1^{[i-1]}(s),\cdots,x_n^{[i-1]}(s))ds}x_{10}\\ &\cdots\\ x_n^{[i]}(t) &= e^{\int_0^t f_n(x_1^{[i-1]}(s),\cdots,x_n^{[i-1]}(s))ds}x_{n0},\end{aligned}$$

where $x^{[i]}(0) = x_0$ for all i. This representation can be used to prove stability and periodicity results for logistic systems.

1.4 Control of Nonlinear Systems

Consider a general nonlinear system

$$\begin{aligned}\dot{x} &= f(x,u)\\ y &= h(x,u).\end{aligned} \tag{1.7}$$

There are many techniques for controlling such a system, but they are mainly local, such as differential geometric (Lie type) methods (see [3]), local linearisations, Lyapunov-like methods ([4]), etc. To introduce a truly global method, we assume without loss of generality, that f and h have zeros at $(x,u) = (0,0)$. Then we can write the system in the form

$$\begin{aligned} \dot{x} &= A(x,u)x + B(x,u)u,\ x(0) = x_0 \\ y &= C(x,u)x + D(x,u)u \end{aligned} \tag{1.8}$$

(assuming that $\lim_{x\to 0}\frac{f(x,u)}{x}$ exists for all u, $\lim_{u\to 0}\frac{f(x,u)}{u}$ exists for all x and similar conditions for h).

Remark 1.1. *Given a dynamical system*

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

we can always make it linear in the control at the expense of increasing the dimension of the state-space by the dimension of the control u. Thus, if we put

$$\dot{u} = v$$

then we have the system

$$\frac{d}{dt}\begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} f(x,u) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ I \end{pmatrix} v,$$

which is linear in the new control v. This assumes, of course, that u is differentiable. However, the method will still be valid if u is not differentiable and we are prepared to consider distributional systems.

Thus, returning to the system (1.2), we shall introduce a sequence of controls $u^{[i]}(t)$ and states $x^{[i]}(t)$, where we can take $u^{[1]}(t) = 0$, $x^{[1]}(t) = x_0$, such that

$$\begin{aligned} \dot{x}^{[i]}(t) &= A(x^{[i-1]}(t), u^{[i-1]}(t))x^{[i]}(t) + B(x^{[i-1]}(t), u^{[i-1]}(t))u^{[i]}(t) \\ y^{[i]}(t) &= C(x^{[i-1]}(t), u^{[i-1]}(t))x^{[i]}(t) + D(x^{[i-1]}(t), u^{[i-1]}(t))u^{[i]}(t). \end{aligned}$$

Since each of these systems is linear and time-varying, we can apply linear control techniques to control these systems, and ultimately derive a controller for the nonlinear system. A number of such approaches to nonlinear control will be given later in the book.

1.5 Vector Fields on Manifolds

We next consider the case of global systems defined on compact differentiable manifolds. Thus, let M denote an n-dimensional compact differentiable (C^∞) manifold with tangent bundle TM and let X be a C^∞ section of TM, *i.e.* a smooth vector field. Cover M with a set $\{\varphi_i, U_i\}_{1\le i\le K}$ of local coordinate systems and suppose that X takes the form

$$\dot{x} = f_i(x),\ x \in \varphi_i(U_i),\ 1 \le i \le K. \tag{1.9}$$

(For simplicity, we shall assume that each U_i is equal to $\mathbb{R}^n$, as we may do.) Fix in M a set of K points $\xi_1, \cdots, \xi_K$ so that $\xi_i \in U_i = \mathbb{R}^n$. Then the systems (1.9) may be written in the form

$$\dot{x} = \widetilde{f}_i(x) + \eta_i \tag{1.10}$$

where

$$\widetilde{f}_i(x) = f_i(x) - f_i(\xi_i)$$

and

$$\eta_i = f_i(\xi_i).$$

Now each system (1.10) can be written as

$$\dot{x} = A_i(x)x + \eta_i \tag{1.11}$$

where

$$\widetilde{f}_i(x) = A_i(x)x.$$

Therefore we may regard a vector field X as equivalent to a section of a bundle $\mathfrak{M}$ of $n \times n$ matrices on M together with a specification of K vectors $\{\eta_1, \cdots, \eta_K\}$ at the fixed points $\{\xi_1, \cdots, \xi_K\}$. We can then obtain a sequence of linear, time-varying systems

$$\dot{x}^{[k]}(t) = A_i(x^{[k-1]}(t))x^{[k]}(t) + \eta_i \tag{1.12}$$

in each region $U_i = \mathbb{R}^n$. These systems can be 'pieced together' by the transition functions connecting different overlapping regions U_i, U_j in an obvious way. On a compact manifold this will give rise to global solutions given by the limits of compatible local systems of the form (1.11).

1.6 Nonlinear Partial Differential Equations

We can also apply the method to nonlinear partial differential (evolution) equations of certain types. In particular, we shall see that we can effectively reduce systems of the form

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

(where $A(t,x,u)$ is a linear elliptic operator for each t,x,u) to a sequence of linear, time-varying parabolic systems of the form

$$\frac{\partial u^i}{\partial t} = A(t, x, u^{i-1})u^i.$$

Much of the apparatus for studying these equations in a variety of Sobolev and similar type spaces exists (see, for example [5,6]) and so rather than repeat these theories, which would take us too far from the main ideas of the book, we shall approach these problems by discretization, thus reducing them to finite-dimensional problems. Since these approximations are known to converge to the true solution (under fairly mild conditions) this will not cause any loss of generality. We shall illustrate the ideas using two problems which have considerable difficulties from the classical viewpoint. These are the Stefan problem and the control of solitons. The Stefan problem is concerned with heat flow in a material which is partially solid and partially liquid and so there is an unknown moving boundary. The classical (see [7]) can be written in the form

$$\frac{\partial T}{\partial t} = \alpha_L \nabla^2 T,\ (x,t) \in \Gamma_1 \doteq \Omega_1 \times (0,\tau)$$
$$\frac{\partial T}{\partial t} = \alpha_S \nabla^2 T,\ (x,t) \in \Gamma_2 \doteq \Omega_2 \times (0,\tau)$$

where

$$\Omega = \Omega_1 \cup \Omega_2$$

is some open region in $\mathbb{R}^n$, α_L, α_S are the thermal conductivities in the liquid and solid regions of the material and the phase change takes place at the boundary of Ω_1 (and Ω_2):

$$\partial\Omega_1 = \partial\Omega_2.$$

The problem can be solved by replacing the two coupled linear equations by a single nonlinear equation

$$\frac{\partial T}{\partial t} = \alpha(T)\nabla^2 T,\ (x,t) \in \Gamma \doteq \Omega \times (0,\tau).$$

Because of the unknown boundary this problem is difficult to solve classically. We shall show that the iteration scheme provides an effective method of solution since the coupled diffusion equations describing the liquid and solid regions have been combined into a single nonlinear diffusion equation, which can be easily solved by our method.

Similarly the nonlinear wave equation for soliton dynamics can be reduced to a sequence of linear, time-varying wave equations, again leading to a fairly simple solution to the problem of boundary control of the system. The original approach to solitary waves derived by an analytical method (see [8]) can be derived in a more abstract setting in the following way. For a linear wave equation

$$\varphi_{tt} - c^2\varphi_{xx} = 0$$

we have a set of solutions of the form

$$\varphi = e^{i(\omega t - kx)},$$

where the wave number k and frequency ω are related by

$$\omega^2 = c^2k^2.$$

In the case of general linear wave equations with higher order derivatives, we may have a dispersion relation of the form

$$\omega^2 = f(k^2)$$

where f may not be linear. For example, the linear wave equation

$$\varphi_t + c\varphi_x - \frac{\varepsilon}{2c}\varphi_{xxx} = 0$$

has the dispersion relation

$$\omega = ck + \frac{1}{2}\frac{\varepsilon k^3}{c}.$$

If the wave speed depends on the amplitude φ then the wave equation becomes nonlinear. If the dependence is linear then we obtain an equation of the form

$$\varphi_t + (c - a\varphi)\varphi_x = 0$$

and if we include the above dispersion term we get

$$\varphi_t + c\varphi_x - a\varphi\varphi_x - (\varepsilon/2c)\varphi_{xxx} = 0.$$

If we now choose coordinates moving to the right with speed c we get the Korteweg-de Vries equation

$$\varphi_t = 6\varphi\varphi_x - \varphi_{xxx} = 0$$

up to scaling constants, since the φ_x term drops out. Later in the book we shall consider the more general soliton equation

$$\varphi_t + \varphi_x + k(\varphi)\varphi_x + \varphi_{xxx} = 0$$

together with boundary control to stabilise the nonlinear waves.

Of course, it is possible to use the technique on many other nonlinear partial differential equations, such as nonlinear Schrodinger models:

$$i\frac{\partial\psi}{\partial t} = -\frac{1}{2m}\Delta\psi - g|\psi|^2\psi$$

or the Hamilton-Jacobi equation

$$\frac{\partial S}{\partial t} + H\left(x, \frac{\partial S}{\partial x}\right) = 0.$$

1.7 Conclusions and Outline of the Book

We have described above the main ideas associated with the ‘iteration technique’ with which this book is concerned. It will be seen to be a powerful method for studying general nonlinear systems and, in particular, for obtaining global solutions to nonlinear control problems.

The contents of the book will now be outlined for the convenience of the reader. In Chapter 2 we discuss the general method and prove the basic convergence theorem which underlies the theory of the iteration scheme. As we shall see, it is a very general method and requires the mildest of conditions on the vector field of the system. In fact, we require nothing more than local Lipschitz continuity. (A function $f:\mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz if

$$\|f(x-y)\| \leq K\,\|x-y\|$$

for all $x,y \in \mathbb{R}^n$, for some constant K. If K depends on x and y then we say that f is locally Lipschitz.) Note that Lipschitz continuity is the minimum condition for uniqueness of solutions of a differential equation, so our condition is indeed very mild (unlike many differential geometric conditions). In Chapter 3, we shall outline the theory of linear, time-varying systems, since the whole basis of the method consists of reducing nonlinear systems to linear, time-varying ones and hence their theory is extremely important for us. Thus, we shall determine some explicit solutions, and consider their spectral theory, including a discussion of Oseledec's theorem, the Sacker-Sell spectrum and exponential dichotomies. In the following chapter, we shall show how to generalize these theorems to nonlinear systems. The systems theoretical concept of spectral assignability will be discussed in detail in Chapter 5, including a general theory of pole assignment for nonlinear systems. Chapter 6 shows that the method can be used very powerfully to determine (sub-) optimal controllers for general nonlinear systems and forms one of the main applications of the method. The ideas can, however, be applied to many problems in nonlinear control theory, including nonlinear sliding mode theory. This is done in Chapter 7 where we demonstrate that the method gives rise to a convergent sequence of moving, linear sliding surfaces for the linear, time-varying approximations which converge to an effective nonlinear sliding surface for the nonlinear system. In Chapter 8 we shall show how the method is related to fixed-point theories and use it to derive existence results for periodic solutions of nonlinear systems. The method easily generalizes to partial differential (and functional differential) equations and in Chapter 9 we show, largely by the use of two examples, how this is achieved. Finally in Chapter 10 we consider the future prospects for the method and discuss its possible future applications.

References

1. Coddington, E.A., Levinson, N.: Theory of Ordinary Differential Equations. McGraw-Hill, New York (1955)
2. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields. Springer, New York (1983)
3. Banks, S.P.: Mathematical Theories of Nonlinear Systems. Prentice-Hall, London (1988)
4. Banks, S.P.: Control Systems Engineering. Prentice-Hall, London (1986)
5. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, vol. 19. AMS (1991)

6. Hormander, L.: The Analysis of Linear Partial Differential Operators, vol. 1-4. Springer, New York (1983-1985)
7. Alexiades, V., Solomon, A.D.: Mathematical Modeling of Melting and Freezing Processes. Hemisphere Pub. Co., New York (1993)
8. Korteweg, D.J., de Vries, G.: On the Change of Form of Long Waves Advancing in a Rectangular Canal, and on a New Type of Long Stationary Waves. Philos. Mag. 39, 422–443 (1895)

Chapter 2
Linear Approximations to Nonlinear Dynamical Systems

2.1 Introduction

In this chapter the iteration approach to nonlinear systems under study is explained in detail. This technique is based on the replacement of the original nonlinear system by a sequence of linear time-varying systems, whose solutions will converge to the solution of the nonlinear problem. The only condition required for its application is a mild Lipschitz condition which must be satisfied by a matrix associated with the nonlinear system.

This approach will allow many of the classical results in linear systems theory to be applied to nonlinear systems. There are many approaches to the study of nonlinear dynamical systems, including local linearisations in phase space [14], global linear representations involving the Lie series solution [3,4,5], Lie algebraic methods [11] and global results based on topological indices [2,14]. Linear systems, on the other hand, are very well understood and there is, of course, a vast literature on the subject (see, for example, [7]). The simplicity of linear mathematics relative to nonlinear theory is evident and forms the basis of much of classical mathematics and physics. It is therefore attractive to try to attack nonlinear problems by linear methods, which are not local in their applicability. We shall study a recently introduced approach to nonlinear dynamical systems based on a representation of the system as the limit of a sequence of linear, time-varying approximations which converge in the space of continuous functions to the solution of the nonlinear system, under a very mild local Lipschitz condition. This approach will be seen later to be useful in optimal control theory [8], in the theory of nonlinear delay systems [9] and can also be applied in the theory of chaos [10]. In this chapter we shall prove basic convergence results and give a number of examples.

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 11–28.
springerlink.com

2.2 Linear, Time-varying Approximations

The iteration scheme we introduce here is based on the replacement of the original nonlinear equation by a sequence of linear, time-varying equations whose solutions converge in the space of continuous functions to the solution of the nonlinear system under a mild Lipschitz condition [16]. This technique, in a sense, is an adaptation of Picard iteration which is used in the theory of general nonlinear differential equations.

Given a nonlinear system of the form:

$$\dot{x} = A(x)x,\ x(0) = x_0 \in \mathbb{R}^n. \tag{2.1}$$

if it can be written in the SDC (state dependent coefficient) form:

$$\dot{x} = A(x)x,\ x(0) = x_0 \in \mathbb{R}^n \tag{2.2}$$

in which the origin $x = 0$ is an equilibrium point and assuming that $A(x)$ is locally Lipschitz – it is the usual minimum assumption for the existence and uniqueness of solutions – then the nonlinear system (2.2) can be approximated by the following sequence of linear, time-varying approximations:

$$\dot{x}^{[1]}(t) = A(x_0)x^{[1]}(t),\ x^{[1]}(0) = x_0 \tag{2.3}$$

$$\vdots$$

$$\dot{x}^{[i-1]}(t) = A(x^{[i-2]}(t))x^{[i-1]}(t),\ x^{[i-1]}(0) = x_0 \tag{2.4}$$

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t))x^{[i]}(t),\ x^{[i]}(0) = x_0 \tag{2.5}$$

for $i \geq 1$, where the initial function $x^{[0]}(t)$ is usually taken to be the initial conditions x_0, being possible to generalize and use other functions.

The solutions of this sequence, $\{x^{[i]}(t)\}_{i\geq 1}$, each of which satisfies a linear, time-varying system, can be found numerically (or exceptionally explicitly) and converge to the solution of the nonlinear system given in (2.2), in the sense that

$$lim_{i\to\infty}\{x^{[i]}(t)\} \to x(t)$$

uniformly for t in any compact interval, $[0,\tau]$.

Note that the first approximation, $\dot{x}^{[1]}(t) = A(x_0)x^{[1]}(t)$, $x^{[1]}(0) = x_0$, is a linear time-invariant system because the replacement of $x(t)$ by the initial condition x_0, produces a constant matrix for this first equation. The remainder of the equations in the sequence (2.3) to (2.5) are linear time-varying systems.

Remark 2.1. *The SDC form is an instantaneous parametrisation of the original nonlinear system, $\dot{x} = f(x)$ with state dependent coefficients $A(x)$. Infinite numbers* *of such realisations clearly exist. However only those parametrisations for which $A(x)$ satisfies the above mentioned Lipschitz condition will be considered here.*

Remark 2.2. *In the case when the origin is not an equilibrium point, this can be achieved by a suitable change of coordinates.*

Remark 2.3. *Unless otherwise stated, the initial conditions for each iterated differential equation, $x^{[i]}(0)$, are the same as the initial conditions given for the original nonlinear problem, $x(0)$.*

In the following section, we shall first prove a local convergence result for this system. This proof appears in a slightly different form in [16]; It has been included here for the convenience of the reader and to set the notation for the global result.

Lemma 2.1. *Suppose that $A : \mathbb{R}^n \to \mathbb{R}^{n^2}$ is locally Lipschitz. Then the sequence of functions $x^{[i]}(t)$ defined by (2.3–2.5) converges uniformly on $[0,T]$, for some $T > 0$ in the space $C([0,T];\mathbb{R}^n)$.*

Proof. Let $\Phi^{[i-1]}(t,t_0)$ denote the transition matrix of $A(x^{[i-1]}(t))$ so that we have ([16]):

$$\left\|\Phi^{[i-1]}(t,t_0)\right\| \leq e^{\left(\int_{t_0}^{t} \mu(A(x^{[i-1]}(\tau)))d\tau\right)},$$

where $\mu(A)$ is the logarithmic norm of A. By the local Lipschitz condition on $A(x)$, we have

$$\|A(x) - A(y)\| \leq \alpha(K)\,\|x-y\|,$$

for $x,y \in B(K,x_0)$ (the ball, centre x_0, radius K) for some $K > 0$ and some $\alpha(K)$. We have

$$\begin{aligned} x^{[i]}(t) - x_0 &= e^{A(x_0)t}x_0 \\ -x_0 \quad &+ \int_0^t e^{A(x_0)(t-s)}[A(x^{[i-1]}(s)) - A(x_0)]ds \end{aligned}$$

and so, for any $T > 0$,

$$\begin{aligned} \left\|x^{[i]}(t) - x_0\right\| \leq & \sup_{t\in[0,T]} \left\|e^{[A(x_0)t]} - I\right\| \cdot \|x_0\| \\ & + \sup_{t\in[0,T]} \left\{e^{[\|A(x_0)t\|]}\alpha(K)\right\} \times T \sup_{t\in[0,T]} \left\|x^{[i-1]}(t) - x_0\right\|. \end{aligned}$$

Hence, if $x^{[i-1]}(t) \in B(K,x_0)$, then $x^{[i]}(t) \in B(K,x_0)$ (for $t \in [0,T]$) if T is small enough and by the continuity of $e^{A(x_0)t}$ in t.

Since $x^{[0]}(t) \in B(K,x_0)$ for small enough T, all the solutions $x^{[i]}(t)$ are bounded for $i \geq 0$ and $t \in [0,T]$. Also,

$$\left\|A(x^{[i-1]}(t))\right\| \le \alpha(K)\left\|x^{[i-1]}(t)-x_0\right\| + \|A(x_0)\|$$

and since

$$\mu(A) = \frac{1}{2}\max[\sigma(A+A^T)]$$

in the standard matrix norm, we have that $\mu(A(x^{[i-1]}(t)))$ is bounded for all i, say $\mu(A(x^{[i-1]}(t))) \le \mu$, for all $t \in [0,T]$ and all i. Hence, by (2.3),

$$\begin{aligned}\dot{x}^{[i]}(t) - \dot{x}^{[i-1]}(t) &= A(x^{[i-1]}(t))x^{[i]}(t) - A(x^{[i-2]}(t))x^{[i-1]}(t)\\ &= A(x^{[i-1]}(t))(x^{[i]}(t)-x^{[i-1]}(t)) + \\ &\quad (A(x^{[i-1]}(t)) - A(x^{[i-2]}(t)))x^{[i-1]}(t)\end{aligned}$$

and so if we put

$$\xi^{[i]}(t) = \sup_{s\in[0,T]}\left\|x^{[i]}(s)-x^{[i-1]}(s)\right\|,$$

then

$$\xi^{[i]}(t) \le \int_0^t \left\|\Phi^{[i-1]}(t,s)\right\|\cdot\left\|A(x^{[i-1]}(s)) - A(x^{[i-2]}(s))\right\|\cdot\left\|x^{[i-1]}(s)\right\| ds.$$

Hence,

$$\xi^{[i]}(t) \le \int_0^t e^{[\mu(t-s)]}\alpha(K)\xi^{[i-1]}(s)Kds$$

so

$$\begin{aligned}\xi^{[i]}(T) &\le \sup_{s\in[0,T]}\{e^{[\mu(T-s)]}\}\alpha(K)TK\xi^{[i-1]}(T)\\ &\le \lambda\xi^{[i-1]}(T),\end{aligned}$$

where $\lambda = \sup_{s\in[0,T]}\{e^{[\mu(T-s)]}\}\alpha(K)TK$. If T is small enough, then $\lambda < 1$. In this case we have, for any $i \ge j$,

$$\begin{aligned}\left\|x^{[i]}(s)-x^{[j]}(s)\right\| &\le \left\|x^{[i]}(s)-x^{[i-1]}(s)\right\| + \left\|x^{[i-1]}(s)-x^{[i-2]}(s)\right\|\\ &\quad + \cdots + \left\|x^{[j+1]}(s)-x^{[j]}(s)\right\|\end{aligned}$$

so

$$\begin{aligned}\sup_{s\in[0,T]}\left\|x^{[i]}(s)-x^{[j]}(s)\right\| &\le \lambda^{i-j}\xi^{[j]}(T) + \lambda^{i-j-1}\xi^{[j]}(T) + \cdots + \lambda\xi^{[j]}(T)\\ &= \lambda\left(\frac{1-\lambda^{i-j}}{1-\lambda}\right)\xi^{[j]}(T).\end{aligned}$$

Hence, if N is a fixed positive integer and $i \ge j > N$, then

$$\sup_{s\in[0,T]}\left\|x^{[i]}(s)-x^{[j]}(s)\right\|\le\lambda^{j-N+1}\left(\frac{1-\lambda^{i-j}}{1-\lambda}\right)\xi^{[N]}(T).$$

Since $\xi^{[N]}(T)$ is bounded, the right hand side is arbitrarily small if j is large and so $\{x^{[i]}(t)\}$ is a Cauchy sequence in $C([0,T];\mathbb{R}^n)$. □

Having shown the local convergence of the sequence $\{x^{[i]}(t)\}$ in $C([0,T];\mathbb{R}^n)$ we now proceed to prove the global convergence in the sense that if the solution of the nonlinear equation exists and is bounded in the interval $[0,\tau]\subseteq\mathbb{R}$,then the sequence of approximations converges uniformly on $[0,\tau]$ to the solution of the nonlinear equation.

Theorem 2.1. *Suppose that the nonlinear Equation (2.1) has a unique solution on the interval $[0,\tau]$ and assume that $A:\mathbb{R}^n\to\mathbb{R}^{n^2}$ is locally Lipschitz. Then the sequence of functions $\{x^{[i]}(t)\}$ defined in (2.3) to (2.5) converges uniformly on $[0,\tau]$.*

Proof. We know from the previous lemma that, given any initial state x_0, the sequence (2.3) to (2.5) converges uniformly on some interval $[0,T]$, where T may depend on x_0. However, it is clear from the proof of the lemma that T can be chosen to be locally constant; *i.e.* for any $\bar{x}$ there exists a neighbourhood $B_{\bar{x}}$ of $\bar{x}$ such that the sequence in (2.3) to (2.5) with initial state $x_0\in B_{\bar{x}}$ converges uniformly on some interval $[0,T_{\bar{x}}]$, where $T_{\bar{x}}$ is independent of x_0.

Now suppose that the result is false, so that there is a maximal time interval $[0,\bar{T})$ such that, for any $T<\bar{T}$, the sequence (2.3) to (2.5) converges uniformly on $[0,T]$. Now consider the solution trajectory $x(t;x_0)$ of the original nonlinear system (2.1) on the interval $[0,\tau]$*;* define the set

$$S=\{x(t;x_0):t\in[0,\tau]\}.$$

For each $\bar{x}\in S$, choose a neighbourhood $B_{\bar{x}}$ as above; *i.e.* the sequence of approximations converges uniformly on the interval $[0,T_{\bar{x}}]$ for any $x_0\in B_{\bar{x}}$ for $T_{\bar{x}}$ independent of x_0. Since S is compact and $\cup_{\bar{x}\in S}B_{\bar{x}}$ is an open cover of S, there exists a finite subcover $\{B_{\bar{x}_1},\cdots,B_{\bar{x}_p}\}$ with corresponding times $\{T_{\bar{x}_1},\cdots,T_{\bar{x}_p}\}$. Let

$$T_m=\min\{T_{\bar{x}_1},\cdots,T_{\bar{x}_p}\}.$$

Now the sequence (2.3) to (2.5) converges uniformly on $[0,\bar{T}-T_m/2]$, by assumption. Let

$$x_{0,i}=x^{[i]}(\bar{T}-T_m/2).$$

Since these converge to $x(\bar{T}-T_m/2)$ we can assume that they belong to $B_{\bar{x}_p}$, so that we get a sequence of solutions given by the Equations (2.3) to (2.5) from the initial states $x_{0,i}$ and converging uniformly on the interval $[0,T_m]$ to the corresponding solutions of the nonlinear equation given by (2.1). (See Figure 2.1.)

These solutions will be denoted by $x^{[i,j]}(t)$. Now, using a Cantor-like diagonal argument, consider the functions

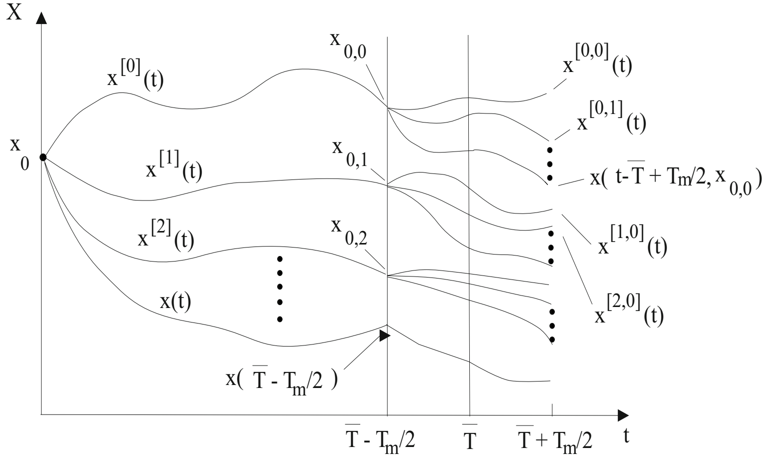


Fig. 2.1 Approximations to the nonlinear solution

$$y^{[i]}(t) = \begin{cases} x^{[i]}(t) & , 0 \le t \le \bar{T} - T_m/2 \\ x^{[i,i]}(t) & , \bar{T} - T_m/2 \le t \le \bar{T} + T_m/2. \end{cases}$$

Then $y^{[i]}(t)$ converges uniformly to $x(t)$ on $[0, \bar{T} + T_m/2]$ and is arbitrarily close to $x^{[i]}(t)$ on $[0, \bar{T}]$ which contradicts the assumption that $\{x^{[i]}(t)\}$ is not uniformly convergent on $[0, \bar{T}]$. □

2.3 The Lorenz Attractor

In this section the dynamical equations of the Lorenz Attractor are considered and it will be shown how its nonlinear solution vector $x(t) = [x_1, x_2, x_3]$ can be approximated by using the iteration technique previously introduced:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} 10 & 10 & 0 \\ 25 & -1 & -x_1 \\ 0 & x_1 & -2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

with initial condition: $[x_{01}, x_{02}, x_{03}]^T = [1, 1, 1]^T$.

Applying the iteration technique just by replacing the previous solution $x^{[i-1]}(t)$ into the actual matrix $A(x^{[i-1]}(t))$ a sequence of linear, time-varying approximations is obtained:

$$\begin{pmatrix} \dot{x}_1^{[1]} \\ \dot{x}_2^{[1]} \\ \dot{x}_3^{[1]} \end{pmatrix} = \begin{pmatrix} 10 & 10 & 0 \\ 25 & -1 & -x_{01} \\ 0 & x_{01} & -2 \end{pmatrix} \cdot \begin{pmatrix} x_1^{[1]} \\ x_2^{[1]} \\ x_3^{[1]} \end{pmatrix}, \quad \begin{pmatrix} x_1^{[1]}(0) \\ x_2^{[1]}(0) \\ x_3^{[1]}(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\vdots$$

$$\begin{pmatrix} \dot{x}_1^{[i]} \\ \dot{x}_2^{[i]} \\ \dot{x}_3^{[i]} \end{pmatrix} = \begin{pmatrix} 10 & 10 & 0 \\ 25 & -1 & -x_1^{[i-1]} \\ 0 & x_1^{[i-1]} & -2 \end{pmatrix} \cdot \begin{pmatrix} x_1^{[i]} \\ x_2^{[i]} \\ x_3^{[i]} \end{pmatrix}, \quad \begin{pmatrix} x_1^{[i]}(0) \\ x_2^{[i]}(0) \\ x_3^{[i]}(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Each of these linear time-varying equations can be solved successively by just replacing the previous solution $x^{[i-1]}(t)$ into the system matrix $A(x^{[i-1]}(t))$. This will generate a sequence of linear solutions, $x^{[i]}(t)$, for each equation in the sequence. In Figure 2.2, the true solution of the nonlinear system is represented and some of these iterations have also been plotted. It is easy to see that in this case too, the convergence of these linear solutions towards the nonlinear solution is clear, since after 12 iterations they converge to the true solution of the nonlinear system. The norm of the error, $\max_{t\in[0,T]}||x(t)-x^{[i]}(t)||$, is shown in Figure 2.3. Clearly the technique works effectively for the Lorenz Attractor.

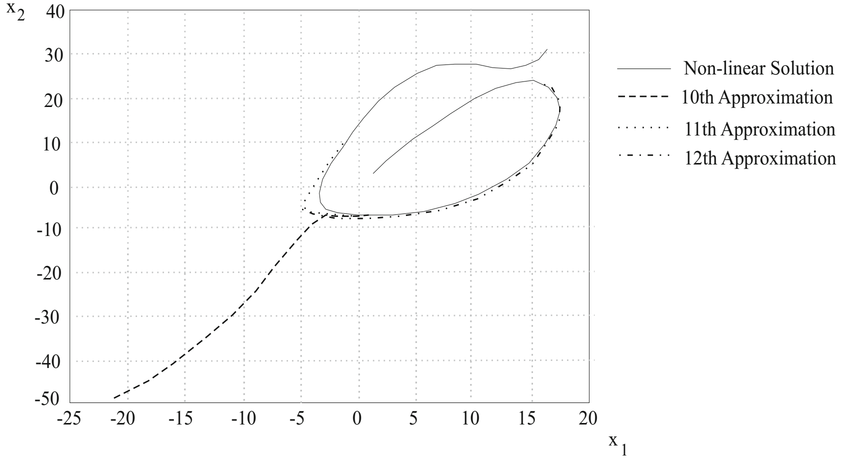


Fig. 2.2 Approximations to the nonlinear solution. Lorenz Attractor

2.4 Convergence Rate

During the study of the application of this iteration technique, some questions related to the rate of convergence and the different representation of a nonlinear

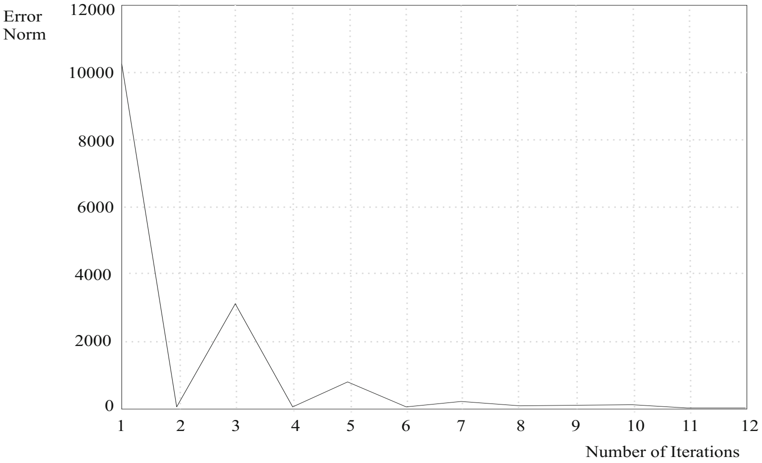


Fig. 2.3 Convergence error for the Lorenz Attractor

system of the form $\dot{x} = A(x)x$ arose in a natural way. In this section, these two issues are briefly addressed. These two subjects are important since they lead to a better understanding of the technique and closely related to its optimal use.

Having discussed both local and global convergence of the iteration technique above, the next logical step is to study the rate of convergence in terms of the system dynamics, and a maximum desired error between the actual solution and the iterated solutions. In other words, given a nonlinear system,

$$\dot{x} = f(x),\ x(0) = x_0$$

and writing it in the SDC form

$$\dot{x} = A(x)x,\ x(0) = x_0,$$

the following question arises: *Is it possible to estimate the number of iterations $x^{[i]}(t)$ needed to be computed in order to achieve certain maximum degree of error ε such that $max_{t\in[0,T]}||x(t) - x^{[i]}(t)|| \leq \varepsilon$?*

Here, an initial approach to this question is presented assuming the following conditions are satisfied (where R, λ and T are any fixed positive numbers):

- The initial condition is inside a ball of radius $R/2$ centered at 0:
$$x_0 \in B(0, R/2).$$
- The solution of the nonlinear system is bounded in $B(0,R)$ for a finite time interval $[0,T]$:
$$x(t,x_0) \in B(0,R), \forall t \in [0,T].$$
- $\mu(A(x)) \leq \mu, \forall x \in B(0,R)$, so that if $x^{[i]}(t) \in B(0,R), \forall i$ then

$$||\Phi^{[i]}(t,t_0)|| \leq e^{\int_{t_0}^{t}\mu(A(x^{[i]}(\tau)))d\tau} \leq e^{\mu(t-t_0)},$$

where $\Phi^{[i]}(t,t_0)$ is the transition matrix of $A(x^{[i]}(t))$, see [17].

The error between the true solution and the i^{th} iteration is $x(t)-x^{[i]}(t)$, so from (2.5):

$$\begin{aligned}\dot{x}(t)-\dot{x}^{(i)}(t) &= A(x(t))x(t)-A(x^{(i-1)}(t))x^{(i)}(t)\\ &= A(x(t))\Big[x(t)-x^{(i-1)}(t)\Big]+\Big[A(x(t))-A(x^{(i-1)}(t))\Big]x^{(i)}(t)\end{aligned}$$

Since

$$x^{[i]}(t)=\Phi^{[i-1]}(t,0)x_0,$$

it follows that $||x^{[i]}(t)|| \leq e^{\mu t}\cdot||x_0||$ if $x^{[i]}(t)\in B(0,R)\forall i$.

Setting $\xi^{i}(t)=sup_{s\in[0,T']}||x(s)-x^{[i]}(s)||$, the estimate

$$\xi^{[i]}(t) \leq \int_0^{T'} \Phi(t,s)\cdot||A(x(s))-A(x^{[i]}(s))||\cdot||x^{[i]}(s)||ds$$

follows, for any $T'>0$. Hence,

$$\xi^{[i]}(t) \leq \int_0^{T'} e^{(t-s)\mu}\cdot M(\alpha)\cdot||x(s)-x^{[i-1]}(s)||\cdot e^{\mu t}\cdot||x_0||\cdot ds$$

and so,

$$\xi^{[i]}(T') \leq sup_{s\in[0,T']}\Big[e^{(T'-s)\mu}\Big]\cdot M(\alpha)\cdot\xi^{[i-1]}(T')\cdot e^{\mu T'}\cdot||x_0||\cdot T'.$$

Choosing T so that,

$$\lambda := sup_{s\in[0,T']}e^{\mu(T'-s)}\cdot M(\alpha)\cdot T'\cdot e^{\mu T'}\cdot||x_0||<1$$

then, $\xi^{[i]}(T') \leq \lambda^{i}\cdot\xi^{[0]}(T') \leq \lambda^{i}\cdot sup_{s\in[0,T']}||e^{A(x_0)s}-I||\cdot||x_0|| = \lambda^{i}\cdot\rho(x_0)$, where

$$\rho(x_0)=sup_{s\in[0,T']}||e^{A(x_0)s}-I||\cdot||x_0||.$$

Let $\varepsilon>0$ be given, and let $N=\left[\frac{T}{T'}\right]+1$, where $[v]$ is the integer part of v. Choose i so that

$$\lambda^{i}\cdot\rho(x_0)\leq\frac{\varepsilon}{N}\longrightarrow i\geq\frac{log\Big(\frac{\varepsilon}{N\cdot\rho(x_0)}\Big)}{log\lambda}.$$

Then, an estimate for the approximation $x^{[i]}(t)$ to be within an error of ε of the nonlinear solution $x(t)$ on the interval $(0,T)$ can be obtained in terms of the initial conditions:

$$i \geq N \times \left[\frac{log\left(\frac{\varepsilon}{N\rho(x_0)}\right)}{log\lambda} \right] + 1. \tag{2.6}$$

It is necessary at this point to say that this is a very conservative approach because of the use of norm bounds, and the assumption that convergence rates are similar at each small interval.

2.5 Influence of the Initial Conditions on the Convergence

In this section the choice of initial conditions will be discussed and studied. It is important to note the difference between

(a) the *initial conditions* $x(0)$, that are used as the initial value of the solution when solving a linear differential equation and

(b) the *initial chosen solution* $x^{[0]}(t)$ included directly in the system's matrix on the first iteration $\dot{x}^{[1]}(t) = A(x^{[0]}(t))x^{[1]}(t)$. The speed of convergence will depend on this latter choice which until now had been chosen for simplicity to be equal to the initial condition x_0 of the differential equation:

$$\dot{x}^{[1]}(t) = A\underbrace{(x_0)}_{(b)}x^{[1]}(t),\ \underbrace{x_0^{[1]} = x(0)}_{(a)}.$$

In this section the dependence of the convergence rate on the *initial chosen solution* will be introduced and in the following section a formal approach to this convergence rate will be given.

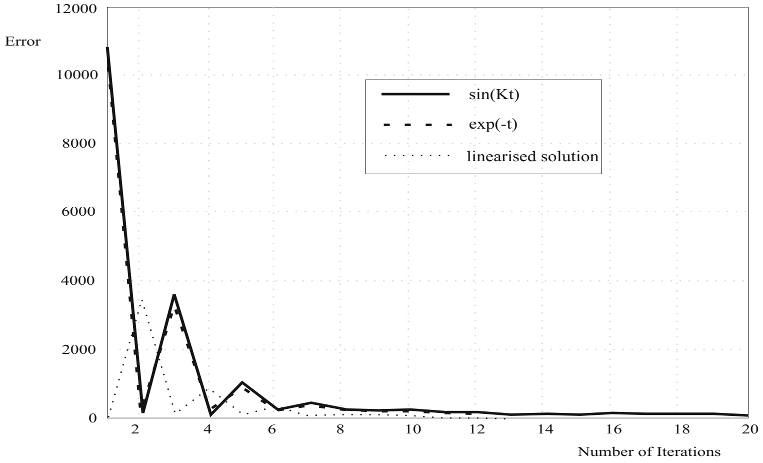


Fig. 2.4 Comparison of the norm of the error for different initial conditions

Figure 2.4 shows that in the Lorenz Attractor example, depending on the choice of the *initial chosen solution*, the error between the real nonlinear solution $x(t)$ and the approximated solutions $x^{[i]}(t)$, converges to zero at a different rate: the number of iterations needed in order to achieve a satisfactory approximation to the nonlinear solution of the original system changes.

The main motivation for choosing time-varying initial functions $x^{[0]}(t)$ is based on the idea of the iteration technique itself: after *i* iterations, the iterated solutions are close to each other and the $x^{[i]}(t)$ solution from $\dot{x}^{[i]}(t) = A(x^{[i-1]})x^{[i]}(t)$ tends to the true nonlinear solution $x(t)$ so this means that the closer the *initial chosen solution* is to the true nonlinear solution, the sooner this convergence will be achieved, (see Figure 2.5). Therefore the closer one iteration is to the true nonlinear solution, the better the next iteration...and so on. Keeping this idea in mind, the next logical step is to suggest that the most accurate choice of initial conditions could be the linearised solution around the equilibrium point.

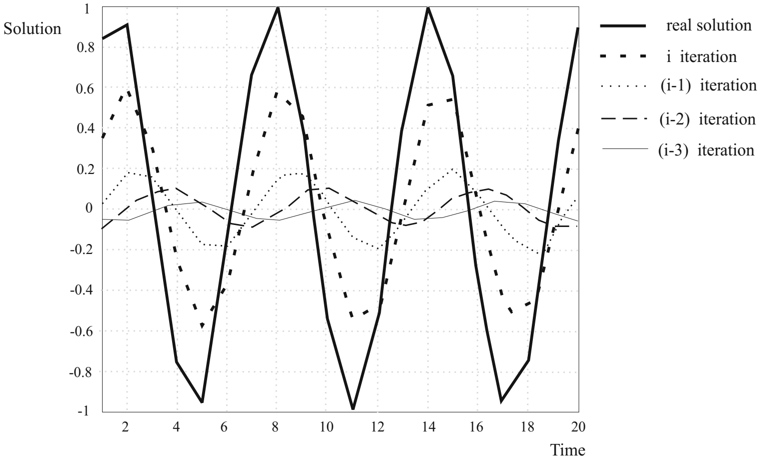


Fig. 2.5 General schema for the iteration technique

In Figure 2.4 the dynamical equations of the Lorenz Attractor were simulated for different choices of initial conditions, it could be seen that the overall error decreases considerably in the case where the linearised solution has been used as initial function. These experimental results confirm the statement above: there exist a difference in the convergence rate of the iteration technique depending on the choice of the initial conditions substituted in the original matrix for the first iteration and this choice could be optimised by choosing as initial function a function similar to the linearised solution of the nonlinear system.

2.6 Notes on Different Configurations

The different ways a nonlinear system of the form $\dot{x} = A(x)x$ can be represented has been the object of interest before by different authors and depending on the area of research, it could be of great importance to understand the effects of using different representations. In this section, attention is focussed on the different rates of convergence achievable using one configuration or another. Here, it will be assumed that $A(x)$ is controllable and/or stabilisable (since this will be required later in its application to control) and it will be shown how this occurs in the particular case of the Lorenz Attractor by writing its dynamics in two different ways, and after this applying the iteration technique in order to compare the convergence rate for each configuration.

The dynamical equations of the Lorenz Attractor are given by:

$$\dot{x}_1 = -10x_1 + 10x_2$$

$$\dot{x}_2 = 25x_1 - x_3x_1 - x_2$$

$$\dot{x}_3 = x_1x_2 - 2x_3.$$

They can be written on the form $\dot{x} = A(x)x$ in different various ways, *i.e.*:

First configuration:

$$\dot{x} = \begin{pmatrix} -10 & 10 & 0 \\ 25 - x_3 & -1 & 0 \\ x_2 & 0 & -2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Second configuration:

$$\dot{x} = \begin{pmatrix} -10 & 10 & 0 \\ 25 & -1 & -x_1 \\ 0 & x_1 & -2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Comparison:

If now both first and second configurations are iterated with the technique presented in this chapter, for initial conditions $[x_{01}, x_{02}, x_{03}] = [1, 1, 1]^T$ and plotting the error between the iterated solutions $x^{[i]}(t)$ and the real solution $x(t)$ it can be seen in figure 2.6 that there is a difference in the convergence rate according to the representation of the system.

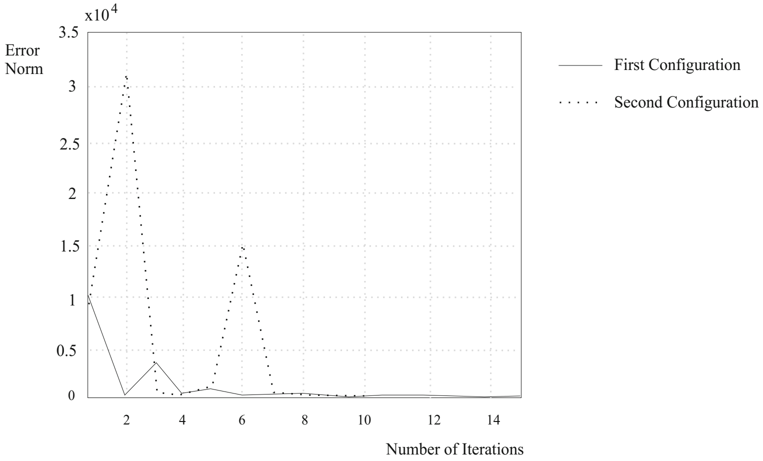


Fig. 2.6 Convergence for the different configurations

The 'shape' of this graph shows a great difference especially in the initial iterations where the error is much greater in the case of the first representation. This leads to some interesting questions:

- *Could it be possible that in the limit the integral of this error is the same in both cases?*
  It seems that in the example of the Lorenz Attractor this is not the case as the first iteration for the first representation produces a quantitative bigger error compared with the second representation.
- *Does this behaviour happen for all the systems with more than one representation?*
  This divergence of behaviour between the different representations, is caused to some extent by the different eigenvalues the first iterated matrices $A(x_0)$ have, so if different representations happen to have the same eigenvalues at the first iteration, it is expected that this difference will diminish.

The question of non-uniqueness of the representation will not be discussed further in this book, but it is an important topic for further research.

2.7 Comparison with the Classical Linearisation Method

Different methods of linearisation and approximation have been presented in the past by many authors: all of them have in common the intention to provide a model which is locally equivalent to the nonlinear system. In fact, it turns out in many situations that nonlinear systems can be approximated in some regions of operation by linear systems.

The most classical linearisation method is based on the Taylors expansion. Of course, the classical linearisation near an equilibrium point x_e has many drawbacks already enumerated in the introduction and so all the efforts made in approaching a nonlinear system by other methods are of significant importance.

Next, a comparison between the linearised solution obtained when a classical linearisation is applied to the nonlinear system and the solution obtained when applying this iteration technique is presented. The relationship of this iteration method with the classical linearisation shows the effectiveness of this method.

Example 2.1. Consider a nonlinear system modelling an over-damped bead on rotating hoop. The system is modelled by the following equation:

$$mr\frac{d^2\phi}{dt^2}+b\frac{d\phi}{dt}=-mgsin(\phi)+mrw^2sin(\phi)cos(\phi),\ x_0=x(0), \tag{2.7}$$

where ϕ is the angle, b is the damping constant, w is a constant angular velocity and r is a radius. If $\varepsilon=\frac{m^2gr}{b^2}$, $\gamma=\frac{rw^2}{g}$ and $\tau=\frac{mg}{b}$, (2.7) can be written as a dimensionless system:

$$\frac{d^2\phi}{dt^2}+\frac{1}{\varepsilon}\frac{d\phi}{dt}=\frac{1}{\varepsilon}\Big[sin(\phi)+\gamma sin(\phi)cos(\phi)\Big],\ x_0=x(0). \tag{2.8}$$

By making the change of variables $\phi=x_1$, $\dot{\phi}=x_2$, the state-space representation of (2.8) is:

$$\begin{aligned}\dot{x_1}&=x_2\\ \dot{x_2}&=\tfrac{sin(x_1)}{\varepsilon}+\tfrac{\gamma}{\varepsilon}sin(x_1)cos(x_1)-\tfrac{x_2}{\varepsilon},\ x_0=x(0).\end{aligned} \tag{2.9}$$

with $\varepsilon=0.2$, $\gamma=-0.1$, and $[x_{01},x_{02}]^T=[1.5,1.5]^T$. The aim of this section is to compare the linearisation and the iterative technique. In the following, both methods are applied and results are compared.

Linearisation:

This is a method by which a nonlinear system is replaced by a linear approximation about its equilibrium point/s or a given trajectory. It is required that the nonlinear system should have a convergent Taylor's series expansion about the equilibrium point:

$$f(a+h)=f(a)+\frac{\partial f}{\partial x}\Big|_{x=a}h+h^T\frac{\partial^2 f}{\partial x^2}\Big|_{x=a}h+\cdots$$

A good approximation to the nonlinear system can be obtained when the variation 'h' from the equilibrium point 'a' is small enough so that the higher order terms and above can be neglected, that is, if $h\to 0$, then:

$$f\to f(a)+\frac{\partial f}{\partial x}\Big|_{x=a}h.$$

The nonlinear system can be then approximated by its Taylor series expansion neglecting the high order terms. In this case, the system (2.7) is replaced by the matrix of its first derivatives (Jacobian) evaluated at the origin and substituting $\varepsilon = 0.2$ and $\gamma = -0.1$ the following linear system is obtained:

$$\begin{pmatrix} \dot{x_1} \\ \dot{x_2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -4.5 & -5 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}. \tag{2.10}$$

The solutions $x_1(t)$ and $x_2(t)$ can now be calculated from (2.10). These solutions have been plotted respectively in Figures 2.7 and 2.8.

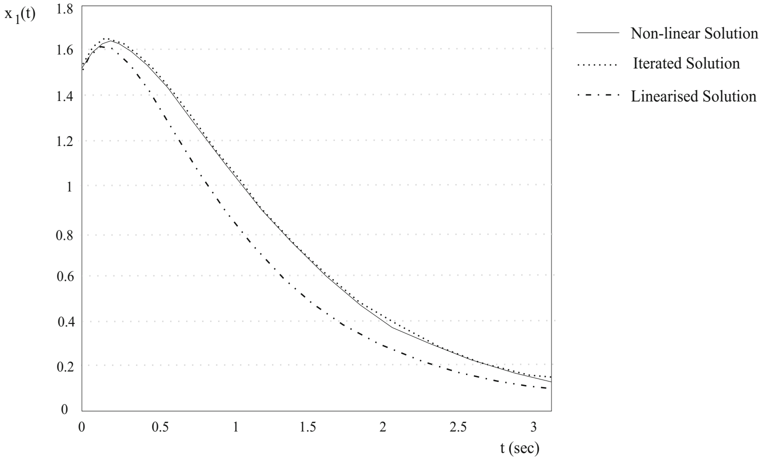


Fig. 2.7 $x_1(t)$ of the nonlinear, 16^{th} iteration and linearised solution

Iteration Technique:

On the other hand, by applying the iteration technique, and simulating up to 50 iterations in Figure 2.9 it can be seen that the norm of the error between the nonlinear solution and the iterated one converges almost to zero. It is clear that the 16^{th} iteration is a good approximation and converges to the nonlinear solution (Figures 2.7 and 2.8), confirming in this way that in this case, the iteration technique is better in approximating the nonlinear solution.

In this section a comparison between the classical method of linearisation and the iteration technique has been presented for a standard physical example. The results of the simulations prove the ability of the iteration technique to approach the nonlinear solution more accurately than classical linearisation methods.

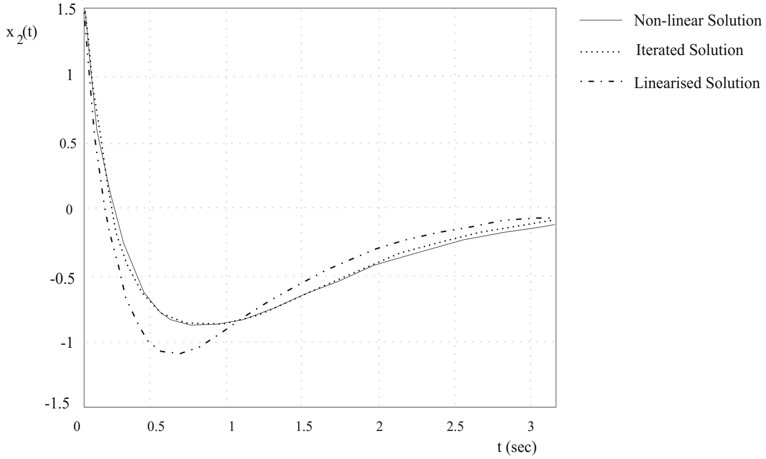


Fig. 2.8 $x_2(t)$ of the nonlinear, 16^{th} iteration and linearised solution

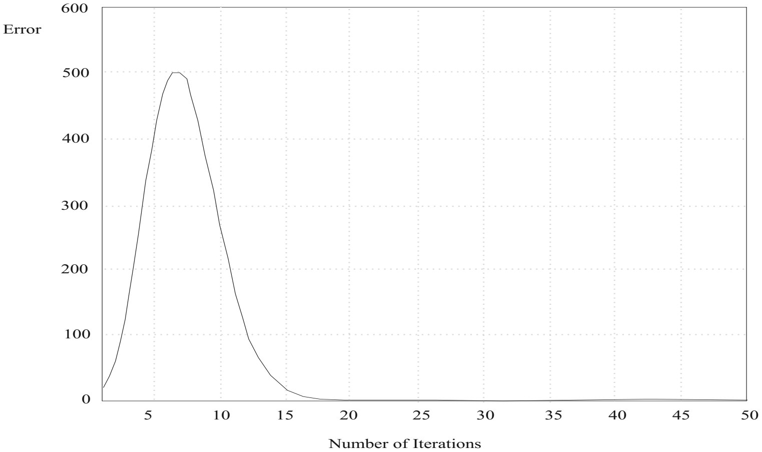


Fig. 2.9 Error between $x(t)$ and $x^{[i]}(t)$

2.8 Conclusions

In this chapter, a technique to approach the solution of a broad class of nonlinear systems has been introduced.

This technique replaces a nonlinear system by a sequence of linear time-varying systems whose solutions converge to the solution of the nonlinear system. Unlike many other approaches to nonlinear systems, this is a *global* technique since under the assumption of existence and uniqueness of the solution, the only prerequisite needed to be satisfied is that the system's matrix should be locally Lipschitz. Both local and global convergence theorems have been included in this chapter.

Based on the convergence of the sequence, linear control ideas can be now applied to nonlinear control systems by studying each of these linear time-varying systems with linear tools.

In this chapter, an estimation of this convergence rate has been studied; that is, would it be possible to know how many iterations need to be computed in order to achieve certain degree of convergence? This question has been studied and a mathematical expression for this convergence rate has been developed in function of the nonlinear system itself and the initial conditions. However, it is important at this point to recognise that this approach is a very conservative one as it relies on the maximum value the norm of the matrix can have in the given time interval, so the exact estimation provided here is expected to be higher than the actual number of iterations needed to achieve a given accuracy for any given system.

In the last part of the chapter, the optimal choice of initial conditions has been discussed. Clearly the best choice is the actual solution through the initial point. Of course, the solution of the system is not known *a priori*, so the use of any method to find a first approximation which is *close* to the true solution is clearly desirable.

However, the question of the best choice of initial conditions and the different ways to represent a nonlinear system in order to minimise the error are topics still open to discussion and probably further research in this area could be done in the future.

The infinity of different ways of representing the nonlinear system in the form $\dot{x} = A(x)x$ has been object of discussion too, it being clear from examples that this initial representation has an important effect on the speed of convergence.

In the last example it has been shown how this technique works in a better way than the widely applied linearisation method as for an adequate number of iterations it *converges* to the true nonlinear solution of the system under study.

References

1. Banks, S.P.: Mathematical Theories of Nonlinear Systems. Prentice-Hall, London (1988)
2. McCaffrey, D., Banks, S.P.: Lagrangian Manifolds, Viscosity Solutions and Maslov Index. J. Convex Analysis 9, 185–224 (2002)
3. Banks, S.P., Iddir, N.: Nonlinear Systems, the Lie Series and the Left Shift Operator: Application to Nonlinear Optimal Control. IMA J. Math.Contr. Inf. 9, 23–34 (1992)
4. Banks, S.P.: Infinite-Dimensional Carleman Linearisation, the Lie Series and Optimal Control of Nonlinear PDEs. Int. J. Sys. Sci. 23, 663–675 (1992)
5. Banks, S.P., Moser, A., McCaffrey, D.: Lie Series and the Realization Problem. Comp. and App. Maths 15, 37–54 (1996)
6. Banks, S.P., Riddalls, C., McCaffrey, D.: The Schwartz' Kernel Theorem and the Frequency-Domain Theory of Nonlinear Systems. Arch. Cont. Sci. 6, 57–73 (1997)
7. Banks, S.P.: Control Systems Engineering: Modelling and Simulation, Control Theory and Microprocessor Implementation. Prentice-Hall, Englewood Cliffs (1986)
8. Banks, S.P., Dinesh, K.: Approximate Optimal Control and Stability of Nonlinear Finite and Infinite-Dimensional Systems. Ann. Op. Res. 98, 19–44 (2000)

9. Banks, S.P.: Nonlinear delay systems, Lie algebras and Lyapunov transformations. IMA J. Math. Cont. & Inf. 19, 59–72 (2002)
10. Banks, S.P., McCaffrey, D.: Lie Algebras, Structure of Nonlinear Systems and Chaotic Motion. Int. J. Bifurcation & Chaos 8(7), 1437–1462 (1998)
11. Banks, S.P.: The Lie Algebra of a Dynamical System and its Application to Control. Int. J. Sys. Sci. 32, 220–238 (2001)
12. Bredon, G.: Sheaf Theory. Springer, New York (1998)
13. Kalman, R.E., Bertram, C.: Control System Analysis and Design via the Second Method of Lyapunov. ASME J. Basic Eng. 82, 371–393 (1960)
14. Perko, L.: Differential Equations and Dynamical Systems. Springer, New York (1991)
15. Sacker, R.J., Sell, G.: A Spectral Theory for Linear differential Systems. J. Diff. Eqn. 27, 320–358 (1978)
16. Tomás-Rodríguez, M., Banks, S.P.: Linear Approximations to Nonlinear Dynamical Systems with Applications to Stability and Spectral Theory. IMA Journal of Math. Control and Inf. 20, 89–103 (2003)
17. Brauer, F.: Perturbations of nonlinear systems of differential equations II. J. Math. Analysis App. 17, 418–434 (1967)

Chapter 3
The Structure and Stability of Linear, Time-varying Systems

3.1 Introduction

In view of the basic approximation theory in Chapter 2, nonlinear dynamical systems can be approximated uniformly on compact intervals by linear, time-varying systems. It is therefore important to study the general questions of existence, uniqueness, etc. for dynamical systems of this type. In this chapter we shall consider the general theory of linear time-varying dynamical systems first from the point of view of existence and uniqueness, and then we shall determine a number of explicit solutions, based on the theory of Lie algebras.

The remainder of the chapter is concerned, essentially with stability theory. After considering the classical theory, we shall introduce the ideas of Lyapunov numbers and describe Oseledec's theorem on decomposition of the state-space into invariant subbundles, which generalises the hyperbolic splitting of the state-space for time-invariant systems. Finally we shall consider the theory of exponential dichotomies and its generalisation to invariant subbundles.

3.2 Existence and Uniqueness

Consider the linear, time-varying system of equations

$$\dot{x} = A(t)x, x(0) = x_0 \in \mathbb{R}^n, \tag{3.1}$$

where $A : [0,\infty) \to \mathbb{R}^n$ is assumed to be continuous. The first simple result shows that such a system of equations has a unique solution on the whole of the interval $[0,\infty)$.

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 29–60.
springerlink.com

Lemma 3.1. *The system of equations (3.1) has a unique solution on* $[0,\infty)$*, provided that the matrix function* $A:[0,\infty)\to\mathbb{R}^n$ *is continuous.*

Proof. First we prove local existence. Let $L_{t_1}: C[0,t_1]\to C[0,t_1]$ denote the linear operator

$$L_{t_1}(x)=\int_0^{t_1} A(s)x(s)ds$$

where $C[0,t_1]$ is given the usual sup norm. Thus

$$L_{t_1}(x)-L_{t_1}(y)=\int_0^{t_1} A(s)(x(s)-y(s))ds$$

so that

$$||L_{t_1}(x)-L_{t_1}(y)||\le\int_0^{t_1} A(s)ds\cdot||(x-y||\le t_1\cdot K\cdot||x-y||$$

for some constant K, by continuity of A. Hence,

$$||L_{t_1}(x)-L_{t_1}(y)||\le\lambda\cdot||x-y||$$

where $\lambda<1$, if t_1 is small enough. The usual fixed point theorem for continuous mappings on the Banach space $C[0,t_1]$ (see [5]) now proves local existence on $[0,t_1]$, for some $t_1>0$. Hence, either the solution exists for all $t\in[0,\infty)$ or on a maximal time interval $[0,T)$, say, where $T<\infty$. By continuity of A, the function $A:[0,T]\to\mathbb{R}^n$ is uniformly continuous and hence is bounded on $[0,T)$, say

$$||A(t)||\le M, t\in[0,T].$$

Now from (3.1), we have

$$x(t)=x_0+\int_0^t A(s)x(s)ds$$

so that

$$||x(t)||\le||x_0||+M\int_0^t||x(s)||ds,$$

for $t\in[0,T)$. By Gronwall's inequality (see [3]) we have

$$||x(t)||\le e^{Mt}||x_0||, t\in[0,T)$$

so that the solution exists on $t\in[0,T]$ and is bounded and continuous, so that $lim_{t\to T}x(t)$ exists, giving a contradiction. Uniqueness follows easily from the fact that the solution is linear in x_0. □

Remark 3.1. *Clearly from the above proof, continuity is not necessary; boundedness of* $||A(\cdot)||$ *on any compact interval is sufficient.*

We can find approximate solutions to (3.1) by using the so-called Duhamel principle:

Lemma 3.2. *If $A : [0,\infty) \to \mathbb{R}^n$ is continuous, then the solution of (3.1) is given by*

$$x(t) = lim_{\ell\to\infty} e^{A((\ell-1)h))h} \cdots e^{A(2h)h} e^{A(h)h} e^{A(0)h} x_0$$

where $h = t/\ell$. (The limit is taken in the standard Euclidean norm on $\mathbb{R}^n$.)

Proof. From (3.1) we have

$$x(t) = x_0 + \int_0^t A(s)x(s)ds.$$

By continuity, the integral on the right is a standard Riemann integral and so it can be approximated arbitrarily closely by rectangles:

$$x(t) = x_0 + \sum_{i=0}^{\ell-1} A(ih) \int_{ih}^{(i+1)h} x(s)ds + \varepsilon \tag{3.2}$$

for any $\varepsilon > 0$, where ℓ depends on ε. Consider the ℓ systems

$$\begin{aligned} \dot{\xi}_1(t) &= A(0)\xi_1, \quad \xi_1(0) = x_0 + \varepsilon \\ \dot{\xi}_2(t) &= A(h)\xi_2, \quad \xi_2(0) = \xi_1(h) \\ &\vdots \\ \dot{\xi}_\ell(t) &= A((\ell-1)h)\xi_\ell(t), \quad \xi_\ell(0) = \xi_{\ell-1}(h). \end{aligned}$$

Then

$$\begin{aligned} \xi_1(t) &= e^{A(0)t}\xi_1(0), \\ \xi_2(t) &= e^{A(h)t}\xi_1(h), \\ \xi_3(t) &= e^{A(2h)t}\xi_2(h), \\ &\vdots \end{aligned}$$

and so

$$\xi_\ell(h) = e^{A((\ell-1)h))h} e^{A((\ell-2)h))h} \cdots e^{A(2h)h} e^{A(h)h} e^{A(0)h} \xi_0. \tag{3.3}$$

But we also have

$$\begin{aligned}
\xi_1(t) &= \xi_1(0) + A(0)\int_0^t \xi_1(s)ds \\
\xi_2(t) &= \xi_2(0) + A(h)\int_0^t \xi_2(s)ds \\
&= \xi_1(0) + A(0)\int_0^t \xi_1(s)ds + A(h)\int_0^t \xi_2(s)ds \\
&\cdots \\
\xi_\ell(t) &= \xi_1(0) + A(0)\int_0^t \xi_1(s)ds + A(h)\int_0^t \xi_2(s)ds + \cdots + A((\ell-1)h)\int_0^t \xi_\ell(s)ds
\end{aligned} \tag{3.4}$$

The result now follows from (3.2), (3.3) and (3.4). □

3.3 Explicit Solutions

In this section we shall give a number of explicit solutions for linear, time-varying systems of equations of the form

$$\dot{x} = A(t)x, \quad x(t) = x_0. \tag{3.5}$$

By integrating (3.5) we have

$$x(t) = x_0 + \int_0^t A(s)x(s)ds.$$

This is an integral equation which can be iterated, *i.e.*

$$\begin{aligned}
x(t) &= x_0 + \int_0^t A(\tau_1)\left(x_0 + \int_0^{\tau_1} A(\tau_2)x(\tau_2)d\tau_2\right)d\tau_1 \\
&= x_0 + \int_0^t A(\tau_1)d\tau_1 x_0 + \int_0^t \int_0^{\tau_1} A(\tau_1)A(\tau_2)x(\tau_2)d\tau_2 d\tau_1.
\end{aligned}$$

Proceeding in this way, this suggests that the (formal) solution of (3.5) is given by

$$x(t) = \sum_{n=0}^{\infty}\left(\int_{t\ge\tau_1\ge\cdots\ge\tau_n\ge0} A(\tau_1)\cdots A(\tau_n)d\tau_n\cdots d\tau_1\right)x_0.$$

To show that it is the actual solution, it is sufficient to show that the sum converges. Clearly, the nth term is bounded in norm by $t^nK^n||x_0||/n!$, where

$$K = sup_{\tau\in[0,t]}||A(\tau)||,$$

and so the result follows.

Remark 3.2. *Physicists write*

$$\int_{t\geq\tau_1\geq\cdots\geq\tau_n\geq 0} A(\tau_1)\cdots A(\tau_n)d\tau_n\cdots d\tau_1 = \frac{1}{n!}\int_{\tau_1\in[0,t]} P[A(\tau_1)\cdots A(\tau_n)]d\tau_n\cdots d\tau_1$$

where

$$P[A(\tau_1)\cdots A(\tau_n)]$$

is the 'time-ordered' product of the factors; i.e. with the factors permuted in the form $A(\tau_{\sigma(1)})\cdots A(\tau_{\sigma(n)})$ so that the larger values of τ_i appear first:

$$t_{\sigma(1)} \geq t_{\sigma(2)} \geq \cdots \geq t_{\sigma(n)}.$$

Hence we can write the path-ordered exponential as

$$P\left[e^{-\int_0^t A(s)ds}\right] = \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}P\left(\int_0^t A(s)ds\right)^n$$

and so the solution appears in the form

$$x(t) = P\left[e^{-\int_0^t A(s)ds}\right] = x_0.$$

However, care must be taken in interpreting this result since the solution of the Equation 3.1 is given by

$$x(t) = e^{-\int_0^t A(s)ds}x_0$$

only if the matrices $A(t)$ commute for all t. (A general solution for non-commuting A's will be given later.)

Next we discuss the general solution given by Wei and Norman [12], which generalises a result of Magnus [7]. Thus, consider again Equation 3.5 and write $A(t)$ in the form

$$A(t) = \sum_{i=1}^{\ell} a_i(t)X_i \tag{3.6}$$

where $\{X_1,\cdots,X_\ell\} \subseteq g\ell(n)$ is a basis of the Lie algebra $\mathfrak{L}_A$ generated by the set $\{A(t), t\geq 0\}$. (We shall use the elementary theory of Lie algebras – for general results, see, *e.g.* .) We shall need the Campbell-Baker-Hausdorff formula which states that, for any Lie algebra $\mathfrak{L}$, if $X,Y\in\mathfrak{L}$ then $exp(X)Yexp(-X)\in\mathfrak{L}$ is given by

$$exp(X)Yexp(-X) = Y + [X,Y] + [X,[X,Y]]/2! + [X,[X,[X,Y]]]/3! + \cdots$$

i.e.

$$exp(X)Yexp(-X) = (e^{adX})Y,$$

where

$$(adX)Y = [X,Y].$$

For the basis $\{X_i\}$ in (3.6), let γ_{ij}^k be the structure constants of the Lie algebra, *i.e.*

$$[X_i, X_j] = \sum_{k=1}^{\ell} \gamma_{ij}^k X_k, \quad 1 \leq i, j \leq \ell.$$

It follows that

$$\left(\prod_{j=1}^{r} exp(g_j X_j) \right) X_i \left(\prod_{j=r}^{1} exp(-g_j X_j) \right)$$

belongs to $\mathfrak{L}_A$ for any $r \in \{1, \cdots, \ell\}$ and any variables g_i. Hence we can write it in terms of the basis $\{X_i\}$:

$$\left(\prod_{j=1}^{r} exp(g_j X_j) \right) X_i \left(\prod_{j=r}^{1} exp(-g_j X_j) \right) = \sum_{k=1}^{\ell} \xi_{ki} X_k \tag{3.7}$$

and it is easy to see that each ξ_{ki} is an analytic function of the g_j's. We shall show that the solution of (3.5) can be written in the form

$$x(t) = exp(g_1(t)X_1) exp(g_2(t)X_2) \cdots exp(g_\ell(t)X_\ell) x_0, \tag{3.8}$$

in a neighbourhood of $t = 0$. In fact, we shall show that (3.7) satisfies (3.5) for certain g_i's and so by uniqueness of the solution, it must be given by (3.8). To do this note that if we write

$$\Phi(t) = exp(g_1(t)X_1) exp(g_2(t)X_2) \cdots exp(g_\ell(t)X_\ell)$$

then

$$\frac{d\Phi}{dt}(t) = \sum_{i=1}^{\ell} \frac{dg_i(t)}{dt} \left(\prod_{j=1}^{i-1} exp(g_j X_j) X_i \prod_{j=i}^{\ell} exp(g_j x_J) \right)$$

and the right hand side equals $A(t)\Phi(t) = \sum_{i=1}^{\ell} a_i(t) X_i \cdot \Phi(t)$ if (and only if)

$$\begin{aligned} \sum_{i=1}^{\ell} a_i(t) X_i &= \sum_{i=1}^{\ell} \frac{dg_i(t)}{dt} \left(\prod_{j=1}^{i-1} exp(g_j X_j) X_i \prod_{j=i-1}^{1} exp(-g_j x_J) \right) \\ &= \sum_{i=1}^{\ell} \frac{dg_i(t)}{dt} \left(\prod_{j=1}^{i-1} exp(g_j adX_j) X_i \right) \\ &= \sum_{i=1}^{\ell} \sum_{k=1}^{\ell} \frac{dg_i(t)}{dt} \xi_{ki} X_k \end{aligned}$$

by (3.7). Since $\{X_i\}$ is a basis we have

$$\xi g' = a$$

where

$$\xi = (\xi_{ki}), \quad g' = (g_i'), \quad a = (a_i).$$

Since $\xi(0) = I$, ξ is invertible in a neighbourhood of $t = 0$, so that

$$g' = \xi^{-1} a$$

and the result follows by analyticity of ξ and ξ^{-1} near $t = 0$.

This result is local, but it can be shown (see [12]) that it is global if $\mathfrak{L}_A$ is a solvable Lie algebra. However, there is a simpler and more useful way to see this. Thus, suppose that $\mathfrak{L}_A$ is solvable. Then by Lie's theorem (see Appendix B), the matrices in $\mathfrak{L}_A$ are simultaneously triangularizable, so that we may assume (after a change of coordinates) that Equation 3.5 may be written in the form

$$\dot{x}_i = \sum_{j=i}^{n} \widetilde{a}_{ij}(t) x_j, \quad 1 \le i \le n, \quad x(0) = x_0$$

and so the solution is given inductively by

$$\begin{aligned} x_n(t) &= e^{\int_0^t \widetilde{a}_{nn}(s)ds} x_{0n} \\ x_i(t) &= e^{\int_0^t \widetilde{a}_{ii}(s)ds} x_{0i} + \sum_{j=i+1}^{n} \int_0^t e^{\int_s^t \widetilde{a}_{ij}(\tau)d\tau} x_j(s)ds, \quad 1 \le i < n. \end{aligned} \tag{3.9}$$

Finally, in this section, we present briefly another explicit expression for the solution of (3.5) (for a full discussion, see [1]). We shall need the following results (see, *e.g.* [8]):

Theorem 3.1. *If* A, B *are sufficiently close to* 0*, then* $C = ln(e^a e^B)$ *is given by*

$$C = B + \int_0^1 g[exp(tAdA)exp(AdB)](A)dt$$

where

$$g(z) = \frac{lnz}{z-1} = 1 + \frac{1}{2}(1-z) + \frac{1}{3}(1-z)^2 + \cdots = \sum_{\ell=0}^{\infty} \frac{1}{\ell+1}(-1)^{\ell}(z-1)^{\ell}.$$

Corollary 3.1. *If A, B are as in Theorem 3.1, then*

$$C = B + \sum_{\ell=0}^{\infty} \frac{(-1)^{\ell}}{\ell+1} \sum_{\substack{i_1=0, j_1=0 \\ (i_1,j_1)\neq(0,0)}}^{\infty} \sum_{\substack{i_2=0, j_2=0 \\ (i_2,j_2)\neq(0,0)}}^{\infty} \cdots \sum_{\substack{i_\ell=0, j_\ell=0 \\ (i_\ell,j_\ell)\neq(0,0)}}^{\infty} \frac{1}{i_1!i_2!\cdots i_\ell!j_1!j_2!\cdots j_\ell!(|\boldsymbol{i}|+1)}$$
$$\times (AdA)^{i_1}(AdB)^{j_1}(AdA)^{i_2}(AdB)^{j_2}\cdots(AdA)^{i_\ell}(AdB)^{j_\ell}\cdot A, \qquad (3.10)$$

where $|i| = i_1+\cdots+i_\ell$.

Theorem 3.2. *Given k matrices $A_1,\cdots,A_k$, in a sufficiently small neighbourhood of 0, then $C_k = ln(e^{A_k}e^{A_{k-1}}\cdots e^{A_1})$ is given by*

$$C_k = \int_0^1 g[exp(tAdA_k)exp(tAdA_{k-1})exp(tAdA_{k-2})\cdots exp(tAdA_1)](A_k)dt + C_{k-1}$$

where

$$C_{k-1} = ln(e^{A_{k-1}}e^{A_{k-2}}\cdots e^{A_1}).$$

Proof. Let

$$\Gamma(t) = ln(e^{tA_k}e^{A_{k-1}}\cdots e^{A_1})$$

so that

$$e^{\Gamma(t)} = e^{tA_k}e^{A_{k-1}}\cdots e^{A_1}.$$

Then,

$$\begin{aligned}(exp[Ad\Gamma(t)])H &= e^{\Gamma(t)}He^{-\Gamma(t)} \\ &= e^{tA_k}e^{A_{k-1}}\cdots e^{A_1}He^{-A_1}\cdots e^{-A_{k-1}}e^{-tA_k}\end{aligned}$$

for any matrix H, and so

$$exp[Ad\Gamma(t)] = exp(tAdA_k)exp(AdA_{k-1})\cdots exp(AdA_1).$$

Also,

$$\begin{aligned}e^{\Gamma(t)}\frac{d}{dt}e^{-\Gamma(t)} &= e^{tA_k}e^{A_{k-1}}\cdots e^{A_1}\frac{d}{dt}(e^{-A_1}\cdots e^{-A_{k-1}}e^{-tA_k}) \\ &= -A_k\end{aligned}$$

and so

$$f(Ad\Gamma(t))\dot{\Gamma}(t) = A_k.$$

However,

$$f(lnz)g(z)=1, \quad \text{for} \quad |1-z|<1$$

and so

$$f(lnF)g(F)=I, \quad \text{or} \quad g(F)=(f(lnF))^{-1}$$

for any matrix F with $||I-F||<1$. Setting $F=exp(AdtA_k)exp(AdA_{k-1})\cdots exp(AdA_1)$ gives

$$\Gamma(t)=\int_0^t g[exp(AdtA_k)exp(AdA_{k-1})\cdots exp(AdA_1)](A_k)dt+\text{ constant.}$$

The constant is given by $\Gamma(0)=ln(e^{A_{k-1}}\cdots e^{A_1})=C_{k-1}$. □

Corollary 3.2. *If $A_1,\cdots,A_k$ are as in the theorem, then*

$$C_k=\sum_{\ell=0}^{\infty}\frac{(-1)^\ell}{\ell+1}\cdot\sum_{\substack{\boldsymbol{i}(1)=0\\|\boldsymbol{i}(1)|\neq 0}}^{\infty}\cdots\sum_{\substack{\boldsymbol{i}(\ell)=0\\|\boldsymbol{i}(\ell)|\neq 0}}^{\infty}\frac{1}{\boldsymbol{i}(1)!\cdots\boldsymbol{i}(\ell)!(i_1(1)+\cdots+i_1(\ell)+1)}\cdot$$
$$(AdA_k)^{i_1(1)}(AdA_{k-1})^{i_2(1)}\cdots(AdA_1)^{i_k(1)}$$
$$(AdA_k)^{i_1(2)}(AdA_{k-1})^{i_2(2)}\cdots(AdA_1)^{i_k(2)}\cdot$$
$$(AdA_k)^{i_1(\ell)}(AdA_{k-1})^{i_2(\ell)}\cdots(AdA_1)^{i_k(\ell)}\cdot A_k$$

where

$$\boldsymbol{i}(p)=(i_1(p),\cdots,i_k(p)), \quad \boldsymbol{i}(p)!=i_1(p)!(i_2(p)!\cdots i_k(p)!.$$

(If $\ell=0$ then we interpret the value as A_k.)

Now by lemma 3.2, we have:

Lemma 3.3. *The system*

$$\dot{x}=A(t)x, \quad x(0)=x_0 \tag{3.11}$$

has solution given by

$$x(t)=lim_{h\to 0}e^{A((m-1)h)h}e^{A((m-2)h)h}\cdots e^{A(2h)h}e^{A(h)h}e^{A(0)h}x_0 \tag{3.12}$$

for any $t>0$, where $mh=t$.

From Corollary 3.3 and Lemma 3.1, we have:

Lemma 3.4. *The solution of the system*

$$\dot{x}=A(t)x, \quad x(0)=x_0$$

is given by

$$x(t)=lim_{h\to 0}e^{C_m}x_0$$

where $mh=t$ and

$$C_m = \sum_{p=2}^{m} \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell+1} \cdot \sum_{\substack{\boldsymbol{i}(1)=0 \\ |\boldsymbol{i}(1)|\neq 0 \\ \boldsymbol{i}(1)\in\mathbb{N}^p}}^{\infty} \cdots \sum_{\substack{\boldsymbol{i}(\ell)=0 \\ |\boldsymbol{i}(\ell)|\neq 0 \\ \boldsymbol{i}(\ell)\in\mathbb{N}^p}}^{\infty} \frac{1}{\boldsymbol{i}(1)!\cdots\boldsymbol{i}(\ell)!(i_1(1)+\cdots+i_1(\ell)+1)} \times$$
$$(AdA_p)^{i_1(1)}(AdA_{p-1})^{i_2(1)}\cdots(AdA_1)^{i_p(1)} \times$$
$$(AdA_p)^{i_1(2)}(AdA_{p-1})^{i_2(2)}\cdots(AdA_1)^{i_k(2)} \times$$
$$(AdA_p)^{i_1(\ell)}(AdA_{p-1})^{i_2(\ell)}\cdots(AdA_1)^{i_p(\ell)} \times A_p + A_1 \tag{3.13}$$

where

$$A_q = A((q-1)h)h. \tag{3.14}$$

Combining Lemmas 3.1 and 3.2, we have

Theorem 3.3. *The solution of the non-autonomous differential equation in (3.5) is given by*

$$x(t;x_0) = exp\left(\int_0^t A(\tau)d\tau + \sum_{k=2}^{\infty} \sum_{\sigma^{k-1}\in S_{k-1}} \mu(\sigma^{k-1}) \int_0^t \int_0^{\tau_k} \cdots \int_0^{\tau_3} \int_0^{\tau_2} \right.$$
$$\left. [A(\tau_{\sigma^{k-1}(1)}),[A(\tau_{\sigma^{k-1}(2)}),[\cdots,[A(\tau_{\sigma^{k-1}(k-1)}),A(\tau_k)]\cdots]]]d\tau_1\cdots d\tau_k\right)x_0, \tag{3.15}$$

where S_{k-1} is the set of all permutations of $1,\cdots,k-1$ and $\mu(\sigma^{k-1})$ is a number, depending on k and the permutation, to be determined below.

Proof. This follows from (3.12) and (3.13) since each multiple integral in (3.15) is the limit of typical terms in (3.13) where each $i_j(k) = 1$. The latter condition follows from the fact that, for a sequence

$$(AdA_p)^{i_1(1)}(AdA_{p-1})^{i_2(1)}\cdots(AdA_1)^{i_p(1)}(AdA_p)^{i_1(2)}\cdots A_p$$

of a given total degree $k = \sum_{j=1}^{\ell}(i_1(j)+\cdots+i_p(j))$, any repeated factors will converge to a zero integral since they are multiplied by h^k and there are at most $O(1/(h^{k-1}))$ of such terms. □

The only remaining thing, therefore, is to find the multipliers $\mu(\sigma^{k-1})$. This will be done in three steps. Consider first the case of $k=2$. Clearly for terms with brackets of the form $[A_i,A_j]$ we must have $\ell=1$ in the expression (3.13); thus we must choose these terms from the expression

$$-\frac{1}{2}\sum_{p=2}^{m}\sum_{\substack{|\mathbf{i}(1)|\neq 0\\ \mathbf{i}(1)\in\mathbb{N}^p}}\frac{1}{\mathbf{i}(1)!(i_1(1)+1)}(AdA_p)^{i_1(1)}(AdA_{p-1})^{i_2(1)}\cdots(AdA_1)^{i_p(1)}A_p.$$

Since we do not have to consider terms of the form $[A_i,A_i]=0$, we must have $i_1(1)=0$ and some $i_1(j)\neq 0, j\neq 1$. In this case, all the factors $\frac{1}{\mathbf{i}(1)!(i_1(1)+1)}$ equal 1, so we have

Lemma 3.5. $\mu(\sigma^1)=-\frac{1}{2}$, *i.e. the second order term in (3.15) is*

$$-\frac{1}{2}\int_0^t\int_0^\tau[A(\rho),A(\tau)]d\rho d\tau.$$

Next, terms of order 3 come from (3.13) with $\ell\leq 2$, *i.e.* from the expressions

$$-\frac{1}{2}\sum_{p=2}^{m}\sum_{\substack{|\mathbf{i}(1)|\leq 2\\ \mathbf{i}(1)\in\mathbb{N}^p}}\frac{1}{\mathbf{i}(1)!(i_1(1)+1)}(AdA_p)^{i_1(1)}(AdA_{p-1})^{i_2(1)}\cdots(AdA_1)^{i_p(1)}A_p$$

$$+\frac{1}{3}\sum_{p=2}^{m}\sum_{\substack{|\mathbf{i}(1)|=1,|\mathbf{i}(2)|=1\\ \mathbf{i}(1),\mathbf{i}(2)\in\mathbb{N}^p}}\frac{1}{\mathbf{i}(1)!\mathbf{i}(2)!(i_1(1)+i_1(2)+1)}(AdA_p)^{i_1(1)}$$
$$(AdA_{p-1})^{i_2(1)}\cdots(AdA_1)^{i_p(1)}\cdot$$
$$(AdA_p)^{i_1(2)}(AdA_{p-1})^{i_2(2)}\cdots(AdA_1)^{i_p(2)}A_p.$$

We will obtain brackets of the form $[A_i,[A_j,A_k]]$ where (i) $k>i>j$ or (ii) $k>j>i$. Terms of type (i) (*i.e.* for $k>i>j$) can come from both the series above for any given fixed i,j,k we get a factor of $-\frac{1}{2}$ from the first and a factor of $\frac{1}{3}$ from the second, *i.e.* a factor of $-\frac{1}{6}$. Terms of type (ii) (i.e. for $k>j>i$), however, can only come from the second series because the terms

$$(AdA_p)^{i_1(1)}(AdA_{p-1})^{i_2(1)}\cdots(AdA_1)^{i_p(1)}A_p$$

in the first series are ordered so we must have $k>i>j$. Hence for any term of the second type we have a factor of $\frac{1}{3}$, and so we have:

Lemma 3.6. *The third order term in (3.15) is*

$$-\frac{1}{6}\int_0^t\int_0^{\tau_3}\int_0^{\tau_2}[A(\tau_2),[A(\tau_1),A(\tau_3)]]d\tau_1d\tau_2d\tau_3+$$
$$\frac{1}{3}\int_0^t\int_0^{\tau_3}\int_0^{\tau_2}[A(\tau_1),[A(\tau_2),A(\tau_3)]]d\tau_1d\tau_2d\tau_3.$$

Consider next the case of the k^{th} order terms. As before, each factor

$$\frac{1}{\mathbf{i}(1)!\cdots\mathbf{i}(\ell)!(i_1(1)+\cdots+i_1(\ell)+1)}$$

will reduce to 1 and we will get only k^{th} order terms for $\ell \le k-1$. Hence we must choose k^{th} order terms from

$$\begin{aligned}
&-\frac{1}{2}\sum_{p=k-1}^{m}(AdA_p)(AdA_{p-1})\cdots(AdA_1)A_p\\
&+\frac{1}{3}\sum_{p=k-1}^{m}(AdA_p)\cdots(adA_1)(AdA_p)\cdots(AdA_1)A_p\\
&-\cdots\\
&+\frac{(-1)^{k-1}}{k}\sum_{p=k-1}^{m}((AdA_p)\cdots(AdA_1)^{k-1}A_p.
\end{aligned}\tag{3.16}$$

Consider first the term $B_{i_1}\cdots B_{i_{k-1}}A_{i_k}$ where $i_k > i_1 > i_2 > \cdots > i_{k-1}$ and $B_{i_j} = AdA_v$ for some v depending on i_j. This can be chosen in only one way from the first term in (3.16) and in $k-2$ ways from the second term in (3.16). (We must choose at least one B_v from each group of terms $(AdA_p)\cdots(AdA_1)$, so we could choose the first one, B_{i_1} from the first group and the remaining $k-2$ from the second, or the first two, B_{i_1}, B_{i_2} from the first group and the remaining $k-3$ from the second, etc.) In the r^{th} term in (3.16) we will have r groups $(AdA_p)\cdots(AdA_1)$, *i.e.*

$$\underbrace{(AdA_p)\cdots(AdA_1)\cdots(AdA_p)\cdots(AdA_1)\cdots(AdA_p)\cdots(AdA_1)}_{r}A_p\ . \tag{3.17}$$

Suppose there are $\rho(s,t)$ ways of selecting terms of the form (AdA_v) from t groups. Then the number of ways of selecting $k-1$ from r, *i.e.* $\rho(k-1,r)$ is

$$\rho(k-1,r) = \sum_{i=r-1}^{k-2}\rho(i,r-1)$$

since we can choose 1 from the first group and $k-2$ from the remaining, *i.e.* $\rho(k-2,r-1)$ or 2 from the first group and $k-3$ from the remaining, *i.e.* $\rho(k-3,r-1)$, etc.

Lemma 3.7. *We have*

$$\rho(k-1,r) = \frac{1}{(r-1)!}(k-r)(k-r+1)\cdots(k-2), \quad r \ge 2.$$

Proof. Note that $\rho(\nu,1) = 1$ for all ν and $\rho(\nu,2) = \nu-1$ for all ν. Hence the formula is correct for $r=2$. Suppose it is true for $r-1$, *i.e.*

$$\rho(k-1,r-1)=\frac{1}{(r-2)!}(k-r+1)(k-r+2)\cdots(k-2),\quad r\geq 2.$$

Then,

$$\begin{aligned}\rho(k-1,r) = &= \sum_{i=r-1}^{k-2}\rho(i,r-1)\\ &= 1+\rho(r,r-1)+\rho(r+1,r-1)+\cdots+\rho(k-2,r-1)\\ &= 1+\frac{1}{(r-2)!}(r+1-r+1)(r+1-r+2)\cdots(r+1-2)+\\ &\qquad \frac{1}{(r-2)!}(r+2-r+1)(r+2-r+2)\cdots(r+2-2)+\cdots\\ &\qquad +\frac{1}{(r-2)!}(k-r)\cdots(k-3)\\ &= \frac{1}{(r-2)!}(1.2\cdots(r-2)+2\cdots(r-1)+3\cdots r+\cdots+\\ &\qquad (k-r)\cdots(k-3))\\ &= \frac{1}{(r-2)!}\sum_{i=1}^{k-r} i(i+1)\cdots(i+(r-2)-1)\\ &= \frac{1}{(r-2)!}\frac{(k-r)(k-r+1)\cdots(k-r+r-1)}{(r-2)+1}.\end{aligned}$$

□

Corollary 3.3. *The total number of terms of the form* $[B_{i_i},[B_{i_2},[\cdots,[B_{i_{k-1}},A_{i_k}]\cdots]]]$ *which can be chosen, where the indices* $i_1,i_2,\cdots,i_{k-1}$ *are increasing, is given by* $-\frac{1}{k(k-1)}$.

Proof. The required number is given by

$$\begin{aligned}\sum_{\ell=2}^{k}\frac{(-1)^{\ell-1}}{\ell}\rho(k-1,\ell-1) &= \sum_{\ell=2}^{k}\frac{(-1)^{\ell-1}}{\ell}\frac{1}{(\ell-2)!}(k-\ell+1)(k-\ell+2)\cdots(k-2)\\ &= \sum_{\ell=2}^{k}\frac{(-1)^{\ell-1}}{\ell}\frac{1}{(\ell-2)!}\frac{\Gamma(k-1)}{\Gamma(k-\ell+1)}\\ &= -\frac{1}{k(k-1)}.\end{aligned}$$

□

For the general case, let σ^{k-1} be a permutation of the set $\{1,\cdots,k-1\}$ and write it as $\sigma^{k-1}=(i_1\cdots i_{k-1})$. We can partition the permutation in the form $(\mathbf{i}^1,\mathbf{i}^2,\cdots,\mathbf{i}^\gamma)$ where

$$\mathbf{i}^1=(i_1,\cdots,i_{v_1}),\mathbf{i}^2=(i_{v_1+1},\cdots,i_{v_1+v_2}),\cdots$$

such that

$$\begin{aligned}&\mathbf{i}^{\alpha}\text{is a decreasing sequence for } \alpha \in \mathscr{A} \subseteq \{1,\cdots,\gamma\}\\&\mathbf{i}^{\beta}\text{is a decreasing sequence for } \beta \in \{1,\cdots,\gamma\}\backslash\mathscr{A} = \mathscr{B}\end{aligned}$$

i.e. if $\mathbf{i}^{\alpha} = (i_{\ell_1},\cdots,i_{\ell_2})$, then $i_{\ell_1} > i_{\ell_1+1}\cdots > i_{\ell_2}$. Moreover, we choose the partition so that the sets $\mathbf{i}^{\alpha}$ for $\alpha \in \mathscr{A}$ are maximal. Let

$$\varepsilon = \sum_{\beta\in\mathscr{B}} |\mathbf{i}^{\beta}| + \aleph(\mathscr{A})$$

where $\aleph(\mathscr{A})$ denotes the cardinality of $\mathscr{A}$ and if $\mathbf{i}^{\beta} = (i_{k_1},\cdots,i_{k_j})$, then $|\mathbf{i}^{\beta}| = \sum_{v=1}^{j} i_{k_v}$. If we are selecting from a term of the form (3.17) with ℓ repeated strings $(AdA_p)\cdots(AdA_1)$, then we require $\varepsilon \le \ell$. Put $\zeta = \ell - \varepsilon$. If $\zeta > 0$ let P_ℓ be the set of distinct partitions of ζ into $\aleph(\mathscr{A})$ pieces, *i.e.*

$$\zeta = \sum_{\alpha\in\mathscr{A}} \zeta_\alpha$$

where $\zeta_\alpha \ge 0$. Then the number of possible selections in the ℓ^{th} term is

$$\sum_{\zeta\in P_\ell}\prod_{\alpha\in\mathscr{A}} \rho(|\mathbf{i}^{\alpha}|,\zeta_\alpha+1)$$

where we take $\rho(k,r) = 0$ if $r > k$. Hence we have proved:

Theorem 3.4. *The number $\mu(\sigma^{k-1})$ is given by*

$$\begin{aligned}\mu(\sigma^{k-1}) &= \sum_{\ell=\varepsilon}^{k-1}\frac{(-1)^{\ell}}{\ell+1}\sum_{\zeta\in P_\ell}\prod_{\alpha\in\mathscr{A}}\rho(|\boldsymbol{i}^{\alpha}|,\zeta_\alpha+1)\\&= \sum_{\ell=\varepsilon}^{k-1}\frac{(-1)^{\ell}}{\ell+1}\sum_{\zeta\in P_\ell}\prod_{\alpha\in\mathscr{A}}\frac{1}{\zeta_\alpha!}(|\boldsymbol{i}^{\alpha}|-\zeta_\alpha)(|\boldsymbol{i}^{\alpha}|-\zeta_\alpha+1)\cdots(|\boldsymbol{i}^{\alpha}|-1).\end{aligned}$$

Example 3.1. Consider, for example, the permutation of $\{1,2,3,4,5\}$ given by $\sigma^5 =$ (5 2 3 4 1). Here we have

$$(5\ 2\ 3\ 4\ 1) = (\mathbf{i}^1,\mathbf{i}^2,\mathbf{i}^3)$$

where

$$\mathbf{i}^1 = (5,2),\ \mathbf{i}^2 = (3),\ \mathbf{i}^3 = (4,1)$$

so $\mathscr{A} = \{1,2,3\}, \mathscr{B} = \emptyset$ and $\varepsilon = 3$. For $\ell = 3$ there is only one choice, so the contribution to $\mu(\sigma^5)$ is $-\frac{1}{4}$ in this case. For $\ell = 4$ we have $\zeta = 1$ and the partitions are $(0,1)$ and $(1,0)$, so the contribution from this term is

$$\frac{1}{5}(\rho(2,1)\cdot\rho(2,2)+\rho(2,2)\cdot\rho(2,1)) = 2/5.$$

Finally, for $\ell = 5$ we have $\zeta = 2$ and the partitions are $(2,0,)(0,2)$ and $(1,1)$. Hence the contribution here is

$$-\frac{1}{6}(\rho(2,3)\cdot\rho(2,1)+\rho(2,1)\cdot\rho(2,3)+\rho(2,2)\cdot\rho(2,2)) = -1/6$$

since $\rho(2,3) = 0$. Hence we have $\mu(\sigma^5) = -\frac{1}{4}+\frac{2}{5}-\frac{1}{6} = -\frac{1}{60}$.

Remark 3.3. *We obtain the same answer if we regard the singleton $\boldsymbol{i}^2 = (3)$ as increasing or decreasing. We have regarded it as increasing in this example.*

The explicit formula (3.15) in Theorem 3.3 for the solution of a general non-autonomous differential equation of the form (3.5) will now be applied to obtain some general results about such systems. Of course, the closer the matrices $A(t)$ are to commuting, the simpler will be the expressions for the solution. We shall see that in the nilpotent case, we can obtain finite, closed-form solutions. It can be shown [2] that, if $A(t)$ is analytic, so that we can write $A(t) = \sum_{i=0}^{\infty} t^i A_i$ for some matrices A_i, then $\mathfrak{L}_A$ is equal to the Lie algebra generated by the matrices $\{A_i : 0 \le i < \infty\}$. Suppose that $\{E_k : 1 \le k \le r\}$ is a basis of $\mathfrak{L}_A$, so that

$$A(t) = \sum_{k=1}^{r} g_k(t)E_k \tag{3.18}$$

for some functions $g_k, 1 \le k \le r$. Let c_{ij}^k be the structure constants of $\mathfrak{L}_A$, so that

$$[E_i, E_j] = \sum_{k=1}^{r} c_{ij}^k E_k$$

and so

$$[A(t), A(\tau)] = \sum_i \sum_j \sum_k c_{ij}^k g_i(t) g_j(\tau) E_k.$$

Then from Theorem 3.3 we have the following result which gives a simpler form for the general structure of the explicit solution:

Theorem 3.5. *If $A(t)$ is given by (3.18) then the solution of Equation 3.5 is given by*

$$
\begin{aligned}
x(t;x_0) &= exp\Bigg\{\sum_{k=1}^{r}\int_0^t g_k(\tau)d\tau E_k + \sum_{k=2}^{\infty}\sum_{\sigma^{k-1}\in S_{k-1}}\mu(\sigma^{k-1})\int_0^t\int_0^{\tau_k}\cdots\int_0^{\tau_3}\int_0^{\tau_2} \\
&\qquad \sum_{i_1}\cdots\sum_{i_k}\sum_{w}\sum_{v_{k-2}}\cdots\sum_{v_1} c_{i_1 v_{k-2}}^{w} c_{i_2 v_{k-3}}^{v_{k-2}}\cdots c_{i_{k-3}v_2}^{v_3} c_{i_{k-2}v_1}^{v_2} c_{i_{k-1}i_k}^{v_1} \\
&\qquad g_{i_1}(\tau_{\sigma^{k-1}(1)}) g_{i_2}(\tau_{\sigma^{k-1}(2)})\cdots g_{i_{k-1}}(\tau_{\sigma^{k-1}(k-1)}) g_{i_k}(\tau_k) E_w d\tau_1\cdots d\tau_k\Bigg\} x_0 \\
&= exp\Bigg\{\sum_{k=1}^{r}\int_0^t g_k(\tau)d\tau E_k + \sum_{k=2}^{\infty}\sum_{\sigma^{k-1}\in S_{k-1}}\mu(\sigma^{k-1})\int_0^t\int_0^{\tau_k}\cdots\int_0^{\tau_3}\int_0^{\tau_2} \\
&\qquad \sum_{i_1}\cdots\sum_{i_k}\sum_{w} C(w,i_1,\cdots,i_k) \\
&\qquad g_{i_1}(\tau_{\sigma^{k-1}(1)}) g_{i_2}(\tau_{\sigma^{k-1}(2)})\cdots g_{i_{k-1}}(\tau_{\sigma^{k-1}(k-1)}) g_{i_k}(\tau_k) E_w d\tau_1\cdots d\tau_k\Bigg\} x_0
\end{aligned}
\tag{3.19}
$$

where

$$
C(w,i_1,\cdots,i_k) = \sum_{v_{k-2}}\cdots\sum_{v_1} c_{i_1 v_{k-2}}^{w} c_{i_2 v_{k-3}}^{v_{k-2}}\cdots c_{i_{k-3}v_2}^{v_3} c_{i_{k-2}v_1}^{v_2} c_{i_{k-1}i_k}^{v_1}.
$$

As a specific example, consider the system with $so(3)$ as its Lie algebra:

$$
\begin{aligned}
\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= \begin{pmatrix} 0 & -g_3(t) & -g_2(t) \\ g_3(t) & 0 & -g_1(t) \\ g_2(t) & g_1(t) & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\
&= (g_1(t)M_1 + g_2(t)M_2 + g_3(t)M_3)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}
\end{aligned}
$$

where

$$
M_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.
$$

Here,

$$
[M_1,M_2] = M_3, \quad [M_2,M_3] = M_1, \quad [M_3,M_1] = M_2
$$

and we have the structure constants

$$
\begin{aligned}
c_{12}^3 &= c_{23}^1 = c_{31}^2 = -c_{21}^3 = -c_{32}^1 = -c_{13}^2 = -1 \\
c_{ij}^k &= 0 \text{ if } \{i,j,k\} \text{ is not a permutation of } 1,2,3.
\end{aligned}
$$

Hence,

$$c^i_{jk} = \varepsilon_{ijk}$$

where

$$\varepsilon_{ijk} = \begin{cases} 1 & \text{if } i,j,k \text{ is an even permutation of } 1,2,3 \\ -1 & \text{if } i,j,k \text{ is an odd permutation of } 1,2,3 \\ 0 & \text{otherwise} \end{cases}$$

(the standard tensorial ε-function), and so from the theorem, we have

$$\begin{aligned} x(t;x_0) &= exp\Bigg\{ \sum_{k=1}^{r} \int_0^t g_k(\tau) d\tau E_k + \sum_{k=2}^{\infty} \sum_{\sigma^{k-1} \in S_{k-1}} \mu(\sigma^{k-1}) \int_0^t \int_0^{\tau_k} \cdots \int_0^{\tau_3} \int_0^{\tau_2} \\ &\quad \sum_w \sum_{i_1=1}^{3} \cdots \sum_{i_k=1}^{3} \sum_{v_{k-2}=1}^{3} \cdots \sum_{v_1=1}^{3} \varepsilon_{wi_1v_{k-2}} \varepsilon_{v_{k-2}i_2v_{k-3}} \cdots \varepsilon_{v_3i_{k-3}v_2} \varepsilon_{v_2i_{k-2}v_1} \varepsilon_{v_1i_{k-1}i_k} \\ &\quad g_{i_1}(\tau_{\sigma^{k-1}(1)}) g_{i_2}(\tau_{\sigma^{k-1}(2)}) \cdots g_{i_{k-1}}(\tau_{\sigma^{k-1}(k-1)}) g_{i_k}(\tau_k) E_w d\tau_1 \cdots d\tau_k \Bigg\} x_0 \\ &= exp\Bigg\{ \sum_{k=1}^{r} \int_0^t g_k(\tau) d\tau E_k + \sum_{k=2}^{\infty} \sum_{\sigma^{k-1} \in S_{k-1}} \mu(\sigma^{k-1}) \int_0^t \int_0^{\tau_k} \cdots \int_0^{\tau_3} \int_0^{\tau_2} \\ &\quad \sum_{i_1} \cdots \sum_{i_k} \sum_w \Xi^w(i,v) \\ &\quad g_{i_1}(\tau_{\sigma^{k-1}(1)}) g_{i_2}(\tau_{\sigma^{k-1}(2)}) \cdots g_{i_{k-1}}(\tau_{\sigma^{k-1}(k-1)}) g_{i_k}(\tau_k) d\tau_1 \cdots d\tau_k \Bigg\} E_w x_0 \end{aligned}$$

where

$$\begin{aligned} \Xi^w(i,v) &= \varepsilon_{wi_1v_{k-2}} \varepsilon_{v_{k-2}i_2v_{k-3}} \cdots \varepsilon_{v_3i_{k-3}v_2} \varepsilon_{v_2i_{k-2}v_1} \varepsilon_{v_1i_{k-1}i_k} \\ &= \pm 1. \end{aligned}$$

In the case of systems with nilpotent Lie algebra, we get an explicit closed form

$$\begin{aligned} x(t;x_0) = exp\Bigg\{ &\sum_{k=1}^{r} \int_0^t g_k(\tau) d\tau E_k + \sum_{k=2}^{K} \sum_{\sigma^{k-1} \in S_{k-1}} \mu(\sigma^{k-1}) \int_0^t \int_0^{\tau_k} \cdots \int_0^{\tau_3} \int_0^{\tau_2} \\ &\sum_{i_1} \cdots \sum_{i_k} \sum_w \sum_{v_{k-2}} \cdots \sum_{v_1} c^w_{i_1v_{k-2}} c^{v_{k-2}}_{i_2v_{k-3}} \cdots c^{v_3}_{i_{k-3}v_2} c^{v_2}_{i_{k-2}v_1} c^{v_1}_{i_{k-1}i_k} \\ &g_{i_1}(\tau_{\sigma^{k-1}(1)}) g_{i_2}(\tau_{\sigma^{k-1}(2)}) \cdots g_{i_{k-1}}(\tau_{\sigma^{k-1}(k-1)}) g_{i_k}(\tau_k) E_w d\tau_1 \cdots d\tau_k \Bigg\} x_0 \end{aligned}$$

where K is the degree of nilpotency.

Example 3.2. Consider the system

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \left[\begin{pmatrix} -4 & -3 & 2 \\ 12 & 8 & 0 \\ 0 & 0 & 2 \end{pmatrix} cos\, t + \begin{pmatrix} -2 & -1 & 0 \\ 4 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} sin\, t \right.$$
$$\left. + \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} t^2 \right] \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad x(0) = x_0.$$

Put

$$F_1 = \begin{pmatrix} -4 & -3 & 2 \\ 12 & 8 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad F_2 = \begin{pmatrix} -2 & -1 & 0 \\ 4 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad F_3 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then F_1, F_2, F_3 form a basis of a nilpotent Lie algebra with

$$[F_1, F_2] = -2F_3$$

and all the other commutators zero. Hence

$$c_{12}^3 = c_{21}^3 = -2$$

are the only non-zero structure constants. It follows that the solution of the system is

$$x(t;x_0) = exp\left(\int_0^t \{F_1 cos\, \tau + F_2 sin\, \tau + F_3 \tau^2\} d\tau \right.$$
$$\left. - \frac{1}{2} \int_0^t \int_0^\tau (-2 cos\, \rho\, sin\, \tau - 2 sin\, \rho\, cos\, \tau) F_3 d\rho\, d\tau \right)$$
$$= exp\left(sin\, t F_1 + (1 - cos\, t) F_2 + \frac{t^3}{3} F_3 + sin\, t (1 - cos\, t) F_3 \right) x_0.$$

3.4 Stability Theory

In this section we shall discuss the stability of equations of the form (3.5). The most important point here is that, unlike the time-invariant case, the eigenvalues of $A(t)$ do not determine the stability of the system. Later we shall see to what extent one can use the eigenvalues and eigenvectors of $A(t)$ to prove stability. First, we begin with Lyapunov theory – we have the following simple generalisation of the time-invariant Lyapunov theorem:

Theorem 3.6. *The linear, time-varying system of Equations 3.5 is Lyapunov stable if for any positive definite, symmetric matrix Q, there is a positive definite matrix function $P(t)$ which satisfies the equation*

$$\dot{P}(t) = -A^T(t)P(t) - P(t)A(t) - Q \tag{3.20}$$

(for some positive definite initial matrix $P(0)$). Moreover, conversely, if $A(t)$ is continuous and the system is L^2-stable in the sense that $x(\cdot;x_0) \in L^2[0,\infty)$, then the Equation 3.20 has a positive definite solution $P(t)$, for some initial matrix $P(0)$.

Proof. The first part follows directly from Lyapunov's main stability theorem (see, *e.g.*, [3]), since

$$V(t) = x^T P(t)x$$

is clearly a Lyapunov function (for

$$\begin{aligned}\dot{V} &= \dot{x}^T Px + x^T P\dot{x} + x^T \dot{P}x = x^T(A^T P + PA + \dot{P})x \\ &= -x^T Qx).\end{aligned}$$

For the converse, note that by lemma 3.2, Equation 3.20 has a unique symmetric solution and so it remains to show that $P(t)$ is positive definite. Let $x(t;x_0) \in L^2[0,\infty)$ be a solution and take the inner product of (3.20) on the left and right with x:

$$x^T(t)\dot{P}(t)x(t) = -x^T(t)A^T(t)P(t)x(t) - x^T(t)P(t)A(t)x(t) - x^T(t)Qx(t).$$

Then

$$\frac{d}{dt}\langle x(t), P(t)x(t)\rangle = -\langle x(t), Qx(t)\rangle,$$

whence

$$\begin{aligned}\langle x(t), P(t)x(t)\rangle &= \langle x(0), P(0)x(0)\rangle - \int_0^t \langle x(s), Qx(s)\rangle ds \\ &\geq \langle x(0), P(0)x(0)\rangle - ||Q|| \int_0^t ||x(s)||^2 ds \\ &\geq \langle x(0), P(0)x(0)\rangle - ||Q|| \cdot ||x||_{L^2[0,\infty)}\end{aligned}$$

so the result follows if $P(0)$ is large enough. (A similar inequality using the positivity of Q shows that $P(t)$ is bounded. □

Unlike the time-invariant case, however, this result is not easy to apply in general, so we consider other approaches to stability. We can give a very simple sufficient condition in the case where $A(t)$ is continuous and $A(\infty) \doteq lim_{t\to\infty}A(t)$ exists.

Lemma 3.8. *The system (3.5) is asymptotically stable if the eigenvalues of $A(\infty)$ have negative real parts.*

Proof. Write the equation in the form

$$\dot{x} = A(\infty)x + (A(t) - A(\infty))x.$$

By continuity of $A(\infty)$, there exists $T > 0$ such that for $t \geq T$, we have

$$||A(t) - A(\infty)|| < \frac{\delta}{2}$$

where

$$||e^{A(\infty)}t|| \leq Me^{-\delta t}$$

for some $M > 0$ (by the eigenvalue condition).Hence we have

$$||x(t)|| \leq Me^{-\delta(t-T)}||x(T)|| + \int_T^t e^{-\delta(t-s)} \frac{\delta}{2}||x(s)||ds, \quad t \geq T,$$

and by Gronwall's inequality,

$$||x(t)|| \leq Me^{-\delta(t-T)}||x(T)|| \to 0$$

as $t \longrightarrow \infty$. The result now follows since $||A(t)||$ is bounded on $[0,T]$, by continuity. □

The converse to Lemma 3.8 is, of course, false, as shown by the example

$$\dot{x} = -\frac{1}{1+t}x, \quad x(0) = x_0$$

which has the solution

$$x(t) = \frac{1}{1+t}x_0.$$

Clearly, therefore, if some of the eigenvalues of $A(\infty)$ are zero, we may still have asymptotic stability. A partial converse can be given in the case of exponential stability, the proof of which is elementary:

Lemma 3.9. *If the system (3.5) is exponentially asymptotically stable, then the eigenvalues of $A(\infty)$ have negative real parts.*

Now assume that at least one eigenvalue of $A(\infty)$ is zero or, indeed, that $A(\infty)$ does not exist at all, and let, as before, $\mathscr{L}_A$ be the Lie algebra generated by the set

$$\{A(t) : t \in \mathbb{R}\}.$$

Let $\mathscr{C}$ be a Cartan subalgebra of $\mathscr{L}_A$ (which always exists; see [2]). Thus we can write

$$A(t) = A_1(t) + A_2(t), \quad t \in \mathbb{R}$$

where

$$A_1(t) \in \mathscr{C} \text{ and } A_2(t) \in \mathscr{L}_A \ominus \mathscr{C}, \quad t \in \mathbb{R}.$$

In particular, if $A(\infty)$ exists, then we have

$$A(\infty) = A_1(\infty) + A_2(\infty),$$

where

$$A_1(\infty) \in \mathscr{C} \text{ and } A_2(\infty) \in \mathscr{L}_A \ominus \mathscr{C}.$$

Since $\mathscr{C}$ is a maximal abelian subalgebra of $\mathscr{L}_A$ we can simultaneously diagonalise the elements of $\mathscr{C}$. (Since all Cartan subalgebras are related by a similarity transformation, the choice of $\mathscr{C}$ does not affect the study of stability of (3.5).) Let P be a diagonalising matrix for $\mathscr{C}$, *i.e.*

$$P^{-1}CP = \Lambda_C, \text{ for all } C \in \mathscr{C}$$

where Λ_C is the diagonal matrix of eigenvalues of C. Of course, P is independent of C.Thus we may write the Equation 3.5 in the form

$$\dot{y} = \Lambda_{A_1(t)} y + B_2(t) y \tag{3.21}$$

where

$$B_2(t) = P^{-1} A_2(t) P.$$

Write

$$\Lambda_{A_1(t)} = \begin{pmatrix} \lambda_1(t) & & \\ & \ddots & \\ & & \lambda_n(t) \end{pmatrix}.$$

Of course, if $Re\lambda_i(\infty) < 0$ for $i \in \{1, \cdots, n\}$ and $B_2(t)$ is 'small' enough then the Equation 3.21 is clearly stable. However, if $\lambda_i(\infty) = 0$ for some i, then stability is more difficult. The transition matrix of (3.21) is given by

$$\Phi(t,s) = \Phi(t,0)(\Phi(s,0))^{-1}$$

where

$$\Phi(t,0) = \begin{pmatrix} e^{\int_0^t \lambda_1(s)ds} & & \\ & \ddots & \\ & & e^{\int_0^t \lambda_n(s)ds} \end{pmatrix}.$$

The solution of (3.21) is

$$y(t) = \Phi(t,0) y_0 + \int_0^t \Phi(t,s) B_2(s) y(s) ds$$

and so

$$||y(t)|| \le ||\Phi(t,0)|| \cdot ||y_0|| + \int_0^t ||\Phi(t,s)|| \cdot ||B_2(s)|| \cdot ||y(s)|| ds.$$

Now

$$\begin{aligned}||\Phi(t,s)|| &= \left\| \begin{pmatrix} e^{\int_s^t Re\lambda_1(s)ds} & & \\ & \ddots & \\ & & e^{\int_s^t Re\lambda_n(s)ds} \end{pmatrix} \right\| \\ &= max_{1\le i\le n}\left(e^{\int_s^t Re\lambda_i(s)ds}\right) \\ &\le e^{\int_s^t max_{1\le i\le n}(Re\lambda_i(s))ds}.\end{aligned}$$

Let

$$\mu(s) = max_{1\le i\le n}(Re\lambda_i(s)).$$

Then

$$||y(t)|| \le e^{\int_0^t}\mu(s)ds||y_0|| + \int_0^t e^{\int_s^t \mu(\tau)d\tau}||B_2(s)||\cdot||y(s)||ds.$$

Using a standard argument now shows that

$$||y(t)|| \le e^{\int_0^t(\mu(s)+||B_2(s)||)ds}||y_0||,$$

and we get stability if

$$e^{\int_0^t(\mu(s)+||B_2(s)||)ds} \longrightarrow -\infty \text{ as } t \longrightarrow \infty.$$

Hence we have proved:

Theorem 3.7. *Let $\mathfrak{L}_A$ denote the Lie algebra generated by the matrices $\{A(t)\}_{t\ge 0}$ and let $\mathfrak{C}$ be a Cartan subalgebra of $\mathfrak{L}_A$. Let $\mu(t) = max_{1\le i\le n}(Re\lambda_i(t))$ where $\lambda_i(t), 1\le i\le n$ are the eigenvalues of the matrices of $\mathfrak{C}$ and suppose that*

$$\int_0^t (\mu(s) + ||B(s)||)ds \longrightarrow -\infty \text{ as } t \longrightarrow \infty$$

where $B(s) = P^{-1}A_2(t)P$. Then the system is asymptotically stable. Here, P diagonalizes $\mathfrak{C}$ and $A(t) = A_1(t) + A_2(t)$, where $A_1 \in \mathfrak{C}$.

Another approach to stability is via the logarithmic norm. If $||\cdot||$ denotes any induced norm on $n\times n$ matrices, we define the *measure* of A by

$$\mu(A) = lim_{h\to 0+}(||I+hA||-1)/h.$$

To show that $\mu(A)$ exists, for any matrix A, note that if we write

$$f(h) = (||I+hA||-1)/h,$$

then $f(kh)\le f(h)$, for any $k\in(0,1)$, so that f is decreasing with h. Also, $f(h)\ge (|1-h||A|||-1)/h \ge -||A||$, so f is bounded below. The measure $\mu(A)$ of a matrix A has a number of useful and elementary properties, listed below:

(a) $\mu(I) = 1, \mu(-I) = -1, \mu(0) = 0.$
(b) $-||A|| \leq -\mu(-A) \leq \mu(A) \leq ||A||.$
(c) $\mu(\alpha A) = \alpha\mu(A), \quad$ for all $\alpha \geq 0.$
(d) $\mu(A + \alpha I) = \mu(A) + \alpha, \quad$ for all $\alpha \in \mathbb{R}.$
(e) $\mu : \mathbb{C}^{n\times n} \to \mathbb{R}$ is a convex function :
$\mu(\lambda A + (1-\lambda)B) \leq \lambda\mu(A) + (1-\lambda)\mu(B), \quad 0 \leq \lambda \leq 1.$
(f) $|\mu(A-B)| \leq ||A-B||.$
(g) $-\mu(-A) \leq Re\lambda(A) \leq \mu(A),$ for any eigenvalue $\lambda(A)$ of A.
(h) $-\mu(-A)||x|| \leq ||Ax||, \quad -\mu(A)||x|| \leq ||Ax||, \quad x \in \mathbb{C}^n.$
(i) If $detA \neq 0,$
$-\mu(-A) \leq (||A||^{-1})^{-1} \leq ||A||.$

The importance of the measure of A is that, unlike a norm, it can be negative, so its main application is the following:

Theorem 3.8. *If $t \to A(t) : \mathbb{R} \to \mathbb{C}^{n\times n}$ is continuous, then the solution of Equation 3.5 satisfies the inequalities*

$$||x_0||exp\left\{-\int_{t_0}^{t}\mu(-A(t'))dt'\right\} \leq ||x(t)|| \leq ||x_0||exp\left\{\int_{t_0}^{t}\mu(A(t'))dt'\right\}.$$

Proof. Let $D^+||x(t)||$ denote the right-hand derivative. Then

$$\begin{aligned} D^+||x(t)|| &= lim_{h\to 0+}[||x(t+h)|| - ||x(t)||]/h \\ &= lim_{h\to 0+}[||x(t) + hA(t)x(t)|| - ||x(t)||]/h \\ &= lim_{h\to 0+}[||I + hA(t)|| \cdot ||x(t)|| - ||x(t)||]/h \\ &\to \mu(A(t)) \cdot ||x(t)||, \end{aligned}$$

and the second inequality follows by integration. The first is similar. □

Corollary 3.4. *If*

$$\int_{t_0}^{t}\mu(A(t'))dt' \to -\infty$$

as $t \to \infty$, then the system is asymptotically stable.

3.5 Lyapunov Exponents and Oseledec's Theorem

Lyapunov's characteristic numbers and exponents are important generalisations of the real parts of eigenvalues for linear, time-invariant systems. Here we give a brief summary of the ideas which lead to Oseledec's theorem on the general invariant decomposition of the state-space in terms of the 'eigenspaces' associated with the Lyapunov exponents. More details can be found in [9].

Let $f(t)$ be a continuous function defined on $[t_0,\infty)$. We define the *characteristic number* of f to be

$$\lambda = \lambda(f) = -\overline{lim}_{t\to\infty}\frac{ln|f(t)|}{t}. \tag{3.22}$$

It follows easily from the definition that

$$\begin{aligned}\overline{lim}_{t\to\infty}|f(t)|e^{(\lambda+\varepsilon)t} &= +\infty\\ lim_{t\to\infty}|f(t)|e^{(\lambda-\varepsilon)t} &= 0\end{aligned}$$

for any $\varepsilon > 0$. Note that, for any real number c,

$$\lambda(cf) = \begin{cases}\lambda(f) & \text{if } c \neq 0\\ +\infty & \text{if } c = 0.\end{cases} \tag{3.23}$$

We also need the following simple result; if the numbers $\lambda(f_i)$, $1 \le i \le n$ are distinct, then

$$\lambda(\sum_{i=1}^{n} f_i) = min_i(\lambda(f_i)). \tag{3.24}$$

We now consider the Lyapunov numbers associated with a linear, time-varying system of the form (3.5). For any vector $x(t)$ of continuous functions, we define

$$\lambda(x) = min_{1\le i\le n}\lambda(x_i).$$

The most important result is due to Perron:

Theorem 3.9. *The set of all possible Lyapunov numbers $\{\lambda(x)\}$ of all solutions of (3.5) contains at most n distinct values.*

Proof. We first show that, for any vector functions $y^{(i)}$, $1 \le i \le k$, if $\lambda_i \doteq \lambda(y^{(i)})$ are distinct, then the vector functions $y^{(i)}$ are linearly independent. Suppose they are not, so that

$$\sum_{i=1}^{k}\alpha_i y^{(i)}(t) = 0$$

for some scalars α_i, not all zero. Let $\lambda_q = min_i\lambda(\alpha_i y^{(i)})$. If we write $y^{(i)} = (y_{i1},\cdots,y_{in})$, in terms of the components, we may assume that

$$= \lambda(y_{q1})$$

and since the λ_i's are distinct we have

$$\lambda_q < \lambda(\alpha_i y_{i1}), \quad i \neq q.$$

Now, $\sum_{i=1}^{k} \alpha_i y_{i1}(t) = 0$, and so

$$\lambda_q = \lambda(\alpha_q y_{q1}) = \lambda\left(\sum_{i=1}^{k} \alpha_i y_{i1}(t)\right) = \lambda(0) = \infty$$

by (3.24) and (3.24), which contradicts the fact that λ_q is finite. The theorem now follows from the fact that there are at most n linearly independent solutions to a linear, time-varying system of equations. □

Corollary 3.5. *For the system (3.5), if $\Phi(t)$ denotes the transition function, then the numbers*

$$\lambda(x_0) = \overline{lim}_{t\to\infty} \frac{1}{t} ln||\Phi(t)x_0||$$

for all $x_0 \in \mathbb{R}^n$ take only k distinct values $\lambda_1, \cdots, \lambda_k$, where $1 \le k \le n$.

We can order the values in Corollary 3.5 as follows:

$$-\infty < \lambda_1 < \lambda_2 < \cdots < \lambda_k < \infty,$$

and we define the linear subspaces

$$V_i = \{x \in \mathbb{R}^n : x = 0 \text{ or } \lambda(x_0) \le \lambda_i\}$$

of $\mathbb{R}^n$. Then we clearly have

$$\{0\} \subseteq V_1 \subseteq \cdots \subseteq V_k = \mathbb{R}^n$$

is a filtration of $\mathbb{R}^n$. A system for which

$$\underline{lim}_{t\to\infty} \frac{1}{t} \ln det\Phi(t) = \sum_{i=1}^{k} d_i \lambda_i$$

where $d_i = \dim V_i - \dim V_{i-1}$ is called *Lyapunov regular*. This allows us to associate a decomposition of $\mathbb{R}^n$ with a system which generalises the eigendecomposition for a time-invariant system. To state the full version of Oseledec's theorem, we need first to introduce the ideas of integral invariants and ergodic measures. Let

$$\dot{x} = f(x,t) \tag{3.25}$$

be a differential equation defined in some region $\Omega \subseteq \mathbb{R}^n$ which has unique solutions, depending continuously on the initial conditions. A function $M : \mathbb{R}^{n+1} \to \mathbb{R}$ is called an *integral invariant* for (3.25) if

$$\int_{\Delta_t} M(x,t)dx$$

is constant, where Δ_0 is any domain in Ω and $\Delta_t = \phi(t;\Delta_0,0)$ where $\phi(t;x_0,t_0)$ is the solution of (3.25) through (x_0,t_0). It is easy to check that a necessary and sufficient condition for this is that

$$\frac{\partial M}{\partial t} + \text{div } (Mf) = 0. \tag{3.26}$$

Of course, if (3.25) is autonomous, then $M = M(x)$ is independent of t and then (3.26) becomes

$$div(Mf) = 0.$$

To extend this to invariant measures we shall assume a knowledge of standard measure theory (see, *e.g.* [6]). Let μ be a measure defined on Ω and assume that $\mu(\Omega) = 1$. We say that the measure μ is invariant for (3.25) if, for any μ-measurable set $S \subseteq \Omega$, we have

$$\mu\phi(t;S,t_0) = \mu(S), \text{ for all } t, t_0.$$

We then have Poincaré's invariance theorem:

Theorem 3.10. *Under the assumptions above, if $S \subseteq \Omega$ is measurable, with $\mu(S) > 0$, then for any $T > 0$, there exists $\tau > T$ such that*

$$\mu(S \cap \phi(t;S,t_0)) > 0.$$

The theorem of Khintchine also gives a lower limit to the inequality in theorem 3.10:

Theorem 3.11. *Continuing with the above assumptions, for any measurable set $S \subseteq \Omega$ with $\mu(S) = \sigma > 0$, we have*

$$\mu(S \cap \phi(t;S,t_0)) > \lambda\sigma^2$$

for a dense set of points $t \in \mathbb{R}$ and for any $\lambda > 1$.

(For a proof, see [9]).

Birkhoff's ergodic theorem shows that the time-average of a summable function evaluated on the trajectories of a system exist for all initial states (apart possibly from a set of measure zero).

Theorem 3.12. *If μ is an invariant measure on Ω with $\mu(\Omega) = 1$, then for any absolutely summable function g on Ω, the limit*

$$lim_{T\to\infty}\frac{1}{T}\int_0^T g(\phi(t;x_0,t_0))dt$$

exists for almost all x_0.

(For a proof, see [9]).

The results above assume the existence of a normalised invariant measure. The next theorem of Kryloff and Bogoliuboff shows that this always holds for any compact metric space.

Theorem 3.13. *In a compact metric phase space Ω, there exists an invariant (normalised) measure for any continuous dynamical system $\phi(t;x_0,t_0)$.*

Proof. Let $\tilde{\mu}$ be any normalised measure on Ω (we can assume that $\tilde{\mu}(\Omega) = 1$ by compactness). Consider the linear functional

$$L_\tau g = \frac{1}{\tau}\int_0^\tau dt \int_\Omega g(\phi(t;x_0,t_0))\tilde{\mu}(dx_0)$$

defined on $C(\Omega)$. By the Radon-Nikodym theorem, any positive linear functional can be represented by a measure:

$$\begin{aligned} L_\tau g &= \int_\Omega g(x)\mu_\tau(dx) \\ &= \frac{1}{\tau}\int_0^\tau dt \int_\Omega g(\phi(t;x_0,t_0))\tilde{\mu}(dx). \end{aligned}$$

The set of measures is compact and so there is a convergent subsequence of measures $\{\mu_{\tau_i}\}$ where $\tau_i \to \infty$. Let

$$\mu = lim_{i\to\infty}\mu_{\tau_i}.$$

The measure μ can be seen to be invariant as follows. We must show that

$$lim_{i\to\infty}\frac{1}{\tau_i}\int_0^{\tau_i} dt \int_\Omega g(\phi(t;x_0,t_0))\tilde{\mu}(dx_0) =$$
$$lim_{i\to\infty}\frac{1}{\tau_i}\int_0^{\tau_i} dt \int_\Omega g(\phi(t+t';x_0,t_0))\tilde{\mu}(dx_0)$$

for any $g \in C(\Omega)$ and all t, t'. This follows by Fubini's theorem and a simple norm estimate. □

Now let $G\ell(n)$ denote the set of all $n \times n$ (real) matrices with the usual operator norm and let $L^\infty(\mathbb{R}; G\ell(n))$ denote the set of bounded, measurable functions defined on $\mathbb{R}$ with values in $G\ell(n)$, with the weak* topology. If $A(\cdot) \in L^\infty(\mathbb{R}; G\ell(n))$, let A_s be the 'translate' by s:

$$A_s(t) = A(t+s), \quad \text{for all } t, s \in \mathbb{R}.$$

Clearly, this is a linear operator on $L^\infty(\mathbb{R}; G\ell(n))$ which we denote by T_s :

$$T_s(A)(t) = A(t+s). \tag{3.27}$$

Let $A(\cdot) \in L^\infty(\mathbb{R}; G\ell(n))$. Then we define the set

$$\Psi_A = \overline{\{A_s(\cdot) : s \in \mathbb{R}\}}^*$$

where the closure is in the weak*-topology. By Alaoglu's theorem, the set Ψ_A is compact and it is clearly an invariant set, so $\{T_s : s \in \mathbb{R}\}$ defines a flow on Ψ_A given by (3.27). For each $A(\cdot) \in \Psi_A$ we let $\Phi_A(t)$ denote the fundamental solution of the equation

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

Then we obtain a flow $\mathfrak{T}_s$ on $\Psi_A \times \mathbb{R}^n$ by setting

$$\mathfrak{T}_s(A(\cdot),x) = (T_s(A), \Phi_A(s)x)$$

called the *linear, skew-product flow* (see [11]).

Now let $\mathfrak{A} \subseteq L^\infty(\mathbb{R}; G\ell(n))$ be a shift-invariant weak*-compact set, let $\mathbb{G}(n,m)$ denote the Grassmann manifold of m-dimensional subspaces of $\mathbb{R}^n$ and let μ be an invariant normalised measure on $\mathfrak{A}$ (*i.e.* $\mu(\mathfrak{A}) = 1$). Then we say that a non-empty subset $X \subseteq \mathfrak{A} \times \mathbb{R}^n$ is a *measurable subbundle* of $\mathfrak{A} \times \mathbb{R}^n$ if:

(a) there exists $\mathfrak{A}_1 \subseteq \mathfrak{A}$ such that $\mu(\mathfrak{A}_1) = 1$.

(b) each fibre $\mathfrak{B}_A = \mathfrak{A}_1 \bigcap (\{A\} \times \mathbb{R}^n)$ is a subspace of $\mathbb{R}^n$, for all $A \in \mathfrak{A}_1$ of constant dimension m.

(c) the map $A \to \mathfrak{A}_1 \bigcap (\{A\} \times \mathbb{R}^n) = \mathfrak{B}_A$ from $\mathfrak{A}_1$ to $\mathbb{G}(n,m)$ is μ-measurable.
If $\Phi_A(t)\mathfrak{B}_A \subseteq \mathfrak{B}_{T_t(A)}$ for all $A \in \mathfrak{A}_1$ and all $t \in \mathbb{R}$, then $\mathfrak{A}$ is said to be *invariant*. We can now state Oseledec's theorem, which follows from the above considerations (see [10]):

Theorem 3.14. *Consider the family of linear, time-varying differential equations*

$$\dot{x} = A(t)x, \quad A \in \mathfrak{A} \subseteq L^\infty(\mathbb{R}; G\ell(n)), \tag{3.28}$$

and suppose that μ is a normalised invariant measure on $\mathfrak{A}$. Then there exist real numbers $\beta_1 < \beta_2 < \cdots < \beta_k$, $1 \le k \le n$, such that:

(a) there exists a basis $\{e_i\}$ of $\mathbb{R}^n$ such that the numbers

$$lim_{t\to\infty}\frac{1}{t}ln||\Phi_A(t)e_r||, \quad lim_{t\to-\infty}\frac{1}{t}ln||\Phi_A(t)e_r||$$

exist for all A in some shift-invariant set $\mathfrak{A}_1 \subseteq \mathfrak{A}$ with $\mu(\mathfrak{A}) = 1$ and belong to the set $\{\beta_i\}$.

(b) Let $\mathfrak{B}_r$ denote the set

$$\mathfrak{B}_r = \{(A,x) \in \mathfrak{A}_1 \times \mathbb{R}^n : x = 0 \ or \ lim_{t\to\infty}\frac{1}{t}ln||\Phi_A(t)x|| = \beta_r\}, \quad 1 \le r \le k.$$

Then each $\mathfrak{B}_r$ is a measurable invariant subbundle of $\mathfrak{A} \times \mathbb{R}^n$, and

$$\mathfrak{A}_1 \times \mathbb{R}^n = \mathfrak{B}_1 \oplus \mathfrak{B}_2 \oplus \cdots \oplus \mathfrak{B}_k.$$

Also, the limits are uniform in x.

(c) For each $A \in \mathfrak{A}_1$, the Equation 3.5 is regular in the sense of Lyapunov, i.e.

$$lim_{t\to\pm\infty} ln\ det\Phi_A(t) = \sum_{r=1}^{k} \beta_r \textit{ for all } A \in \mathfrak{A}_1.$$

(d) For each $A \in \mathfrak{A}_1$, the maximal Lyapunov exponent is β_k.

3.6 Exponential Dichotomy and the Sacker-Sell Spectrum

Consider again the linear, time-varying differential equation, with a locally integrable function $A(\cdot)$:

$$\dot{x} = A(t)x. \tag{3.29}$$

We say that the Equation 3.29 has an *exponential dichotomy* on an interval $I \subseteq \mathbb{R}$ if there exists a projection $P : \mathbb{R}^n \to \mathbb{R}^n$ and positive constants C and ε such that

$$\begin{aligned} ||\Phi(t)P\Phi(s)^{-1}|| &\leq Cexp(-\varepsilon(t-s)), \quad (t \geq s) \\ ||\Phi(t)(I-P)\Phi(s)^{-1}|| &\leq Cexp(\varepsilon(t-s)), \quad (t \leq s) \end{aligned} \tag{3.30}$$

for all $s,t \in I$. We shall state a number of results, the proofs of which can be found, for example, in [4]. First note that if $A(\cdot) = A$ is constant then (3.29) has an exponential dichotomy if and only if $\sigma(A)$ (the spectrum of A) has no elements with zero real part. Also, Equation 3.29 for time-varying $A(\cdot)$ has an exponential dichotomy with $P = I$ (=identity) if and only if it is uniformly asymptotically stable. (If $\varepsilon = 0$ in (3.30) we speak of an *ordinary exponential dichotomy*.) An important aspect of the spectral theory of systems is reducibility, such as the existence of invariant subbundles studied above. In the present case, we say that the system (3.29) is *reducible* if it is similar to a system of the form

$$\dot{y} = \begin{pmatrix} B_1(t) & 0 \\ 0 & B_2(t) \end{pmatrix} y = B(t)y \tag{3.31}$$

where B_1, B_2 are of lower order than B. By 'similar' here we mean that there exists an invertible continuously differentiable and bounded matrix function $S(t)$ such that the change of variable $x = S(t)y$ maps (3.29) into (3.30). It is easy to see that $S(t)$ must satisfy the differential equation

$$\dot{S}(t) = A(t)S(t) - S(t)B(t).$$

The main criterion for exponential dichotomy is given by:

Theorem 3.15. *([4]) Suppose that $A(t)$ is a bounded continuous matrix function defined on the interval $I \subseteq \mathbb{R}$ such that it has k eigenvalues with negative real part $\alpha < 0$ and $n-k$ eigenvalues with positive real part $\beta > 0$ for all $t \in I$. Then for any positive constant $\varepsilon < min(-\alpha, \beta)$, there exists $\delta > 0$ (depending on M, α, β and ε, where $M = sup_{t\in I}||A(t)||$) such that if*

$$||A(t_2)-A(t_1)|| \leq \delta \, for \; |t_1-t_2| \leq length(I)$$

the fundamental matrix $X(t)$ of (3.29) satisfies

$$||X(t)PX^{-1}(s)|| \leq Ke^{-(-\alpha-\varepsilon}(t-s), \quad t \geq s$$
$$||X(t)(I-P)X^{-1}(s)|| \leq Le^{-(\beta-\varepsilon}(t-s), \quad s \geq t$$

where K, L are constants and P is the projection

$$P = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}.$$

Exponential dichotomy of (3.29) is related to the existence of bounded solutions of the inhomogeneous equation

$$\dot{y} = A(t)y + f(t). \tag{3.32}$$

Let

$$\mathfrak{M} = \{f : f \text{ is locally integrable and } \int_t^{t+1} |f(s)|ds \text{ is bounded in } t\}$$

with norm

$$||f||_{\mathfrak{M}} = \text{sup}_{t\geq 0} \int_t^{t+1} |f(s)|ds.$$

Then $\mathfrak{M}$ is a Banach space and we have:

Theorem 3.16. *([4]) The inhomogeneous equation in (3.6) has at least one bounded solution for every $f \in \mathfrak{M}$ if and only if the homogeneous equation (3.29) has an exponential dichotomy.*

Note also that exponential dichotomy is stable in the sense that small perturbations of the coefficient matrix do not change the exponential dichotomy.

The concept of dichotomy can be extended to sets of differential equations. Thus if the system $\dot{x} = A(t)x$ has an exponential dichotomy on all of $\mathbb{R}$, then so does the equation $\dot{y} = B(t)y$ for all $B \in \Psi_A = \overline{\{A_s(\cdot) : s \in \mathbb{R}\}}^*$. Thus we say, more generally, that if $\Psi \in L^\infty(\mathbb{R}; G\ell(n))$ is a weak* compact, translation invariant set, then the family of equations $\dot{y} = B(t)y$ for $B \in \Psi$ has an *exponential dichotomy* if there exist constants K, ε and a continuous projection-valued function $B \to P_B$ such that

$$||\Phi_B(t) \cdot P_B \cdot \Phi_B(s)^{-1}|| \leq K\text{exp}(-\varepsilon(t-s)), \quad t \geq s$$
$$||\Phi_B(t) \cdot (I-P_B) \cdot \Phi_B(s)^{-1}|| \leq K\text{exp}(\varepsilon(t-s)), \quad t \leq s$$

for all $t, s \in \mathbb{R}$, $B \in \Psi$. From the definition we have

$$\Psi_A \times \mathbb{R}^n = W^s \oplus W^u$$

where the invariant subbundles W^s and W^u are given by

$$W^{s(u)} = \{(A,x) \in \Psi_A \times \mathbb{R}^n : x \in \mathrm{Im} P_A \, (x \in \mathrm{Ker}\ P_A)\}.$$

We now introduce the Sacker-Sell spectrum of the system of equations

$$\dot{x} = A(t)x \tag{3.33}$$

for $A(\cdot) \in \Psi \subseteq L^\infty(\mathbb{R}; G\ell(n))$ (a weak*-compact translation invariant set). Then we say that a point $\lambda \in \mathbb{R}$ is in the *Sacker-Sell spectrum* of the equations (3.33) for $A \in \Psi$ if the associated equations

$$\dot{x} = (\lambda I + A(t))x$$

do not admit an exponential dichotomy over Ψ.

Theorem 3.17. *(See [11].) The Sacker-Sell spectrum of equations (3.33) is a finite union of compact intervals (some of which may be single points).*

Let $\cup_{i=1}^{L}[a_i, b_i]$ be the Sacker-Sell spectrum of (3.33). Define the sets

$$W_i = \{(A,x) \in \Psi_A \times \mathbb{R}^n : x = 0 \text{ or } \lim_{t \to \pm\infty} \begin{matrix}\sup\\ \inf\end{matrix} \left(\frac{1}{t} \ln ||\Phi_A(t)x|| \right) \in [a_i, b_i]\}.$$

Then it can be shown (see [11]) that

$$\Psi \times \mathbb{R}^n \cong \oplus_{i=1}^{L} W_i$$

in the sense that W_i is a continuous invariant subbundle of $\Psi \times \mathbb{R}^n$. It will be seen later that we can use the iteration theory of Chapter 1 to generalise these results to nonlinear systems.

3.7 Conclusions

In this chapter we have considered the general theory of linear, time-varying systems, including existence and uniqueness of solutions, explicit expressions for the solutions and stability theory. The general theory of eigendecomposition leading to Oseldelec's theorem and exponential dichotomy has also been covered. In the following chapters we will extend the ideas of this chapter to nonlinear systems using the iteration scheme developed in Chapter 2.

References

1. Banks, S.P.: Nonlinear Delay Systems, Lie Algebras and Lyapunov Transformations. IMA J. Math. Contr. Inf. 19, 59–72 (2002)
2. Banks, S.P., McCaffrey, D.: Lie Algebras, Structure of Nonlinear Systems and Chaotic Motion. Int. J. Bifurcation and Chaos 8(7), 1437–1462 (1998)
3. Coppel, W.A.: Stability and Asymptotic Behaviour of Differential Equations. Heath, Boston (1965)
4. Coppel, W.A.: Dichotomies in Stability Theory, vol. 629. Springer, New York (1978)
5. Granas, A., Dugundji, J.: Fixed Point Theory. Springer, New York (2003)
6. Halmos, P.R.: Measure Theory. Van Nostrand, New York (1950)
7. Magnus, W.: On the Exponential Solution of Differential Equations for a Linear Operator. Comm. Pure App. Math. 7, 649–673 (1954)
8. Miller, W.: Symmetry Groups and their Applications. Academic Press, New York (1972)
9. Nemytskii, V.V., Stepanov, V.V.: Qualitative Theory of Differential Equations. Dover Pubs., New York (1989)
10. Oseledec, V.: A Multiple Ergodic Theorem, Lyapunov Characteristic Exponents for Dynamical Systems. Trans. Moscow Math. Soc. 19, 197–231 (1968)
11. Sacker, R., Sell, G.: Dichotomies and Invariant Splittings for Linear Differential Equations I. J. Diff. Eqns. 15, 429–458 (1974)
12. Wei, J., Norman, E.: On Global Representations of the Solutions of Linear Differential Equations as a Product of Exponentials. Phys. Lett. 52, 327–334 (1964)

Chapter 4
General Spectral Theory of Nonlinear Systems

4.1 Introduction

The spectral theory of linear (time-invariant) systems is, of course, the most important aspect of control theory in classical feedback design, and historically this was the approach taken by control engineers until the introduction of state-space theory. The techniques developed in the past include Nyquist and Bode diagrams, pole assignment and root locus methods. Frequency domain methods are therefore extremely important, especially for the suppression of resonant vibrations in mechanical systems.

In this chapter we shall outline a general spectral theory for nonlinear systems. First, a generalised transform theory is developed which can generate Volterra series kernels directly using Schwartz' kernel theorem, without the need of a (somewhat arbitrary) definition of multi-dimensional Laplace transforms. This theory can then be directly applied, by using the iteration scheme developed in this book to derive a general spectral theory for nonlinear systems. We shall then briefly show how to apply the same ideas to generalise exponential dichotomies and the Sacker-Sell spectrum to nonlinear systems. We shall assume a basic knowledge of functional analysis and distribution theory.

4.2 A Frequency-domain Theory of Nonlinear Systems

In this section we outline the methods of [1] and the resulting frequency-domain theory of nonlinear systems. The systems which we consider initially are the bilinear ones of the form

$$\dot{x} = Ax + uDx + bu, x(0) = x_0.$$

These systems are usually studied by means for the Volterra series (see, *e.g.* [2]), a typical term of which is of the form

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 61–74.
springerlink.com

$$\int_0^1 \int_0^{\tau_1} \cdots \int_0^{\tau_k} K_k(t,\tau_1,\cdots,\tau_k)u(\tau_1)\cdots u(\tau_k)d\tau_1\cdots d\tau_k,$$

where the kernel K_k has a number of different representations. In order to make this look like a k-dimensional convolution, a number of subtle transformations are made and then k-dimensional Laplace transforms are taken (see [2]). In this section we shall use a direct way of obtaining these results which are valid for a much more general class of inputs.

We begin by defining a very general frequency domain theory for nonlinear systems as given in [3]. Let $L_T^2[0,\infty)$ be the Hilbert space of all measurable, square-integrable, real-valued functions defined on $[0,\infty)$ and which are zero for $t > T$. This is clearly a direct subspace of $L^2[0,\infty)$. Let S be a nonlinear causal system, *i.e.* S maps $L_T^2[0,\infty)$ to itself for all $T > 0$. (For input-output stable systems, we can take $T = \infty$, *i.e.* $[0,\infty)$.) Let $\mathfrak{F}$ denote the Fourier transform, so that $\mathfrak{F}$ is an isomorphism

$$\mathfrak{F} : L^2(-\infty,\infty) \longrightarrow L^2(-\infty,\infty)$$

and $\mathfrak{F}$ maps $L_T^2[0,\infty)$ one-to-one and isometrically onto a subspace $\widetilde{L}_T^2[0,\infty)$ of $L^2(-\infty,\infty)$. We define the *transformed system* $\widetilde{S}$ by

$$\widetilde{S}(v) = \mathfrak{F}S\mathfrak{F}^{-1}(v).$$

Thus, $\widetilde{S}$ makes the diagram

$$\begin{tikzcd}
L_T^2[0,\infty) \arrow[r, "S"] \arrow[d, "\mathfrak{F}"'] & L_T^2[0,\infty) \arrow[d, "\mathfrak{F}"] \\
\widetilde{L}_T^2[0,\infty) \arrow[r, "\widetilde{S}"] & \widetilde{L}_T^2[0,\infty)
\end{tikzcd}$$

commute. Note that if S is an analytic function on $\widetilde{L}_T^2[0,\infty)$ (*i.e.* has a convergent Taylor series consisting of S and its Fréchet derivatives), then $\widetilde{S}$ is analytic on $\widetilde{L}_T^2[0,\infty)$, since $\mathfrak{F}$ is linear and invertible. Thus we can expand $\widetilde{S}$ in a Taylor series:

$$\widetilde{S} = \sum_{i=0}^{\infty} M_i(v)$$

where M_i is an i form defined on $\widetilde{L}_T^2[0,\infty)$, *i.e.* $M_i = L_i(v,\cdots,v)$ for some multilinear form $L_i : \bigoplus_{j=1}^{i} \widetilde{L}_T^2[0,\infty)$.

Example 4.1. Consider the linear system given by

$$S : y(t) = \int_0^t g(t-\tau)u(\tau)d\tau.$$

Then,

$$\widetilde{S}: Y(i\omega) = G(i\omega)U(i\omega)$$

so that $\widetilde{S}$ is simply multiplication by $G(i\omega)$ or in integral form

$$Y(i\omega) = \int_{\infty}^{\infty} \delta(\omega - \omega')G(i\omega')U(i\omega')d\omega'$$

with kernel distribution $\delta(\omega - \omega')G(i\omega')$.

Example 4.2. Consider the causal scalar distributed bi-linear system

$$\dot{x} = Ax + uDx + bu, x(0) = x_0, x \in L^2(\Omega). \tag{4.1}$$

The solution is given by the Volterra series

$$\begin{aligned} x(t) = {} & e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)ds + \\ & \sum_{i=1}^{\infty} \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_{i-1}} e^{A(t-\tau_1)}De^{A(\tau_1-\tau_2)}D \cdots \\ & \cdots De^{A(\tau_{i-1}-\tau_{i-2})}De^{A\tau_i}x_0u(\tau_1)u(\tau_2)\cdots u(\tau_i)d\tau_1\cdots d\tau_i \\ & + \sum_{i=1}^{\infty} \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_{i-1}} \int_0^{\tau_i} e^{A(t-\tau_1)}De^{A(\tau_1-\tau_2)}D \cdots \\ & \cdots De^{A(\tau_{i-1}-\tau_{i-2})}De^{A(\tau_i-\tau_{i+1})}bu(\tau_1)u(\tau_2)\cdots u(\tau_{i+1})d\tau_1\cdots d\tau_{i+1} \end{aligned}$$

where e^{At} is the semigroup generated by A. The usual procedure for generating the frequency-domain kernels is then quite subtle and requires consideration of different types of kernel (*e.g.* triangular ones) – see [2]. Here we show that the theory can be derived in a more general way and such that the resulting expression is true for any input. Thus, consider the system (4.1) again and take the one-sided Fourier transform (assuming the system is causal):

$$i\omega X(i\omega) = AX(i\omega) + \frac{1}{2\pi}(U * DX)(i\omega) + bU(i\omega) + x(0),$$

where $*$ denotes convolution. Hence, if $i\omega$ does not belong to $\sigma(A)$, we have

$$\left(I - \frac{1}{2\pi}(i\omega - A)^{-1}U * D\right)X(i\omega) = (i\omega - A)^{-1}(bU(i\omega) + x(0)).$$

We shall suppose that A is a closed sectorial operator (see [4]), with dense domain, so that

$$||(i\omega - A)^{-1}|| \le \frac{M}{|i\omega - a|}$$

for some real a and for all $i\omega$ in the sector

$$S_{a,\phi} = \lambda : \phi \le |arg(\lambda - a)| \le \pi, \lambda \ne a$$

of the complex plane, where $\phi \in (0,\pi/2)$. Thus, if $U \in L^1(0,\infty)$, then the operator V defined by

$$VX = \frac{1}{2\pi}(i\omega - A)^{-1} \circ (U * (DX))$$

satisfies

$$\begin{aligned}
||VX||_{L^2(0,\infty,L^2(\Omega))} &\le \frac{1}{2\pi}\left\| ||(i\omega - A)^{-1}||_{\mathscr{W}} ||U * (DX)||_{L^2(\Omega)} \right\|_{L^2(0,\infty)} \\
&\le \frac{1}{2\pi}\left\| ||(i\omega - A)^{-1}||_{\mathscr{W}} |U| * ||D||_{\mathscr{W}} ||X||_{L^2(\Omega)} \right\|_{L^2(0,\infty)} \\
&\le \frac{1}{2\pi}\left\| \frac{M}{|i\omega - A|}\left(|U| * \|D\|_{\mathscr{W}} \|X\|_{L^2(0,\infty)}\right) \right\| \\
&\le \frac{M}{2\pi}\left\| \frac{1}{|i\omega - A|} \right\|_{L^2(0,\infty)} \|U\|_{L^1(0,\infty)} \|D\|_{\mathscr{W}} \|X\|_{L^2(0,\infty,L^2(\Omega))} \\
&\le \frac{M}{4|a|} \|U\|_{L^1(0,\infty)} \|D\|_{\mathscr{W}} \|X\|_{L^2(0,\infty,L^2(\Omega))}
\end{aligned}$$

by Young's inequality [1] (where $\mathscr{W} = \mathscr{L}(L^2(\Omega))$), so that we obtain the following norm estimate for V:

$$\|V\| \le \frac{M}{4|a|} \|U\|_{L^1(0,\infty)} \|D\|_{\mathscr{L}(L^2(\Omega))} .$$

Hence, if we assume that the input is bounded in the L^1 norm in by a constant depending on the operator D and the number a,

$$\|U\|_{L^1(0,\infty)} \le \frac{4|a|}{M\|D\|} \tag{4.2}$$

then

$$\|V\| \le 1.$$

Hence, by the Neumann series (see [4]), we have

$$X = (I + V + V^2 + V^3 + \cdots)(i(\cdot) - A)^{-1}(bU(\cdot) + x(0)).$$

Now we have the following equations which are to be interpreted in the sense of distributions

[1] Young's inequality is $\|f * g\|_{L_r} \le | f\|_{L_p}\|g\|_{L_q}$ where $1/r = 1/p + 1/q - 1$.

$$\begin{aligned}
V^p(i\omega - A)^{-1}bU &= \frac{1}{2\pi}V^{p-1}(i\omega - A)^{-1}\int_{-\infty}^{\infty} U(\omega-\omega_1)D(i\omega_1 - A)^{-1}bU(\omega_1)d\omega_1 \\
&= \frac{1}{(2\pi)^2}V^{p-2}(i\omega - A)^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(\omega-\omega_2)U(\omega_2-\omega_1)U(\omega_1) \\
&\quad \times (i\omega_2 - A)^{-1}D(i\omega_1 - A)^{-1}bd\omega_1 d\omega_2 \\
&\quad \cdots \\
&= \frac{1}{(2\pi)^p}(i\omega - A)^{-1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} U(\omega-\omega_p)U(\omega_p-\omega_{p-1})\cdots \\
&\quad U(\omega_2-\omega_1)U(\omega_1)\times D(i\omega_p - A)^{-1}D\cdots \\
&\quad D(i\omega_1 - A)^{-1}bd\omega_1\cdots d\omega_p \\
&= \frac{1}{(2\pi)^p}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \delta(\omega-\omega_p)U(\omega_p-\omega_{p-1})U(\omega_{p-1}-\omega_{p-2}) \\
&\quad \cdots U(\omega_2-\omega_1)U(\omega_1)(i\omega_p - A)^{-1}D \\
&\quad \times (i\omega_{p-1} - A)^{-1}D\cdots(i\omega_1 - A)^{-1}bd\omega_1\cdots d\omega_{p+1}.
\end{aligned}$$

This gives the p^{th} kernel

$$\begin{aligned}
K_p(\omega_1,\cdots,\omega_{p+1}) &= \frac{1}{(2\pi)^p}\delta(\omega_{p+1}-\omega_p)(i\omega_p - A)^{-1}D(i\omega_{p-1} - A)^{-1} \\
&\quad \times D\cdots D(i\omega_1 - A)^{-1}b.
\end{aligned} \tag{4.3}$$

Of course, this approach required the condition (4.2); however, the solution may be extended by infinite-dimensional analytic continuation. It does not require any specific structure on the inputs (such as step functions, sinusoids, etc.).

In fact, the above result is a special case of a general theorem on the kernel representations of distributions which follows from the classical Schwartz' kernel theorem:

Theorem 4.1. *(See [5]) The space of distributions $\mathscr{D}'(X\times Y)$ on the product space $X\times Y\subseteq\mathbb{R}^m\times\mathbb{R}^n$ is isomorphic to the set of continuous linear maps $\{K : C_c^\infty(Y)\longrightarrow \mathscr{D}'(X)\}$.*

The distribution $K(x,y)\in\mathscr{D}'(X\times Y)$ corresponds to $\overline{K} : C_c^\infty(Y)\longrightarrow\mathscr{D}'(X)$ under this isomorphism, where $\overline{K}v$ is the distribution on X given by

$$C_c^\infty(X)\ni u\longrightarrow\langle K(x,y),u(x)v(y)\rangle,$$

which is usually written

$$(\overline{K}v)(x) = \int K(x,y)v(y)dy.$$

(For example, $\delta(x-y) \in \mathscr{D}'(X \times Y)$ and

$$\int \delta(x-y)v(y)dy = v(x)$$

and so $\delta(x-y)$ is associated with the natural injection $C_c^\infty(Y) \hookrightarrow \mathscr{D}'(X)$.)

Now consider the analytic input-output map

$$S : L_T^2[0,\infty) \longrightarrow L_T^2[0,\infty)$$

and its (analytic) Fourier transform

$$\widetilde{S} : \widetilde{L}_T^2[0,\infty) \longrightarrow \widetilde{L}_T^2[0,\infty).$$

Writing $\widetilde{S}$ in a Taylor series gives

$$\widetilde{S}(v) = \sum_0^\infty M_i(v)$$

where

$$M_i(v) = \frac{F^i\widetilde{S}(0)}{i!} v^{(i)}$$

and F is the Fréchet derivative. Here, $v^{(i)} = (v, \cdots, v)$ (i components). Thus, M_i determines a symmetric multi-linear form

$$\widetilde{M}_i(v_1, \cdots, v_i) \in \mathscr{L}(\Lambda, \mathscr{L}(\Lambda, \cdots, \mathscr{L}(\Lambda, \Lambda) \cdots))$$

where $\Lambda = \widetilde{L}_T^2[0,\infty)$.

By induction from Theorem 4.1, we have:

Theorem 4.2. *$\widetilde{M}_i$ can be represented by a kernel distribution*

$$K_i \in \mathscr{D}'([0,\infty) \times [0,\infty) \times \cdots \times [0,\infty)).$$

For example, M_2 is given by

$$\widetilde{M}_2(v_1, v_2)(\omega) = \int K_2(\omega, \omega_1, \omega_2) v_1(\omega_1) v_2(\omega_2) d\omega_1 d\omega_2,$$

for some $K_2 \in \mathscr{D}'([0,\infty) \times [0,\infty) \times [0,\infty))$, since $K(\omega_1, \omega_2; v)$ induces a map $K' : C_c^\infty([0,\infty)) \longrightarrow \mathscr{D}'([0,\infty) \times [0,\infty))$ given by

$$K'(v)(\omega_1, \omega_2) = K(\omega_1, \omega_2; v).$$

In the linear case, we have

$$K_1(\omega_1, \omega_2) = \delta(\omega_1 - \omega_2) G(i\omega_1)$$

and in the bi-linear one, (2.3) gives the kernel

$$K_p(\omega_1,\cdots,\omega_{p+1})=\delta(\omega_{p+1}-\omega_p)(i\omega_p-A)^{-1}D(i\omega_{p-1}-A)^{-1}D\cdots D(i\omega_1-A)^{-1}b.$$

(This differs by the factor $\frac{1}{(2\pi)^p}$ which can be absorbed into the input terms u.)

We can now see how this works in the case of time-varying bi-linear systems of the form

$$\dot{x}=A(t)x+uD(t)x+b(t)u,x(0)=x_0\in L^2(\Omega),$$

where $A(t)$ is a sectorial operator for each $t\geq 0$ and $D(t)$ and $b(t)$ are bounded operators. Since the convolution algebra is associative we can take the Fourier transform of the equation and write it in the form

$$i\omega X(i\omega)=(\widetilde{A}*X)(i\omega)+\frac{1}{(2\pi)^2}(U*\widetilde{D}*X)(i\omega)+\frac{1}{2\pi}(\widetilde{b}*U)(i\omega)+x(0) \quad (4.4)$$

where

$$\widetilde{A}=\mathfrak{F}(A(t)),\widetilde{D}=\mathfrak{F}(D(t))\text{ and }\widetilde{b}=\mathfrak{F}(b(t)).$$

We assume that $\widetilde{A}$ exists in the strong sense, *i.e.* that

$$\int_0^\infty A(t)ve^{-i\omega t}dt\in L^2(\Omega)$$

for all $v\in\bigcap_{t\geq 0}\mathscr{D}(A(t))$. Consider first the equation

$$i\omega X(i\omega)-(\widetilde{A}*X)(i\omega)=S(i\omega)$$

for some given $S(i\omega)$. Define the operator Γ by

$$(\Gamma X)(i\omega)=i\omega X(i\omega)-(\widetilde{A}*X)(i\omega).$$

Of course, in the time-domain we have

$$\dot{x}=A(t)x+S(t),x(0)=0,$$

so that

$$x(t)=\int_0^t\Phi(t,\tau)s(\tau)d\tau$$

where Φ is the convolution operator generated by $A(t)$. Thus,

$$\|x(t)\|_{L^2(\Omega)}\leq\int_0^t\|\Phi(t,\tau)\|_{\mathscr{L}(L^2(\Omega))}\|s(\tau)\|d\tau.$$

We assume that Φ is exponentially bounded, *i.e.*

$$\|\Phi(t,\tau)\|_{\mathscr{L}(L^2(\Omega))}\leq Ce^{\theta(t-\tau)},$$

for some $C>0$ and some real θ. If $\varepsilon>\theta$, we have

$$e^{-\varepsilon t}\|x(t)\|_{L^2(\Omega))} \le C\int_0^t e^{(-\varepsilon+\theta)(t-\tau)}e^{-\varepsilon\tau}\|s(\tau)\|d\tau.$$

Let $w(t)=e^{-\varepsilon t}$ and let $H_w = L^2_w([0,\infty);L^2(\Omega))$, the weighted L^2 space of all measurable functions $x(t)$ such that

$$\|x\|_{H_w} = \|x(t)\|_{L^2_w([0,\infty);L^2(\Omega))} = \left(\int_0^\infty e^{-2\varepsilon t}\|x(t)\|^2_{L^2(\Omega)}dt\right)^{1/2} < \infty.$$

Then $\mathfrak{F}$ is an isomorphism from H_w to $\widetilde{H}_w$ and by Parseval's theorem,

$$\|x\|_{H_w} = \left(\int_{-\infty}^\infty X^2_\varepsilon(\omega)d\omega\right)^{1/2} = \|X_\varepsilon\|_{\widetilde{H}_w}$$

where

$$X_\varepsilon(i\omega) = \int_0^\infty e^{-\varepsilon t}\|x(t)\|_{L^2(\Omega)}e^{-i\omega t}d\omega.$$

Hence we have

$$\left\|\Gamma^{-1}\right\|_{\mathscr{L}(\widetilde{H}_w)} \le C\int_0^\infty e^{(-\varepsilon+\theta)t}dt.$$

Returning to (4.1), we have

$$\left(I-\frac{1}{(2\pi)^2}\Gamma^{-1}U*\widetilde{D}*\right)X = \frac{1}{2\pi}\Gamma^{-1}\widetilde{b}*U+\Gamma^{-1}x(0).$$

If $X\in\widetilde{H}_w$ then we have

$$\|KX\|_{\widetilde{H}_w} \le \frac{1}{(2\pi)^2}\left\|\Gamma^{-1}\right\|_{\mathscr{L}(\widetilde{H}_w)}\|U\|_{L^1(-\infty,\infty)}\left\|\widetilde{D}\right\|_{L^1(-\infty,\infty;\mathscr{L}(L^2(\Omega)))}\|X\|_{\widetilde{H}_w}$$

where

$$K = \frac{1}{(2\pi)^2}\Gamma^{-1}U*\widetilde{D}*,$$

and so if

$$\frac{1}{(2\pi)^2}\frac{C}{\varepsilon-\theta}\|U\|_{L^1(-\infty,\infty)}\left\|\widetilde{D}\right\|_{L^1(-\infty,\infty;\mathscr{L}(L^2(\Omega)))} < 1$$

we have

$$\|KX\|_{\widetilde{H}_w} < 1.$$

Hence, as before, we have the Neumann series

$$X = (I+K+K^2+\cdots)\left(\frac{1}{2\pi}\Gamma^{-1}\widetilde{b}*U+\Gamma^{-1}x(0)\right).$$

Consider the general term

$$\xi_p = \frac{1}{2\pi}K^p\left(\Gamma^{-1}\widetilde{b}*U\right).$$

Note that we have the maps

$$C_c^\infty([0,\infty)\otimes L^2(\Omega)) \hookrightarrow L_w^2([0,\infty);L^2(\Omega)) \xrightarrow{\Gamma^{-1}} L_w^2([0,\infty);L^2(\Omega)) \hookrightarrow \\ \mathscr{D}'([0,\infty)\otimes L^2(\Omega)).$$

Each map is continuous and so, by the kernel theorem, their composition is given by a kernel. Since the first and last maps are injections, this gives a kernel representation of Γ^{-1}. We write this as

$$(\Gamma^{-1}X)(\omega) = \int_{-\infty}^{\infty} \gamma(\omega,\omega_1)X(i\omega_1)d\omega_1,$$

where $\gamma \in \mathscr{D}'(([0,\infty)\times[0,\infty))\otimes L^2(\Omega))$. We have

$$\begin{aligned}
\xi_p &= \frac{1}{(2\pi)^3}K^{p-1}\Gamma^{-1}(U*\widetilde{D}*)\Gamma^{-1}\widetilde{b}*U \\
&= \frac{1}{(2\pi)^3}K^{p-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\gamma(\omega,\omega_5)\widetilde{D}(\omega_5-\omega_4-\omega_3)\gamma(\omega_3,\omega_2) \\
&\quad \times\widetilde{b}(\omega_2-\omega_1)U(\omega_1)U(\omega_4)d\omega_1 d\omega_2\cdots d\omega_5 \\
&= \cdots \\
&= \frac{1}{(2\pi)^{3p}}\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{3p+2}\gamma(\omega,\omega_{3p+2})\widetilde{D}(\omega_{3p+2}-\omega_{3p+1}-\omega_{3p})\gamma(\omega_{3p},\omega_{3p-1}) \\
&\quad \widetilde{D}(\omega_{3p-1}-\omega_{3p-2}-\omega_{3p-3})\cdots\gamma(\omega_6,\omega_5)\widetilde{D}(\omega_5-\omega_4-\omega_3)\gamma(\omega_3,\omega_2) \\
&\quad \times\widetilde{b}(\omega_2-\omega_1)U(\omega_{3p+1})U(\omega_{3p-2})\cdots U(\omega_4)U(\omega_1)d\omega_1\cdots d\omega_{3p+2}.
\end{aligned}$$

The case of general nonlinear systems can be approached by Lie theory (see [3]), or by using the sequential approach developed in this book. We shall use the latter course and just present the formal manipulations (for more details see [1]). For simplicity we consider the unforced finite-dimensional case - the general distributed input-output theory is similar. Thus, consider a nonlinear dynamical system of the form

$$\dot{x} = (A_0 + A'(x))x,\ x(0) = x_0$$

for some time-invariant matrix A_0. Introduce the sequence of systems

$$\dot{x}^{[i]}(t) = (A_0 + A'(x^{[i-1]}(t)))x^{[i]}(t),\ x^{[i]}(0) = x_0,\ i \geq 2$$

and

$$\dot{x}^{[1]}(t) = (A_0 + A'(x_0))x^{[1]}(t),\ x^{[1]}(0) = x_0.$$

Taking the Fourier of the i^{th} equation gives

$$i\omega X^{[i]}(i\omega) - x_0 = \left(A_0 + \frac{1}{2\pi}\widetilde{A}'^{[i-1]}(i\omega)*\right)X^{[i]}(i\omega)$$

where $\widetilde{\cdot}$ denotes the Fourier transform and $*$ is the convolution operation. Hence, as before, we can write

$$X^{[i]}(i\omega) = \Bigg((i\omega I - A_0)^{-1} + \frac{1}{2\pi}(i\omega I - A_0)^{-1}\int_{-\infty}^{\infty}\widetilde{A}'^{i-1]}(i\omega')\left(i(\omega-\omega')I - A_0\right)^{-1}$$

$$d\omega' + \left(\frac{1}{2\pi}\right)^2 (i\omega I - A_0)^{-1}\int_{-\infty}^{\infty}\widetilde{A}'^{i-1]}(i\omega'')\left(i(\omega-\omega'')I - A_0\right)^{-1}\cdot$$

$$\int_{-\infty}^{\infty}\widetilde{A}'^{i-1]}(i\omega')\left(i(\omega-\omega''-\omega')I - A_0\right)^{-1}d\omega' d\omega'' + \cdots\Bigg)x(0)$$

provided the operator L given by

$$(LX^{[i]})(i\omega) = \frac{1}{2\pi}(i\omega I - A_0)^{-1}\left[\widetilde{A}'^{i-1]}(i\omega) * X(i\omega)\right]$$

is bounded by 1. Thus we have, for the spectrum $X(i\omega)$ of the nonlinear system,

$$X(i\omega) = lim_{k\to\infty}X^{[k]}(i\omega),$$

where

$$X^{[k]}(i\omega) = \sum_{0}^{\infty}\frac{1}{(2\pi)^k}\left((i\omega I - A_0)^{-1}\widetilde{A}'^{k-1]}(i\omega)*\right)^k (i\omega I - A_0)^{-1}x(0)$$

and

$$\widetilde{A}'^{k-1]}(i\omega) = \mathfrak{F}\left(\widetilde{A}(x^{[k-1]}(t))\right).$$

For the convergence proof of this sequence, see [1].

4.3 Exponential Dichotomies

In this section we shall generalise some results on exponential dichotomies to nonlinear systems. Recall that an *exponential dichotomy* for a linear, time-varying system

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

on an interval $J \subseteq \mathbb{R}$ is given by a projection P and constants K and α,β such that

$$\|X(t)PX^{-1}(s)\| \leq Ke^{-\alpha(t-s)}, t \geq s$$

$$\|X(t)(I-P)X^{-1}(s)\| \leq Ke^{-\beta(s-t)}, s \geq t.$$

We shall state some results from Coppel [6] which we shall generalise here. From these ideas it is clear that many other results can be generalised in a similar way.

Lemma 4.1. *Consider the inhomogeneous equation*

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

This equation has at least one bounded solution, for each locally integrable function f for which $\int_t^{t+1} |f(s)|ds$ is bounded for all $t \geq 0$, if and only if the homogeneous equation

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

has an exponential dichotomy.

The property of exponential dichotomy being preserved under 'small' perturbations is called *roughness*. Then we have

Lemma 4.2. *Suppose that the system*

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

has an exponential dichotomy with constants K and α, then the system

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

also has an exponential dichotomy provided

$$\delta \overset{\Delta}{=} \sup_{t \geq 0} \|B(t)\|$$

satisfies

$$\delta < \alpha/4K^2.$$

Lemma 4.3. *Let $A(t)$ be continuous and bounded (i.e. $\|A(t)\| \leq M$) and suppose that $A(t)$ has k eigenvalues with real part $\leq -\alpha < 0$ and $n-k$ eigenvalues with real part $\geq -\beta > 0$ for all t in some interval J. Assume also that for all $\varepsilon > 0$ ($\varepsilon < \min(\alpha, \beta)$), there exists δ such that*

$$\|A(t_2) - A(t_1)\| \leq \delta \text{ for } |t_2 - t_1| \leq h,\ t_1, t_2 \in J$$

where h is a fixed number $\leq$ length J. Then the linear system

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

has an exponential dichotomy on J with constants $\alpha - \varepsilon, \beta - \varepsilon$.

Thus, if $A(t)$ is continuous, bounded and of bounded variation and has the spectral properties in the statement of the lemma, then the system has an exponential dichotomy.

Now consider the nonlinear equation

$$\dot{x} = A(x)x \tag{4.5}$$

(which we assume has a unique solution defined for all t). We denote, as usual, the solution of this equation through (x_0,t_0) by $x(t;x_0,t_0)$. We say that the system (4.1) has an *exponential dichotomy* on an open set $\mathfrak{O} \subseteq \mathbb{R}^n$ if there exists a projection P and constants $K,\alpha,\beta > 0$ such that

$$\left\|X_{(x_0,t_0)}(t)PX^{-1}_{(x_0,t_0)}(s)\right\| \leq Ke^{-\alpha(t-s)},\ t \geq s,\ s,t \in \mathfrak{T}_{(x_0,t_0)}(\mathfrak{O})$$

$$\left\|X_{(x_0,t_0)}(t)(I-P)X^{-1}_{(x_0,t_0)}(s)\right\| \leq Ke^{-\beta(s-t)},\ s \geq t,\ s,t \in \mathfrak{T}_{(x_0,t_0)}(\mathfrak{O})$$

where $X_{(x_0,t_0)}$ is the fundamental solution of the linear, time-varying system

$$\dot{x} = A(x(t;x_0,t_0))x,\ x(t_0) = x_0 \tag{4.6}$$

and $\mathfrak{T}_{(x_0,t_0)}(\mathfrak{O})$ is the set of times for which the solution through (x_0,t_0) lies in $\mathfrak{O}$. We shall say that the system (4.1) has an exponential dichotomy in some region $\mathfrak{R} \subseteq \mathbb{R}^n$ if there exists an open cover of $\mathfrak{R}$ such that it has an exponential dichotomy on each open set of the cover. Many results of the types of Lemmas 4.1–4.3 can be generalised to nonlinear systems – since the ideas are the same we shall just give a generalisation of lemma 4.3.

Theorem 4.3. *Consider the system*

$$\dot{x} = A(x)x, \tag{4.7}$$

and assume that $A(x)$ has eigenvalues off the imaginary axis for all x in the open set $\mathfrak{O}$. Suppose that $A(x)$ is Lipschitz continuous and that the solutions of () are bounded for initial states in the subset $\mathfrak{O}_1 \subseteq \mathfrak{O}$. Then the system (4.7) has an exponential dichotomy on any compact subset of $\mathfrak{O}_1$.

Proof. This follows directly from the iteration theory together with an application of Lemma 3.3 coupled with the roughness property of dichotomy. □

Finally we also note the following result on the Sacker-Sell spectrum (see Chapter 3 for a discussion of the Sacker-Sell spectrum): suppose that $\Omega \subseteq \mathbb{R}^n$ is a compact invariant set for the flow of this dynamical system and consider the set of approximations

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t),x_0)x^{[i]}(t), \quad x^{[i]}(0) = x_0$$

and

$$\dot{x}^{[0]}(t) = A(x_0)x^{[0]}(t), \quad x^{[0]}(0) = x_0$$

where $x_0 \in \Omega$. (Here we have shown the explicit dependence of the systems on x_0.) We have

Theorem 4.4. *Let $\mathfrak{A} \subseteq L^\infty(\mathbb{R}; G\ell(n))$ be the set*

$$\mathfrak{A} = \{A(x) : x \in \Omega\}$$

and let μ be a normalised invariant measure on $\mathfrak{A}$. Then there exist real numbers $\beta_1 < \beta_2 < \cdots < \beta_k$, $1 \leq k \leq n$, such that:

(a) There exists a basis $\{e_i\}$ of $\mathbb{R}^n$ such that the numbers

$$lim_{i\to\infty}lim_{t\to\infty}\frac{1}{t}ln||\Phi_{A(x^{[i]}(\cdot))}(t)e_r||,$$
$$lim_{i\to\infty}lim_{t\to-\infty}\frac{1}{t}ln||\Phi_{A(x^{[i]}(\cdot))}(t)e_r||$$

exist for all A in some shift-invariant set $\mathfrak{A}_1 \subseteq \mathfrak{A}$ with $\mu(\mathfrak{A}) = 1$ and belong to the set $\{\beta_i\}$.

(b) Let $\mathfrak{B}_r$ denote the set

$$\begin{aligned}\mathfrak{B}_r = \{(A,x) \in \mathfrak{A}_1 \times \mathbb{R}^n : x = 0 \ or \\ lim_{i\to\infty}lim_{t\to\infty}\frac{1}{t}ln||\Phi_{Ax^{[i]}(\cdot)}(t)x|| = \beta_r\}, \\ 1 \leq r \leq k.\end{aligned}$$

Then each $\mathfrak{B}_r$ is a measurable invariant subbundle of $\mathfrak{A} \times \mathbb{R}^n$, and

$$\mathfrak{A}_1 \times \mathbb{R}^n = \mathfrak{B}_1 \oplus \mathfrak{B}_2 \oplus \cdots \oplus \mathfrak{B}_k.$$

Also, the limits are uniform in x.

(c) For each $A(\cdot) \in \mathfrak{A}_1$, the equation

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

is regular in the sense of Lyapunov, i.e.

$$lim_{t\to\pm\infty}ln\ det\Phi_{A(\cdot)}(t) = \sum_{r=1}^{k} \beta_r\ for\ all\ A(\cdot) \in \mathfrak{A}_1.$$

(d) For each $A(\cdot) \in \mathfrak{A}_1$, the maximal Lyapunov exponent is β_k.

4.4 Conclusions

In this chapter we have outlined a general spectral theory for nonlinear systems, based on the iteration scheme. We have seen that many results for linear spectral theory can be generalised to nonlinear systems. In particular, the theory of Volterra kernels has been shown to be derivable easily from a function expansion and then generalised to nonlinear systems. Moreover, the theory of exponential dichotomy and the Sacker-Sell spectrum have also been generalised to nonlinear systems.

The Volterra kernels can be used, along with some elementary algebraic geometry to define the poles and zeros of a nonlinear system (see [7,8]), but they are algebraic manifolds in an arbitrarily high dimensional complex manifold and so are difficult to apply practically. They do have some theoretical importance and,as shown in [7] we can define a nonlinear root locus technique. The present methods described in this chapter, however, are much more easily applied and so are likely to be more useful in general systems theory.

References

1. Banks, S.P., Riddalls, C., McCaffrey, D.: The Schwartz Kernel Theorem and the Frequency-Domain Theory of Nonlinear Systems. Arch. Contr. Sci. 6, 29–45 (1997)
2. Mitzel, S.J., Rugh, W.J.: On Transfer Function Representations for Homogeneous Nonlinear Systems. IEEE Trans. Aut. Contr. AC-24, 242–249 (1979)
3. Chanane, B., Banks, S.P.: Realization and Generalized Frequency Response for Nonlinear Input-Output Maps. Int. J. Sys. Sci. 20, 2161–2170 (1989)
4. Yosida, K.: Functional Analysis. Springer, New York (1970)
5. Treves, F.: Topological Vector Spaces, Distributions and Kernels. Academic Press, New York (1978)
6. Coppel, W.A.: Dichotomies in Stability Theory. LNM, vol. 629. Springer, New York (1978)
7. Banks, S.P.: On Nonlinear Systems and Algebraic Geometry. Int. J. Control 42, 333–352 (1985)
8. Banks, S.P.: State-Space and Frequency Domain Methods in the Control of Distributed Parameter Systems. Peter Peregrinus, London (1983)

Chapter 5
Spectral Assignment in Linear, Time-varying Systems

5.1 Introduction

Pole placement is a well-known tool in designing control for linear systems. It belongs to the class of so-called *feedback stabilisation methods*. Basically, the aim of this approach is to design a controller so that the closed-loop poles of the plant are assigned to some desired locations chosen according to some specific stability and performance criteria.

Several authors have approached the pole placement idea for general nonlinear systems in the past. Most of these techniques have in common the idea of linearising the nonlinear system about a countable set of equilibrium points and finding a single controller that will stabilise each member of the finite countable set (see [2] for an example). Of course, these approaches present some difficulties as for example the impossibility of deciding *a priori* whether or not a set of three or more systems is simultaneously stabilisable [3]. On the other hand, in the area of nonlinear systems too, and having its origins in the geometric control theory, exact feedback linerisation with pole placement is achieved by following a two-step design method [37, 38]: first, a simultaneous implementation of a nonlinear coordinate transformation and a state feedback law is obtained in order to transform the original nonlinear system into a linear one in Brunowsky canonical form. The second step is the application of the already established pole placement methods for linear time-invariant (LTI) systems. However, the first step is very restrictive as involutivity conditions arise and these are hardly met by any physical system of order higher than two. There have been later attempts to obtain both feedback linearisation and pole placement objectives in just one step [40].

In this chapter, the classical idea of pole placement for LTI systems is extended to a general pole placement technique applicable to time-varying systems and hence with the aid of the iteration technique presented in here, ultimately to nonlinear systems in a general form.

Most systems arising in practice have time-varying parameters which will affect the performance of a system which has been designed for some nominal parameter

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 75–100.
springerlink.com

set; this happens for example, in the aerodynamic coefficients of high-speed aircrafts, circuit parameters in electronic circuits or diffusion coefficients in chemical processes. Time variation may arise too as the result of linearising a nonlinear system about a family of operating points and/or about a time-varying operating point [12].

In our case, the time-varying characteristics arise once the iteration technique is used: the application of this technique to a nonlinear system of the form $\dot{x} = A(x)x + B(x)u, x_0 = x(0)$, leads to a sequence of time-varying systems that will be individually treated with linear techniques of pole placement. Pole placement for LTI systems has been for a long time the object of diverse studies: some of them were based on the well-known Ackerman's formula [13], some others approached the pole assignment problem by using a periodic output feedback ([14, 15]) for second order systems or even arbitrary order as in [16].

The original pole placement method for LTI single-input, single-output (SISO) systems was first extended to time-varying systems by [18] using a canonical representation of the original system. Since then, it had been several contributions by different authors to develop pole placement techniques for linear time-varying (LTV) systems: some of them are complex in their formulation or rely on transformations of the original system into controllable canonical forms (*i.e.* Frobenius and Hessenberg forms) with an external input and subsequently the employment of linear pole placement techniques ([18, 19, 20, 21, 22, 23, 24, 25], to embed the pole placement problem within the more general problem of eigenstructure assignment as in [34, 35, 36] or derived pole placement algorithms via Sylvester's equation [29].

Also, many pole placement techniques for LTV systems require the computation of the characteristic polynomial coefficients for either the original or the new state matrices or the eigenvalues of the original system matrix: in [26] Blanchini introduced a method that removes this requirement for SISO systems by using an intermediate transformation to a bi-diagonal Frobenius form. This was then extended to LTV systems in [27]. Nguyen in [28] introduced the Frobenius transformation for time-varying systems; however, this treatment requires that the characteristic polynomial coefficients of the desired behaviour be computed as well as the complete Frobenius transformation of the system. Most recently, Valasek *et al.* ([30, 31, 32]) initiated a series of publications related to the eigenvalue placement problem based on the extension of Ackerman's formula to time-varying SISO and later to time-invariant and time-varying multi-input, multi-output (MIMO) systems in which the eigenvalue placement was based on the equivalence of the closed-loop original system via a Lyapunov transformation to a LTI system with poles at prescribed locations.

It should be pointed out that an important limitation of the pole placement algorithm is the lack of guaranteed tracking performance. This topic is treated in more general output feedback approaches. A typical remedy for this involves the incorporation of the *internal model principle* into the control law design ([41] and [43]) or the inclusion of integrators into the loop. This issue will not be addressed in this chapter, since pole placement design is the main interest here.

The contents of this chapter are based on a combination of the classical pole placement approach for LTI systems, extended to LTV system such that certain stability conditions are satisfied and the iteration technique. The objective is to develop a methodology such that the stability of a nonlinear system of the form (5.1), can be achieved by using pole placement tools on a sequence of LTV systems that represent the original nonlinear system:

$$\begin{aligned} \dot{x} &= A(x)x(t) + B(x)u(t), \quad x(0) = x_0 \\ y &= C(x)x(t) + D(x)u(t). \end{aligned} \tag{5.1}$$

Replacing the nonlinear system (5.1) by a sequence of LTV systems, a sequence of feedback laws of the form $u^{[i]}(t) = K^{[i]}(t)x^{[i]}(t)$ can be generated and for each of them the closed-loop poles for the i^{th} LTV system at each time of the time interval are allocated to some desired location $\sigma = (\lambda_{1d}, \cdots \lambda_{nd})$ where each λ_i can be time-varying or constant.

It is well-known that LTV systems can be unstable despite having left-half plane poles; that is, for time-varying systems, poles do not have the same stability meaning as in the time-invariant case, so the allocation of the pole in the left-hand side plane does not guarantee the stability of the closed-loop plant. In order to overcome this problem an approach to stability using Duhamel's principle is presented in Section 5.3 and some conditions based on differentiability of the eigenvector matrix are derived.

From the convergence of the sequence of LTV solutions, by choosing the $K^{[i]}(t)$ feedback gain corresponding to the i^{th} iteration and applying the limiting value to the closed-loop nonlinear system, the pole placement and stability objectives are achieved for a wide variety of nonlinear cases. This generalisation to nonlinear systems is given in Section 5.4, followed by a numerical example. Section 5.5 contains a practical application of this theory.

5.2 Pole Placement for Linear, Time-invariant Systems

Consider an open-loop LTI system of the form:

$$\dot{x} = Ax(t) + Bu(t), \quad x(0) = x_0 \tag{5.2}$$

$$y = Cx + Du$$

where $x(t) \in \mathbb{R}^n$ is the vector of the measurable states, $u(t) \in \mathbb{R}^m$ is the control signal and A, B, C are constant matrices of appropriate dimensions. If the pair (A, B) is controllable, then the LTI system (5.2) can be stabilised by designing a linear state feedback control law:

$$u(t) = r(t) - Kx(t), \tag{5.3}$$

where $r(t)$ is a vector of desired states and the matrix $K \in \mathbb{R}^{nxm}$ is chosen so that the spectrum of the closed-loop system matrix is allocated arbitrarily on the left-hand side of the complex plane.

The closed-loop representation of the system (5.2) is of the form:

$$\dot{x} = Ax(t) - BKx(t) + Br(t) = (A - BK)x + Br(t). \tag{5.4}$$

The eigenvalues of a system determine the decay/growth rate of the response in LTI cases, *i.e.*:

$$x(t) = \sum_{i=1}^{N} e^{\lambda_i t} x_i(0) \tag{5.5}$$

so by ensuring they are placed in the left-hand side (or inside the unit circle in discrete cases, $|z| < 1$), the response $x(t)$ is guaranteed to decay to zero (only in time-invariant cases). The pole placement control problem is to determine the value of K such that the desired set of closed-loop poles is achieved. In the regulator problem the reference signal is set to be zero, $r(t) = 0$ and therefore the feedback control (5.3) is now $u(t) = -Kx(t)$. The aim is to allocate the poles of the closed-loop matrix $\big(A - BK\big)$ on the left-hand side such that the state $x(t)$ is now driven to zero:

$$if \quad Re\{eig\big(A - BK\big)\} < 0 \rightarrow lim_{t\rightarrow\infty}\{x(t)\} = 0.$$

We next consider the algebraic solution of the pole placement problem. For example, if the system is two-dimensional ($n = 2$) with $r(t) = 0$, the feedback gain $K = [k_1, k_2]$ can be designed conveniently in order to place the closed-loop poles anywhere in the left-half plane (or inside the unit circle $|z| < 1$; in the discrete-time case) by using the eigenvalue placement theorem (known as well as Ackerman's formula). Thus the control $u(t)$ designed in this way will drive the system response from a set of initial conditions $x(0)$ to zero as $t \longrightarrow \infty$.

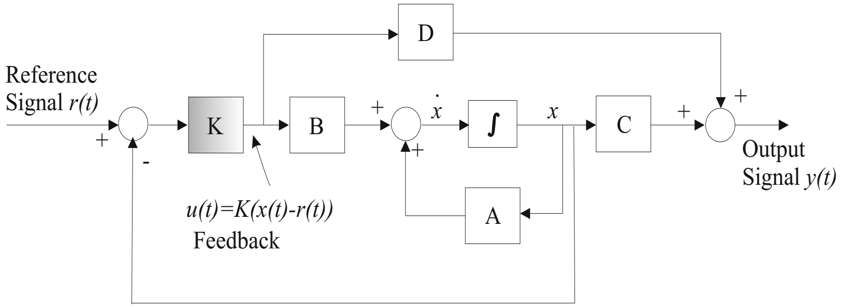


Fig. 5.1 Diagram of the pole placement design

Theorem 5.1. *For a state controllable system* (A, B, C, D)*, given any* n^{th} *order polynomial,*

$$P^d(\lambda) = (\lambda - \lambda_1^d)(\lambda - \lambda_2^d)\cdots(\lambda - \lambda_n^d) = \lambda^n + a_{n-1}^d \lambda^{n-1} + \cdots + a_1^d \lambda + a_0^d$$

with real coefficients a_i^d, there exists a real matrix K such that the closed-loop matrix $(A - BK)$ has $P^d(\lambda)$ as its characteristic polynomial. Hence, the feedback gain matrix K is determined by the condition:

$$|\lambda \cdot I - A + BK| = P^d(\lambda) = 0$$

where λ_i^d are the desired eigenvalues for a i^{th}-dimensional system.

Proof. Let $\gamma(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0$ be the characteristic polynomial of (5.2). As (A, B) is controllable, there exist a non-singular transform matrix T_c so that

$$\tilde{A} = T_c^{-1} A T_c = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix}, \quad \tilde{B} = T_c^{-1} B T_c = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

Then $K = \left[a_0 - a_0^d, a_1 - a_1^d, \cdots, a_{n-1} - a_{n-1}^d\right] T_c$ ensures the characteristic polynomial of $(A - BK)$ is $\lambda^n + a_{n-1}^d \lambda^{n-1} + \cdots + a_1^d \lambda + a_0^d$. □

For an adequate selection of the desired eigenvalues λ_i^d, it is convenient to keep in mind the fact that controllability slips away as the poles are moved closer to the zeros producing zero-pole cancellations inside the transfer function. The design strategy should be done in a way that improves only the undesirable aspects of the open-loop response and avoids large increases in bandwidth produced by too large negative eigenvalues (or too far inside the unit circle). This approach will be used in the following section to determine a control law $u^{[i]}(t)$ for each linear time-varying system of the sequence of linear time-varying equations, so that each of their solutions $x^{[i]}(t)$ will converge to zero.

This pole placement method for LTI systems is a well-known technique in control due to its simplicity and versatility of applications: it has been used in areas such as control and stability of aircrafts ([44, 45]), helicopters [49], missile autopilots [47] and more recently some of these ideas have been applied to the correct and precise positioning of storing devices as CDs or DVDs [50] where linearised models of the plant were used.

5.3 Pole Placement for Linear, Time-varying Systems

In this section the LTI pole placement method will be extended to LTV systems of the form:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0 \tag{5.6}$$

$$y(t) = C(t)x(t) + D(t)u(t).$$

where $x(t) \in \mathbb{R}^n$ is the vector of the measurable states, $u(t) \in \mathbb{R}^m$ is the control signal and $A(t)$, $B(t)$, $C(t)$ are time-varying matrices of appropriate dimensions.

As summarised in the introduction, there have been several attempts from different authors to design pole placement techniques for LTV systems, see as an example ([18], [15], [16]) and references within. In our case, the pole placement method for LTV systems has been kept as similar as possible to the standard LTI case; that is, given a set of desired stable eigenvalues, $\sigma = \left(\lambda_{1d} \cdots \lambda_{nd}\right)$ and a time interval $[0,t]$, the aim is to place the closed-loop eigenvalues of (5.6) at those desired points at each time of the interval by using a convenient state feedback control $u(t) = -K(t)x(t)$ where the feedback gain $K(t)$ is a time-dependent function.

Remark 5.1. *For simplicity, it is assumed that* $C(t) = I$ *and* $D(t) = 0$*; it is possible to generalize to other cases.*

Given that the pair $\left[A(t), B(t)\right]$ is controllable for all $t \in [0,T]$, the eigenvalue placement theorem is applied to (5.6) as in Section 5.2.

$$\det\left|\lambda \cdot I - [A(t) - B(t)K(t)]\right| = (\lambda - \lambda_{1d}) \cdots (\lambda - \lambda_{nd}) \tag{5.7}$$

and by solving this algebraic expression (5.7), the time-varying feedback gain $K(t)$ can be determined so that the closed-loop form of the system (5.6) will now be of the form

$$\dot{x}(t) = [A(t) - B(t)K(t)]x(t) = \tilde{A}(t)x(t) \tag{5.8}$$

with stable eigenvalues on the left-half plane at $\left(\lambda_1 \cdots \lambda_n\right) = \left(\lambda_{1d} \cdots \lambda_{nd}\right)$. In order to guarantee stability of the system (5.6), further issues should be taken into account. It is well-known that for linear time-varying systems, the fact of having the closed-loop poles on the left-half side of the plane is not a sufficient condition for stability.

We next consider necessary and sufficient conditions for exponential stability of LTV systems with negative eigenvalues will be derived and these results will be extended for the nonlinear case. In most LTV cases, stability is not achieved only by allocating the closed-loop poles at the desired values $\sigma = (\lambda_{1d}, \cdots, \lambda_{nd})$ on the time interval $[0,t]$ but also by satisfying some additional conditions.

The need of additional conditions for stability of LTV systems is a topic frequently covered in the literature; some of the approaches rely on the slow variation of time parameters [17], some other in the existence of a triangular transformation of the system's matrix, just to mention a few.

In our case, having in mind that the matrix $\widetilde{A}(t)$ already has negative eigenvalues due to the solution of the algebraic equation in (5.7), some other conditions for stability of the closed-loop system (5.8) should be satisfied, these conditions can be summarised in the following theorem:

Theorem 5.2. *Given the open-loop LTV system $\dot{x} = A(t)x(t) + B(t)u(t)$, $x_0 = x(0)$, whose closed-loop matrix $\widetilde{A}(t) = A(t) - B(t)K(t)$ has designed left-hand plane eigenvalues $\sigma = (\lambda_{1d}, \ldots, \lambda_{nd})$ via the feedback signal $u(t) = -K(t)x(t)$ and assuming the following conditions to be satisfied:*

(a) λ_{1d} is the real part of the greatest of the eigenvalues of $\widetilde{A}(t)$

(b) The matrix of eigenvectors, $P(t)$, is differentiable

(c) $||P^{-1}(t)\dot{P}(t)|| < \beta$

for $\beta < Re(\lambda_{1d})$ the closed-loop system $\dot{x} = [A(t) - B(t)K(t)]x(t) = \widetilde{A}(t)x(t)$ is exponentially stable.

Proof. The system (5.8) can be solved over any given time interval $[0,t]$ using Duhamel's principle: the time interval can be divided into N subintervals of length h, such that $h = t/N \to 0$ when $N \to \infty$, such that:

$$x(t) = Lim_{h \to 0}\left(e^{\tilde{A}[Nh]h} \cdot e^{\tilde{A}[(N-1)h]h} \ldots e^{\tilde{A}[h]h} \cdot I \cdot x_0\right) \tag{5.9}$$

Now, the similarity transform $M(t) = P(t)\Lambda(t)P^{-1}(t)$ can be applied to the closed-loop matrix $\widetilde{A}(t)$:

$$x(t) = \left(P_N e^{\Lambda[Nh]h} P_N^{-1}\right) \cdot \left(P_{N-1} e^{\Lambda[(N-1)h]h} P_{N-1}^{-1}\right) \ldots \left(P_1 e^{\Lambda[h]h} P_1^{-1}\right) \cdot I \cdot x_0 \tag{5.10}$$

where $\Lambda(t) \in \mathbb{C}^{nxn}$ is a diagonal matrix of desired eigenvalues and $P(t) \in \mathbb{C}^{nxn}$ is the time-varying matrix of correspondent eigenvectors. □

Remark 5.2. *P_N is $P(t)$ at time $t = Nh$.*

Remark 5.3. *$\Lambda(t)$ is considered to be time-varying to generalise the results of Theorem 5.2. In here, it is considered to be constant as the desired eigenvalues were taken to be constant.*

The second assumption was that $P(t)$ was differentiable, therefore its Taylor expansion will be of the form:

$$P(t+h) = P(t) + h\frac{dP(t)}{dt} + \frac{h^2}{2!}\frac{d^2P(t)}{dt} + \cdots \tag{5.11}$$

Neglecting high order terms and noting that $\frac{dP(t)}{dt} = \dot{P}(t)$:

$$P(t+h) = P(t) + h\dot{P}(t) \tag{5.12}$$

Then, multiplying (5.12) by $P(t+h)^{-1}$:

$$P(t+h)^{-1}P(t+h) = I = P(t+h)^{-1}P(t) + hP(t+h)^{-1}\dot{P}(t)I$$
$$= P(t+h)^{-1}\left[P(t) + h\dot{P}(t)\right]$$

we have that

$$P(t+h)^{-1} = \left[P(t) + h\dot{P}(t)\right]^{-1}.$$

Therefore,

$$P(t+h)^{-1}P(t) = \left[P(t) + h\dot{P}(t)\right]^{-1}P(t) = \left[P(t)^{-1}\left(P(t) + h\dot{P}(t)\right)\right]^{-1}$$

and so,

$$P(t+h)^{-1}P(t) = \left[I + P(t)^{-1}h\dot{P}(t)\right]^{-1}$$

and as $(1+a)^{-1} = 1 - a + \cdots$, this can be written as:

$$P(t+h)^{-1}P(t) = \left[I - P(t)^{-1}h\dot{P}(t)\right] \tag{5.13}$$

and,

$$\left[P(t+h)^{-1} \cdot P(t)\right] = I + \varepsilon(h) \tag{5.14}$$

where $\varepsilon = o(h)$, so that $\varepsilon \to 0$ as $h \to 0$.

Thus (5.10) becomes:

$$x(t) = P_N \cdot e^{\Lambda[Nh]h} \cdot (I+\varepsilon) \cdot e^{\Lambda[(N-1)h]h} \cdot (I+\varepsilon) \cdots e^{\Lambda[h]h} P_1 \cdot I \cdot x_0 \tag{5.15}$$

Taking norms of the above a bound on the norm of $x(t)$ can be estimated by,

$$||x(t)|| \leq ||P_N|| \cdot ||(1+\varepsilon)||^N \cdot ||P_1|| \cdot ||e^{(\Lambda t)}|| \cdot ||x_0|| \tag{5.16}$$

and taking into account that

$$||e^{\Lambda t}|| \leq e^{Re(-\lambda_{max})t}$$

where λ_{max} is the eigenvalue of the matrix Λ with largest real part, then for $\lambda_{max} = \lambda_{1d}$

$$||x(t)|| \leq ||P_N|| \cdot ||(I+\varepsilon)||^N \cdot ||P_1|| \cdot e^{-\lambda_{1d}t} \cdot ||x_0||. \tag{5.17}$$

Now, it was shown in (5.14) that $\varepsilon(h) = -hP^{-1}(t)\dot{P}(t)$, so

$$||I+\varepsilon|| = ||I - hP^{-1}\dot{P}(t)|| \leq 1 + h||P^{-1}(t)\dot{P}(t)||.$$

By assumption 3 in Theorem 5.2, the expression $||P^{-1}(t)\dot{P}(t)||$ is bounded by β, and so

$$||I+\varepsilon(h)|| \leq 1 + h\beta.$$

Thus,

$$||x(t)|| \leq ||P_N|| \cdot (1+h\beta)^N \cdot ||P_1|| \cdot e^{-\lambda_{1d}t} \cdot ||x_0||$$

and

$$(1+h\beta)^N = \left(1+\frac{\beta t}{N}\right)^N \to e^{\beta t}$$

so

$$||x(t)|| \leq ||P_N|| \cdot e^{(\beta-\lambda_{1d})t} \cdot ||P_1|| \cdot ||x_0||.$$

Analysing this for exponential stability, P_N, P_1, and x_0 are constant values, so $e^{t(\beta-\lambda_{1d})} \to 0$ is required:

$$e^{(\beta-\lambda_{1d})t} \to 0, \quad (\beta-\lambda_{1d}) < 0 \to \beta < |\lambda_{1d}|. \tag{5.18}$$

That is, for exponential stability, the closed-loop eigenvalues λ_d should be chosen so that the greatest of them λ_{1d} satisfies (5.18) which represents a compromise between the upper bound of the rate of change of $P(t)$ and λ_{1d}.

Example 5.1. Consider the dynamical equation of the single-input LTV system:

$$\dot{x}(t) = \begin{pmatrix} e^{-t} & 2 & 0 \\ 0 & 2 & e^{-t} \\ 1 & 0 & 1 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} u(t). \tag{5.19}$$

This example was first introduced in [30], [31] and [42] where a different method of pole placement by state feedback and derivative state feedback respectively was applied. In this example the time interval will be divided in N steps of size h ($h \to 0$), and Duhanmel's principle applied. The set of desired eigenvalues is $\sigma = \big(-8,-5,-7\big)$ and the initial conditions for the states vector is $x(0) = [0.3,-0.2,0.1]^T$. The transient components of the response are shown in Figure 5.2. These transient responses show the same characteristics as the results obtained in [31] where state feedback was used and the gain of the feedback $K(t)$ was designed via Lyapunov transformation of the original LTV system. The control law $u(t) = -K(t)x(t)$ that sets the poles at the indicated location is shown in Figure 5.3. In addition to these results, the matrix of eigenvectors $P(t)$ was calculated at each step of the simulation, and an estimation of $\varepsilon(h) = \big[P(t+h)^{-1} \cdot P(t)\big] - I$ was obtained as in (5.14) as a way of testing differentiability of $P(t)$: Figure 5.4 shows the performance of ε at each step of the time interval and it is shown how it goes to zero as expected according with the theory stated in this section.

In this example it has been shown that once the pole placement algorithm has been applied, exponential stability is attained as the conditions in Theorem 5.2 are satisfied. In fact, in order to strengthen this idea, the differentiability of $P(t)$ condition 2 in Theorem 5.2 can be estimated by the quantity $\frac{P(t+h)-P(t)}{h}$, and Figure 5.5 shows how this condition is satisfied, confirming in this way the exponential stability of the system. Now consider a necessary condition for the differentiability of $P(t)$. It was shown above how the exponential stability properties of the closed-loop system relied upon the satisfaction of conditions 1-3 of in Theorem 5.2. These conditions were sufficient conditions for stability. In this section, a *necessary condition* for

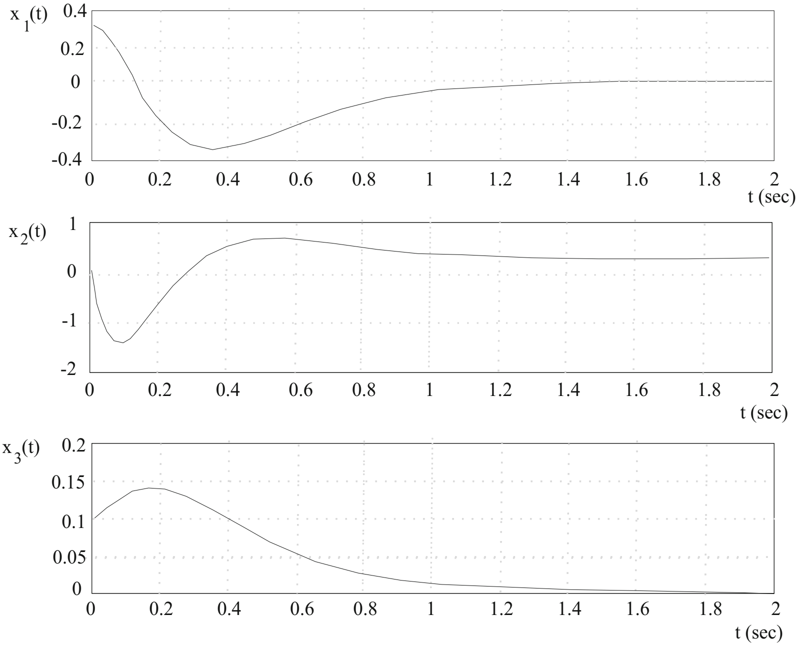


Fig. 5.2 Controlled components $x_1(t)$, $x_2(t)$ and $x_3(t)$

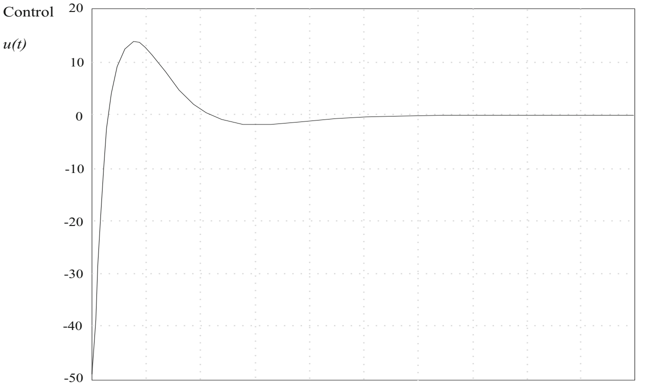


Fig. 5.3 Control $u(t) = -K(t)x(t)$

stability will be derived. This condition is given in terms of a differential equation which places restrictions on $\Lambda(t)$, $K(t)$ and $P(t)$. A necessary condition for stability is stated in the following lemma:

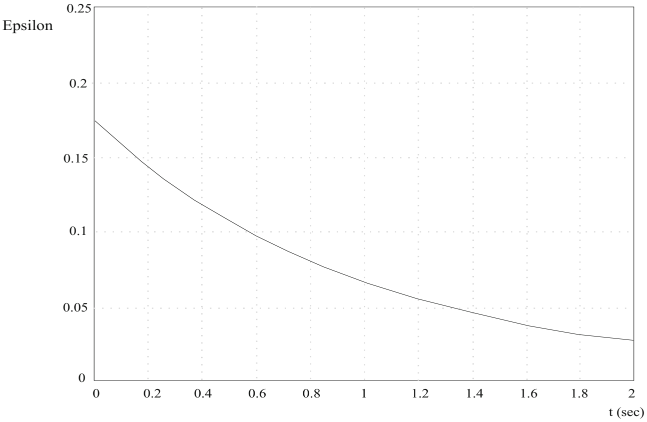


Fig. 5.4 $\varepsilon(h)$ at each point of the time interval

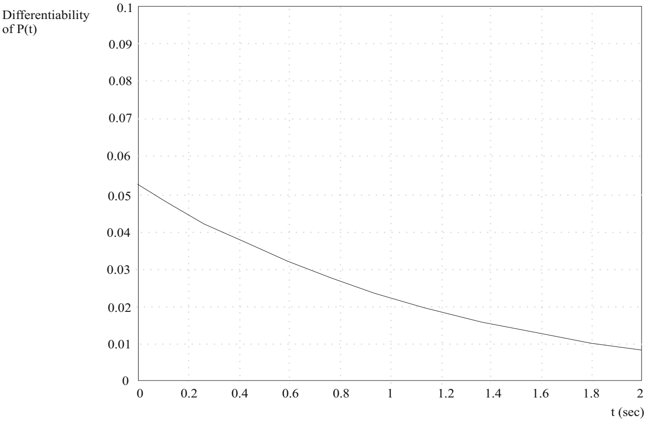


Fig. 5.5 Differentiability $(P(t+h)-P(t))/h$ at each point of the time interval

Lemma 5.1. *The differentiability of the matrix of eigenvectors $P(t)$ (and $A(t)$, $B(t)$, $K(t)$ and $\Lambda(t)$) imply that the following equation is satisfied:*

$$\begin{aligned}&\left[\dot{A}(t)-\dot{B}(t)K(t)-B(t)\dot{K}(t)\right]\\&\quad=P(t)\left[\dot{\Lambda}(t)-\Lambda(t)P^{-1}(t)\dot{P}(t)+P^{-1}(t)\dot{P}(t)\Lambda(t)\right]P^{-1}(t),\end{aligned}\tag{5.20}$$

where $\Lambda(t)$ is the diagonal matrix of eigenvalues of $A(t)$ and $K(t)$ is the feedback gain designed for stable closed-loop poles.

Proof. Consider two *nearby* time points, t and $t+h$, and evaluate the similarity transforms at those points keeping in mind that the matrix $\Lambda(t)$ is a diagonal matrix containing the eigenvalues of the matrix $\left[A(t)-B(t)K(t)\right]=\tilde{A}(t)$:

$$P^{-1}(t)\,[A(t)-B(t)K(t)]\,P(t)=\Lambda(t) \tag{5.21}$$

$$P^{-1}(t+h)\,[A(t+h)-B(t+h)K(t+h)]\,P(t+h)=\Lambda(t+h) \tag{5.22}$$

as in (5.11) and assuming the differentiability of $A(t), B(t)$ and $K(t)$. Then,

$$\begin{aligned} A(t+h) &= A(t)+h\dot{A}(t)+\ldots \\ B(t+h) &= B(t)+h\dot{B}(t)+\ldots \\ K(t+h) &= K(t)+h\dot{K}(t)+\ldots \end{aligned}$$

and by the differentiability of $P(t)$ and (5.13);

$$P^{-1}(t+h)=P^{-1}(t)-hP^{-1}(t)\dot{P}(t)P^{-1}(t)$$

it follows that (5.22) can be written as

$$\begin{gathered} \Lambda(t+h)=\left[P^{-1}(t)-hP^{-1}(t)\dot{P}(t)P^{-1}(t)\right]\cdot\left[A(t)+h\dot{A}(t)\right. \\ \left.-\big(B(t)+h\dot{B}(t)\big)\cdot\big(K(t)+h\dot{K}(t)\big)\right]\cdot\left[P(t)+h\dot{P}(t)\right]=\Lambda(t)+h\dot{\Lambda}(t). \end{gathered}$$

Expanding and rejecting high order terms yields:

$$\begin{gathered} \Lambda(t)+h\dot{\Lambda}(t)=\left[P^{-1}(t)-hP^{-1}(t)\dot{P}(t)P^{-1}(t)\right]\cdot\left[A(t)+h\dot{A}(t)-B(t)K(t)\right. \\ \left.-h\dot{B}(t)K(t)-hB\dot{K}(t)\right]\cdot\left[P(t)+h\dot{P}(t)\right] \end{gathered}$$

i.e.,

$$\begin{aligned} \Lambda(t)+h\dot{\Lambda}(t) = {} & P^{-1}(t)\big[A(t)-B(t)K(t)\big]P(t)+P^{-1}(t)h\big[A(t)-B(t)K(t)\big]\dot{P}(t) \\ & P^{-1}(t)h\big[\dot{A}(t)-\dot{B}(t)K(t)-B(t)\dot{K}(t)\big]P(t) \\ & -hP^{-1}(t)\dot{P}(t)P^{-1}(t)\big[A(t)-B(t)K(t)\big]P(t). \end{aligned} \tag{5.23}$$

Taking into account that

$$P^{-1}(t)\big[A(t)-B(t)K(t)\big]P(t)=\Lambda(t)$$

and

$$P^{-1}(t)h\big[A(t)-B(t)K(t)\big]\dot{P}(t)=h\Lambda(t)P^{-1}(t)\dot{P}(t),$$

(5.23) can be written as:

$$\begin{gathered} \Lambda(t)+h\dot{\Lambda}(t)=\Lambda(t)+h\Lambda(t)P^{-1}(t)\dot{P}(t)-hP^{-1}(t)\dot{P}(t)\Lambda(t) \\ +hP^{-1}(t)\big[\dot{A}(t)-\dot{B}(t)K(t)-B(t)\dot{K}(t)\big]P(t). \end{gathered} \tag{5.24}$$

Dividing by h on both sides a differential equation in $\Lambda(t)$ is obtained:

$$\dot{\Lambda}(t)=P^{-1}(t)\big[\dot{A}(t)-\dot{B}(t)K(t)-B(t)\dot{K}(t)\big]P(t)+\Lambda(t)P^{-1}(t)\dot{P}(t)-P^{-1}(t)\dot{P}(t)\Lambda(t)$$

or

$$P^{-1}(t)\left[\dot{A}(t)-\dot{B}(t)K(t)-B(t)\dot{K}(t)\right]P(t)=\dot{\Lambda}(t)-\Lambda(t)P^{-1}(t)\dot{P}(t)+P^{-1}(t)\dot{P}(t)\Lambda(t).$$

Multiplying on the left by $P^{-1}(t)$ and on the right by $P(t)$, then:

$$\begin{aligned}\left[\dot{A}(t)-\dot{B}(t)K(t)-B(t)\dot{K}(t)\right] &= P(t)\dot{\Lambda}(t)P^{-1}(t)-P(t)\Lambda(t)P^{-1}(t)\dot{P}(t)P^{-1}(t)\\ &\quad +P(t)P^{-1}(t)\dot{P}(t)\Lambda(t)P^{-1}(t) \qquad (5.25)\end{aligned}$$

so,

$$\begin{aligned}\left[\dot{A}(t)-\dot{B}(t)K(t)-B(t)\dot{K}(t)\right] &= P(t)\left[\dot{\Lambda}(t)-\Lambda(t)P^{-1}(t)\dot{P}(t)\right.\\ &\quad \left.+P^{-1}(t)\dot{P}(t)\Lambda(t)\right]P^{-1}(t). \qquad (5.26)\end{aligned}$$

□

To summarise: if $P(t)$, $A(t)$, $B(t)$, $K(t)$ and $\Lambda(t)$ are differentiable (which is required in order to prove Theorem 5.2, then (5.25) must be satisfied. If it is not, then Theorem 5.2 does not strictly apply. However, as shown in the following example, $P(t)$ may not be differentiable at a discrete set of points of the time interval $t \in [0,t]$ and the result will still hold.

Example 5.2. Given the following LTV open-loop system:

$$\dot{x}(t)=\begin{pmatrix} e^{cos(t)} & log\left[\frac{1}{1+t^2}\right] \\ t^2 & t \end{pmatrix}x(t)+\begin{pmatrix}1\\1\end{pmatrix}u(t), \quad x(0)=\begin{pmatrix}0.5\\0.5\end{pmatrix}$$

The aim is to set the closed-loop poles at $\sigma=(-8,-6)$. When the pole placement method is applied as in the previous example, it can be seen in Figure 5.6 that despite the poles being placed at the designed location, the shape of the response shows a discontinuity along the time interval and so does the designed control $u(t)=-K(t)x(t)$, Figure 5.7.

Plotting the profile of $\varepsilon(h)$ as in the previous example, it can be seen it reflects the two discontinuities at times $t=1.1$s and $t=2.68$s, where the condition for differentiability of $P(t)$ fails. In Figure 5.9 an estimate of the differentiability of $P(t)$ is shown, it is represented by the quantity $\frac{P(t+h)-P(t)}{h}$ calculated at each step h of the time interval. As expected it shows two discontinuities along the interval $[0,t_f]$, the first one happening at $t=1.1$s and the second one at $t=2.68$s, it does not show a smooth decreasing profile. On the other hand, if now the location of the poles is changed to be *i.e.* $\sigma=(-12,-10)$, Figures 5.10 and 5.11 show the components of the response and the control law for this choice of left-hand side poles.

This time it can be seen how the discontinuities in the stable responses and the control after the pole placement are smoother than in the previous case. The plots of epsilon $\varepsilon(h)$ in Figure 5.12 and differentiability of $P(t)$, $\frac{P(t+h)-P(t)}{h}$ in Figure 5.13 clearly show smaller discontinuities too, verifying the existence of the relation

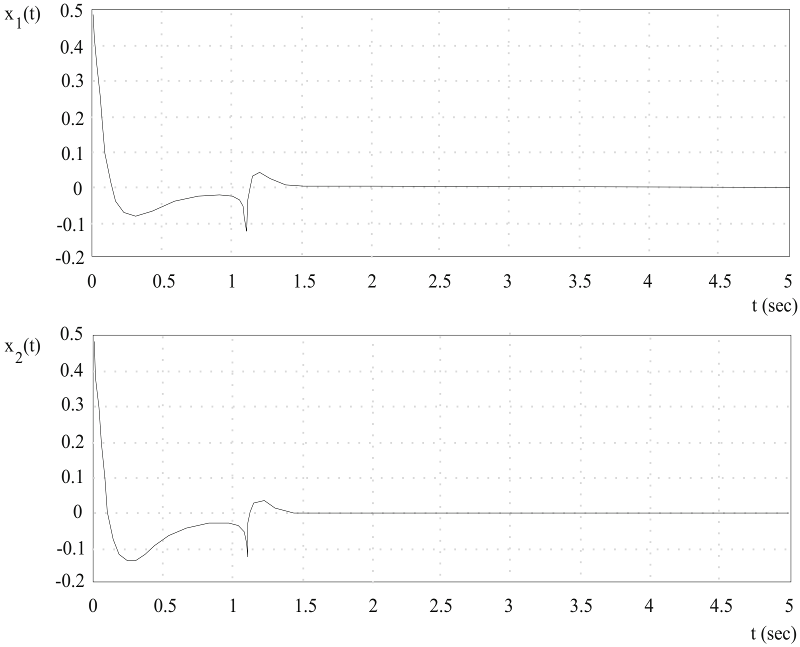


Fig. 5.6 Components of the response $x_1(t), x_2(t)$

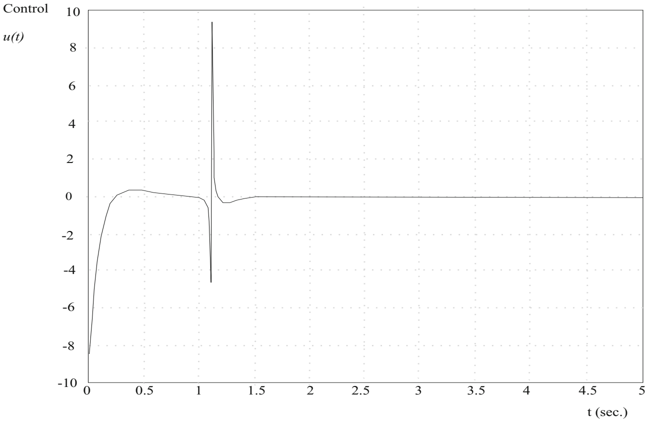


Fig. 5.7 Control $u(t) = -Kx(t)$

between $P(t)$, $\Lambda(t)$, $A(t)$, B(t) and $K(t)$ as indicated in (5.25). As the desired poles have change, so did $\Lambda(t)$ and consequently $K(t)$ and $P(t)$ and its differentiability.

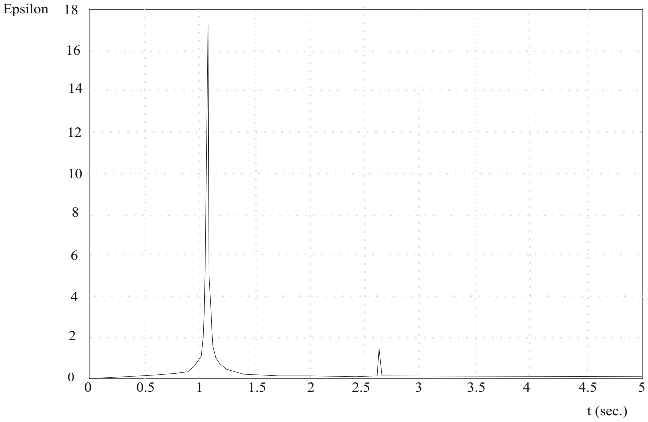


Fig. 5.8 $\varepsilon(h)$

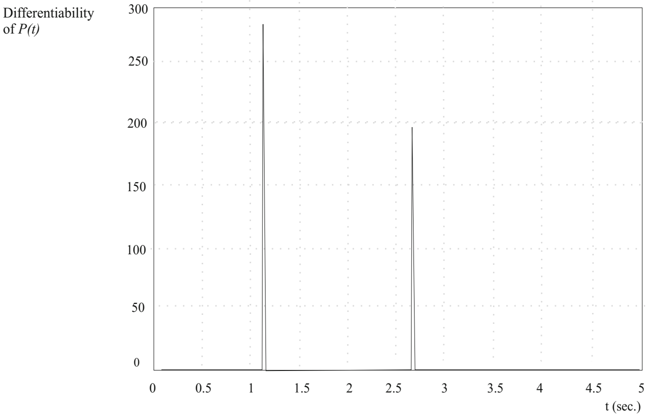


Fig. 5.9 Differentiability of $P(t)$

5.4 Generalisation to Nonlinear Systems

In this section an approach to the problem of pole placement when the system under consideration is nonlinear is presented. A nonlinear system of the form:

$$\dot{x} = A(x)x(t) + B(x)u(t), \quad x(0) = x_0 \tag{5.27}$$

where $A(x) \in \mathbb{R}^{nxn}, B(x) \in \mathbb{R}^{nxn}$, $u(t)$ is the control signal and $x(0) = x_0$ are some given initial conditions. (5.27) can be written as a sequence of LTV systems:

$$\dot{x}^{[1]} = A(x_0)x^{[1]}(t) + B(x_0)u^{[1]}(t), \quad x^{[1]}(0) = x_0$$

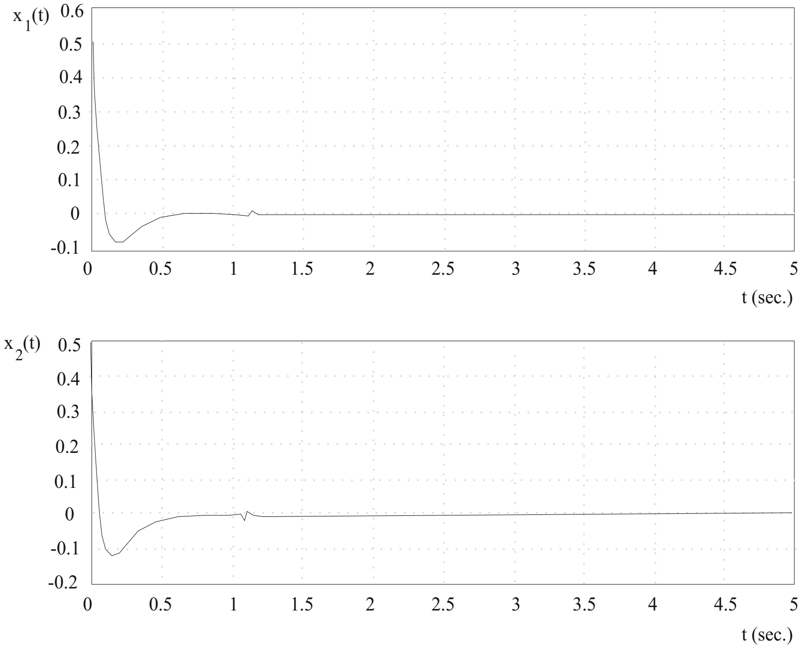


Fig. 5.10 Components of the response $x_1(t), x_2(t)$

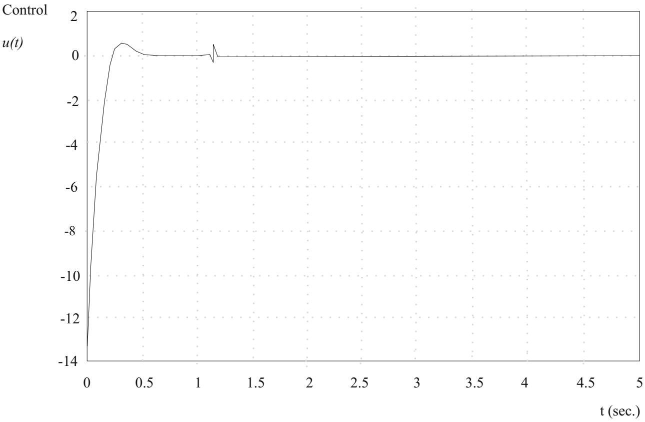


Fig. 5.11 Control $u(t) = -Kx(t)$

$$\vdots \tag{5.28}$$

$$\dot{x}^{[i]} = A(x^{[i-1]}(t))x^{[i]}(t) + B(x^{[i-1]}(t))u^{[i]}(t), \quad x^{[i]}(0) = x_0.$$

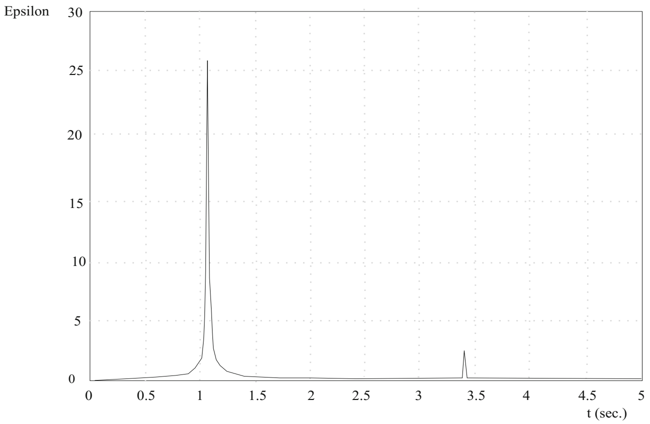


Fig. 5.12 $\varepsilon(h)$

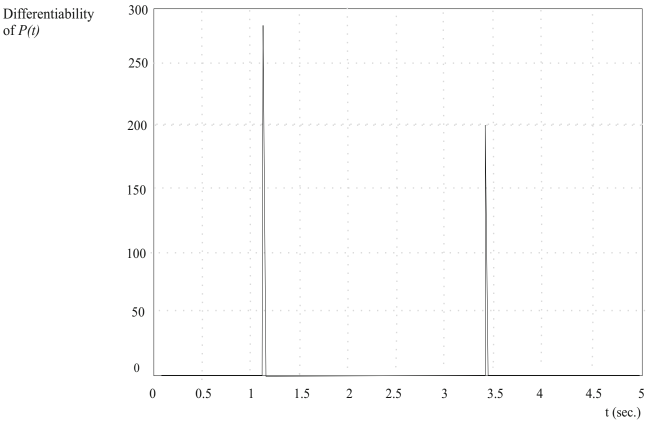


Fig. 5.13 Differentiability of $P(t)$

Applying the ideas introduced previously in Section 5.3, for some given choice of closed-loop poles, *i.e.* $\sigma = (\lambda_{1d}, \cdots \lambda_{nd})$, a sequence of feedback control laws of the form $u^{[i]}(t) = -K^{[i]}(t)x^{[i]}(t)$ can be generated at each iteration i, each of these $K^{[i]}$ is the feedback gain obtained to ensure stability on each of the iterates closed-loop forms:

$$\begin{aligned} \dot{x}^{[1]} &= [A(x_0) - B(x_0)K^{[1]}(t)]x^{[1]}(t), \quad x^{[1]}(0) = x_0. \\ &\vdots \\ \dot{x}^{[i]} &= [A(x^{[i-1]}(t)) - B(x^{[i-1]}(t))K^{[i]}(t)]x^{[i]}(t), \quad x^{[i]}(0) = x_0. \end{aligned} \tag{5.29}$$

As before, the eigenvalue placement theorem can be applied to each of these systems (5.29) being the set of desired poles $\sigma = (\lambda_{1d}, \cdots \lambda_{nd})$ the same for each iteration:

$$\begin{gathered}\det[\lambda \cdot I - A(x_0) + B(x_0)K^{[1]}(t)] = (\lambda - \lambda_{1d})\cdots(\lambda - \lambda_{nd}) \\ \vdots \\ \det[\lambda \cdot I - A(x^{[i-1]}(t)) + B(x^{[i-1]}(t))K^{[i]}(t)] = (\lambda - \lambda_{1d})\cdots(\lambda - \lambda_{nd}).\end{gathered} \tag{5.30}$$

Therefore each iterated equation of the sequence of LTV closed-loop systems (5.29) will be exponentially stable provided the conditions from Section 5.3, are satisfied. After a finite number of iterations, the solution $x^{[i]}(t)$ converges to the nonlinear solution $x(t)$. Then, the correspondent feedback $K^{[i]}(t)$ that stabilises the 'i^{th}' system, can be applied to the original nonlinear system in order to satisfy the stability requirements for this nonlinear closed-loop:

$$\dot{x} = [A(x(t)) - B(x(t))K^{[i]}(t)]x(t), \quad x(0) = x_0$$

provided that the desired eigenvalues $\sigma = (\lambda_1, \cdots \lambda_n)$ are chosen to be far on the left-half plane as stated in Section 5.3.

The exponential stability of the nonlinear system achieved as indicated here can summarised as follows:

Theorem 5.3. *Given a nonlinear system of the form (5.27) where the matrices $A(x)$ and $B(x)$ are Lipschitz and the pair (A,B) is controllable $\forall x(t), \quad \forall t \in [0,T]$, there exists a feedback control $u(t)$ given by:*

$$lim_{i\to\infty} u^{[i]}(t) = lim_{i\to\infty} K^{[i]}(t)x(t) \to u(t)$$

where $K^{[i]}(t)$ is Lipschitz, such that the solution $x(t)$ of the nonlinear system is exponentially stable in $[0,T]$.

Proof. If $K^{[i]}(t)$ is Lipschitz for each iteration, (in fact differentiability is a necessary condition for exponential stability of the LTV systems on the sequence) and assuming $A(x)$ and $B(x)$ are Lipschitz, the iteration technique can be applied. $K^{[i]}(t)$ can be written in canonical form:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ k_1(t) & k_2(t) & \ldots & k_n(t) \end{pmatrix}$$

from the pole placement algorithm an algebraic equation needs to be solved in order to obtain the elements of the feedback gain matrix $K^{[i]}(t)$;

$$\lambda^n + \Gamma_{n-1}^{[i]} \lambda^{n-1} + \Gamma_{n-2}^{[i]} \lambda^{n-2} + \ldots + \Gamma_1^{[i]} \lambda^1 + \lambda_0 = (\lambda - \lambda_1)(\lambda - \lambda_2)\ldots(\lambda - \lambda_n). \tag{5.31}$$

The coefficients $\Gamma_j^{[i]}$ contain a linear combination of the linear elements of $K^{[i]}(t) = (k_1^{[i]}(t), \ldots, k_n^{[i]}(t))$ so the identification of parameters can be done by equating the coefficients of both sides of Equation 5.31:

$$\begin{aligned} \Gamma_{n-1}^{[i]} &= \alpha_{n-1}^{[i]} + \beta_{n-1}^{[i]} \cdot k_{n-1}^{[i]}(t) = \phi_{n-1}(\lambda_1, \ldots, \lambda_n) \\ \Gamma_{n-2}^{[i]} &= \alpha_{n-2}^{[i]} + \beta_{n-2}^{[i]} \cdot k_{n-2}^{[i]}(t) = \phi_{n-2}(\lambda_1, \ldots, \lambda_n) \\ &\vdots \\ \Gamma_1^{[i]} &= \alpha_1^{[i]} + \beta_1^{[i]} \cdot k_1^{[i]}(t) = \phi_1(\lambda_1, \ldots, \lambda_n). \end{aligned} \tag{5.32}$$

Therefore, the elements of $K^{[i]}(t)$ of the feedback gain can be obtained by solving (5.32):

$$k_{n-1}^{[i]}(t) = \frac{\phi_{n-1}^{[i]}(\lambda_1,\ldots,\lambda_n) - \alpha_{n-1}^{[i]}}{\beta_{n-1}^{[i]}}, \ldots, k_1^{[i]}(t) = \frac{\phi_1^{[i]}(\lambda_1,\ldots,\lambda_n) - \alpha_1^{[i]}}{\beta_1^{[i]}}. \tag{5.33}$$

The functions $\alpha^{[i]}$ and $\beta^{[i]}$ at each iteration depend on those elements of $A(x^{[i-1]}(t))$ and $B(x^{[i-1]}(t))$ which are non-zero due to the pole placement so $K^{[i]}(t)$ is a Lipschitz function. Therefore, provided $K^{[i]}(t)$, $A(x)$ and $B(x)$ are Lipschitz then it follows from Theorem 5.2 that the sequence of exponentially stable solutions of (5.29) converges to the exponentially stable solution of the original nonlinear problem. □

Example 5.3. Given the following nonlinear system:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} cos(x_1) & -1 \\ 1 & -cos(x_1) \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot u(t) \tag{5.34}$$

with initial conditions $x(0) = [0.1, 0.1]^T$, the task is to obtain a state feedback control law $u(t) = -K(x(t))x(t)$ that places the closed-loop poles at $\lambda = (-10, -8)$ so that the transient response converges to zero. For this particular example, a time interval of $T = t_f = 3$s was used and the step length was $h = 0.01$.

The iteration technique as explained in Section 5.4 is applied following the next steps:

- Take as starting point the nonlinear system (5.34) and generate a sequence of open-loop LTV systems:

$$\begin{pmatrix} \dot{x}_1^{[1]} \\ \dot{x}_2^{[1]} \end{pmatrix} = \begin{pmatrix} cos(x_{01}) & -1 \\ 1 & cos(x_{01}) \end{pmatrix} \cdot \begin{pmatrix} x_1^{[1]} \\ x_2^{[1]} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot u^{[1]}(t). \tag{5.35}$$

$$\vdots$$

$$\begin{pmatrix} \dot{x}_1^{[i]} \\ \dot{x}_2^{[i]} \end{pmatrix} = \begin{pmatrix} cos(x_1^{[i-1]}) & -1 \\ 1 & cos(x_1^{[i-1]}) \end{pmatrix} \cdot \begin{pmatrix} x_1^{[i]} \\ x_2^{[i]} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot u^{[i]}(t)$$

In this particular example, only 4 iterations were needed in order to have good approach to the nonlinear system.

- Applying the pole placement method for LTV systems as shown in Section 5.3 for each of the iterated open-loop systems, the corresponding feedback gains $K^{i^{th}}(t)$ are generated according to the choice of left-half plane poles, in this case, $\sigma = (-10, -8)$.
- By selecting the last $K(t)$ of all the iterations generated (in this example, $K^{4^{th}}$) and applying this feedback gain to the closed-loop form of the system (5.34), the output (or solution) of the system is stabilised. The components of the output $x_1(t)$ and $x_2(t)$ have been plotted in Figure 5.14 proving to be stable after the state feedback procedure has been applied. The profile of the designed feedback control law can be observed in Figure 5.15.

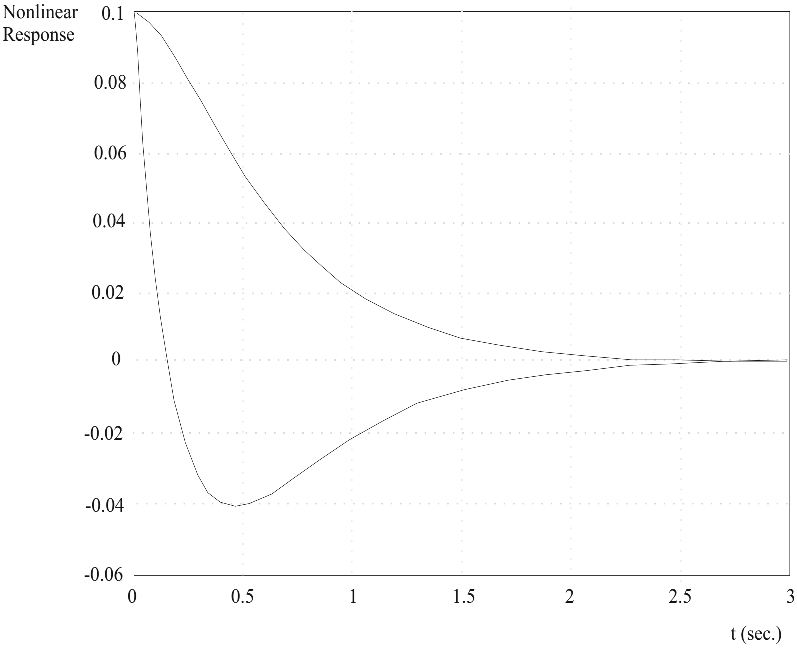


Fig. 5.14 Controlled components $x_1(t)$ and $x_2(t)$

5.5 Application to F-8 Crusader Aircraft

In this section the pole placement technique will be applied to the nonlinear equations of the F-8 aircraft in a level trim, unaccelerated flight at Mach=0.85 and

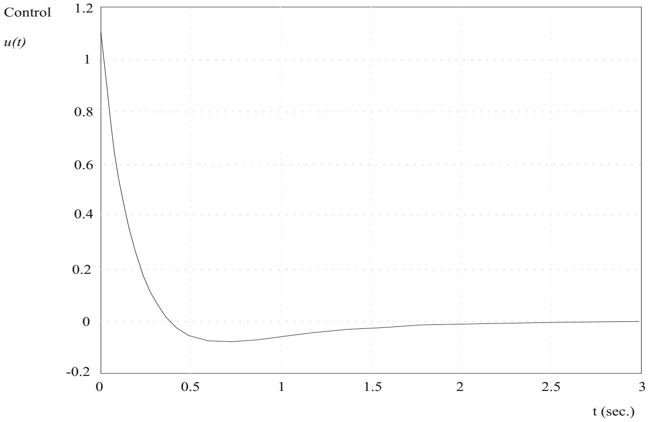


Fig. 5.15 Control laws $u(t) = -K^{(4^{th})}(t)x(t)$

altitude of $30,000$ ft (9000m), for which the nonlinear equations of motion representing the dynamics of the aircraft are [51]:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} -0.877 & 0 & 1 \\ 0 & 0 & 1 \\ -4.208 & 0 & -0.396 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{5.36}$$

$$+ \begin{pmatrix} -x_1^2 x_3 - 0.088 x_1 x_3 - 0.019 x_2^2 + 0.47 x_1^2 + 3.846 x_1^3 \\ 0 \\ -0.47 x_1^2 - 3.564 x_1^3 \end{pmatrix} + \begin{pmatrix} -0.215 \\ 0 \\ -20.967 \end{pmatrix} u(t)$$

where $x_1(t)$ is the angle of attack (rad), $x_2(t)$ the pitch angle (rad), $x_3(t)$ the pitch rate (rad s^{-1}) and $u(t)$ the control input.

The control objective is to place the *desired poles of the nonlinear system* on the left-hand side of the complex plane by applying simultaneously the iteration technique and the placement algorithm introduced in Section 5.3 for LTV plants.

The set of desired poles is $\sigma = \big(-10, -1.7108, -0.5129\big)$. This choice of poles corresponds to the closed-loop poles of the linearised and stabilised system when the control $\mu = -0.053x_1 + 0.5x_2 + 0.521x_3$ is applied [51].

The first step was to write Equation 5.36 on the form

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

and generate a sequence of LTV systems

$$\dot{x}^{[1]}(t) = \begin{pmatrix} \alpha_{11}^{[1]} & \alpha_{12}^{[1]} & \alpha_{13}^{[1]} \\ 0 & 0 & 1 \\ \alpha_{31}^{[1]} & 0 & -0.396 \end{pmatrix} x^{[1]}(t) + \begin{pmatrix} -0.215 \\ 0 \\ -20.967 \end{pmatrix} u^{[1]}(t)$$

$$\vdots$$

$$\dot{x}^{[i]}(t) = \begin{pmatrix} \alpha_{11}^{[i]} & \alpha_{12}^{[i]} & \alpha_{13}^{[i]} \\ 0 & 0 & 1 \\ \alpha_{31}^{[i]} & 0 & -0.396 \end{pmatrix} x^{[i]}(t) + \begin{pmatrix} -0.215 \\ 0 \\ -20.967 \end{pmatrix} u^{[i]}(t)$$

with

$$\begin{aligned} \alpha_{11}^{[1]} &= -0.877 + 0.47 x_1^{[i-1]} + 3.846 (x_1^{[i-1]})^2 \\ \alpha_{12}^{[1]} &= -0.019 x_2 \\ \alpha_{13}^{[1]} &= -x_1^2(0) - 0.088 x_1(0) \\ \alpha_{31}^{[1]} &= -4.208 - 0.47 x_1(0) - 3.564 x_1^2(0) \end{aligned}$$

$$\begin{aligned} \alpha_{11}^{[i]} &= -0.877 + 0.47 x_1^{[i-1]} + 3.846 (x_1^{[i-1]})^2 \\ \alpha_{12}^{[i]} &= -0.019 x_2^{[i-1]} \\ \alpha_{13}^{[i]} &= -(x_1^{[i-1]})^2 - 0.088 x_1^{[i-1]} \\ \alpha_{31}^{[i]} &= -4.208 - 0.47 x_1^{[i-1]} - 3.564 (x_1^{[i-1]})^2 \end{aligned}$$

where the initial condition vector is $x(0) = [0.5253, 0, 0]^T$.

At each iteration, a feedback law $u^{[i]}(t) = -K^{[i]}(t) x^{[i]}(t)$ is designed following the specifications: the closed-loop poles at each iteration should be allocated at $\lambda_d = \big(-10, -1.7108, -0.5129\big)$,

$$\dot{x}^{[i]}(t) = \big[A(x^{[i-1]}(t)) - B(x^{[i-1]}(t)) K^{[i]}(t)\big] x^{[i]}(t) = \tilde{A}(x^{[i-1]}(t)) x^{[i]}(t)$$

where $\tilde{A}(x^{[i-1]}(t))$ is the closed-loop matrix for the ^{ith}-iteration. Now, using Ackerman's formula:

$$\det\big[\lambda \cdot I - \tilde{A}(x^{[i-1]}(t))\big] = \big(\lambda - \lambda_1\big)\big(\lambda - \lambda_2\big)\big(\lambda - \lambda_3\big), \tag{5.37}$$

the feedback matrix $K^{[i-1]}(t)$ at each iteration is obtained. The simulations were done for $t_f = 15$ sec. After 30 iterations, the sequence of LTV systems converge to the nonlinear system; taking the 30^{th} feedback control and applying this to the nonlinear system,

$$\dot{x}(t) = A(x)x(t) - B(x)K^{(30)}(t)x(t)$$

it can be seen in figure 5.16, how the states of the nonlinear system converge to zero. The control law applied to the nonlinear system is shown in figure 5.17, where it can be seen how it presents an isolated discontinuity in the differentiability of the matrix of eigenvalues $P(t)$; this does not affect the states as seen in Figure 5.16.

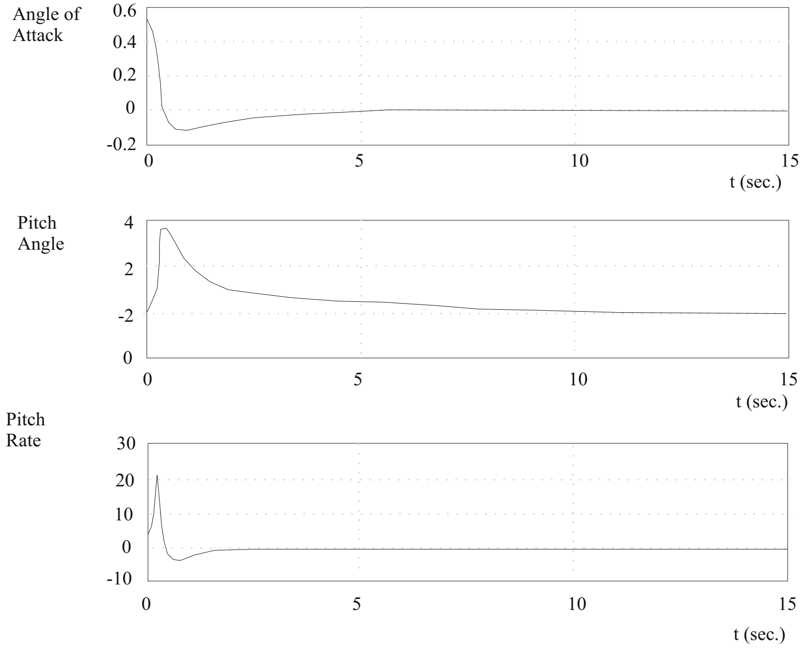


Fig. 5.16 Closed-loop response and the states $x_1(t)$, $x_2(t)$ and $x_3(t)$

5.6 Conclusions

In this chapter a pole placement algorithm for nonlinear systems has been presented. The method is based on the application of an iteration technique that replaces the nonlinear system by a sequence of LTV systems.

Once this sequence of linear time varying systems has been obtained, a standard pole-placement procedure is applied for each of the LTV systems by dividing the interval in N steps of length h and applying Duhamel's principle. It has been shown how this method alone does not guarantee stability for linear, time-varying systems and therefore additional requirements for stability were developed in Section 5.4:

If the matrices $A(t)$, $B(t)$, $P(t)$ and $K(t)$ are differentiable, then writing Equation 5.25 in the form:

$$\dot{\Lambda} = P^{-1}(t)\left(\dot{A}(t) - \dot{B}(t)K(t) - B(t)\dot{K}(t)\right)P(t) + \Lambda(t)P^{-1}(t)\dot{P}(t) - P^{-1}(t)\dot{P}(t)\Lambda(t)$$

gives a coupled equation relating $P(t)$, $K(t)$ and $\Lambda(t)$ which states that these are not independent. Hence, in general, it may not be possible (in some cases) to choose Λ constant. Thus, Equation 5.6 is an important condition for the exponential stability of the already pole placed LTV system.

The restriction it places on $P(t)$, $K(t)$ and $\Lambda(t)$ could be the object of further research.

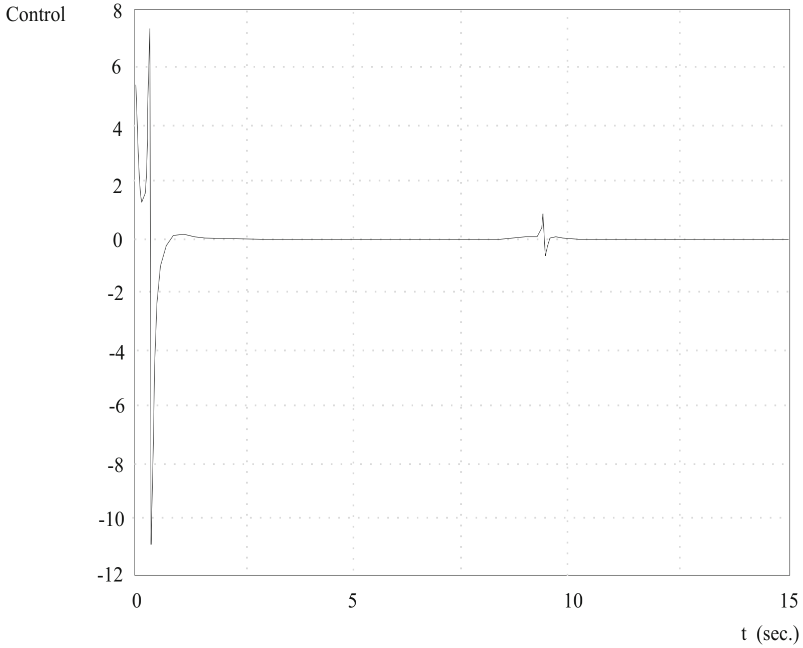


Fig. 5.17 Control law $u(t)$

These results were extended to nonlinear systems by the convergence of the iteration technique, thus the feedback gain designed for the last of the LTV iterated systems is applied to the nonlinear system and achieving in this way exponential stability. Due to the accurate approach of the iteration technique to the original nonlinear plant, this pole placement method results in a more robust method than those relying on the linearisation of the original system, at least the uncertainties of the unmodelled original dynamics do not exist in this case.

Some numerical examples were presented showing how the technique works and showing that, even in the case where differentiability of $P(t)$ is not satisfied at every point of the time interval $[0,t]$, the nonlinear system can be stabilised using this technique.

References

1. Slotine, E., Li, W.: Applied Nonlinear Control. Prentice Hall, United States (1991)
2. Chow, J.H.: A Pole-Placement Design Approach for Systems with Multiple Operating Conditions. IEEE Transactions on Automatic Control 35(3), 278–288 (1990)
3. Blondel, V.: Simultaneous Stabilisation of Linear Systems. Springer, London (1994)
4. Yan, X.G., Edwards, C., Spurgeon, S.K.: Decentralised robust sliding mode control for a class nonlinear interconnected systems by static output feedback. Automatica 40, 613–620 (2004)

5. Lu, X.Y., Spurgeon, S.K.: Output feedback stabilisation of SISO nonlinear systems via dynamic sliding modes. Int. J. Control 70, 735–759 (1998)
6. Tomás-Rodríguez, M., Banks, S.P.: Linear Approximations to Nonlinear Dynamical Systems with Applications to Stability and Spectral Theory. IMA J. Math. Cont. and Inf. 20, 89–104 (2003)
7. Frasca, R., Iannelli, L., Vasca, F.: Boundary Layer Using Dithering in Sliding Mode Control. In: 16th IFAC World Congress, Prague, Czech Republic (July 2005)
8. Tomás-Rodríguez, M., Navarro Hernandez, C., Banks, S.P.: Parametric Approach to Optimal Nonlinear Control Problem using Orthogonal Expansions. In: 16th IFAC World Congress, Prague, Czech Republic (July 2005)
9. Navarro Hernandez, C., Banks, S.P., Aldeen, M.: Observer Design for Nonlinear Systems using Linear Approximations. IMA J. Math. Cont. and Information 20, 359–370 (2003)
10. Tomás-Rodríguez, M., Banks, S.P.: Pole placement for nonlinear systems. In: NOLCOS 2004, Stuttgart, Germany (2004)
11. Cimen, T., Banks, S.P.: Nonlinear optimal tracking control with application to supertankers for autopilot design. Automatica 40, 1845–1863 (2004)
12. Jae, W.C., Ho, C.L., Zhu, J.J.: Decoupling and tracking control using eigenstruture assignment for LTV systems. Int. J. Control 74, 453–464 (1990)
13. Ackerman, J.: Abtastregelung. Springer, New York (1972)
14. Greschak, J.P., Verghese, G.C., Verghese, W.: Periodically varying compensation of time-invariant systems. Syst. Control Lett. 2, 88–93 (1990)
15. Aeyels, D., Willems, J.L.: Pole assignment for linear time-invariant second-order systems by periodic static output feedback. IMA J. of Mathematical Control and Information 8, 267–274 (1991)
16. Aeyels, D., Willems, J.L.: Pole assignment for linear time-invariant systems by periodic memoryless output feedback. Automatica 28, 1159–1168 (1992)
17. Peuteman, J., Aeyels, D.: Exponential Stability of Slowly Time-Varying Nonlinear Sytems. Math. Control Signals Systems 15, 202–228 (2002)
18. Silverman, L.M.: Transformation of Time-variable Systems to Canonical (Phase-variable) Form. IEEE Trans. Autom. Control 11, 300–303 (1966)
19. Tuel, W.G.: On the Transformation to (Phase-variable) Canonical Form. IEEE Trans. Autom. Control 12, 607 (1967)
20. Luenberger, D.G.: Canonical forms for linear multivariable systems. IEEE Trans. Autom. Control 12, 290–292 (1967)
21. Kailath, T.: Linear Systems. Prentice Hall, Englewood Cliffs
22. Varga, A.: A Schur method for pole assignment. IEEE Trans. Autom. Control 26, 517–519 (1981)
23. Miminis, G.S., Paige, C.C.: An algorithm for pole assigment of time-invariant linear systems. Int. J. Control 45(5), 341–354 (1982)
24. Petkov, P.H., Christov, N.D., Konstantinov, M.M.: Computational algorithms for linear control systems: a brief review. Int. J. Syst. Sci. 16, 465–477 (1985)
25. Kautsky, J., Nichols, N.K., Van Dooren, P.: Robust pole assignment in linear state feedback. Int. J. Control 41, 1129–1155 (1985)
26. Blanchini, F.: New canonical form for pole placement. IEE Proc., Pt D 136, 314–316 (1989)
27. Tsai, J.S.H., Chiang, H.K., Sun, Y.Y.: Novel canonical forms for LTV multivariable systems and their application. In: Proc. American Control Conferences, Boston, vol. 1, pp. 337–342 (1991)
28. Nguyen, C.: Arbitrary eigenvalue assignments for linear time-varying multivariable control systems. Int. J. Control 45, 1051–1057 (1987)

29. Bhattacharyya, S.P., de Souza, E.: Pole assignment via Sylvester′s equation. Systems and Control Letters 1, 261–263 (1982)
30. Valasek, M., Olgac, N.: Efficient eigenvalue assignments for general Linear MIMO systems. Automatica 11, 1605–1617 (1995)
31. Valasek, M., Olgac, N.: An efficient pole placement technique for LTI SISO systems. IEE Control Theory Applic., Proc. D 142(5), 451–458 (1995)
32. Valasek, M., Olgac, N.: Pole placement for LTV non-lexicographically fixed MIMO systems. Automatica 35, 101–108 (1999)
33. Nguyen, C.: Arbitrary eigenvalue assignments for linear time-varying multivariable control systems. Int. J. Control 45, 1051–1057 (1987)
34. Choi, J.W., Lee, L.G., Kim, Y., Kang, T.: Design of an effective controller via disturbance accomodating left eigenstructure assingment. AIAA Journal of Guidance, Control and Dynamics 18, 347–354 (1995)
35. Choi, J.W., Lee, L., Suzuki, H., Suzuki, T.: Comments on Matrix Method of Eigenstructure Assignment: The multi-input Case with application. AIAA Journal of Guidance, Control and Dynamics 19, 983 (1996)
36. Choi, J.W.: Left eigenstructure assignment via the Sylvester equation. KSME International Journal 12, 1034–1040 (1998)
37. Isidori, A.: Nonlinear Control Systems: An introduction. Springer, New York (1989)
38. Sontang, E.: Mathematical Control theory: Deterministic Finite Dimensional Systems, 2nd edn. Springer, New York (1998)
39. Slotine, E., Li, W.: Applied Nonlinear Control. Prentice Hall, New Jersey (1991)
40. Kazantzis, N., Costas, K.: Singular PDEs and the single step formulation of feedback linearisation with pole placement. Systems and Control Letters 39, 115–122 (2000)
41. Francis, B.A., Wonham, W.M.: The internal model principle of control theory. Automatica 12, 457–465 (1976)
42. Abdelaziz, T., Valasek, M.: Pole placement for SISO linear systems by state-derivative feedback. IEE Proc.-Control Theory Appl. 151(4) (2004)
43. Bengtsson, G.: Output regulations and internal models- a frequency domain approach. Automatica 13, 333–345 (1997)
44. Ackermann, J.: Robustness Against Sensor Failures. Automatica 20, 211–215 (1984)
45. Franklin, S.N.: Design of robust flight control systems, Msc Thesis, University of Illinois, Urbana-Champaign, Coordinates Sciences Laboratory (1980)
46. Franklin, S.N., Ackermann, J.: Robust flight control- a design example. AIAA J. of Guidance and Control 4, 597 (1981)
47. Menon, P.K., Iragavarapu, V.R., Ohlmeyer, E.J.: Software Tools for Nonlinear Missile Autopilot Design. AIAA Optimal synthesis Inc., American Institute of Aeronautics and Astronautics (1999)
48. MIL-F-8785B, Flying Qualities of Piloted Airplanes, AIAA Optimal synthesis Inc., ASG (1969)
49. Botto, M.A., Babuska, R., Sa da Costa, J.: Discrete Time robust pole-placement design through global optimization. In: Proceedings 15th IFAC World Congress, Barcelona, Spain (2002)
50. Marcel, H., Maarten, S.: Stability and preformance of a variable gain controller with application to a DVD storage drive. Automatica 40, 591–602 (2004)
51. William, L.G., Jordan, J.M.: Design of Nonlinear Automatic Flight Control Systems. Automatica 13, 497–505 (1977)

Chapter 6
Optimal Control

6.1 Introduction

Optimal control is one of the main techniques of modern control design, as it has been for many years. The linear-quadratic theory of optimal control design is well established and has various forms including the receding horizon approach for a robust, easily implementable variation of the theory. It is also useful in H^{∞} control in the well-known state-space game theoretic formulation. Obtaining the 'best' controller in any given circumstance is clearly important, but for general nonlinear systems, one is led to the solution of an extremely difficult (in general, non-smooth) partial differential equation. This makes the existing general nonlinear theory very difficult to apply.

In this chapter we shall show how to use the iteration technique developed above to solve nonlinear optimal control problems. In the next section we shall outline the classical linear quadratic regulator theory and derive the optimal feedback control in terms of the solution of a Riccati equation. We shall also indicate the modifications necessary to solve the linear tracking problem. The generalisation to nonlinear systems will be given in Section 6.3 and some examples will be presented in Section 6.4. Some comments on viscosity solutions of the Hamilton-Jacobi-Bellman (HJB) equation and the optimality of the solution will be given in Section 6.5.

6.2 Calculus of Variations and Classical Linear Quadratic Control

We shall first derive Lagrange's variational equations in the simpler case of the Lagrange problem, *i.e.* minimise the cost functional

$$J(x) = \int_{t_0}^{t_f} \phi(x(t), \dot{x}(t), t)dt \tag{6.1}$$

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 101–121.
springerlink.com

for some function $\phi : \mathbb{R}^{2n+1} \to \mathbb{R}^+$, over all twice continuously differentiable (C^2) functions $x : [t_0, t_f] \to \mathbb{R}^n$. The classical necessary condition for an optimum (maximum or minimum) of a function $f : \mathbb{R}^n \to \mathbb{R}$ is, of course, that

$$\frac{\partial f}{\partial x_i} = 0,\ 1 \le i \le n.$$

In order to extend this to functionals, note that

$$\frac{\partial f}{\partial x_i}(\overline{x}) = \left.\frac{d}{d\varepsilon} f(\overline{x} + \varepsilon e_i)\right|_{\varepsilon=0}$$

where e_i is the standard i^{th} unit basis vector. Thus, the classical necessary condition becomes

$$\left.\frac{d}{d\varepsilon} f(\overline{x} + \varepsilon e_i)\right|_{\varepsilon=0} = 0 \ \text{ at } \overline{x}.$$

We define the *directional derivative* $\delta f(\overline{x}; y)$ of f at $\overline{x}$ in the direction y ($\in \mathbb{R}^n$) as

$$\delta f(\overline{x}; y) = \left.\frac{d}{d\varepsilon} f(\overline{x} + \varepsilon e_i)\right|_{\varepsilon=0}$$

and so we have the necessary condition

$$\delta f(\overline{x}; y) = 0,\ \text{for all } y \in \mathbb{R}^n.$$

Hence, we can immediately generalise this to functionals J as in (6.1) and obtain the necessary condition

$$\delta J(x; y) = \left.\frac{d}{d\varepsilon} J(x + \varepsilon y)\right|_{\varepsilon=0} = 0$$

where we now have $x, y \in C^2[t_0, t_f]$. We have, by (6.1),

$$J(x + \varepsilon y) = \int_{t_0}^{t_f} \varphi(x + \varepsilon y, \dot{x} + \varepsilon \dot{y}, t)\,dt$$

and so the necessary condition becomes

$$\begin{aligned} 0 = \delta J(x; y) &= \int_{t_0}^{t_f} \left.\frac{d}{d\varepsilon} \varphi(x + \varepsilon y, \dot{x} + \varepsilon \dot{y}, t)\right|_{\varepsilon=0} dt \\ &= \int_{t_0}^{t_f} (\varphi_x y + \varphi_{\dot{x}} \dot{y})\,dt \\ &= \int_{t_0}^{t_f} \left(\varphi_x - \frac{d}{dt}\varphi_{\dot{x}}\right) y\,dt + [\varphi_{\dot{x}} y]|_{t_0}^{t_f}. \end{aligned}$$

Since $y \in C^2[t_0, t_f]$ is arbitrary we must then have

$$\phi_x - \frac{d}{dt}\phi_{\dot{x}} = 0 \text{ Euler-Lagrange equation}$$

and

$$[\phi_{\dot{x}} y]|_{t_0}^{t_f} \text{ transversality.}$$

If we also have a vector constraint

$$G(x,\dot{x},t) = 0,$$

then, again as in the finite-dimensional case, we form the augmented cost

$$J_\lambda = \int_{t_0}^{t_f} \left\{\varphi(x,\dot{x},t) + \lambda^T G(x,\dot{x},t)\right\} dt$$

and again obtain the Euler-Lagrange equation

$$(\varphi + \lambda^T G)_x - \frac{d}{dt}(\varphi + \lambda^T G)_{\dot{x}} = 0.$$

In order to solve the linear-quadratic regulator problem, it is easiest to consider it as a general Bolza problem of the form

$$\min J = [\theta(x(t),t)]_{t_0}^{t_f} + \int_{t_0}^{t_f} \varphi(x(t),u(t),t)dt$$

subject to the dynamic constraint

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

Thus, as above, we consider the augmented cost functional

$$J_\lambda = [\theta(x(t),t)]_{t_0}^{t_f} + \int_{t_0}^{t_f} \left\{\varphi(x(t),u(t),t) + \lambda^T [f(x,u,t) - \dot{x}]\right\} dt$$

Next, we introduce the *Hamiltonian* function

$$H(x,u,\lambda,t) = \varphi(x,u,t) + \lambda^T f(x,u,t).$$

Then J_λ becomes

$$\begin{aligned} J_\lambda &= [\theta(x(t),t)]_{t_0}^{t_f} + \int_{t_0}^{t_f} \{H(x,u,\lambda,t) - \lambda^T \dot{x}\}dt \\ &= \left[\theta(x(t),t) - \lambda^T(t)x(t)\right]_{t_0}^{t_f} + \int_{t_0}^{t_f} \{H(x,u,\lambda,t) + \dot{\lambda}^T x\}dt. \end{aligned}$$

Proceeding as in the Lagrange problem, we take the directional derivative of J_λ at an assumed minimum point (x,u) in the direction (y,v), giving

$$\begin{aligned}\delta J_\lambda(x,u;y,v) = &\left\{ y^T \left(\frac{\partial \theta}{\partial x} - \lambda \right) \right\}\Bigg|_{t_0}^{t_f} \\ &+ \int_{t_0}^{t_f} \left\{ y^T \left(\frac{\partial H}{\partial x} + \dot{\lambda} \right) + v^T \frac{\partial H}{\partial u} \right\} dt.\end{aligned}$$

By the arbitrariness of (y,v), we therefore obtain

$$\left.\begin{aligned} \dot{x} = \tfrac{\partial H}{\partial \lambda} = f(x,u,t) \\ \dot{\lambda} = -\tfrac{\partial H}{\partial x} \end{aligned}\right\} \text{Hamilton's equations}$$

$$\frac{\partial H}{\partial u} = 0 \text{ control equation}$$

$$y^T \left(\frac{\partial \theta}{\partial x} - \lambda \right) = 0 \text{ at } t_0 \text{ and } t_f \text{ (transversality).}$$

We can now apply these equations to the special case of the linear, quadratic regulator problem:

$$\min J = \frac{1}{2} x^T(t_f) F x(t_f) + \frac{1}{2} \int_{t_0}^{t_f} \left(x^T(t) Q(t) x(t) + u^T(t) R(t) u(t) \right) dt$$

subject to the dynamic constraint

$$\dot{x} = Ax + Bu, \ \ x(t_0) = x_0.$$

The Hamiltonian for this problem is given by

$$H = \frac{1}{2} x^T(t) Q(t) x(t) + \frac{1}{2} u^T(t) R(t) u(t) + \lambda^T(t) A(t) x(t) + \lambda^T(t) B(t) u(t),$$

and so we obtain the necessary conditions

$$\begin{aligned} \dot{\lambda} &= -\frac{\partial H}{\partial x} = -Q(t)x(t) - A^T(t)\lambda(t) \\ \dot{x}(t) &= A(t)x(t) - B(t)R^{-1}(t)B^T(t)\lambda(t), \ \ x(t_0) = x_0 \end{aligned} \tag{6.2}$$

since the control is given by

$$\frac{\partial H}{\partial u} = 0$$

i.e.

$$u = -R^{-1}(t) B^T(t) \lambda(t).$$

The transversality condition gives the final value for λ:

$$\lambda(t_f) = \frac{\partial\theta}{\partial x}(t_f) = Fx(t_f).$$

The equations (6.2) represent a two-point boundary value problem, which can be solved by assuming that $\lambda(t)$ is of the form

$$\lambda(t) = P(t)x(t).$$

Thus,

$$\begin{aligned}\dot{\lambda}(t) &= \dot{P}(t)x(t) + P(t)\dot{x}(t)\\ &= \dot{P}(t)x(t) + P(t)\left(A(t)x(t) - B(t)R^{-1}(t)B^T(t)\lambda(t)\right)\end{aligned}$$

and so

$$-Q(t)x(t) - A^T(t)P(t)x(t) = \dot{P}(t)x(t) + P(t)(A(t)x(t) - B(t)R^{-1}(t)B^T(t)P(t)x(t))$$

i.e.

$$(\dot{P}(t) + P(t)A(t) + A^T(t)P(t) + Q(t) - P(t)B(t)R^{-1}(t)B^T(t)P(t))x(t) = 0.$$

Since the solution is unique, it is clearly given by solving the differential Riccati equation

$$\dot{P}(t) = -(P(t)A(t) + A^T(t)P(t) + Q(t) - P(t)B(t)R^{-1}(t)B^T(t)P(t)),\ \ P(t_f) = F.$$

Then we have the optimal control

$$u(t) = -R^{-1}(t)B^T(t)P(t)x(t),$$

and the controlled dynamics become

$$\dot{x}(t) = A(t)x(t) - B(t)R^{-1}(t)B^T(t)P(t)x(t).$$

If we now consider the tracking problem

$$\begin{aligned}\min J = &\ \frac{1}{2}\left(x^T(t_f) - x_d^T(t_f)\right)F\left(x(t_f) - x_d(t_f)\right)\\ &+\frac{1}{2}\int_{t_0}^{t_f}\left(\left(x^T(t) - x_d^T(t)\right)Q(t)\left(x(t) - x_d(t)\right) + u^T(t)R(t)u(t)\right)dt,\end{aligned}$$

where $x_d(t)$ is some desired trajectory (rather than the regulator problem), then we can modify the above solution as follows. First note that the Hamiltonian in this case is

$$\begin{aligned}H(x,u,\lambda,t) = &\ \frac{1}{2}\left(x^T(t) - x_d^T(t)\right)Q(t)\left(x(t) - x_d(t)\right) +\\ &u^T(t)R(t)u(t) + \lambda^T(A(t)x(t) + B(t)u(t))\end{aligned}$$

and so we have

$$u(t) = -R^{-1}B^T(t)\lambda(t),$$

$$\dot{\lambda} = -Q(t)\,(x(t) - x_d(t)) - A^T(t)\lambda(t)$$

and

$$\dot{x}(t) = A(t)x(t) - B(t)R^{-1}(t)B^T(t)\lambda(t), \;\; x(t_0) = x_0.$$

The transversality condition gives

$$\lambda(t_f) = F\left(x(t_f) - x_d(t_f)\right)$$

and so this time the clue is to take

$$\lambda(t) = P(t)x(t) + s(t)$$

for some $s(t)$ with $s(t_f) = -Fx_d(t_f)$. As before, we see that

$$\dot{\lambda}(t) = \dot{P}(t)x(t) + P(t)\dot{x}(t) + \dot{s}(t)$$

i.e.

$$\begin{aligned} -Q(t)(x(t) - x_d(t)) - A^T(t)\lambda(t) \;&=\; \dot{P}(t)x(t) + P(t)(A(t)x(t) - \\ &\qquad B(t)R^{-1}(t)B^T(t)\lambda(t)) + \dot{s}(t) \end{aligned}$$

or

$$\begin{gathered} -Q(t)(x(t) - x_d(t)) - A^T(t)\,(P(t)x(t) + s(t)) = \dot{P}(t)x(t) + \\ P(t)\left(A(t)x(t) - B(t)R^{-1}(t)B^T(t)(P(t)x(t) + s(t))\right) + \dot{s}(t) \end{gathered}$$

and taking P to satisfy the same Riccati equation as above, we obtain for $s(t)$:

$$Q(t)x_d(t) - A^T(t)s(t) = -P(t)B(t)R^{-1}(t)B^T(t)s(t) + \dot{s}(t)$$

i.e.

$$\dot{s}(t) = \left(P(t)B(t)R^{-1}(t)B^T(t) - A^T(t)\right)s(t) + Q(t)x_d(t), \quad s(t_f) = -Fx_d(t_f).$$

This is a time-varying differential equation for $s(t)$ with forcing term $Q(t)x_d(t)$ (the feedforward term).

6.3 Nonlinear Control Problems

We now come to the case of nonlinear control problems with possibly non-quadratic cost functionals. Consider a nonlinear system of the form

$$\dot{x}(t) = A(x(t),u(t))x(t) + B(x(t),u(t))u(t), \quad x(t_0) = x_0. \tag{6.3}$$

(We could also consider an output equation

$$y(t) = C(x(t),u(t))x(t) + D(x(t),u(t))u(t)$$

in a similar way, but for simplicity, we shall just develop the case of full observation – for the application to nonlinear observers, see [1].)

Together with the dynamical constraint (6.3), we consider the non-quadratic cost functional

$$\begin{aligned} J = \frac{1}{2}x^T(t_f)Fx(t_f) + \frac{1}{2}\int_{t_0}^{t_f} \big(x^T(t)Q(x(t),u(t))x(t) + \\ u^T(t)R(x(t),u(t))u(t)\big)\, dt \end{aligned} \tag{6.4}$$

where $Q(\cdot,\cdot)$ $(R(\cdot,\cdot))$ is a positive semi-definite (definite) matrix-valued function. We rewrite Equations (6.3) and (6.4) as a sequence of linear, quadratic (time-varying) problems:

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t),u^{[i-1]}(t))x^{[i]}(t) + B(x^{[i-1]}(t),u^{[i-1]}(t))u^{[i]}(t), \quad x^{[i]}(t_0) = x_0 \tag{6.5}$$

and

$$\begin{aligned} J^{[i]} = \frac{1}{2}x^{[i]T}(t_f)Fx^{[i]}(t_f) + \frac{1}{2}\int_{t_0}^{t_f} \Big(x^{[i]t}(t)Q(x^{[i-1]}(t),u^{[i-1]}(t))x^{[i]}(t) + \\ u^{[i]T}(t)R(x^{[i-1]}(t),u^{[i-1]}(t))u^{[i]}(t)\Big)\, dt. \end{aligned} \tag{6.6}$$

To start the process we can choose a zero control and put

$$x^{[1]}(t) = x_0,$$

so that we get the first approximation as a solution to

$$\dot{x}^{[1]}(t) = A(x^{[1]},0)x^{[1]}(t) + B(x^{[1]},0)u^{[1]}(t), \quad x^{[1]}(t_0) = x_0$$

and

$$\begin{aligned} J^{[1]} = \frac{1}{2}x^{[1]T}(t_f)Fx^{[1]}(t_f) + \frac{1}{2}\int_{t_0}^{t_f} \Big(x^{[1]T}(t)Q(x_{[1]},0)x^{[1]}(t) + \\ u^{[1]T}(t)R(x_0,0)u^{[1]}(t)\Big)\, dt. \end{aligned}$$

However, it may be better numerically to take

$$x^{[1]}(t) = e^{-t}x_0$$

since we are trying to stabilise the system. The solution of the problem (6.5), (6.6) is given by

$$u^{[i]}(t) = -R^{-1}(x^{[i-1]}(t), u^{[i-1]}(t))B^T(x^{[i-1]}(t), u^{[i-1]}(t))P^{[i]}(t)x^{[i]}(t)$$

where $P^{[i]}(t)$ satisfies the Riccati equation

$$\begin{aligned}\dot{P}^{[i]}(t) = {} & -A^T(x^{[i-1]}(t), u^{[i-1]}(t))P^{[i]}(t) - P^{[i]}(t)A(x^{[i-1]}(t), u^{[i-1]}(t)) - \\ & Q(x^{[i-1]}(t), u^{[i-1]}(t)) + \\ & P^{[i]}(t)B(x^{[i-1]}(t), u^{[i-1]}(t))R^{-1}(x^{[i-1]}(t), u^{[i-1]}(t)) \times \\ & B^T(x^{[i-1]}(t), u^{[i-1]}(t))P^{[i]}(t), \;\; P^{[i]}(t_f) = F.\end{aligned}$$

If these Riccati equations have solutions on $[0, t_f]$, then the convergence theory of Chapter 2 shows that these equations converge to a solution

$$u^*(t) = -R^{-1}(x^*(t), u^*(t))B^T(x^*(t), u^*(t))P^*(t)x^*(t)$$

which will control the nonlinear system, giving the controlled dynamics

$$\begin{aligned}\dot{x}^*(t) = {} & A(x^*(t), u^*(t))x^*(t) - \\ & B(x^*(t), u^*(t))R^{-1}(x^*(t), u^*(t))B^T(x^*(t), u^*(t))P^*(t)x^*(t), \; x^*(t_0) = x_0.\end{aligned}$$

Now consider the case of a tracking problem of the form

$$\dot{x}(t) = A(x(t), u(t))x(t) + B(x(t), u(t))u(t), \;\; x(t_0) = x_0$$

together with the cost functional

$$\begin{aligned}J = {} & \frac{1}{2}\left(x^T(t_f) - x_d^T(t_f)\right)F\left(x(t_f) - x_d(t_f)\right) + \\ & \frac{1}{2}\int_{t_0}^{t_f}\left(\left(x^T(t) - x_d^T(t)\right)Q(x(t), u(t))\left(x(t) - x_d(t)\right) + \right. \\ & \left. u^T(t)R(x(t), u(t))u(t)\right)dt.\end{aligned}$$

As before, we introduce the sequence of approximations

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t), u^{[i-1]}(t))x^{[i]}(t) + B(x^{[i-1]}(t), u^{[i-1]}(t))u^{[i]}(t), \;\; x^{[i]}(t_0) = x_0$$

and

$$\begin{aligned}J^{[i]} = {} & \frac{1}{2}\left(x^{[i]T}(t_f) - x_d^T(t_f)\right)F\left(x^{[i]}(t_f) - x_d(t_f)\right) + \\ & \frac{1}{2}\int_{t_0}^{t_f}\left(\left(x^{[i]T}(t) - x_d^T(t)\right)Q(x^{[i-1]}(t), u^{[i-1]}(t))\left(x^{[i]}(t) - x_d(t)\right) + \right. \\ & \left. u^{[i]T}(t)R(x^{[i-1]}(t), u^{[i-1]}(t))u^{[i]}(t)\right)dt\end{aligned}$$

with the first approximation

$$\dot{x}^{[0]}(t) = A(x_0,0)x^{[0]}(t) + B(x_0,0)u^{[0]}(t), \quad x^{[0]}(t_0) = x_0$$

and

$$\begin{aligned} J^{[0]} = & \frac{1}{2}\left(x^{[0]T}(t_f) - x_d^T(t_f)\right) F \left(x^{[0]}(t_f) - x_d(t_f)\right) + \\ & \frac{1}{2}\int_{t_0}^{t_f} \left(\left(x^{[0]T}(t) - x_d^T(t)\right) Q(x_0,0) \left(x^{[0]}(t) - x_d(t)\right) + \right. \\ & \left. u^{[0]T}(t)R(x_0,0)u^{[0]}(t)\right) dt. \end{aligned}$$

The solution is

$$u^{[i]}(t) = -R^{-1}(x^{[i-1]}(t),u^{[i-1]}(t))B^T(x^{[i-1]}(t),u^{[i-1]}(t))\left(P^{[i]}(t)x^{[i]}(t) + s^{[i]}(t)\right)$$

where

$$\begin{aligned} \dot{P}^{[i]}(t) = & -A^T(x^{[i-1]}(t),u^{[i-1]}(t))P^{[i]}(t) - P^{[i]}(t)A(x^{[i-1]}(t),u^{[i-1]}(t)) - \\ & Q(x^{[i-1]}(t),u^{[i-1]}(t)) + \\ & P^{[i]}(t)B(x^{[i-1]}(t),u^{[i-1]}(t))R^{-1}(x^{[i-1]}(t),u^{[i-1]}(t)) \times \\ & B^T(x^{[i-1]}(t),u^{[i-1]}(t))P^{[i]}(t), \ \ P^{[i]}(t_f) = F. \end{aligned}$$

and

$$\begin{aligned} \dot{s}^{[i]}(t) = & \left(P^{[i]}(t)B(x^{[i-1]}(t),u^{[i-1]}(t)))R^{-1}(x^{[i-1]}(t),u^{[i-1]}(t))B^T(x^{[i-1]}(t),u^{[i-1]}(t)) - \right. \\ & \left. A^T(x^{[i-1]}(t),u^{[i-1]}(t))s^{[i]}(t) + Q(x^{[i-1]}(t),u^{[i-1]}(t))x_d(t)\right) \\ & s^{[i]}(t_f) = -Fx_d(t_f). \end{aligned}$$

In the next section we shall illustrate the above theory with two examples which will show the effectiveness of the method.

6.4 Examples

We now present two examples of very different types which show that the method has very diverse applications. First we consider a spacecraft with a large flexible structure such as a solar array (to provide sustainable energy during space flight). When we require to maneuver the attitude of the flexible spacecraft, the dynamic coupling between the solar panel vibration and the spacecraft attitude varies with the angle of attitude maneuver. The attitude maneuver of the spacecraft will induce vibration in the flexible solar array, which must be suppressed. Quick and precise response to the attitude command, while at the same time maintaining certain levels of suppression of the vibration modes is a major objective of the attitude control system. However, these are conflicting factors and a trade-off between them is necessary. Here we give a brief outline of the results of applying our technique; for more details, see [2].

The equations of motion of the system are given by

$$\begin{aligned} I_x\ddot{\varphi} + [(I_y - I_z - I_x)\omega_0 - h_y]\dot{\psi} + [(I_z - I_y)\omega_0\varphi + h_z]\dot{\theta} \; + \; (I_z - I_y)\psi\dot{\theta} \\ +[(I_y - I_z)\omega_0^2 - h_y\omega_0]\varphi - h_z\omega_0 + \dot{h}_x + \sum_{i=1}^{n} F_{sxi}\ddot{\eta}_{pi} \; = \; M_x^e \end{aligned}$$

$$\begin{aligned} & I_y\ddot{\theta} + [(I_x - I_z)\dot{\varphi} + (I_z - I_x)\omega_0\psi + h_x]\dot{\psi} \\ & \quad +[(I_x - I_z)\omega_0\dot{\varphi} + (I_z - I_x)\omega_0^2\psi + h_x\omega_0]\varphi \\ & \quad -h_z\dot{\varphi} + h_z\omega_0\psi + 3\omega_0^2(I_x - I_z)\theta + \dot{h}_y + \sum_{i=1}^{n} F_{syi}\ddot{\eta}_{pi} = M_y^e \end{aligned}$$

$$\begin{aligned} I_z\ddot{\psi} + [(I_x + I_z - I_y)\omega_0 + h_y]\dot{\varphi} + [(I_x - I_y)\omega_0\psi - h_x]\dot{\theta} \; + \; (I_y - I_x)\dot{\varphi}\dot{\theta} \\ +[(I_y - I_x)\omega_0^2 - h_y\omega_0]\psi + h_x\omega_0 + \dot{h}_z + \sum_{i=1}^{n} F_{szi}\ddot{\eta}_{pi} \; = \; M_z^e \end{aligned}$$

and

$$\begin{aligned} & \ddot{\eta}_{pi} + 2\xi_{pi}\omega_{pi}\dot{\eta}_{pi} + \omega_{pi}^2\eta_{pi} + F_{sxi}^T(\ddot{\varphi} - \omega_0\dot{\psi}) + F_{syi}^T\ddot{\theta} \\ & \quad +F_{szi}^T(\ddot{\psi} + \omega_0\dot{\varphi}) = 0, \quad (i = 1, \cdots, m), \end{aligned}$$

where the I's are principal moments of inertia, φ, θ, ψ are the roll, pitch and yaw, M_x^e, M_y^e, M_z^e are the external active and environment control torques and the F's are the coupling matrices between the attitude and vibration modes. (For the other variables, see [2].) If $u_c = [u_1, u_2, u_3]$ is the inner control torque generated by the control motors, then these equations can be written in the form

$$\begin{aligned} \dot{x} &= A(x)x + Bu_c + Bu_e \\ \dot{x}_w &= A_w(x)x_w + B_w u_c, \end{aligned}$$

where

$$\begin{aligned} x &= [\varphi, \dot{\varphi}, \theta, \dot{\theta}, \psi, \dot{\psi}, \eta_{pi}, \dot{\eta}_{pi}]^T \\ x_w &= [h_x, h_y, h_z]^T \\ u_c &= [u_1, u_2, u_3]^T \\ u_e &= [M_x^e, M_y^e, M_z^e]^T \end{aligned}$$

and A, A_w, B, B_w are appropriate matrices. Typical responses for the body motion and the vibrational modes (using the iteration control theory) are shown in

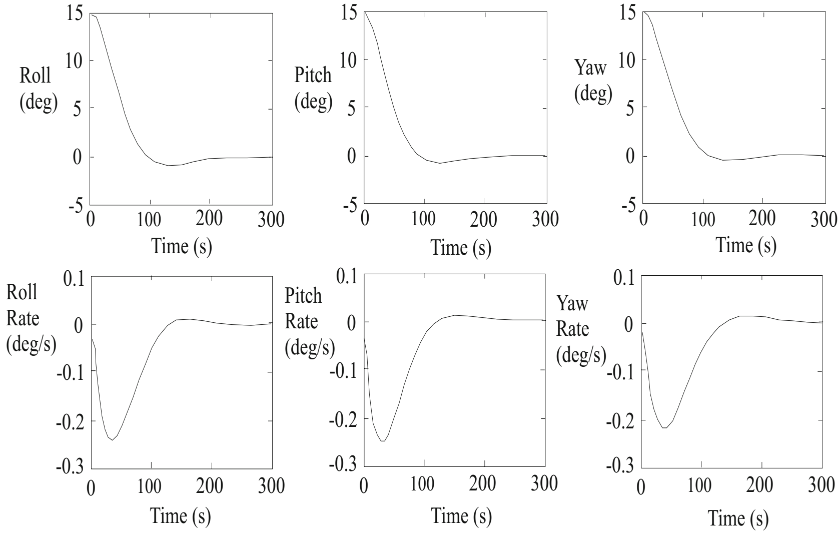


Fig. 6.1 Responses for attitude angle and angle rate

Figures 6.1, 6.2 and 6.3. The second example is the positioning of a large oil tanker (Figure 6.4) with the model given by the equations (for more details see [3]):

surge, sway and yaw dynamics

$$
\begin{aligned}
\dot{u} = {} & \frac{1}{L(1 - X''_{\dot{u}} - X''_{\dot{u}\zeta}\zeta)} \left\{ X''_{uu}u^2 + L(1 + X''_{vr})vr + X''_{vv}v^2 \right. \\
& + X''_{c|c|\delta\delta}|c|c\delta^2 + X''_{c|c|\beta\delta}|c|c\beta\delta + LgT''(1 - t_d) \\
& \left. + X''_{uu\zeta}u^2\zeta + LX''_{vr\zeta}vr\zeta + X''_{vv\zeta\zeta}v^2\zeta^2 \right\}
\end{aligned}
$$

$$
\begin{aligned}
\dot{v} = {} & \frac{1}{L(1 - Y''_{\dot{v}} - Y''_{\dot{v}\zeta}\zeta)} \left\{ Y''_{uv}uv + L(1 + Y''_{v|v|})v|v| + Y''_{c|c|\delta}c|c|\delta \right. \\
& + L(Y''_{ur} - 1)ur + Y''_{c|c|\beta|\beta||\delta|}|c|c\beta|\beta||\delta| + LY''_{T}gT'' \\
& + LY''_{ur\zeta}ur\zeta + Y''_{uv\zeta}uv\zeta + Y''_{v|v|\zeta}|v|v\zeta \\
& \left. + Y''_{c|c|\beta|\beta||\delta|\zeta}c|c|\beta|\beta||\delta|\zeta \right\}
\end{aligned}
$$

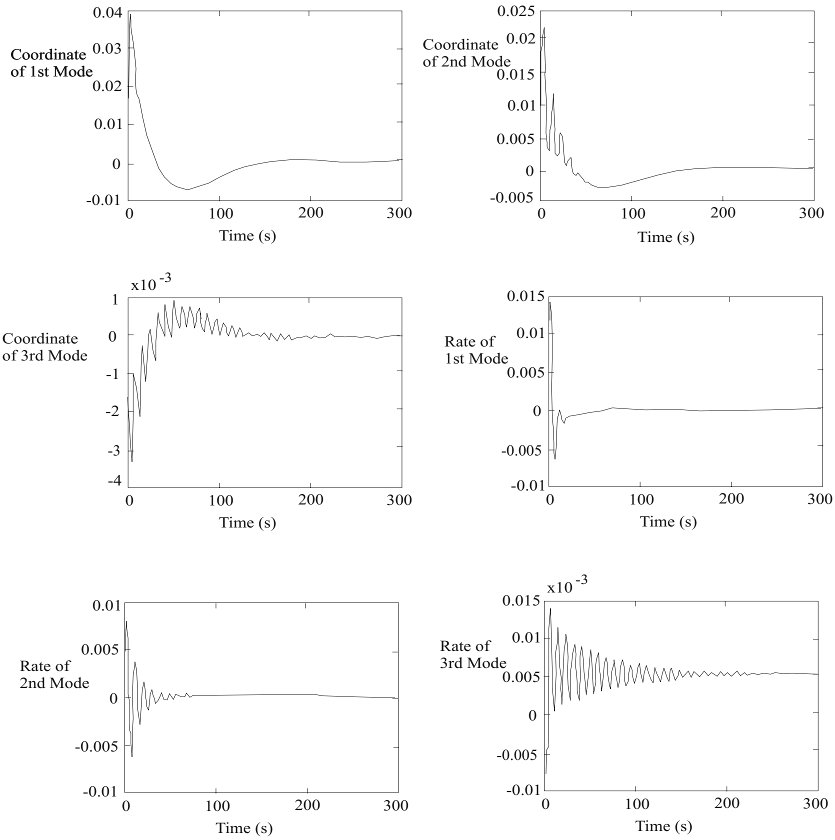


Fig. 6.2 Coordinate and rate responses of 1st – 3rd vibration modes

$$\begin{aligned}\dot{r} = \frac{1}{L^2(k''_{zz}{}^2 - N''_{\dot{r}} - N''_{\dot{r}\zeta}\zeta)} & \Big\{N''_{uv}uv + LN''_{|v|r}|v|r \\ & + N''_{c|c|\delta}c|c|\delta + L(N''_{ur} - x''_G)ur + N''_{c|c|\beta|\beta||\delta|}c|c|\beta|\beta||\delta| \\ & + LN''_T gT'' + LN''_{ur\zeta}ur\zeta + N''_{uv\zeta}uv\zeta + LN''|v|r\zeta|v|r\zeta \\ & + N''_{c|c|\beta|\beta||\delta|\zeta}c|c|\beta|\beta||\delta|\zeta\Big\}.\end{aligned}$$

kinematic equations

$$\begin{aligned}\dot{x}_p &= ucos\psi - vsin\psi, \\ \dot{y}_p &= usin\psi + vcos\psi, \\ \dot{\psi} &= r\end{aligned}$$

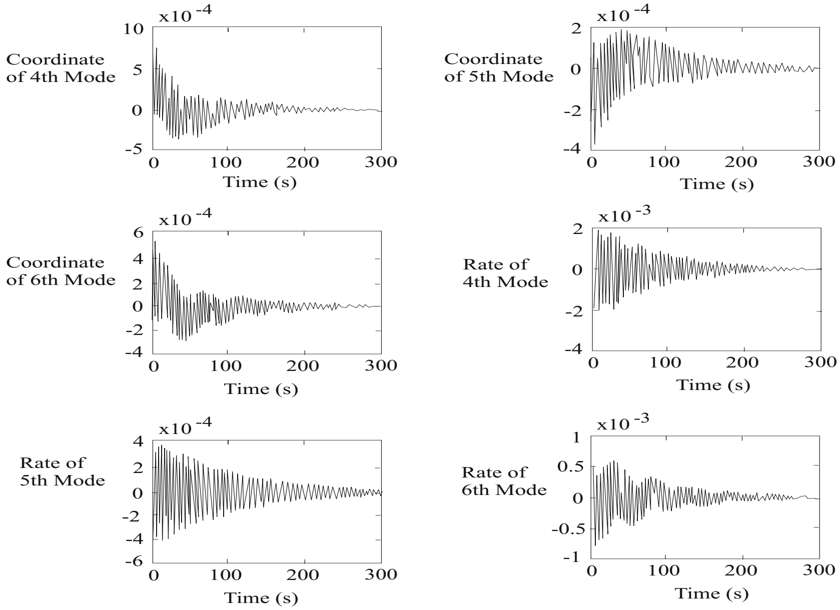


Fig. 6.3 Coordinate and rate responses of 4th –6th vibration modes

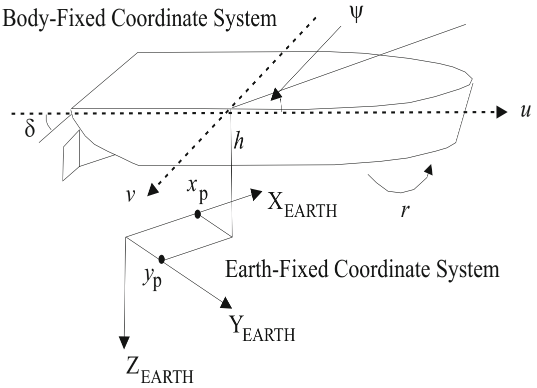


Fig. 6.4 Ship configuration and coordinates

rudder model

$$\dot{\delta} = \delta_c - \delta,$$

propeller model

$$\dot{n} = \frac{1}{T_m}(n_c - n).$$

Here the state vector is

$$\mathbf{x} = (u, v, r, \psi, x_p, y_p, \delta, n)^T$$

and the control is

$$\mathbf{u} = (\delta_c, n_c)^T .$$

As before we can put the equations in the form

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

and using a quadratic cost functional, a typical response of the controlled system is shown in Figure 6.5. These examples demonstrate that the method is very effective even for highly nonlinear and highly coupled systems. In the next section we study the optimality of the method.

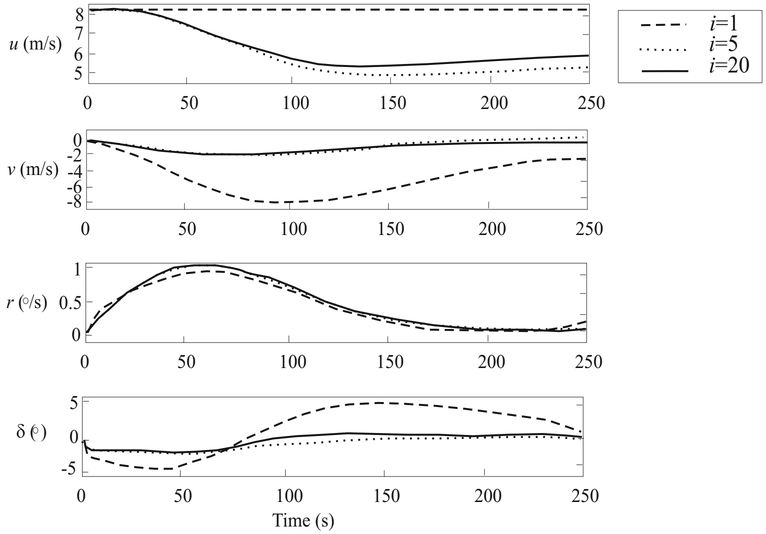


Fig. 6.5 Typical ship response for a heading of 90 degrees

6.5 The Hamilton-Jacobi-Bellman Equation, Viscosity Solutions and Optimality

In this section we study the HJB equation derived from Bellman's dynamic programming principle and its relation to optimal control. Consider the general nonlinear control problem

$$\dot{x}(s) = f(x(s), u(s)), \quad x(s) = x, \tag{6.7}$$

where we allow any bounded measurable control (the *admissible controls*). The cost functional will be taken to be

$$J_{x,t}(u(\cdot)) = g(x(t_f)) + \int_t^{t_f} h(x(s),u(s))ds, \tag{6.8}$$

where g and h are Lipschitz and bounded. We define the *value function*

$$v(x,t) = \inf_u J_{x,t}[u(\cdot)]$$

(where the inf is over all bounded measurable controls).

The next result determines a variational equation satisfied by the value function.

Theorem 6.1. *If δ satisfies $t+\delta \leq t_f$, then the value function $v(x,t)$ satisfies the equation*

$$v(x,t) = \inf_u \left\{ \int_t^{t+\delta} h(x(s),u(s))ds + v(x(t+\delta),t+\delta) \right\},$$

where $x(\cdot)$ is the solution of (6.7) with initial condition x at time t and with the control $u(t)$.

Proof. First choose any admissible control $\widetilde{u}$ and let $\widetilde{x}$ be the solution of

$$\begin{aligned} \dot{\widetilde{x}}(s) &= f(\widetilde{x}(s),\widetilde{u}(s)),\ \ t \leq s \leq t+\delta \\ \widetilde{x}(t) &= x. \end{aligned}$$

Since v is the infimum of J, we can find a control u' such that

$$u'(\widetilde{x}(t+\delta),t+\delta) + \varepsilon \geq \int_{t+\delta}^{t_f} h(x'(s),u'(s))ds + g(x'(t_f))$$

where x' satisfies

$$\begin{aligned} \dot{x}'(s) &= f(x'(s),u'(s)),\ \ t+\delta \leq s \leq t+t_f \\ x'(t+\delta) &= \widetilde{x}(t+\delta). \end{aligned}$$

Splicing the controls $\widetilde{u}$ and u', *i.e.* setting

$$u''(s) = \begin{cases} \widetilde{u}(s), & t \leq s \leq t+\delta \\ u'(s) & t+\delta \leq s \leq t_f \end{cases}$$

and substituting it into (6.7) gives the solution $x''(s)$, which is given by

$$x''(s) = \begin{cases} \widetilde{x}(s), & t \leq s \leq t+\delta \\ x'(s) & t+\delta \leq s \leq t_f \end{cases}$$

and since $\widetilde{u}$ is arbitrary, we must have

$$v(x,t) \leq \inf_u \left\{ \int_t^{t+\delta} h(x(s),u(s))ds + v(x(t+\delta),t+\delta) \right\} + \varepsilon.$$

A similar argument gives the inverse inequality

$$v(x,t) + \varepsilon \geq \inf_u \left\{ \int_t^{t+\delta} h(x(s),u(s))ds + v(x(t+\delta),t+\delta) \right\}$$

and the result follows. □

From the above Lipschitz and boundedness assumptions, we see that $v(x,t)$ is also Lipschitz and bounded. We now consider the HJB equation of the form

$$\begin{cases} v_t + H(v_x,x) = 0, & (x,t) \in \mathbb{R}^n \times (0,t_f) \\ \qquad v(x,t_f) = g(x) \end{cases} \tag{6.9}$$

where $v_x = \partial v/\partial x$. If H is smooth we can define classical solutions in the usual way. However, if H is not smooth, which is often the case in control problems, we need to find more general solutions. Distributional solutions do not work here, but we can find another type of solution called a viscosity solution (for more details see [4], [5]). A bounded, uniformly continuous function $v(x,t)$ is a *viscosity solution* of (6.9) if it satisfies the terminal condition $v(x,t_f) = g(x)$ and the following two properties:

(a) for each $\widetilde{v} \in C^\infty(\mathbb{R}^n \times (0,t_f))$, if $v - \widetilde{v}$ has a local maximum at (x_0,t_0), then

$$\widetilde{v}_t(x_0,t_0) + H(\widetilde{v}_x(x_0,t_0),x_0) \geq 0$$

(b) for each $\widetilde{v} \in C^\infty(\mathbb{R}^n \times (0,t_f))$, if $v - \widetilde{v}$ has a local minimum at (x_0,t_0), then

$$\widetilde{v}_t(x_0,t_0) + H(\widetilde{v}_x(x_0,t_0),x_0) \leq 0.$$

The main result is then:

Theorem 6.2. *The value function v for the problem (6.7),(6.8) is the unique viscosity solution of the HJB equation*

$$v_t + \min_u \{f(x,u)v_x + h(x,u)\} = 0, \quad v(x,t_f) = g(x).$$

(The proof of this theorem can be found in [5].) To find the optimal trajectory from an initial state x_0, we solve the optimal dynamical equation

$$\begin{aligned} \dot{x}^*(t) &= f(x^*(t),u^*(t)), \quad t_0 \leq t \leq t_f \\ x^*(t_0) &= x_0, \end{aligned}$$

where u^* is chosen so that

$$\begin{aligned} f(x^*(t),u^*(t)) \cdot v_x(x^*(t),t) \; + \; h(x^*(t),u^*(t)) \\ = \quad H(v_x(x^*(t),t),x^*(t)), \end{aligned}$$

i.e. so that H is minimised. This will give the optimal trajectory if v and u^* are smooth. (Here, of course, the Hamiltonian is given by

$$H(v_x,x) = \min_u \{f(x,u) \cdot v_x + h(x,u)\}.)$$

If we write the system in the form

$$\dot{x} = A(x,u)x + B(x,u)u$$

together with the cost functional

$$J = \frac{1}{2}x^T(t_f)Fx(t_f) + \frac{1}{2}\int_{t_0}^{t_f} \left(x^T Q(x,u)x + u^T R(x,u)u\right) dt,$$

and take a sequence of approximations, then at each step we are taking the minimum of the Hamiltonian, which must satisfy a Riccati equation. If these systems converge, and the Riccati equations all have solutions on the horizon interval, then the limiting system must also minimise the Hamiltonian along the trajectory and so if the dual variable λ (the value function above) and u are smooth then we will have optimality.

6.6 Characteristics of the Hamilton-Jacobi Equation

In the case when the Hamiltonian is smooth, we can find the solutions to the optimal control problem by the method of characteristics and when these are unique we have a global optimal control. Thus we consider a nonlinear first order partial differential equation

$$G(Du,u,x) = 0$$

where $x \in U$ and U is an open set in $\mathbb{R}^n$, Du is the gradient of u:

$$Du = \left(\frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_n}\right)$$

and

$$G : \mathbb{R}^n \times \mathbb{R} \times \overline{U} \to \mathbb{R}$$

is sufficiently smooth. The boundary condition is given by

$$u = h \text{ on } \Gamma \subseteq \partial U,$$

where h is given on Γ.

To find the characteristics, suppose they are parameterised curves given by $x(s)$. Let

$$z(s) = u(x(s))$$

and set

$$p(s) = Du(x(s)).$$

Then

$$\frac{dp^i}{ds}(s) = \sum_{j=1}^{n} u_{x_i x_j}(x(s)) \frac{dx^j}{ds}(s).$$

From the differential equation, we have

$$\sum_{j=1}^{n} \frac{\partial G}{\partial p^j}(p(s), z(s), x(s)) u_{x_j x_i} + \frac{\partial G}{\partial z}(p(s), z(s), x(s)) p^i(s) + \frac{\partial G}{\partial x_i}(p(s), z(s), x(s)) = 0.$$

If we assume that x_j satisfies

$$\frac{dx_j}{ds} = \frac{\partial G}{\partial p^j}(p(s), z(s), x(s)),$$

then we have

$$\frac{dp^i}{ds} = -\frac{\partial G}{\partial x_i}(p(s), z(s), x(s)) - \frac{\partial G}{\partial z}(p(s), z(s), x(s)) p^i(s), \;\; 1 \leq i \leq n$$

and also

$$\frac{dz}{ds} = \sum_{j=1}^{n} \frac{\partial u}{\partial x_j}(x(s)) \frac{\partial x_j}{ds} = \sum_{j=1}^{n} p^j(s) \frac{\partial G}{\partial p^j}(p(s), z(s), x(s)).$$

Hence the characteristic curves are given by the first order ordinary differential equation

$$\begin{aligned} \dot{p}(s) &= -D_x G(p(s), z(s), x(s)) - D_z G(p(s), z(s), x(s)) p(s) \\ \dot{z}(s) &= D_p G(p(s), z(s), x(s)) p(s) \\ \dot{x}(s) &= D_p G(p(s), z(s), x(s)) \end{aligned}$$

i.e.

$$\dot{y}(s) = F(y(s))$$

where

$$y(s) = (p(s), z(s), x(s))$$

and

$$F(y(s)) = \begin{pmatrix} -D_x G(p(s), z(s), x(s)) - D_z G(p(s), z(s), x(s)) p(s) \\ D_p G(p(s), z(s), x(s)) p(s) \\ D_p G(p(s), z(s), x(s)) \end{pmatrix}.$$

To obtain the initial conditions for this equation note that, since $u = h$ on Γ, we have

$$u_{x_i}(x^0) = h_{x_i}(x^0)$$

for any point $x^0 \in \Gamma$ and so

$$\begin{aligned} p^i(0) &= h_{x_i}(x^0) \\ z(0) &= h(x^0) \\ x(0) &= x^0. \end{aligned}$$

Since these must satisfy the original partial differential equation we must have the compatibility condition

$$G(p(0),z(0),x(0)) = 0$$

and in order for this to be satisfied near x^0, the implicit function theorem says that this is possible if

$$G_p(p(0),z(0),x(0)) \cdot \nu(x(0)) \neq 0$$

where $\nu(x(0))$ is the outward unit normal to Γ at $x(0)$. In this case we say that the point $(p(0),z(0),x(0))$ is *non-characteristic*. In order to solve the ordinary differential equation, we assume that f may be written in the form

$$F(y) = \overline{F}(y)y$$

where $\overline{F}(y)$ is a matrix-valued function of y. (We have seen that this is not a strong condition.) Hence we have the equation

$$\dot{y} = \overline{F}(y)y, \;\; y(0) = y_0 = (p(0),z(0),x(0)).$$

Introducing, as before, the system of linear, time-varying approximations

$$\dot{y}^{[i]}(t) = \overline{F}(y^{[i-1]}(t))y^{[i]}(t), \;\; y^{[i]}(t)(0) = y_0$$

we know from Chapter 2 that the states $y^{[i]}(t)$ defined by this system converge uniformly on compact sets if $\overline{F}$ is locally Lipschitz.

Example 6.1. Consider the linear partial differential equation given by

$$G(Du,u,x) = a(x) \cdot Du(x) + b(x)u(x)$$

where a is a vector-valued function. Then

$$D_pG = a(x)$$

and so

$$\dot{x}(s) = a(x(s)) \cdot p(s) = -b(x(s))z(s)$$

from the equation, and we do not require the 'p' equation. Suppose that a can be written in the form

$$a(x) = \begin{pmatrix} 0 & C(x) \\ -C(x) & 0 \end{pmatrix} x$$

where $C(x)$ and $C(y)$ commute for any x, y. Then the sequence of approximations becomes

$$\dot{x}^{[i]}(s) = \begin{pmatrix} 0 & C(x^{[i-1]}(s)) \\ -C(x^{[i-1]}(s))) & 0 \end{pmatrix} x^{[i]}(s).$$

Since the C's commute, we have

$$\begin{aligned} x^{[i]}(s) &= \exp \begin{pmatrix} 0 & \int_0^s C(x^{[i-1]}(\tau))d\tau \\ -\int_0^s C(x^{[i-1]}(\tau))d\tau & 0 \end{pmatrix} x_0 \\ &= \begin{pmatrix} \cos\left(\int_0^s C(x^{[i-1]}(\tau))d\tau\right) & \sin\left(\int_0^s C(x^{[i-1]}(\tau))d\tau\right) \\ -\sin\left(\int_0^s C(x^{[i-1]}(\tau))d\tau\right) & \cos\left(\int_0^s C(x^{[i-1]}(\tau))d\tau\right) \end{pmatrix} x_0. \end{aligned}$$

Now consider the Hamilton-Jacobi equation in the form

$$u_t + H(Du, x) = 0 \text{ in } \mathbb{R}^n \times [0, \infty)$$

with initial condition

$$u = g \text{ on } \mathbb{R}^n \times \{t = 0\}.$$

The characteristics are given by the nonlinear equations

$$\begin{aligned} \dot{x}(s) &= D_p H(p(s), x(s)) \\ \dot{p}(s) &= -D_x H(p(s), x(s)) \\ \dot{z}(s) &= D_p(p(s), x(s))p(s) - H(p(s), x(s)). \end{aligned}$$

The first two are, of course, Hamilton's equations. If we can write

$$\begin{pmatrix} D_p H(p, x) \\ -D_x H(p, x) \end{pmatrix} = A(p, x) \begin{pmatrix} x \\ p \end{pmatrix}$$

then we can introduce the approximating sequence

$$\begin{pmatrix} \dot{x}^{[i]}(s) \\ \dot{p}^{[i]}(s) \end{pmatrix} = A(p^{[i-1]}(s), x^{[i-1]}(s)) \begin{pmatrix} x^{[i]}(s) \\ p^{[i]}(s) \end{pmatrix}$$

for the first pair of equations.

6.7 Conclusions

In this chapter we have shown that the iteration technique can be applied effectively to solve nonlinear, non-quadratic optimal control problems. By replacing the problem by a sequence of linear, time-varying dynamical constraints together with a quadratic cost, we can solve the nonlinear problem as the limit of this sequence of linear, quadratic ones. This means that we can solve the nonlinear problem by classical means and we obtain an easily computable solution, which in many cases is optimal. We have given two examples, but the method can be applied to many more situations, *e.g.* for the control of chaos and in laser communications [6], nonlinear, high-speed aircraft design [7], and many other types of problems.

References

1. Navarro-Hernandez, C., Banks, S.P., Aldeen, M.: Observer Design for Nonlinear Systems using Linear Approximations. IMA J. Math. Cont and Inf. 20, 359–370 (2003)
2. Zheng, J., Banks, S.P., Alleyne, H.: Optimal Attitude Control for Three-Axis Stabilised Flexible Spacecraft. Acta Astronautica 56, 519–528 (2005)
3. Cimen, T., Banks, S.P.: Nonlinear Optimal Tracking Control with Application to Super-Tankers for Autopilot Design. Automatica 40(11), 1845–1863 (2004)
4. Lions, P.L.: Generalized Solutions of Hamilton-Jacobi Equations. Pitman, Boston (1982)
5. Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Birkhauser, Boston (1997)
6. Hugues-Salas, O., Banks, S.P.: Control of Chaos for Secure Communication. Int. J. Bifur. and Chaos (to appear, 2009)
7. Cimen, T., Banks, S.P.: Global Optimal Feedback Control for General Nonlinear Systems with Non-quadratic Performance Criteria. Sys. Cont. Letts. 53, 327–346 (2004)

Chapter 7
Sliding Mode Control for Nonlinear Systems

7.1 Introduction

In this chapter a method of sliding mode control for nonlinear systems will be presented. Sliding mode techniques are a different approach to solve control problems and are an area of increasing interest. It is well-known that in most of the cases, in the formulation of any control problem, there will appear some differences between the actual plant and the mathematical model developed for the control design. These discrepancies may be due to any number of factors such as unmodelled dynamics, variation in system parameters or the approximation of complex plant behaviour by a simpler model. It is the engineer's responsibility to guarantee some level of performance in spite of the existence of plant/model mismatches. This has led to the development of the so-called *robust methods*.

One way to approach robust control design is the sliding mode control methodology which can be considered to be a particular type of *variable structure control system* (VSCS). VSCSs are characterised by a suite of feedback control laws and a decision rule (called the switching function); it has as its input some measure of the current system behaviour and produces as an output the particular feedback controller which should be used at that instant of time. A VSCS can be regarded as a combination of subsystems where each subsystem has a fixed control structure and is valid for specified regions of system behaviour. The advantage is its ability to combine useful properties of each of the composite structures of the system. Furthermore, the system may be designed to possess new properties not present in any of the composite structures alone. The use of these ideas began in the Soviet Union in the late 1950s and continues up to the present day, see *i.e.* [1], [2] or [3] for more recent examples of work in this field.

In sliding mode control, the controller is designed to drive and then constrain the system state to lie within a neighbourhood of the switching surface. It presents some advantages:

- The dynamic behaviour of the system may be tailored by the choice of switching functions.

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 123–139.
springerlink.com

- The closed-loop response becomes totally insensitive to a particular class of uncertainty.
- Ability to specify performance directly makes sliding mode control attractive from the design perspective.

The chapter reviews briefly, in Section 7.2, the basics of sliding mode control for linear time-invariant (LTI) systems. In Section 7.3 this approach is extended to linear time-varying (LTV) systems by defining a sliding surface that is a function of time. A numerical example is given in this section in order to illustrate the theory. In Section 7.4 the method is generalised to nonlinear systems by using the iteration technique being studied in this book. An example of the sliding mode control for a robotic arm is provided at the end of this section. Finally, Section 7.5 contains conclusions of this chapter.

7.2 Sliding Mode Control for Linear Time-invariant Systems

In this section the basic approach of sliding mode control for a LTI system will be summarised. The concepts are presented for systems with a single control input, which allows to develop intuition about the basic aspects of nonlinear controller design.

Consider a single-input dynamical system of the form:

$$x^n = f(x) + b(x)u, \tag{7.1}$$

where the scalar x is the output of interest (*i.e.*, the position of a mechanical system), the scalar u is the control input (*i.e.*, a motor torque), and $X = [x, \dot{x}, \dots, x^{(n-1)}]^T$ is the state vector. In (7.1), the function $f(x)$ which in general will be nonlinear is not exactly known, but the *degree of imprecision on $f(x)$ is upper bounded by a known continuous function of x*, similarly, the control gain $b(x)$ is not exactly known, but is of known sign and is bounded by known, continuous functions of X [9].

The approach is to define a new so-called *sliding variable* $\sigma(t)$ as an $(n-1)^{th}$ order stable linear system (*reduced model*) so that the closed-loop system is controlled and ultimately follows a trajectory such that $\sigma(t) = 0$ [13]. The relation $\sigma(t) = 0$ defines a surface in state-space and the form of σ is carefully chosen to ensure that the goal of tracking is achieved when the state trajectory remains on this *sliding surface*.

The state trajectories do not always lie on the sliding surface so an additional design need therefore is to ensure that the state vectors move to the sliding surface.

7.3 Sliding Mode Control for Linear Time-varying Systems

In this section the problem of designing a sliding mode controller for a generic LTV system will be presented. In this case, the sliding surface will be defined such that it is a function of time and satisfies the following conditions:

- $\sigma(t) = c_1(t)x_1(t) + c_2(t)x_2(t) + c_3(t)x_3(t) + \ldots + c_n(t)x_n(t) = 0$
- $\dot{\sigma}(t) = -sign(\sigma(t))$

which are required for a successful design of the sliding control of the plant.

Given a generic LTV system of the form:

$$\begin{aligned} \dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0 \\ y(t) = C(t)x(t) + D(t)u(t) \end{aligned} \tag{7.2}$$

where $A(t) \in \mathbb{R}^{mxn}$ is differentiable which is of course a stronger requirement than the usual Lipschitz continuity, $B(t) \in \mathbb{R}^{mx}p$ are of adequate dimensions and $C(t) = \mathbb{I}_{nxm}$ and $D(t) = 0$ for simplicity. The control design strategy here will consist of two steps: first, the stabilisation of the LTV system such that the dynamics of the system on the sliding surface are stable, and secondly the design of the controller such that $u(t)$ forces the dynamics of (7.2) onto the sliding surface $\sigma(t)$. Equation 7.2 can be written on the form:

$$\dot{x}(t) = A(t)x(t) + Bu(t) = \begin{pmatrix} A_{11}(t) & A_{12}(t) \\ A_{21}(t) & A_{22}(t) \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} u(t). \tag{7.3}$$

Now under a change of variables,

$$x(t) = \begin{pmatrix} \xi(t) \\ \nu(t) \end{pmatrix} = \begin{pmatrix} x_1(t) & \ldots & ,x_{n-1}(t) & ,x_n(t) \end{pmatrix}^T$$

where $\xi(t) = (x_1(t), \ldots, x_{n-1}(t))$ and $\nu(t) = x_n(t)$, then this is

$$\begin{aligned} \dot{\xi}(t) = A_{11}(t)\xi(t) + A_{12}(t)\nu(t) \\ \dot{\nu}(t) = A_{21}(t)\xi(t) + A_{22}(t)\nu(t) + u(t). \end{aligned} \tag{7.4}$$

Then, the control strategy begins by defining $\sigma(t) = 0$ so

$$-c_1(t)x_1(t) - c_2(t)x_2(t) - c_3(t)x_3(t) - \ldots - c_{n-1}(t)x_{n-1}(t) = x_n(t) = \nu(t)$$

with $c_n(t) = 1$, *i.e.*, on the sliding surface. Thus the reduced order model becomes:

$$\dot{\xi}(t) = A_{11}(t)\xi(t) + A_{12}(t)\left[-c_1(t)x_1(t) - c_2(t)x_2(t) - \ldots - c_{n-1}(t)x_{n-1}(t)\right]. \tag{7.5}$$

This gives

$$\dot{\xi}(t) = \left[A_{11}(t) - A_{12}(t)\hat{C}(t)\right]\xi(t) = \hat{R}(t)\xi(t) \tag{7.6}$$

where $\hat{C}(t) = [c_1(t), \ldots, c_{n-1}(t)]$ is a time-varying vector containing the parameters that will stabilise the states $\xi(t)$. This will be done by applying the same ideas as in Chapter 5; by dividing the time interval $[0, t_f]$ in N subintervals of length h and using Ackerman's formula to fix the poles of the reduced order model at some prescribed left hand-side locations $(\lambda_1, \ldots, \lambda_{n-1})$:

$$\left|\lambda \cdot I - \hat{R}(t)\right| = (\lambda - \lambda_1) \cdots (\lambda - \lambda_{n-1}) \tag{7.7}$$

then the vector $\hat{C}(t)$ will be obtained in terms of the time-varying coefficients of the reduced order system's matrix $A_{11}(t)$, $A_{12}(t)$ and $\hat{C}(t)$ so that the reduced order system will have left-hand side poles. The application of pole placement method within sliding mode context has been used before by some authors as Zak and Hui [10] where they designed the output feedback sliding mode control based on eigenvector methods or Woodham and Zinober who proposed to position the closed-loop system eigenvalues in a specified sector in the left-hand half plane, involving the solution of a complex continuous Riccati equation [11]. In fact, an explicit form using Ackermans formula for the sliding surface is derived in [12].

It has been shown in Chapter 5 that due to the time dependency of (7.6), exponential stability will be guaranteed if the conditions of Theorem 5.2 are satisfied. Once these conditions are satisfied and the reduced order model has been stabilised by using pole placement techniques, the second sliding mode control condition is taken into account. For this, σ is chosen to satisfy the discontinuous differential equation:

$$\dot{\sigma}(t) = -sign(\sigma(t)).$$

To do this, note that $\dot{\sigma}(t) = \dot{\hat{C}}(t)\xi(t) + \hat{C}(t)\dot{\xi}(t) + \dot{\nu}(t)$ and substituting (7.4) into this equation gives:

$$\begin{aligned}\dot{\sigma}(t) &= \dot{\hat{C}}(t)\xi(t) + \hat{C}(t)\left[A_{11}(t)\xi(t) + A_{12}(t)\nu(t)\right] + A_{21}(t)\xi(t) + A_{22}(t)\nu(t) + u(t)\\ &\overset{\Delta}{=} -sign(\sigma(t)).\end{aligned}$$

Therefore the following sliding control $u(t)$ is obtained:

$$\begin{aligned}u(t) = -\dot{\hat{C}}(t)\xi(t) - \hat{C}(t)\left[A_{11}(t)\xi(t) + A_{12}(t)\nu(t)\right]\\ -A_{21}(t)\xi(t) - A_{22}(t)\nu(t) - sign(\sigma(t))\end{aligned} \tag{7.8}$$

and (7.4) is now:

$$\begin{aligned}\dot{\xi}(t) &= A_{11}(t)\xi(t) + A_{12}(t)\nu(t)\\ \dot{\nu}(t) &= -\dot{\hat{C}}(t)\xi(t) - \hat{C}(t)\left[A_{11}(t)\xi(t) + A_{12}(t)\nu(t)\right] - sign(\sigma(t)).\end{aligned} \tag{7.9}$$

Then, the states $\xi(t) = (x_1(t), \ldots, x_{n-1}(t))$ and $\nu(t) = x_n(t)$ of the closed-loop system:

$$\begin{pmatrix} \dot{\xi}(t) \\ \dot{\nu}(t) \end{pmatrix} = \begin{pmatrix} A_{11}(t) & A_{12}(t) \\ -\dot{\hat{C}}(t) - \hat{C}(t)A_{11}(t) & -\hat{C}(t)A_{12}(t) \end{pmatrix} \cdot \begin{pmatrix} \xi(t) \\ \nu(t) \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} sign(\sigma(t)) \tag{7.10}$$

will be exponentially stable.

The main results of this section could be summarised as follows:

Lemma 7.1. *Given a LTV system of the form (7.3), for a time-varying sliding surface $\sigma = \hat{C}(t)\xi(t) + \nu(t)$ such that $\sigma = 0$ and $\dot{\sigma}(t) = -sign(\sigma)$, it is possible to find a sliding control $u(t)$ so the closed-loop system (7.10) is exponentially stable if:*

- *The eigenvectors $P_i(t)$ of the matrix $\hat{R}(t)$ of the reduced model (7.26) are differentiable and,*
- $||P^{-1}(t) \cdot \dot{P}(t)|| < \beta$, *$P(t)$ matrix of eigenvectors $P_i(t)$*

where $\beta < Re(\lambda_g)$ and λ_g is the greatest of the eigenvalues of $\hat{R}(t)$.

In the following example, this theory will be applied to a LTV system in order to illustrate the above stated.

Example 7.1. Consider the following LTV system:

$$\dot{x}(t) = \begin{pmatrix} t & 0 & 1 \\ 3 & t & 2 \\ 1 & 0 & 1 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} u(t) \tag{7.11}$$

for initial conditions $x(0) = [0.2, 0.3, 0.4]^T$. In this case we assume the matrix B to be constant.

For the first part of the controller design, a pole-placement algorithm as in section 5.3, will be applied in order to stabilise the reduced system on the sliding surface $\sigma(t) = c_1(t)x_1(t) + c_2(t)x_2(t) + c_3(t)x_3(t)$ where in this case we have chosen $c_3 = 1$ for convenience. Thus, under the condition $\sigma = 0$:

$$x_3(t) = -c_1(t)x_1(t) - c_2(t)x_2(t). \tag{7.12}$$

Then, the original plant (7.11) can be reduced to:

$$\begin{aligned} \dot{x}_1 &= tx_1 + x_3 \\ \dot{x}_2 &= 3x_1 + tx_2 + 2x_3 \end{aligned} \tag{7.13}$$

and substituting (7.12) above gives:

$$\begin{aligned} \dot{x}_1 &= tx_1 - c_1x_1 - c_2x_2 \\ \dot{x}_2 &= 3x_1 + tx_2 + 2[-c_1x_1 - c_2x_2]. \end{aligned} \tag{7.14}$$

This reduced representation will be stabilised by pole placement methods for time-varying systems, the choice of closed-loop poles is $\lambda = (-1, -2)$. Therefore:

$$\left|\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} - \begin{pmatrix} t-c_1 & -c_2 \\ 3-2c_1 & t-2c_2 \end{pmatrix}\right| = (\lambda - \lambda_1)\cdot(\lambda - \lambda_2). \tag{7.15}$$

Thus,

$$\lambda^2 + \lambda[c_1 - t - t + 2c_2] + t^2 - c_1 t - 2tc_2 + 3c_2 = \lambda^2 + \lambda[-\lambda_1 - \lambda_2] + \lambda_1\lambda_2 \tag{7.16}$$

and equation coefficients this gives the values of the parameters of the sliding surface:

$$\begin{array}{c} c_1 = -p_1 - p_2 + 2t - 2c_2 \\ c_2 = \frac{1}{3}[p_1 p_2 - t^2 - tp_1 - tp_2 + 2t^2]. \end{array} \tag{7.17}$$

Now, if $\sigma(t) = c_1(t)x_1(t) + c_2(t)x_2(t) + c_3(t)x_3(t)$, then:

$$\dot{\sigma}(t) = \dot{c}_1(t)x_1(t) + c_1(t)\dot{x}_1(t) + \dot{c}_2(t)x_2(t) + c_2(t)\dot{x}_2(t) + \dot{c}_3(t)x_3(t) + c_3(t)\dot{x}_3(t).$$

Substituting the original plant's dynamics:

$$\dot{\sigma}(t) = c_1(tx_1 + x_3) + \dot{c}_1 x_1 + c_2(3x_1 + tx_2 + 2x_3) + \dot{c}_2 x_2 + x_1 + x_3 + u(t).$$

Therefore the control will be designed as the composition of two parts: The first one will be the equivalent control u_{eq} which is continuous and it is based on the obtained parameters $\hat{C}(t)$:

$$u(t) = -\left[c_1(tx_1 + x_3) + \dot{c}_1 x_1 + c_2(3x_1 + tx_2 + 2x_3) + \dot{c}_2 x_2 + x_1 + x_3\right] \tag{7.18}$$

and a second part, consisting of the signum function, it is the discontinuous part of the control law, that requires infinite switching on the part of the control signal and actuator at the intersection of the error state trajectory and sliding surface. In this way the trajectory is forced to move always towards the sliding surface:

$$sign(s) = \left[\begin{array}{rll} -1 & if & s < 0 \\ 0 & if & s = 0 \\ 1 & if & s > 1 \end{array}\right. . \tag{7.19}$$

So the final control applied will be

$$u(t) = -sign(\sigma) - \left[c_1(tx_1 + x_3) + \dot{c}_1 x_1 + c_2(3x_1 + tx_2 + 2x_3) + \dot{c}_2 x_2 + x_1 + x_3\right]. \tag{7.20}$$

Then, the dynamics of the closed-loop form of the original plant (7.11) once the sliding mode control is applied will have a stable behaviour as shown in figure 7.1. It can be seen how the sliding control shows the expected chattering around the sliding surface.

This simulation was carried out for a final time $t_f = 30$s using a step size of $h = 0.01$. It can be seen how the states are successfully stabilised and converge to zero.

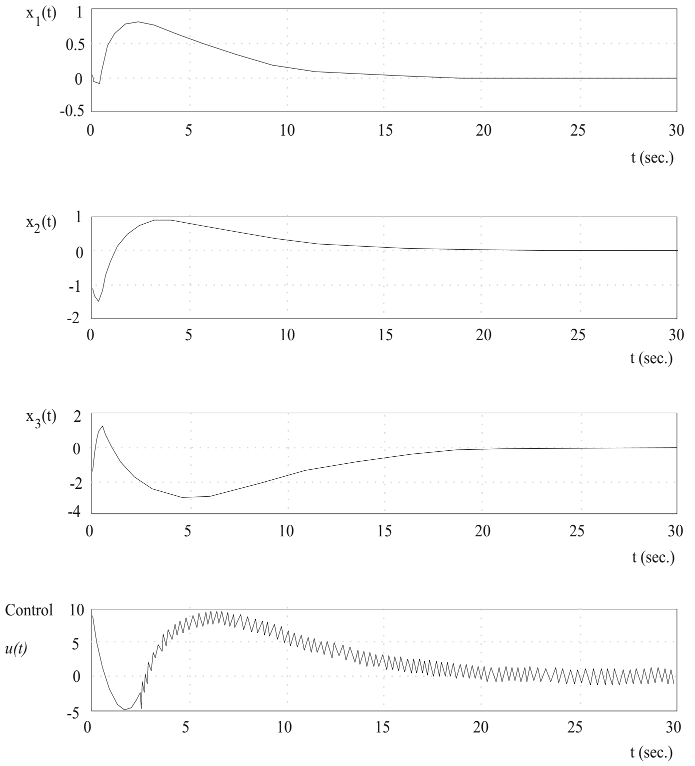


Fig. 7.1 Controlled states and equivalent control law $u(t)$

7.4 Generalisation to Nonlinear Systems

In this section, the sliding control method presented in the previous section will be generalised to the nonlinear case, provided the original nonlinear system can be approximated by a sequence of LTV systems.

Consider now a general n-dimensional nonlinear control problem of the form:

$$\begin{aligned} \dot{x} &= A(x)x(t) + B(x)u(t), \quad x(0) = x_0 \\ y(x) &= C(x)x(t) + D(x)u(t) \end{aligned} \tag{7.21}$$

where $A(x) \in \mathbb{R}^{nxm}$, $B(x) \in \mathbb{R}^{nxp}$ are of appropriate dimensions and satisfy Lipschitz continuity requirement, $B(x)$ is a constant matrix (for simplicity, but could be generalized to the no constant case), $C(x) = I$ and $D = 0$.

Applying the iteration technique, the following sequence of LTV systems is obtained:

$$\dot{x}^{[1]} = A(x_0)x^{[1]}(t) + Bu^{[1]}(t), \quad x^{[1]}(0) = x_0 \tag{7.22}$$

$$\vdots$$

$$\dot{x}^{[i]} = A(x^{[i-1]})x^{[i]}(t) + Bu^{[i]}(t), \quad x^{[i]}(0) = x_0. \tag{7.23}$$

For each of these LTV equations a sliding mode surface can be designed such that $\sigma^{[i]}(t) = 0$ and $\dot{\sigma}^{[i]}(t) = -sign(\sigma^{[i]})(t)$ where

$$\sigma^{[i]}(t) = \hat{C}^{[i]}(t)\xi^{[i]}(t) + \nu^{[i]}(t) \tag{7.24}$$

where $\xi^{[i]}(t) = (x_1^{[i]}(t), \ldots, x_{n-1}^{[i]}(t))$ and $\nu^{[i]}(t) = x_n^{[i]}(t)$, then under this change of variables, equations (7.22) – (7.23) become:

$$\begin{aligned} \dot{\xi}^{[1]}(t) &= A_{11}(\xi_0)\xi^{[1]}(t) + A_{12}(\nu_0)\nu^{[1]}(t) \\ \dot{\nu}^{[1]}(t) &= A_{21}(\xi_0)\xi^{[1]}(t) + A_{22}(\nu_0)\nu(t)^{[1]} + u^{[1]}(t), \end{aligned} \quad \begin{pmatrix} \xi^{[1]}(0) \\ \nu^{[1]}(0) \end{pmatrix} = \begin{pmatrix} \xi_0 \\ \nu_0 \end{pmatrix} \tag{7.25}$$

$$\vdots$$

$$\begin{aligned} \dot{\xi}^{[i]}(t) &= A_{11}(\xi^{[i-1]}(t))\xi^{[i]}(t) + A_{12}(\nu^{[i-1]}(t))\nu^{[i]}(t) \\ \dot{\nu}^{[i]}(t) &= A_{21}(\xi^{[i-1]}(t))\xi^{[i]}(t) + A_{22}(\nu^{[i-1]}(t))\nu(t)^{[i]} + u^{[i]}(t), \end{aligned} \quad \begin{pmatrix} \xi^{[i]}(0) \\ \nu^{[i]}(0) \end{pmatrix} = \begin{pmatrix} \xi_0 \\ \nu_0 \end{pmatrix}.$$

Now, the control strategy begins by defining $\sigma^{[i]}(t) = 0$ such that

$$\begin{gathered} -c_1^{[1]}(t)x_1^{[1]}(t) - c_2^{[1]}(t)x_2^{[1]}(t) - \ldots - c_{n-1}^{[1]}(t)x_{n-1}^{[1]}(t) = x_n^{[1]}(t) = \nu^{[1]}(t) \\ \vdots \\ -c_1^{[i]}(t)x_1^{[i]}(t) - c_2^{[i]}(t)x_2^{[i]}(t) - \ldots - c_{n-1}^{[i]}(t)x_{n-1}^{[i]}(t) = x_n^{[i]}(t) = \nu^{[i]}(t) \end{gathered}$$

with $c_n^{[i]}(t) = \ldots = c_n^{[1]}(t) = 1$, so a sequence of reduced order models can be found by substituting the above expressions into the iterated systems (7.25):

$$\begin{gathered} \dot{\xi}^{[i]}(t) = \left[A_{11}(\xi^{[i-1]}(t)) - A_{12}(\xi^{[i-1]}(t))\hat{C}^{[i]}(t)\right]\xi(t) = \hat{R}(\xi^{[i-1]}(t))\xi^{[i]}(t) \\ \vdots \\ \dot{\xi}^{[1]}(t) = \left[A_{11}(x_0) - A_{12}(x_0)\hat{C}^{[1]}(t)\right]\xi^{[1]}(t) = \hat{R}(x_0)\xi^{[1]}(t). \end{gathered} \tag{7.26}$$

Then a pole placement method like in Section 7.3 can be carried out for an appropiate choice of left-hand side poles $(\lambda_1, \cdots, \lambda_{n-1})$, in order to obtain a vector of stabilising parameters $\hat{C}^{[i]}(t)$ for each of the reduced models.

Remark 7.1. *In here it is assumed that $\lambda_1, \cdots, \lambda_{n-1}$ to be constant and the same for each iteration being straightforward to generalise this and choose a different set of left-hand side eigenvalues at each iteration.*

Remark 7.2. *$\hat{C}^{[1]}(t)$ is written as a time-varying vector because it depends on $A_{11}(x_0)$, $A_{12}(x_0)$ which are constant for the first iteration but also depends on the choice of eigenvalues that could be time-dependent.*

Now, conditions from Lemma 7.1 (or from Theorem 5.2) are assumed to be satisfied for each iterated reduced model in (7.26) in order to guarantee exponential stability of each of them. The second sliding mode condition shall be studied now for each iteration:

$$\dot{\sigma}^{[i]}(t) = -sign(\sigma^{[i]}(t)), \ldots, \dot{\sigma}^{[1]}(t) = -sign(\sigma^{[1]}(t)).$$

So

$$\begin{gathered}
\dot{\sigma}^{[1]}(t) = \dot{\hat{C}}^{[1]}(t)\xi^{[1]}(t) + \hat{C}^{[1]}(t)\dot{\xi}^{[1]}(t) + \dot{\nu}^{[1]}(t) \\
\vdots \\
\dot{\sigma}^{[i]}(t) = \dot{\hat{C}}^{[i]}(t)\xi^{[i]}(t) + \hat{C}^{[i]}(t)\dot{\xi}^{[i]}(t) + \dot{\nu}^{[i]}(t)
\end{gathered}$$

and substituting (7.26) on the above expressions:

$$\begin{gathered}
\dot{\sigma}^{[1]}(t) = \dot{\hat{C}}^{[1]}(t)\xi^{[1]}(t) + \hat{C}^{[1]}(t)\left[A_{11}(\xi_0)\xi^{[1]}(t) + A_{12}(\nu_0)\nu^{[1]}(t)\right] \\
+A_{21}(\xi_0)\xi^{[1]}(t) + A_{22}(\nu_0)\nu^{[1]}(t) + u^{[1]}(t) \doteq -sign(\sigma^{[1]}(t)) \\
\vdots \\
\dot{\sigma}^{[i]}(t) = \dot{\hat{C}}^{[i]}(t)\xi^{[i]}(t) + \hat{C}^{[i]}(t)\left[A_{11}(\xi^{[i-1]}(t))\xi^{[i]}(t) + A_{12}(\nu^{[i]}(t))\nu^{[i]}(t)\right] \\
+A_{21}(\xi^{[i-1]}(t))\xi^{[i]}(t) + A_{22}(\nu^{[i-1]}(t))\nu^{[i]}(t) + u^{[i]}(t) \doteq -sign(\sigma^{[i]}(t)).
\end{gathered}$$

Therefore a sequence of sliding controls $u^{[i]}(t)$ is obtained:

$$\begin{gathered}
u^{[1]}(t) = -\dot{\hat{C}}^{[1]}(t)\xi^{[1]}(t) - \hat{C}^{[1]}(t)\left[A_{11}(\xi_0)\xi^{[1]}(t) + A_{12}(\nu_0)\nu^{[1]}(t)\right] \\
-A_{21}(\xi_0)\xi^{[1]}(t) - A_{22}(\nu_0)\nu^{[1]}(t) - sign(\sigma^{[1]}(t)) \\
\vdots \\
u^{[i]}(t) = -\dot{\hat{C}}^{[i]}(t)\xi^{[i]}(t) - \hat{C}^{[i]}(t)\left[A_{11}(\xi^{[i]}(t))\xi^{[i]}(t) + A_{12}(\nu^{[i]}(t))\nu^{[i]}(t)\right] \\
-A_{21}(\xi^{[i]}(t))\xi^{[i]}(t) - A_{22}(\nu^{[i]}(t))\nu^{[i]}(t) - sign(\sigma^{[i]}(t))
\end{gathered} \tag{7.27}$$

and the sequence of systems (7.25) is now:

$$\begin{gathered}
\dot{\xi}^{[1]}(t) = A_{11}(\xi_0)\xi^{[1]}(t) + A_{12}(v_0)v^{[1]}(t) \\
\dot{v}^{[1]}(t) = -\dot{\hat{C}}^{[i]}(t)\xi^{[1]}(t) - \hat{C}^{[1]}(t)\left[A_{11}(\xi_0)\xi^{[1]}(t) + A_{12}(v_0)v^{[1]}(t)\right] \\
-sign(\sigma^{[1]}(t)) \\
\vdots \\
\dot{\xi}^{[i]}(t) = A_{11}(\xi^{[i-1]}(t))\xi^{[i]}(t) + A_{12}(v^{[i-1]}(t))v^{[i]}(t) \\
\dot{v}^{[i]}(t) = -\dot{\hat{C}}^{[i]}(t)\xi^{[i]}(t) - \hat{C}^{[i]}(t)\left[A_{11}(\xi^{[i-1]}(t))\xi^{[i]}(t) + A_{12}(v^{[i-1]}(t))v^{[i]}(t)\right] \\
-sign(\sigma^{[i]}(t)).
\end{gathered} \tag{7.28}$$

Thus, the sequence of states $\xi^{[i]}(t) = (x_1^{[i]}(t), \ldots, x_{n-1}^{[i]}(t))$ and $v^{[i]}(t) = x_n^{[i]}(t)$ of the closed-loop systems:

$$\begin{gathered}
\begin{pmatrix} \dot{\xi}^{[1]}(t) \\ \dot{v}^{[1]}(t) \end{pmatrix} = \begin{pmatrix} A_{11}(\xi_0) & A_{12}(v_0) \\ -\dot{\hat{C}}^{[1]}(t) - \hat{C}^{[1]}(t)A_{11}(\xi_0) & -\hat{C}^{[1]}(t)A_{12}(v_0) \end{pmatrix} \cdot \begin{pmatrix} \xi^{[1]}(t) \\ v^{[1]}(t) \end{pmatrix} \\
- \begin{pmatrix} 0 \\ 1 \end{pmatrix} sign(\sigma^{[1]}(t)) \\
\vdots \\
\begin{pmatrix} \dot{\xi}^{[i]}(t) \\ \dot{v}^{[i]}(t) \end{pmatrix} = \begin{pmatrix} A_{11}(\xi^{[i-1]}(t)) & A_{12}(v^{[i-1]}(t)) \\ -\dot{\hat{C}}^{[i]}(t) - \hat{C}^{[i]}(t)A_{11}(\xi^{[i-1]}(t)) & -\hat{C}^{[i]}(t)A_{12}(v^{[i-1]}(t)) \end{pmatrix} \cdot \begin{pmatrix} \xi^{[i]}(t) \\ v^{[i]}(t) \end{pmatrix} \\
- \begin{pmatrix} 0 \\ 1 \end{pmatrix} sign(\sigma^{[i]}(t))
\end{gathered} \tag{7.29}$$

will be exponentially stable by convergence assumption of the iteration technique.

This can be summarised in the following theorem:

Theorem 7.1. *Given a nonlinear system of the form $\dot{x} = A(x)x(t) + B(x)u(t), x_0 = x(0)$, where $A(x)$ and $B(x)$ are Lipschitz, it is possible to find a sequence of sliding mode controls $u^{[i]}(t)$ such that the nonlinear system will be exponentially stabilised by $lim_{i \to \infty} u^{[i]}(t)$ if the coefficients $A_{12}(x^{[i-1]}(t))$ from the reduced order models (7.26) are non-zero and lower bounded $\quad 0 < \Gamma_i < ||A_{12}(x^{[i-1]}(t))||$.*

Proof. In order to prove the theorem, convergence of the solutions of (7.29) should be shown. By taking into account the convergence of the iterated solutions to the solution of the nonlinear systems, convergence in this case is guaranteed if the matrices

$$\begin{gathered}
\begin{pmatrix} A_{11}(\xi^{[i-1]}(t)) & A_{12}(v^{[i-1]}(t)) \\ -\dot{\hat{C}}^{[i]}(t) - \hat{C}^{[i]}(t)A_{11}(\xi^{[i-1]}(t)) & -\hat{C}^{[i]}(t)A_{12}(v^{[i-1]}(t)) \end{pmatrix}, \\
\begin{pmatrix} 0 \\ sign(\sigma^{[i]}(t)) \end{pmatrix}
\end{gathered}$$

are Lipschitz. The first part of this proof deals with the requirements and proof for the first matrix to be Lipschitz: in fact $A_{11}(\xi^{[i]}(t)),A_{12}(v^{[i]}(t))$ are Lipschitz by assumption for the iteration technique. The elements that depend on $\hat{C}^{[i]}(t)$ and $\dot{\hat{C}}^{[i]}(t)$ can be analysed now. In fact, for each iteration, from (7.7):

$$|\lambda \cdot I - A_{11}^{[i]}(t) - A_{12}^{[i]}\hat{C}^{[i]}(t)| = (\lambda - \lambda_1)(\lambda - \lambda_2)\dots(\lambda - \lambda_{n-1})$$

From here, it is easy to see that $\hat{C}^{[i]}(t)$ will be a vector depending on the elements of $\left(A_{12}^{[i]}\right)^{-1}$. Since these elements are Lipschitz, for the Lipschitz continuity property to be satisfy for these functions, then, $det\left|A_{12}^{[i]}(t)\right|$ should have a non-zero lower bound $\forall t \in [0,T]$.

On the other hand, $\dot{\hat{C}}(t)$ will be a rational function of the elements of $A(x)$ and its derivatives and since they are differentiable , $\dot{\hat{C}}(t)$ is differentiable, therefore is Lipschitz.

The second part of this proof deals with the properties of

$$\begin{pmatrix} 0 \\ sign(\sigma^{[i]}(t)) \end{pmatrix}.$$

In fact, $sign(\sigma^{[i]}(t))$ is not Lipschitz, and so we consider a sequence of smooth sigmoid functions which approximate $sign(x) := sgm(x)$, *i.e.*, $tanh(\nu x), \nu \to \infty$. We apply the iteration technique to each system obtained on replacing $sign(\nu x)$ and using the above sliding surface design: this leads to a collection of systems of the form:

$$\dot{\hat{x}}^{[i]} = K\left[\hat{x}^{[i]}(t), c^{[i-1]}(t))\hat{x}^{[i]} + tanh(\nu_k \hat{x}^{[i-1]}\right] \tag{7.30}$$

for a sequence of numbers $\nu_k \to \infty$, and the original sequence containing $\sigma(x)$:

$$\dot{x}^{[i]} = K\left[x^{[i]}(t), c^{[i-1]}(t))x^{[i]} + sgm\sigma(x^{[i-1]})\right]. \tag{7.31}$$

Each system of the form (7.30) converges by the proof given in Theorem 2.1. To prove the sequence (7.31) converges, write it in the form:

$$\dot{x}^{[i]} = K\left[x^{[i]}(t), c^{[i-1]}(t))x^{[i]}(t) + tanh(\nu_k \hat{x}^{[i-1]}\right] + \left[sgm\sigma(x^{[i-1]}) - tanh(\nu_k \hat{x}^{[i-1]}\right]. \tag{7.32}$$

The proof again proceeds as in the case of Theorem 2.1, noting that the integrated form:

$$\begin{aligned} x^{[i]}(t) = \Phi^{[i]}(t,0)x_0 + \textstyle\int_0^t \Phi^{[i]}(t,s)tanh(\nu_k x^{[i-1]}(s))ds \\ + \textstyle\int_0^t \Phi^{[i]}(t,s)(sgm\sigma(x^{[i-1]})) - tanh(\nu_k x^{[i-1]}(s))ds. \end{aligned} \tag{7.33}$$

where $\Phi^{[i]}$ is the transition matrix of $K(x^{[i-1]}(t), c^{[i-1]}(t))$, converges since the last term clearly converges to zero as $\nu_k \to \infty$. □

Example 7.2. Consider the example of a robotic arm consisting of two joints. The dynamic equations of motion for such a system are:

$$\begin{array}{c} I\ddot{q}_1 + MgLsinq_1 + k(q_1 - q_2) = 0 \\ J\ddot{q}_2 - k(q_1 - q_2) = u \end{array} \tag{7.34}$$

for some given initial conditions $x(0)$ and where the parameters of the system are:

$$\begin{array}{c} M = 5kg \quad\ g = 9.81m \cdot s^{-1} \quad\ L = 0.5m \\ I = 2.5kg \cdot m^{-2}\ J = 1.5kg \cdot m^{-2}\ k = 100N \cdot m^{-1}. \end{array}$$

By adopting a change of variables $q_1 = x_1$, $q_2 = x_2$, $\dot{q}_1 = x_3$ and $\dot{q}_2 = x_4$, the second order equations (7.34) can be written in the form of a first order system:

$$\begin{array}{c} \dot{x}_1 = x_3 \\ \dot{x}_2 = x_4 \\ \dot{x}_3 = -\frac{k}{I}(x_1 - x_2) - \frac{Mgl}{I}sin(x_1) \\ \dot{x}_4 = -\frac{u}{J} + \frac{k}{J}(x_1 - x_2) \end{array} \tag{7.35}$$

and this can be transformed in the SCD form:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k}{I} - \frac{Mgl}{I}\frac{sin(x_1)}{x_1} & \frac{k}{I} & 0 & 0 \\ \frac{k}{J} & -\frac{k}{J} & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{J} \end{pmatrix} \cdot u(t). \tag{7.36}$$

Now the iteration technique can be applied and a sequence of LTV system will be generated:

$$\begin{pmatrix} \dot{x}_1^{[1]} \\ \dot{x}_2^{[1]} \\ \dot{x}_3^{[1]} \\ \dot{x}_4^{[1]} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k}{I} - \frac{Mgl}{I}\frac{sin(x_{01})}{x_{01}} & \frac{k}{I} & 0 & 0 \\ \frac{k}{J} & -\frac{k}{J} & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1^{[1]} \\ x_2^{[1]} \\ x_3^{[1]} \\ x_4^{[1]} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{J} \end{pmatrix} \cdot u^{[1]}(t) \tag{7.37}$$

$$\vdots$$

$$\begin{pmatrix} \dot{x}_1^{[i]} \\ \dot{x}_2^{[i]} \\ \dot{x}_3^{[i]} \\ \dot{x}_4^{[i]} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k}{I} - \frac{Mgl}{I}\frac{sin(x_1^{[i-1]})}{x_1^{[i-1]}} & \frac{k}{I} & 0 & 0 \\ \frac{k}{J} & -\frac{k}{J} & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1^{[i]} \\ x_2^{[i]} \\ x_3^{[i]} \\ x_4^{[i]} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{J} \end{pmatrix} \cdot u^{[i]}(t). \tag{7.38}$$

The sliding mode control strategy explained before can be applied now to each of these LTV systems. In fact in order to create a control $u^{[i]}(t)$ at each iteration a sequence of LTV sliding surfaces $(\sigma^{[1]}, \ldots, \sigma^{[i]})$ can be designed such that each of the systems from (7.37) – (7.38) will converge to the corresponding iterated surface after the hitting time and remain on it. Therefore for each of the surfaces:

$$\sigma^{[i]}(t) = c_1^{[i]}(t)x_1^{[i]}(t) + c_2^{[i]}(t)x_2^{[i]}(t) + c_3^{[i]}(t)x_3^{[i]}(t) + c_4^{[i]}(t)x_4^{[i]}(t) \tag{7.39}$$

and taking the case when $c_4^{[i]}(t) = 1$ (constant),

$$x_4^{[i]}(t) = -c_1^{[i]}(t)x_1^{[i]}(t) - c_2^{[i]}(t)x_2^{[i]}(t) - c_3^{[i]}(t)x_3^{[i]}(t) \tag{7.40}$$

Equations 7.37 – 7.38 can be written as:

$$\begin{pmatrix} \dot{x}_1^{[1]} \\ \dot{x}_2^{[1]} \\ \dot{x}_3^{[1]} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ -c_1^{[1]}(t) & -c_2^{[1]}(t) & -c_3^{[1]}(t) \\ -40 - 9.81\left(\frac{sin(x_{01})}{x_{01}}\right) & 40 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1^{[1]} \\ x_2^{[1]} \\ x_3^{[1]} \end{pmatrix} \tag{7.41}$$

$$\vdots$$

$$\begin{pmatrix} \dot{x}_1^{[i]} \\ \dot{x}_2^{[i]} \\ \dot{x}_3^{[i]} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ -c_1^{[i]}(t) & -c_2^{[i]}(t) & -c_3^{[i]}(t) \\ -40 - 9.81\left(\frac{sin(x_1^{[i-1]})}{x_1^{[i-1]}}\right) & 40 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1^{[i]} \\ x_2^{[i]} \\ x_3^{[i]} \end{pmatrix}. \tag{7.42}$$

The idea is to stabilise each of the LTV systems with the same spirit as in the pole placement method in Chapter 5. Choosing a set of desired stable eigenvalues, *i.e.*: $(\lambda_1, \lambda_2, \lambda_3) = (-2, -3, -5)$:

$$\left| \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} - \begin{pmatrix} 0 & 0 & 1 \\ -c_1^{[i]}(t) & -c_2^{[i]}(t) & -c_3^{[i]}(t) \\ a^{[i]}(t) & b^{[i]}(t) & 0 \end{pmatrix} \right| = (\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3) \tag{7.43}$$

where $a^{[i]}(t) = -40 - 9.81\left(\frac{sin(x_1^{[i-1]})}{x_1^{[i-1]}}\right)$ and $b^{[i]}(t) = 40 = b$.

Remark 7.3. *Note that in this example the left-hand side poles $(\lambda_1, \lambda_2, \lambda_3)$ have been selected to be the same for each iteration and during all the time interval $(0, t_f)$, being possible to generalise the method by choosing different values at each iteration and at different times.*

$$\begin{aligned} \left| \begin{pmatrix} \lambda & 0 & -1 \\ c_1^{[i]}(t) & \lambda + c_2^{[i]}(t) & c_3^{[i]}(t) \\ -a^{[i]}(t) & -b & \lambda \end{pmatrix} \right| &= \lambda^3 + \lambda^2 c_2^{[i]}(t) - \lambda a^{[i]}(t) \\ &+ b\lambda c_3^{[i]}(t) - bc_1^{[i]}(t) - a^{[i]}(t)c_2^{[i]}(t). \end{aligned} \tag{7.44}$$

Simple identification of coefficients of same order yields the values for the parameters $c_1^{[i]}(t)$, $c_2^{[i]}(t)$ and $c_3^{[i]}(t)$ of the sliding surface $\sigma^{[i]}(t)$ that stabilise each of the iterated systems (7.41):

$$\begin{aligned} c_2^{[i]}(t) &= -\lambda_1 - \lambda_2 - \lambda_3 \\ c_3^{[i]}(t) &= \tfrac{\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_3\lambda_2+a^{[i]}(t)}{b} \\ c_1^{[i]}(t) &= \tfrac{-\lambda_1\lambda_2\lambda_3+a^{[i]}(t)(-\lambda_1-\lambda_2-\lambda_3)}{b}. \end{aligned} \tag{7.45}$$

So now,

$$\dot{\sigma}^{[i]} = \dot{c}_1^{[i]}x_1^{[i]} + c_1^{[i]}\dot{x}_1^{[i]} + \dot{c}_2^{[i]}x_2^{[i]} + c_2^{[i]}\dot{x}_2^{[i]} + \dot{c}_3^{[i]}x_3^{[i]} + c_3^{[i]}\dot{x}_3^{[i]} + \dot{x}_4^{[i]}.$$

For sliding mode control, it is required that $\sigma^{[i]}(t) = 0$ and $\dot{\sigma}^{[i]} = -sign(\sigma^{[i]})$. Applying these conditions to the systems (7.37) – (7.38) gives the controls $u^{[i]}(t)$:

$$\begin{aligned} u^{[i]} = -1.5sign(\sigma^{[i]}) - 1.5\Big[&\dot{c}_1^{[i]}x_1^{[i]} + c_1^{[i]}x_3^{[i]} + \dot{c}_2^{[i]}x_2^{[i]} + c_2^{[i]}x_4^{[i]} \\ &+\dot{c}_3^{[i]}x_3^{[i]} + a^{[i]}c_3^{[i]}x_1^{[i]} + bc_3^{[i]}x_2^{[i]} + \tfrac{k}{J}x_1^{[i]} - \tfrac{k}{J}x_2^{[i]}\Big]. \end{aligned} \tag{7.46}$$

By iterating the original nonlinear problem, only after three iterations the sequence of LTV problems converges to the nonlinear system. In fact, by taking as control law the third iterated control $u^{(3)}(t)$ and applying it to the nonlinear system, it is shown how the states are stabilised. Figures 7.2 – 7.5 show the stable behaviour of the four output components after applying the sliding control to each of the iterations and original nonlinear problem too. In Figure 7.6 the final iterated control law is shown.

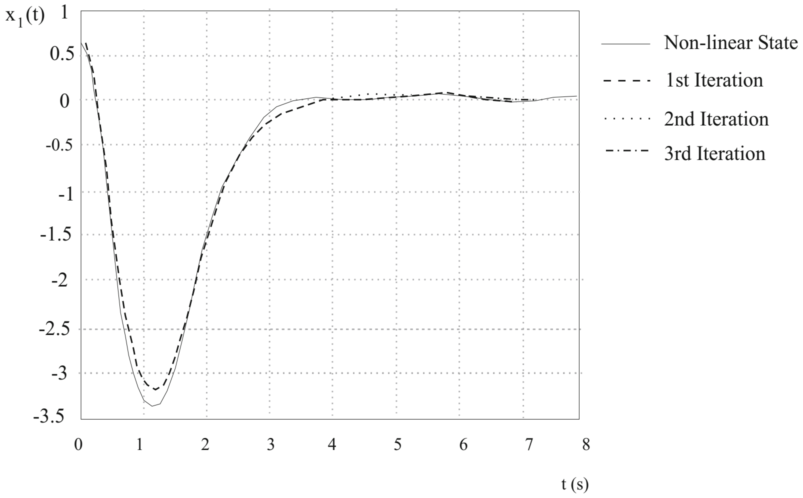


Fig. 7.2 Iterated and nonlinear controlled $x_1(t)$ state

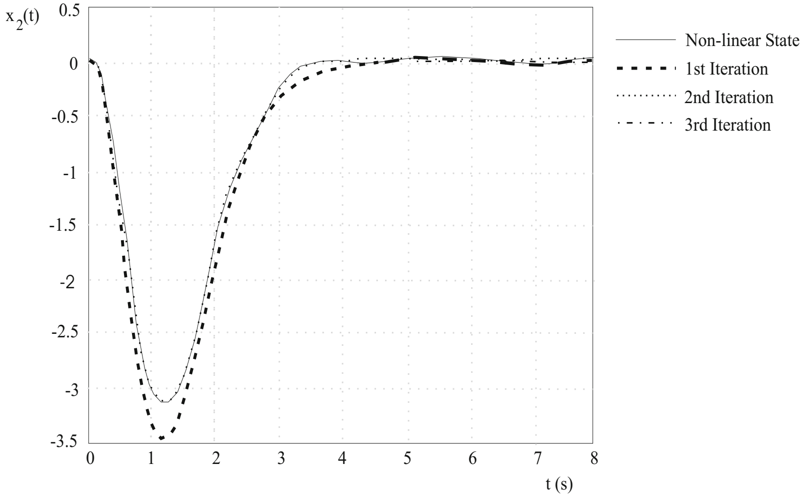


Fig. 7.3 Iterated and nonlinear controlled $x_2(t)$ state

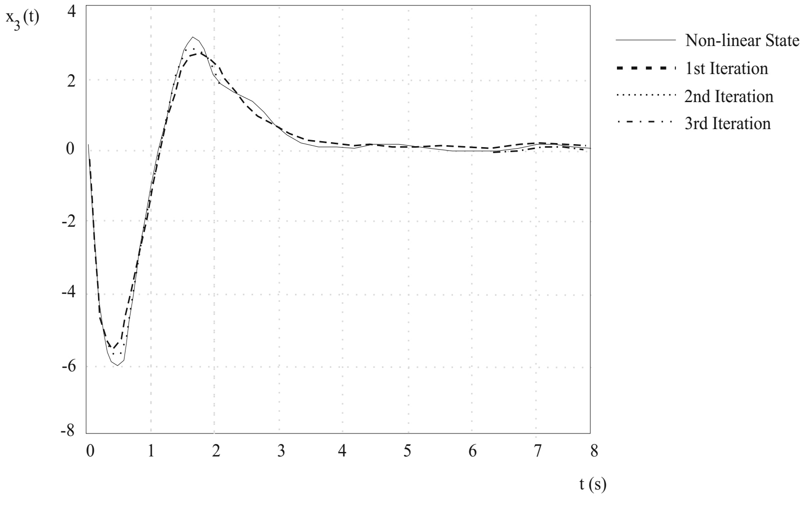


Fig. 7.4 Iterated and nonlinear controlled $x_3(t)$ state

7.5 Conclusions

In this chapter, the problem of designing a sliding mode control for nonlinear systems has been addressed. It is based on the approximation of the nonlinear system by a sequence of LTV ones whose solution converge to the solution of the original nonlinear problem.

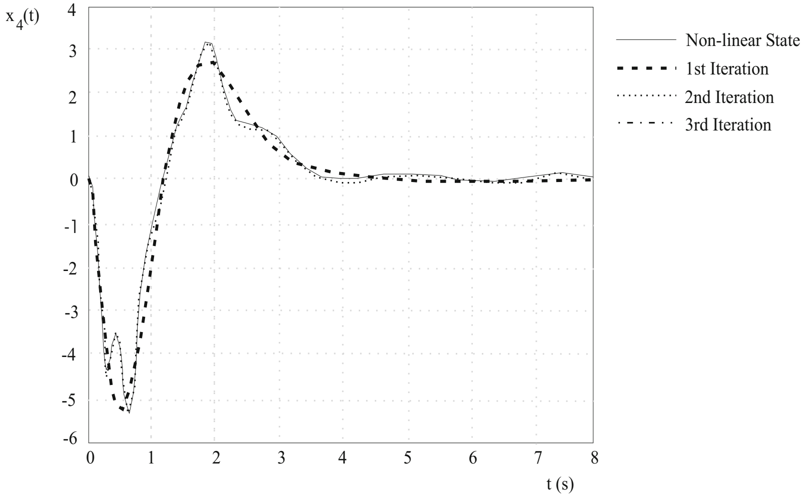


Fig. 7.5 Iterated and nonlinear controlled $x_4(t)$ state

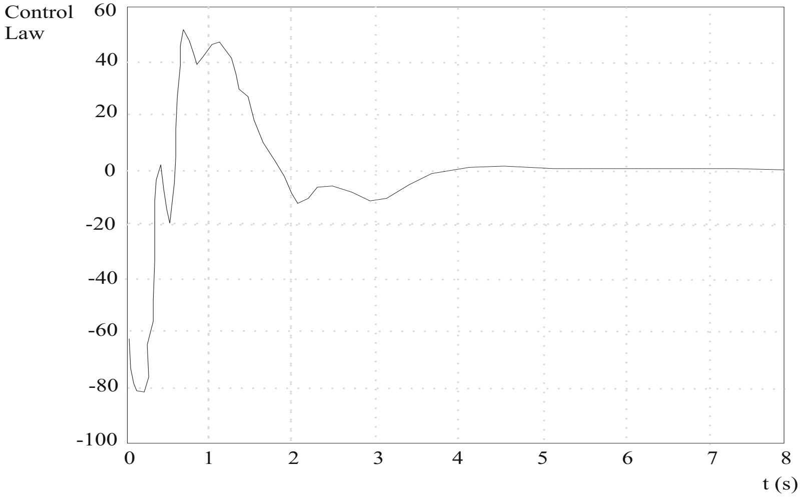


Fig. 7.6 Iterated control

After this, a sequence of time-varying sliding surfaces is designed for each of the LTV problems. This is done by using a LTV pole placement approach combined with additional conditions for stability of the time-varying case. This theory is generalized to general nonlinear systems by the convergence of the LTV sequence. Simulation results on how this is done have also been presented.

References

1. Yan, X.G., Edwards, C., Spurgeon, S.K.: Decentralised robust sliding mode control for a class nonlinear interconnected systems by static output feedback. Automatica 40, 613–620 (2004)
2. Lu, X.Y., Spurgeon, S.K.: Output feedback stabilisation of SISO nonlinear systems via dynamic sliding modes. Int. J. Control 70, 735–759 (1998)
3. Frasca, R., Iannelli, L., Vasca, F.: Boundary Layer Using Dithering in Sliding Mode Control. In: 16th IFAC World Congress, Prague, Czech Republic (July 2005)
4. Tomás-Rodríguez, M., Banks, S.P.: Linear Approximations to Nonlinear Dynamical Systems with Applications to Stability and Spectral Theory. IMA J. Math. Cont. and Inf. 20, 89–104 (2003)
5. Tomás-Rodríguez, M., Navarro-Hernandez, C., Banks, S.P.: Parametric Approach to Optimal Nonlinear Control Problem using Orthogonal Expansions. In: IFAC World Congress, Prague, Czech Republic (July 2005)
6. Navarro-Hernández, C., Banks, S.P., Aldeen, M.: Observer Design for Nonlinear Systems using Linear Approximations. IMA J. Math. Cont. and Inf. 20, 359–370 (2003)
7. Tomás-Rodríguez, M., Banks, S.P.: Pole placement for nonlinear systems. In: NOLCOS 2004, Stuttgart, Germany (2004)
8. Cimen, T., Banks, S.P.: Nonlinear optimal tracking control with application to supertankers for autopilot design. Automatica 40, 1845–1863 (2004)
9. Slotine, E., Li, W.: Applied Nonlinear Control. Prentice Hall, New Jersey (1991)
10. Zak, S.H., Hui, S.: Output feedback in variable structure controllers and state estimators for uncertain nonlinear dynamical systems. Proceedings IEE- Control Theory App. 140, 41–50 (1993)
11. Woodham, C.A., Zinober, S.I.: Eigenvalue placement in a specified sector for variable structure control systems. Int. Journal of Control 5, 1021–1037 (1993)
12. Ackermann, J., Utkin, V.I.: Sliding mode control designed based on Ackermanns formula. IEE Trans. Automatic Control 43, 234–237 (1998)
13. Ha, Q.P., Trinh, H., Nguyen, H.T., Tuan, H.D.: Dynamic output feedback sliding mode control using pole placement and linear functional observers. IEEE Transactions on Industrial Electronics 50(5), 1030 (2003)

Chapter 8
Fixed Point Theory and Induction

8.1 Introduction

In this chapter we shall show that we can obtain results on various aspects of systems of the form

$$\dot{x} = A(x)x,\, x(0) = x_0 \tag{8.1}$$

by using a sequence of approximations

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t))x^{[i]}(t),\, x^{[0]}(0) = x_0 \tag{8.2}$$

as before and a combination of fixed point theorems and induction. The induction will proceed in the following way: suppose we want to prove some property P of Equation 8.1, and assume we can find a function $x^{[0]}(t)$ which has this property. Suppose also that if $x^{[i-1]}(t)$ has the property, then the solution $x^{[i]}(t)$ of Equation 8.2 also has the property. Then if the sequence $\{x^{[i]}(t)\}$ converges on some interval $[0,T]$, it follows by induction that the nonlinear system (8.1) (or the solutions thereof) also have the property P.

We shall see that this can be applied to stability of nonlinear systems and the existence of periodic solutions. The same idea can, however, be applied to many other situations.

8.2 Fixed Point Theory

The most basic fixed point theorem is the Banach contraction principle. Suppose that (M,δ) is a complete metric space. A function $f : M \to M$ is a contraction (or is Lipschitz) with contraction constant $\gamma < 1$ if

$$\delta(f(x), f(y)) \leq \gamma\delta(x,y)$$

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 141–150.
springerlink.com

for all $x, y \in M$. Then the Banach contraction principle states the following:

Theorem 8.1. *If f is a contraction on a metric space (M, δ), with contraction constant γ, then f has a unique fixed point given by*

$$\lim_{n \to \infty} f^n(y)$$

for any $y \in M$.

The most useful form of this principle for differential or integral equations is the following (see [1]):

Corollary 8.1. *If $k : [0,T] \times [0,T] \times \mathbb{R} \to \mathbb{R}$ is a continuous kernel which satisfies the Lipschitz condition*

$$|k(t,s,x) - k(t,s,y)| \le L|x-y|,$$

for all $(s,t) \in [0,T] \times [0,T]$ and $x, y \in \mathbb{R}$, then for all $y \in C[0,T]$, the integral equation

$$x(t) = y(t) + \int_0^t k(t,s,x(s))ds$$

for $0 \le t \le T$, has a unique solution $x \in C[0,T]$.

We can obtain the solution stated in the corollary by choosing any $x_n(\cdot) \in C[0,T]$ and defining inductively

$$x_{n+1}(t) = y(t) + \int_0^t k(t,s,x_n(s))ds.$$

Then $\{x_n\}$ converges uniformly on $[0,T]$.

Of course, to apply the corollary to a (scalar) differential equation

$$\frac{dx}{dt} = f(x,t),\ x(0) = x_0,$$

where f is Lipschitz:

$$|f(x,t) - f(y,t)| \le \gamma|x-y|$$

for all $t \in [0,T]$ and any $x, y \in \mathbb{R}$, we simply write the equation in integral form

$$x(t) = x_0 + \int_0^t f(x(s),s)ds.$$

The generalisation of these results to vector-valued functions $x(t)$ is clear.

We now show that theorem 2.1 (the original convergence theorem for the iteration process) can be proved by fixed point theory, if $A(x)$ is Lipschitz. Thus we have

Theorem 8.2. *If in the nonlinear system*

$$\dot{x} = A(x)x,\, x(0) = x_0,$$

the matrix-valued function $x \to A(x)$ is Lipschitz, then the sequence of functions $x^{[i]}(\cdot)$ given by

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t))x^{[i]}(t),\, x^{[i]}(0) = x_0, \tag{8.3}$$

converges uniformly on any compact interval $[0,T]$.

Proof. Consider the two iteration schemes

$$x_{n+1}(t) = x_0 + \int_0^t A(x_n(s))x_n(s)ds. \tag{8.4}$$

and

$$\widetilde{x}_{n+1}(t) = x_0 + \int_0^t A(\widetilde{x}_n(s))\widetilde{x}_{n+1}(s)ds. \tag{8.5}$$

The first is identical to Picard iteration used in the Banach contraction theorem, while the second is our iteration scheme. Note the subtle difference in the right hand sides. We can easily see that $\widetilde{x}_n(t)$ and $x_n(t)$ are bounded (say by M) on $[0,T]$ for all n and so, by Lipschitz continuity of $A(\cdot)$ we have

$$\|A(\widetilde{x}_n(t))\| \leq L_1, \text{ for all } n, \text{ and } t \in [0,T]$$

for some constant L_1. Let L_2 be the Lipschitz constant of $A(\cdot)$ and let K be a positive number to be specified later. Let X be the Banach space of continuous functions on $[0,T]$ with the norm

$$|||f||| = max_{0\leq t\leq T} e^{-Kt}\|f(t)\|.$$

Then by (8.4) and (8.5), we have

$$\begin{aligned} x_{n+1}(t) - \widetilde{x}_{n+1}(t) &= \int_0^t (A(x_n(s))x_n(s) - A(\widetilde{x}_n(s))\widetilde{x}_n(s)) \\ &= \int_0^t \{(A(x_n(s))x_n(s) - A(\widetilde{x}_n(s))x_n(s)) \\ &\quad + (A(\widetilde{x}_n(s))x_n(s) - A(\widetilde{x}_n(s))x_{n+1}(s)) \\ &\quad + (A(\widetilde{x}_n(s))x_{n+1}(s) - A(\widetilde{x}_n(s))\widetilde{x}_{n+1}(s))\}\, ds \end{aligned}$$

and so

$$|||x_{n+1}(\cdot) - \widetilde{x}_{n+1}(\cdot)||| \le sup_{0\le t\le T} e^{-Kt}\int_0^t \{\|A(x_n(s)) - A(\widetilde{x}_n(s))\|\cdot\|x_n(s)\| \\ +\|A(\widetilde{x}_n(s))\|\cdot\|x_n(s) - x_{n+1}(s)\| \\ +\|A(\widetilde{x}_n(s))\|\cdot\|x_{n+1}(s) - \widetilde{x}_{n+1}(s)\|\}\,ds.$$

Hence,

$$|||x_{n+1}(\cdot) - \widetilde{x}_{n+1}(\cdot)||| = ML_2 sup_{0\le t\le T} e^{-Kt}\int_0^t e^{Ks}e^{-Ks}\|x_n(s) - \widetilde{x}_n(s)\|ds \\ +L_1 sup_{0\le t\le T} e^{-Kt}\int_0^t e^{Ks}e^{-Ks}\|x_n(s) - x_{n+1}(s)\| \\ +L_1 sup_{0\le t\le T} e^{-Kt}\int_0^t e^{Ks}e^{-Ks}\|x_{n+1}(s) - \widetilde{x}_{n+1}(s)\|$$

$$\le ML_2|||x_n(\cdot) - \widetilde{x}_n(\cdot)|||sup_{0\le t\le T} e^{-Kt}\int_0^t e^{Ks}ds \\ +L_1|||x_n(\cdot) - x_{n+1}(\cdot)|||e^{-Kt}\int_0^t e^{Ks}ds \\ +L_1|||x_{n+1}(\cdot) - \widetilde{x}_{n+1}(\cdot)|||e^{-Kt}\int_0^t e^{Ks}ds$$

$$= ML_2|||x_n(\cdot) - \widetilde{x}_n(\cdot)|||\frac{1-e^{-KT}}{K} \\ +L_1|||x_n(\cdot) - x_{n+1}(\cdot)|||\frac{1-e^{-KT}}{K} \\ +L_1|||x_{n+1}(\cdot) - \widetilde{x}_{n+1}(\cdot)|||\frac{1-e^{-KT}}{K}$$

and so

$$\alpha|||x_{n+1}(\cdot) - \widetilde{x}_{n+1}(\cdot)||| \le max(ML_2, L_1)\frac{1-e^{-KT}}{K}(|||x_n(\cdot) - \widetilde{x}_n(\cdot)||| \\ + |||x_n(\cdot) - x_{n+1}(\cdot)|||),$$

where

$$\alpha = 1 - L_1\left(\frac{1-e^{-KT}}{K}\right).$$

Hence, if K is large enough, we have

$$|||x_{n+1}(\cdot) - \widetilde{x}_{n+1}(\cdot)||| \le \beta\,(|||x_n(\cdot) - \widetilde{x}_n(\cdot)||| + |||x_n(\cdot) - x_{n+1}(\cdot)|||)$$

where $\beta < 1$. Thus, since $x_n(\cdot)$ converges in X by the fixed point theorem, it follows easily that $\{\widetilde{x}_{n+1}\}$ is a Cauchy sequence in X and so converges uniformly, for any $T > 0$. □

Returning to Equation 8.3 let

$$\Phi_{x^{[i-1]}}(t)$$

denote the transition matrix generated by the time-varying matrix function $A(x^{[i-1]}(t))$. Then the solution of (8.3) is

$$x^{[i]}(t) = \Phi_{x^{[i-1]}}(t)x_0 \tag{8.6}$$

and theorem 8.2 essentially says that the map $x(\cdot) \longrightarrow \Phi_{x(\cdot)}$ is Lipschitz and so has a fixed point and that the solution is given by the iteration scheme given by Equation 8.6.

8.3 Stability of Systems

In this section we show how to use the induction argument discussed in the introduction to prove stability of a nonlinear system. To do this we first recall the notion of logarithmic norm of a matrix. Thus, for a square matrix A, we define the *logarithmic norm* (or *measure*) of A by

$$\mu(A) = lim_{h \to 0+}(\|I + hA\| - 1)/h,$$

where $\|\cdot\|$ is any induced norm on matrices (see [2]). The main advantage over the usual norm is that $\mu(A)$ can be negative, and as noted in Chapter 3, it can therefore be used to study the stability of systems. The most important property for a linear, time-varying system

$$\dot{x} = A(t)x,\ x(t_0) = x_0$$

is that

$$\|x(t)\| \leq \|x_0\| exp \int_{t_0}^{t} \mu(A(s))ds, \tag{8.7}$$

so that if

$$\int_{t_0}^{t} \mu(A(s))ds \longrightarrow -\infty$$

as $t \longrightarrow \infty$, then the system is stable. It can be shown that if we use the norm

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

on $\mathbb{R}^n$, then

$$\mu(A) = max_j[Re(a_{jj}) + \sum_{i \neq j} |a_{ij}|]$$

while if we use the usual Euclidean norm

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

then we have

$$\mu(A) = max_i(\lambda_i(A + A^*)/2)$$

where $\lambda_i(\cdot)$ is the i^{th} eigenvalues and A^* is the conjugate transpose of A.

Although we only have the result above for linear, time-varying systems, we can use the iteration theory and induction to prove stability for certain nonlinear systems. Thus, suppose that the nonlinear system

$$\dot{x} = A(x)x$$

satisfies

$$\mu(A(x)) < -\varepsilon < 0$$

($\varepsilon > 0$) for all x in a ball B (where $A(\cdot)$ is Lipschitz). Then, as usual, we consider the sequence of approximations

$$\dot{x}^{[i]}(t) = A(x^{[i-1]}(t))x^{[i]}(t), \;\; x^{[i]}(0) = x_0 \in B.$$

This sequence converges uniformly on any compact interval. For the first approximation we take

$$x^{[1]}(t) = e^{-t}x_0.$$

This clearly belongs to B. Now assume that $x^{[k]}(t) \in B$ for all $t \geq 0$. Then

$$\mu(A(x^{[k]}(t))) < -\varepsilon < 0$$

so it follows that $A(x^{[k]}(t) \in B$, for all $t > 0$. Note, however, that this does not immediately imply stability of the nonlinear system since it might be possible for the stable systems, with solutions $x^{[i]}(t)$ to take 'longer and longer' times to stabilise. However, because $\varepsilon > 0$, it is easy to prove that this does not happen. Thus the induction principle and the iteration convergence theorem give:

Theorem 8.3. *If the nonlinear system*

$$\dot{x} = A(x)x$$

satisfies

$$\mu(A(x)) < -\varepsilon < 0$$

for all $x \in B$, *where* $B = \{x : \|x\| \leq K\}$ *for some* $K > 0$, *and* $A(\cdot)$ *is Lipschitz, then it is asymptotically stable in* B.

It is also easy to obtain a result which is similar to LaSalle's invariance principle. Thus we have:

Theorem 8.4. *If the nonlinear system*

$$\dot{x} = A(x)x, \;\; x \in \mathbb{R}^n$$

satisfies

$$\mu(A(x)) < -\varepsilon < 0$$

for all $x \in B \setminus M$ *where* $B = \{x : \|x\| \leq K\}$ *and* M *is a (not necessarily connected) submanifold of* B *of dimension* $\leq n-1$ *such that the nonlinear dynamics are transversal to M, then the conclusion of theorem 8.3 follows also in this case.*

8.4 Periodic Solutions

In order to apply induction and iteration arguments to the problem of periodic solutions of nonlinear systems, we first recall the basic Floquet theory for linear, time-varying systems (see, for example [3] or [4])

$$\dot{x} = A(t)x, \;\; x \in \mathbb{R}^n$$

where $A(\cdot)$ is periodic with period ω:

$$A(t+\omega) = A(t), \text{for all } t.$$

It is well-known that this system always has a fundamental set of linearly independent solutions $u_1, \cdots, u_n$, so that any solution $x(t)$ may be written as a linear combination of the u_i's:

$$x(t) = \sum_{i=1}^{n} \alpha_i u_i(t)$$

for some constants α_i. However, since $A(\cdot)$ is periodic, $u_i(t+\omega)$, $1 \leq i \leq n$, are also solutions of the equation and so they may also be written as linear combinations of the $u_i(t)$'s :

$$u_i(t+\omega) = \sum_{j=1}^{n} \beta_{ij} u_j(t), \;\; 1 \leq i \leq n,$$

for some other constants β_{ij}. Hence the general solution $x(t+\omega)$ becomes

$$x(t+\omega) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_{ij} u_j(t)$$

and so if we assume that

$$x(t+\omega) = \lambda x(t) \tag{8.8}$$

for some λ, then

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\beta_{ij}u_j(t)=\lambda\left(\sum_{i=1}^{n}\alpha_i u_i\right).$$

Since the u_i's are linearly independent, we have

$$det(\beta_{ij}-\lambda\delta_{ij})=0$$

i.e. λ is an eigenvalue of the matrix $B=(\beta_{ij})$. Moreover, if λ is chosen as an eigenvalue of B, then the general solution will satisfy (8.8). If we put $\lambda=e^{\mu\omega}$, then the function $e^{-\mu t}x(t)$ is periodic:

$$\begin{aligned} e^{-\mu(t+\omega)}x(t+\omega) &= e^{-\mu t}e^{-\mu\omega}x(t+\omega)\\ &= e^{-\mu t}\lambda^{-1}x(t+\omega)\\ &= e^{-\mu t}x(t). \end{aligned}$$

Hence the equation has a solution of the form $e^{\mu t}x(t)$, where $x(t)$ is periodic. If $\mu=0$, then the solution of the system is periodic.

Now consider the nonlinear system

$$\dot{x}=A(x)x,\ \ x(0)=x_0$$

and introduce the usual iteration sequence $x^{[i]}(t)$ by

$$\dot{x}^{[i]}(t)=A(x^{[i-1]}(t))x^{[i]}(t),\ \ x^{[i]}(0)=x_0. \tag{8.9}$$

Suppose we choose $x^{[0]}(t)$ to be periodic with period T. Then $x^{[1]}(t)$ is given by

$$\dot{x}^{[1]}(t)=A(x^{[0]}(t))x^{[1]}(t)$$

which is of the form

$$\dot{x}^{[1]}(t)=\widetilde{A}(t)x^{[1]}(t),$$

where $\widetilde{A}(t)$ is periodic with period T. Hence by the Floquet theory above, $x^{[1]}(t)$ is of the form $exp(\mu^{[i]}t)\xi^{[1]}(t)$ where $\xi^{[1]}(t)$ is periodic with period T. Now consider the equation

$$\dot{\overline{x}}^{[2]}(t)=A(\xi^{[1]}(t))\overline{x}^{[2]}(t)$$

instead of (8.9) above. Again this is periodic with period T, and so has a solution of the form

$$\overline{x}^{[2]}(t)=exp(\mu^{[2]}t)\xi^{[2]}(t)$$

where $\xi^{[2]}(t)$ is periodic with period T. Continuing in this way, we obtain a sequence of functions $\xi^{[i]}(t)$ which are periodic with period T and a sequence of numbers $\mu^{[i]}$, such that $\overline{x}^{[i]}(t)=exp(\mu^{[i]}t)\xi^{[i]}(t)$ satisfies the equation

$$\dot{\overline{x}}^{[i]}(t)=A(\xi^{[i-1]}(t))\overline{x}^{[i]}(t).$$

Hence, if $\mu^{[i]} \to 1$ as $i \to \infty$, then the sequence $\xi^{[i]}(t)$ converges to the solution $x^{[i]}(t)$ of (8.9) as $i \to \infty$ and the nonlinear system has a periodic solution.

Now consider the logistic system

$$\begin{aligned}\dot{x}_1 &= f_1(x)x_1 \\ \dot{x}_2 &= f_2(x)x_2 \\ &\vdots \\ \dot{x}_n &= f_n(x)x_n\end{aligned}$$

and introduce the sequence of approximations

$$\begin{aligned}\dot{x}_1^{[i]} &= f_1(x^{[i-1]}(t))x_1^{[i]} \\ \dot{x}_2^{[i]} &= f_2(x^{[i-1]}(t))x_2^{[i]} \\ &\vdots \\ \dot{x}_n^{[i]} &= f_n(x^{[i-1]}(t))x_n^{[i]}.\end{aligned}$$

Hence, if $x^{[i]}(0) = x_0$, we have the solution

$$\begin{aligned}x_k^{[i]}(t) &= e^{\int_0^t f_k(x^{[i-1]}(s))ds} x_k^{[i]}(0) \\ &= e^{\int_0^t f_k(x^{[i-1]}(s))ds} x_{0k}, \quad 1 \le k \le n.\end{aligned}$$

The system has a periodic solution of period T if and only if

$$\int_0^T f_k(x^{[i-1]}(s))ds = 0.$$

Let

$$\mu_k^{[i]}(T) = \int_0^T f_k(x^{[i-1]}(s))ds.$$

If the sequences $\{\mu_k^{[i]}(T)\}_{1 \le k \le n}$ are bounded and bounded away from 0, then they must have a convergent subsequence, which must tend to zero or else the solutions would diverge or tend to zero. The nonlinear system will then have a periodic solution. The conditions clearly hold for Volterra-Lotka type systems.

8.5 Conclusions

In this brief chapter we have shown that the original iteration technique can be regarded as a kind of fixed point theory, although to obtain global convergence from this point of view, we must assume that our systems are globally Lipschitz and so the original theory is more general. However, as we have seen, fixed point theory does provide some useful insights into the method and leads to the idea of using induction arguments, coupled with the iteration technique to prove that nonlinear systems possess certain properties, such as stability or the existence of periodic solutions.

We have shown therefore that the iteration technique is not just a numerical procedure, but coupled with the induction argument, it can be used to prove the existence of certain properties of nonlinear systems. Such properties will carry over from similar properties of linear, time-varying systems in the limit. This demonstrates the usefulness of the method for general nonlinear systems theory and it appears that it will be useful in many other areas (see chapter 12).

References

1. Granas, A., Dugundji, J.: Fixed Point Theory. Springer, New York (2000)
2. Brauer, F.: ‘Perturbations of nonlinear systems of differential equations II. J. Math. Analysis App. 17, 418–434 (1967)
3. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields. Springer, New York (1977)
4. Davis, H.T.: Introduction to Nonlinear Differential and Integral Equations. Dover, New York (1962)

Chapter 9
Nonlinear Partial Differential Equations

9.1 Introduction

In this chapter we shall show how to generalise the results of the previous chapters on finite-dimensional nonlinear systems to partial differential equations. Rather than try to cover any significant part of this vast field, we shall concentrate on two problems, since the ideas then apply to many other nonlinear distributed systems. These two problems are concerned with moving boundaries in heat flow and the motion of solitary nonlinear waves (solitons). The basic idea is as before, *i.e.* to write an evolution equation of the form

$$\frac{\partial \varphi}{\partial t} = N(\varphi)$$

where $\varphi(t,x) \in L^2(0,\infty,L^2(\Omega))$ for some open set $\Omega \subseteq \mathbb{R}^n$ and N is a nonlinear differential operator, as a pseudo-linear one:

$$\frac{\partial \varphi}{\partial t} = A(\varphi)\varphi \tag{9.1}$$

for some nonlinear differential operator $A(\varphi)$. Then we introduce a sequence of approximations $\varphi^{[i]}(t)$ given by

$$\frac{\partial \varphi^{[i]}}{\partial t}(t,x) = A(\varphi^{[i-1]}(t,x))\varphi^{[i]}(t,x). \tag{9.2}$$

The main difference between this and the finite-dimensional problem is the greater technical difficulty in proving convergence of the solutions of (9.2) to those of (9.1). A simpler approach, which we shall take here, is to apply the technique to partial differential equations which have 'regular discretisations' in the sense that a sequence of finite-dimensional discretisations exists which converge pointwise almost everywhere (a.e.) to the solutions of the partial differential equations (so long as they are sufficiently smooth). We then apply the iteration technique to each

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 151–167.
springerlink.com

finite-dimensional system in the discretisation sequence and use a diagonal argument to get the desired convergence of the linear, time-varying approximations to the solutions of the partial differential equation.

9.2 A Moving Boundary Problem

We shall illustrate the general ideas about using linear, time-varying approximations to nonlinear partial differential equations by considering the well-known two-phase Stefan problem. The basic theory of the Stefan problem is given in [1]. Consider a region $\Omega \subseteq \mathbb{R}^n$ which is divided into two (unknown) regions Ω_1, Ω_2 (which may not be connected) in which a material exists in the liquid state (in Ω_1) and the solid state (in Ω_2) such that

$$\Omega = \Omega_1 \cup \Omega_2.$$

The phase change, of course, takes place at the boundary of Ω_1 (and that of Ω_2):

$$\partial\Omega_1 = \partial\Omega_2.$$

In each of the regions Ω_1 and Ω_2 we assume that the process is simply one of heat conduction. Thus we have the equations

$$\frac{\partial T}{\partial t} = \alpha_L \nabla^2 T,\ (x,t) \in \Gamma_1 \doteq \Omega_1 \times (0,\tau) \tag{9.3}$$

$$\frac{\partial T}{\partial t} = \alpha_S \nabla^2 T,\ (x,t) \in \Gamma_2 \doteq \Omega_2 \times (0,\tau)$$

for some time $\tau > 0$, where α_L and α_S are the thermal conductivities associated, respectively, with the liquid and solid phases. In this form, the problem consists of a pair of linear equations defined in some regions in Ω with unknown boundary between them. The moving boundary problem causes considerable difficulty in the theory of partial differential equations and so we reformulate the problem as a single nonlinear partial differential equation which can be tackled easily by our approach, as we shall see. Thus, first note that the classical Stefan boundary condition states that the amount of latent heat at the boundary is given by the difference in thermal gradient there:

$$\rho L V_n = \left[-\kappa \frac{\partial T}{\partial n} \right]_-^+ \text{ on } \partial\Gamma_1,$$

where V_n is the velocity of the moving boundary, ρ is the density, L is the latent heat, κ is the thermal conductivity and n is the normal to the phase boundary. The energy content in the liquid region is given by

$$e^L(T) = L + \int_{T_M}^{T} C_L(\overline{T}) d\overline{T},\ T > T_M$$

and by

$$e^S(T) = L + \int_T^{T_M} C_S(\overline{T})d\overline{T},\ T < T_M$$

where C_L and C_S are the respective specific heats. The corresponding conductivity coefficients (assuming the densities of each phase are equal and constant) are given by

$$\alpha_L(T) = \frac{\kappa_L}{\rho C_L(T)},\ \alpha_S(T) = \frac{\kappa_S}{\rho C_S(T)}$$

and the energy expressions can be unified by defining the specific heat

$$C(T) = \begin{cases} L\delta(T - T_M) + C_L(T), & T \geq T_M \\ \quad C_S(T), & T < T_M, \end{cases}$$

and so the conductivity coefficient becomes

$$\alpha(T) = \begin{cases} \alpha_L(T) = \kappa_L/\rho C_L(T), & T \geq T_M \\ \alpha_S(T) = \kappa_S/\rho C_L(T), & T < T_M. \end{cases}$$

For simplicity we shall assume the conductivity coefficient is constant in each phase, *i.e.*

$$\alpha(T) = \begin{cases} \alpha_L, & T \geq T_M \\ \alpha_S, & T < T_M. \end{cases} \tag{9.4}$$

This gives a reasonable approximation in most cases. We can also use a smooth function α by defining a C^∞ function which is equal to α_S for $T \leq T_M - \varepsilon$ and equal to α_L for $T \geq T_M + \varepsilon$ (see Figure 9.1). This will avoid any technical difficulties in the existence and uniqueness theory of the partial differential equation. Thus, the equations (9.3) can be unified into the equation:

$$\frac{\partial T}{\partial t} = \alpha(T)\nabla^2 T,\ (x,t) \in \Omega \times (0,\tau), \tag{9.5}$$

where $\alpha(T)$ is given by (9.4).

9.3 Solution of the Unforced System

To solve the uncontrolled system (9.5), we introduce a sequence of linear, time-varying problems:

$$\frac{\partial T^{[i]}}{\partial t}(x,t) = \alpha(T^{[i-1]}(x,t))\nabla^2 T^{[i]}(x,t), \tag{9.6}$$

with some initial conditions

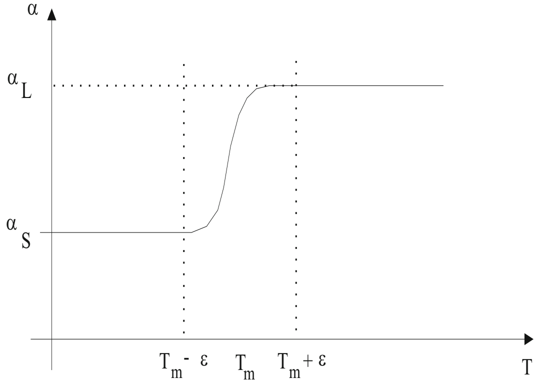


Fig. 9.1 Smooth temperature coefficient

$$\begin{aligned}T^{[i]}(x,0) &= f(x),\, x \in \Omega \\ T^{[0]}(x,t) &= f(x),\, t \in (0,\tau)\end{aligned}$$

and a Dirichlet boundary condition

$$T^{[i]}(\partial\Omega,t) = 0, \text{ say.}$$

Note that we are taking the whole of the zeroth approximation equal to the initial function $f(x)$, for all t. We could take any other (reasonable) function here. This effectively means that we are taking, for the first approximation $T^{[1]}$, the solution of the system

$$\frac{\partial T^{[1]}}{\partial t}(x,t) = \alpha(T^{[0]}(x,t))\nabla^2 T^{[1]}(x,t), \tag{9.7}$$

i.e.

$$\frac{\partial T^{[1]}}{\partial t}(x,t) = \alpha(f(x))\nabla^2 T^{[1]}(x,t).$$

It is well-known that each of the equations (9.7) has a unique solution. To prove the convergence of the sequence of solutions, we approximate each system (9.7) by a finite-dimensional approximation. For simplicity, we shall consider only the one spatial dimension case; the general case follows similarly. Thus, consider a one-dimensional bar with temperature $T(x,t)$, $0 \le x \le \ell$. Write

$$T_j^{[i]}(t) = T^{[i]}(t, j\ell/N),\ j = 1,2,\cdots,N.$$

Then we have the system

$$\frac{d}{dt}\mathfrak{T} = \mathfrak{A}\mathfrak{T}$$

where

$$\mathfrak{T} = \begin{pmatrix} T_1^{[i]} \\ \vdots \\ T_N^{[i]} \end{pmatrix},$$

$$\mathfrak{A} = \frac{1}{(\Delta x)^2} \begin{pmatrix} -2\alpha(T_1^{[i-1]}) & \alpha(T_1^{[i-1]}) & 0 & \cdots & \cdots & 0 \\ \alpha(T_2^{[i-1]}) & -2\alpha(T_2^{[i-1]}) & \alpha(T_2^{[i-1]}) & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & \alpha(T_N^{[i-1]}) & -2\alpha(T_N^{[i-1]}) \end{pmatrix},$$

$$T^{[i]}(0) = T_0,\, T^{[i]} = (T_1^{[i]}, \cdots, T_N^{[i]})^T,$$

and

$$T_0 = (f(\ell/N), f(2\ell/N), \cdots, f((N-1)\ell/N), f(\ell))\,.$$

(f is the initial condition, as above.) Hence we can write the system in the form

$$\frac{\partial T^{[i]}}{\partial t}(t) = A(T^{[i-1]})T^{[i]},\, T(0) = T_0. \tag{9.8}$$

Since A is locally Lipschitz, we have the following theorem ([2], see Chapter 2):

Theorem 9.1. *The sequence of temperatures $T^{[i]}(t)$ defined by (9.8) is uniformly convergent on any compact time interval.*

Combining this with the well-known convergence theory for finite-dimensional approximations to diffusion systems, we have

Theorem 9.2. *The sequence of temperatures $T^{[i]}(t)$ defined by (9.8) converges uniformly almost everywhere on compact time intervals, as $t \to \infty$ and $N \to \infty$, to the solution of (9.6).*

To illustrate the technique in a simple case of heat flow in a two-phase system, the unforced diffusion system has been solved with $f(x) = e^{(4(x-1)^2)}$, so that some regions are liquid and some solid at $t = 0$. Here we have taken $\ell = 2, N = 30$ and $\tau = 0.75$; 300 time steps were used. Two vales of the thermal conductivity coefficients were used and it can be seen that the method correctly predicts the phase boundary and converges in 4–5 iterations (see Figure 9.2).

9.4 The Control Problem

We now consider the problem of controlling the temperature profile in a one-dimensional bar with a pointwise laser heating control. The physical setup is shown in Figure 9.3. The equation of the system is given by

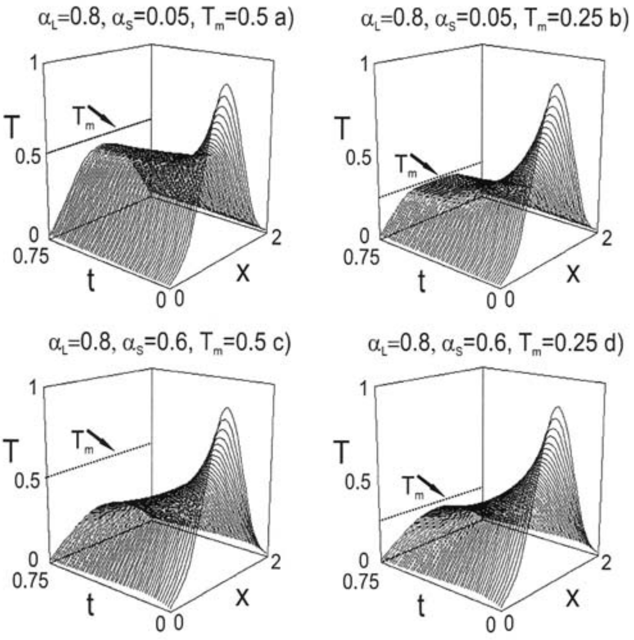


Fig. 9.2 Plots of the solution of the unforced system with $f(x) = e^{4(x-1)^2}$ and various values of α_L, α_S

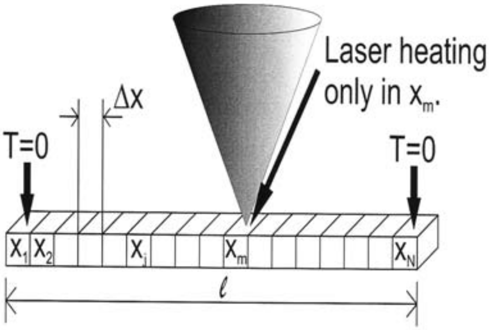


Fig. 9.3 The basic laser heating system

$$\frac{\partial T}{\partial t} = \alpha(T)\nabla^2 T + \delta(x - \overline{x})u,\ T(0) = T_0. \tag{9.9}$$

where u is proportional to the heat input power from the laser and $\overline{x}$ is the point of injection of the heat. We shall apply the iteration method as before to the problem, so that we consider the finite-dimensional approximation

$$\frac{\partial T^{[i]}}{\partial t}(t) = A(T^{[i-1]}(t))T^{[i]}(t) + Bu, \tag{9.10}$$

where A is the matrix defined above and

$$B = (0, 0, \cdots, 1, 0, \cdots, 0)^T \tag{9.11}$$

and the '1' is in the m^{th} place corresponding to the point

$$\overline{x} = m\ell/N. \tag{9.12}$$

If $T_d(t)$ is the desired temperature profile, we shall solve the optimal tracking problem of minimising the cost functional

$$\begin{aligned} J = &\frac{1}{2}(T^{[i]}(t_f) - T_d(t_f))^T F(T^{[i]}(t_f) - T_d(t_f)) \\ &+ \frac{1}{2}\int_0^{t_f} \left\{ (T^{[i]}(t) - T_d(t))^T Q(T^{[i]}(t) - T_d(t)) + u^T Ru \right\} dt \end{aligned}$$

where F and Q are positive semi-definite matrices and R is positive-definite.

In order to consider the 'trackability' of a given desired temperature profile, we shall first look at the general problem in terms of the nonlinear (finite-dimensional) control system

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

Suppose that we desire to track the function $x_d(t)$; then there must exist a control $u_d(t)$ such that

$$\dot{x}_d(t) = f(x_d(t), u_d(t)) \tag{9.13}$$

for all $t \geq \overline{t} > 0$ for some finite $\overline{t}$. Let

$$y(t) = x(t) - x_d(t).$$

Then

$$\begin{aligned} \dot{y}(t) &= \dot{x}(t) - \dot{x}_d(t) \\ &= f(x(t), u(t)) - f(x_d(t), u_d(t)) \\ &= g(y(t), v(t), t) \end{aligned}$$

where

$$v(t) = u(t) = u_d(t)$$

and

$$g(0,0,t)=0,$$

by Taylor's theorem. We can write this equation in the form

$$\dot{y}(t)=A(y(t),v(t),t)y(t)+B(y(t),v(t),t)v(t)$$

for some matrix-valued functions A and B. Hence, we can always rewrite a tracking problem as a regulator one provided the system can track the desired function $x_d(t)$, *i.e.* there is an open-loop control $u_d(t)$ such that (9.13) holds. If x_d is constant then (9.13) becomes

$$f(x_d(t),u_d(t))=0 \tag{9.14}$$

and so, for trackability, there must exist a (constant) control u_d such that (9.14) holds.

Specialising to the heat control problem, if we want to track a given temperature profile T^d, then we must have

$$\frac{1}{(\Delta x)^2}\begin{pmatrix} -2\alpha(T_1^d) & \alpha(T_1^d) & 0 & \cdots & \cdots & 0 \\ \alpha(T_2^d) & -2\alpha(T_2^d) & \alpha(T_2^d) & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & \alpha(T_N^d) & -2\alpha(T_N^d) \end{pmatrix}\begin{pmatrix} T_1^d \\ T_2^d \\ \vdots \\ T_N^d \end{pmatrix}=-\begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}u_d.$$

Hence, if the input is in the m^{th} place, we require

$$\begin{aligned} \frac{\alpha(T_1^d)}{(\Delta x)^2}(-2T_1^d+T_2^d) &= 0 \\ \frac{\alpha(T_2^d)}{(\Delta x)^2}(T_1^d-2T_2^d+T_3^d) &= 0 \\ &\cdots \\ \frac{\alpha(T_{m-1}^d)}{(\Delta x)^2}(T_{m-2}^d-2T_{m-1}^d+T_m^d) &= 0 \\ \frac{\alpha(T_m^d)}{(\Delta x)^2}(T_{m-1}^d-2T_m^d+T_{m+1}^d) &= -u_d \\ &\cdots \\ \frac{\alpha(T_{N-1}^d)}{(\Delta x)^2}(T_{N-2}^d-2T_{N-1}^d+T_N^d) &= 0 \\ \frac{\alpha(T_N^d)}{(\Delta x)^2}(T_{N-1}^d-2T_N^d) &= 0. \end{aligned}$$

An elementary calculation shows that

$$T_m=mT_1,\ ,T_m=(N-m+1)T_N$$

so that

$$T_N = \frac{mT_1}{N-m+1}$$

and

$$u_d = \frac{\alpha(T_m)}{(\Delta x)^2} \frac{N+1}{N-m+1} T_1.$$

Hence, the only constant (in time) temperature profiles which are trackable are piecewise linear about the control point.

In the case where we allow time-varying desired temperature profiles, $T_i^d(t)$, we must satisfy the following equations:

$$\begin{aligned} T_2^d(t) &= \frac{\dot{T}_1^d(t)}{\beta T_1^d(t)} + 2T_1^d(t) \\ T_3^d(t) &= \frac{\dot{T}_2^d(t)}{\beta T_2^d(t)} + 2T_2^d(t) - T_1^d(t) \\ &\cdots \\ T_m^d(t) &= \frac{\dot{T}_{m-1}^d(t)}{\beta T_{m-1}^d(t)} + 2T_{m-1}^d(t) - T_{m-2}^d(t) \end{aligned} \tag{9.15}$$

and

$$\begin{aligned} T_{N-1}^d(t) &= \frac{\dot{T}_N^d(t)}{\beta T_N^d(t)} + 2T_N^d(t) \\ T_{N-2}^d(t) &= \frac{\dot{T}_{N-1}^d(t)}{\beta T_{N-1}^d(t)} + 2T_{N-1}^d(t) - T_N^d(t) \\ &\cdots \\ T_m^d(t) &= \frac{\dot{T}_{m+1}^d(t)}{\beta T_{m+1}^d(t)} + 2T_{m+1}^d(t) - T_{m+2}^d(t) \end{aligned} \tag{9.16}$$

where $\beta = (1/(\Delta x)^2)\alpha(T)$. To solve these equations we write

$$\Gamma_k = \begin{pmatrix} T_{k-1}^d \\ T_k^d \end{pmatrix}.$$

Then the equations become

$$\Gamma_k = \begin{pmatrix} 0 & 1 \\ -1 & N \end{pmatrix} \Gamma_{k-1}$$

where N is the operator defined by

$$N(T) = \frac{1}{\beta(T)} \frac{dT}{dt}.$$

Hence

$$\Gamma_k = K\Gamma_{k-1}$$

where

$$K = \begin{pmatrix} 0 & 1 \\ -1 & N \end{pmatrix}$$

and so

$$\Gamma_k = K^{k-2}\Gamma_2$$

so that

$$T_m^d = \left(K^{m-2} \begin{pmatrix} T_1^d \\ \frac{\dot{T}_1^d(t)}{\beta T_1^d(t)} + 2T_1^d(t) \end{pmatrix} \right)_2$$

where $(\cdot)_2$ means the second component. Similarly, starting with (9.14) we have

$$T_m^d = \left(K^{N-m-1} \begin{pmatrix} T_N^d \\ \frac{\dot{T}_N^d(t)}{\beta T_N^d(t)} + 2T_N^d(t) \end{pmatrix} \right)_1$$

and so we have proved:

Theorem 9.3. *A necessary and sufficient condition for the system (9.9) to be able to track a desired temperature profile is that (9.15) and (9.16) are satisfied and that T_1^d and T_N^d are related by*

$$\left(K^{m-2} \begin{pmatrix} T_1^d \\ \frac{\dot{T}_1^d(t)}{\beta T_1^d(t)} + 2T_1^d(t) \end{pmatrix} \right)_2 = \left(K^{N-m-1} \begin{pmatrix} T_N^d \\ \frac{\dot{T}_N^d(t)}{\beta T_N^d(t)} + 2T_N^d(t) \end{pmatrix} \right)_1 .$$

Moreover, the (open-loop) control is given by

$$u_d = \dot{T}_m^d - \beta(T_m^d)(T_{m-1}^d - 2T_m^d + T_{m+1}^d).$$

Of course, for the case of laser heating, we also have the condition that

$$u_d(t) \geq 0 \text{ for all } t \geq 0.$$

These conditions are highly nonlinear and can be used as a test for any given desired tracking function. In general we will expect perfect tracking only for a very restricted class of functions.

Finally we apply the above results to a real problem. Consider the case of holding a one-dimensional bar such as the one in the above figure at the melt temperature, *i.e.* $T^d(x,t) = T_M$ for all x and t. The parameters used were $N = 30$, $\ell = 2$, $\alpha_L = 0.8$, $\alpha_S = 0.02$, $T_M = 0.25$ and $R = 0.5$. The heating point was taken to be in the middle of the bar, *i.e.* $m = 15$ and the time step was $\delta = 0.001$s. The initial condition was $T^{[0]}(x,0) = 0$. Figure 9.4 shows results obtained with around 5–7 iterations when the approximations have converged. Better tracking can be obtained by reducing α_L/α_S.

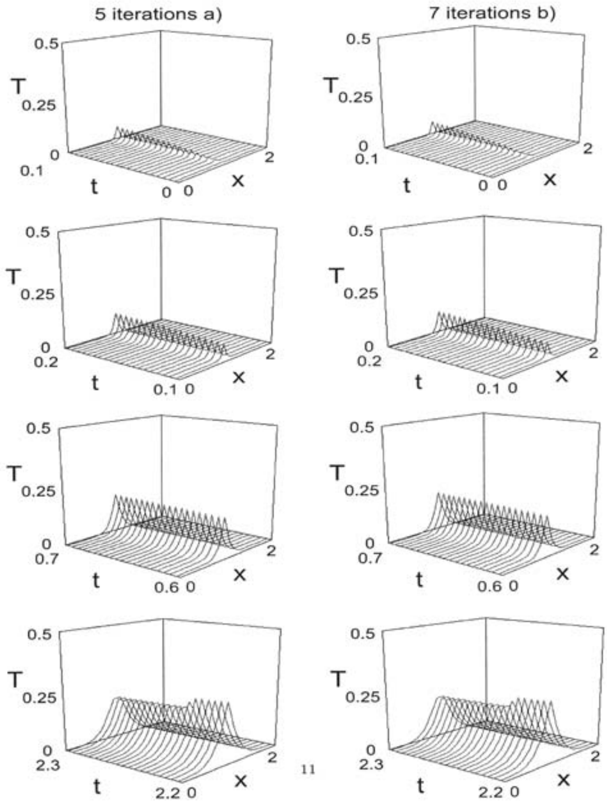


Fig. 9.4 Plots of the controlled laser heating problem

9.5 Solitons and Boundary Control

In the second half of this chapter we shall illustrate the iteration technique for partial differential equations by considering the generalised Korteweg-de Vries (KdV) equation

$$\varphi_t + \varphi_x + k(\varphi)\varphi_x + \varphi_{xxx} = 0 \tag{9.17}$$

defined on the spatial interval (α, β), mentioned in Chapter 1. We shall assume that the system is controlled by the boundary values

$$\varphi(\alpha,t) = u_1(t),\ \varphi(\beta,t) = u_2(t),\ \varphi_x(\beta,t) = u_3(t).$$

The exact boundary control problem for this system can be stated as follows:

Let $T > 0$ and $s \geq 0$. For any $f, g \in H^s(\alpha, \beta)$, find boundary controls u_j, $j = 1,2,3$ such that the solution $\varphi \in C([0,T]; H^s(\alpha, \beta))$ satisfies

$$\varphi(x,0) = f(x),\ \varphi(x,T) = g(x)$$

(in the distributional sense) on the interval (α, β).

We write the system in the form of an abstract equation on some Hilbert space X:

$$\frac{dy}{dt} = Ay + F(y) + Bu \tag{9.18}$$

where A generates a C^0-semigroup $W(t)$ in X and u is the control.

First we study the linearised equation:

Definition 9.1 We say that the linear evolution

$$\frac{dy}{dt} = Ay + Bu$$

is exactly controllable on X if there exists a bounded linear operator $G : X \times X \to L^2(0,T;X)$ such that, for all $f, g \in X$, the unique solution of

$$\frac{dy}{dt} = Ay + BG(f,g),\ y(0) = f$$

satisfies $y(T) = g$.

This simply expresses the control u, if it exists, as a linear function G of the desired starting and ending values f and g. If the linearised system is exactly controllable, then we can prove that the nonlinear system is also exactly controllable in the following way. Write the nonlinear system in integral form:

$$y(t) = W(t)y(0) + \int_0^t W(t-\tau)F(y(\tau))d\tau + \int_0^t W(t-\tau)B(u(\tau))d\tau.$$

Then we define

$$\ell(T,y) = \int_0^t W(t-\tau)F(y(\tau))d\tau$$

and the operator Γ by

$$\Gamma(y) = W(t)f + \int_0^t W(t-\tau)F(y(\tau))d\tau + \int_0^t W(t-\tau)BG(f, g-\ell(T,y))(\tau)d\tau.$$

If we can prove that Γ has a fixed point, then this point is a solution of the nonlinear system with the feedback control

$$u = G(f, g-\ell(T,y))(\tau)$$

such that

$$\Gamma(y)(0) = f$$

and

$$\Gamma(y)(T) = \ell(T,y) + g - \ell(T,y) = g,$$

proving exact controllability. The exact controllability of the linear system is fairly standard and can be found in [3]. In order the use the above approach to prove the exact controllability of the nonlinear system, we consider an equivalent problem on the whole of $\mathbb{R}$ rather than the boundary control problem on (α,β). We then ask if, for any given functions $f,g \in H^s(\mathbb{R})$, we can find a solution φ of the equation

$$\varphi_t + \varphi_x + k(\varphi)\varphi_x + \varphi_{xxx} = 0 \tag{9.19}$$

which satisfies

$$\varphi(x,0) = f(x),\ \varphi(x,T) = g(x)$$

on the interval (α,β). We then get a solution to the original control problem by choosing

$$u_1(t) = \varphi(\alpha,t),\ \varphi u_2(t) = \varphi(\beta,t),\ u_3(t) = \varphi_x(\beta,t)$$

and take the restriction of φ to $[\alpha,\beta]\times[0,T]$. To prove that the above operator Γ has a fixed point we need a number of smoothing properties of the linear system. Let $W(t)$ be the unitary group generated by the operator

$$Af = -f' - f'''$$

from $L^2(\mathbb{R})$ to itself, with domain $\mathfrak{D}(A) = H^3(\mathbb{R})$. Then the solution of the linear KdV equation

$$\begin{aligned} \varphi_t + \varphi_x + \varphi_{xxx} &= 0 \\ \varphi(x,0) &= f(x) \end{aligned}$$

is given by

$$\varphi(t) = W(t)f.$$

If L^2_b denotes the weighted Hilbert space $L^2(e^{2bx}dx)$ for ant $b > 0$, then ([4]) A generates a semigroup $W_b(t)$ in this space, given by

$$W_b(t) = exp(-t(D-b)^3 - t(D-b))$$

where D is a differential operator. Moreover, we have the estimate

$$\|W_b(t)\|_{L(H^s(\mathbb{R}),H^{s'}(\mathbb{R}))} \le ct^{-(s'-s)/2}exp(b^3t)$$

for $s \le s'$ and if

$$\varphi_t + \varphi_x + \varphi_{xxx} = p,\ 0 < t < T$$

and

$$e^{bx}\varphi \in L^\infty([0,T];H^s(\mathbb{R})),\ e^{bx}p \in L^\infty([0,T];H^{s-1}(\mathbb{R}))$$

then

$$e^{bx}\varphi \in C([0,T];H^0) \cap C([0,T];H^s) \text{ for } s' < s+1.$$

Moreover, we have

$$e^{bx}\varphi(t) = W_b(t)\varphi(0) + \int_0^t W_b(t-\tau)e^{bx}p(\tau)d\tau.$$

To determine a better smoothing property we define the space $Y_{s,b}$ $(a,b \in \mathbb{R})$ to be the completion of the space $S(\mathbb{R}^2)$ of tempered distributions with respect to the norm

$$\|f\|^2_{Y_{s,b}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(1+|\tau-\xi-\xi^3|)^{2b}(1+|\xi|)^{2s}|\hat{f}(\xi,\tau)|^2 d\xi d\tau,$$

where $\hat{f}$ is the Fourier transform of f. Then, if $s > -1$ and $b > 1/2$,

$$f \in Y_{s,b} \Rightarrow f \in C^{[1+s],\alpha}_{loc}(\mathbb{R};L^2(\mathbb{R}))$$

for any $0 \le \alpha \le 1+s-[1+s]$, and so

$$f \in L^p_{s,loc}(\mathbb{R};L^2(\mathbb{R})),\ 1 \le p \le \infty.$$

The smoothing property we need is (see [5])

Lemma 9.1. *Let $s > -3/4$, $\sigma \in C_0^\infty(\mathbb{R})$. Then there exists $\beta \in (1/2,1)$ such that for all $b \in (1/2,\beta)$, there exists $c > 0$ such that*

$$\|\sigma(t)\partial_x(\varphi\psi)\|_{Y_{s,b-1}} \le c\|\varphi\|_{Y_{s,b}}\|\psi\|_{Y_{s,b}}$$

for all $\varphi,\psi \in Y_{s,b}$.

Since $Y_{s,b} \subseteq C(\mathbb{R};H^s(\mathbb{R}))$ for $b > 1/2$, it follows that

$$\left\|\sigma_1(t)\int_0^t W(t-\tau)\sigma_2(\tau)(\partial_x(\varphi\psi))(\cdot,\tau)d\tau\right\|_{Y_{s,b}} \le c\|\varphi\|_{Y_{s,b}}\|\psi\|_{Y_{s,b}}$$

for some functions $\sigma_1(t),\sigma_2(t) \in C_0^\infty(\mathbb{R})$.

We can now state a result from which the boundary controllability of the system will follow:

Lemma 9.2. *Consider the nonlinear KdV equation*

$$\begin{aligned}\varphi_t + \varphi_x + k(\varphi)\varphi_x + (a(x,t)\varphi)_x + \varphi_{xxx} &= 0,\ x,t \in H\\ \varphi(x,0) &= h(x),\end{aligned}$$

where $a(x,t) \in Y_{s,b}$ and k is of the form

$$k(\varphi) = p'(\varphi)\varphi + p(\varphi)$$

where p is differentiable and

$$\|p(\varphi)\|_{Y_{s,b}} \leq c\|\varphi\|_{Y_{s,b}}$$

for some constant c.Let $s \geq 0, T > 0$ and $b > 0$ be as in Lemma 9.1. Then there exists $\delta > 0$ such that if $f, g \in H^s(\alpha,\beta)$ with

$$\|f\|_{H^s(\alpha,\beta)} \leq \delta, \|g\|_{H^s(\alpha,\beta)} \leq \delta$$

there exists $h \in H^s(\mathbb{R})$ such that the solution of the equation satisfies

$$\varphi(x,0) = f(x),\ \varphi(x,T) = g(x),\ x \in (\alpha,\beta).$$

Proof. We have

$$\varphi(t) = W_a(t)h - \int_0^t W_a(t-\tau)(k(\varphi)\varphi_x)(\tau)d\tau \tag{9.20}$$

where W_a is the C^0-semigroup introduced above. Let

$$\ell(T,\varphi) = \int_0^t W_a(t-\tau)(k(\varphi)\varphi_x)(\tau)d\tau.$$

Then we can choose

$$h = G(f, g + \ell(T,\varphi))$$

in ([5], Proposition 4.1) to give

$$\varphi(t) = W_a(t)G(f, g + \ell(T,\varphi)) - \int_0^t W_a(t-\tau)(k(\varphi)\varphi_x)(\tau)d\tau$$

and then

$$\varphi(x,0) = f(x),\ \varphi(x,T) = g(x).$$

All that remains is to show that the map

$$\Gamma(\varphi) = W_a(t)G(f, g + \ell(T,\varphi)) - \int_0^t W_a(t-\tau)(k(\varphi)\varphi_x)(\tau)d\tau$$

has a fixed point in $Y_{s,b}$ by demonstrating that it is a contraction. This follows from the inequality

$$\begin{aligned}
&\left\| \int_0^t W_a(t-\tau)(k(\varphi)\varphi_x)(\tau)d\tau \right\|_{H^s(\mathbb{R})} \\
&\leq sup_{t\in\mathbb{R}} \left\| \sigma_1(t) \int_0^T W_a(t-\tau)\sigma_2(t)(k(\varphi)\varphi_x)(\tau)d\tau \right\|_{H^s(\mathbb{R})} \\
&= sup_{t\in\mathbb{R}} \left\| \sigma_1(t) \int_0^T W_a(t-\tau)\sigma_2(t)(\partial_x(p(\varphi)\varphi)\varphi_x)(\tau)d\tau \right\|_{H^s(\mathbb{R})} \\
&\leq c\|\varphi\|^2_{Y_{s,b}}.
\end{aligned}$$

□

We can now state the main result:

Theorem 9.4. *Suppose that $k(\varphi)$ is of the form*

$$k(\varphi) = p'(\varphi)\varphi + p(\varphi)$$

where p is as above and let $T > 0$ and $s \geq 0$ be given. If $[\alpha,\beta] \subseteq (\alpha_1,\beta_1)$, suppose that

$$w(x,t) \in C^\infty[(\alpha_1,\beta_1)\times(-\varepsilon,T+\varepsilon)]$$

for some $\varepsilon > 0$ satisfies

$$w_t + w_x + k(w)w_x + w_{xxx} = 0,\ (x,t) \in (\alpha_1,\beta_1)\times(-\varepsilon,T+\varepsilon).$$

Then there exists a $\delta > 0$ such that for any $f,g \in H^s(\alpha,\beta)$ satisfying

$$\|f(\cdot) - w(\cdot,0)\|_{H^s(\alpha,\beta)} \leq \ and \|g(\cdot) - w(\cdot,0)\|_{H^s(\alpha,\beta)} \leq \delta$$

one can find controls u_1,u_2,u_3 in $L^2(0,T)$ ($u_j \in C[0,T]$, $j = 1,2,3$ if $s > 3/2$) such that the system has the solution

$$\varphi \in C([0,T];H^s(\alpha,\beta)) \cap L^2(0,T;H^s(\alpha,\beta))$$

satisfying

$$\varphi(x,0) = f(x),\ \varphi(x,T) = g(x)$$

on the interval (α,β).

Now that we have controllability we can apply the iteration method for optimal control in a fairly standard way. For the details, see [6].

9.6 Conclusions

In this chapter we have outlined the application of the iteration scheme to nonlinear partial differential equations. We have shown that moving boundary problems and boundary control systems can be effectively treated by our methods. In order to avoid technical difficulties with nonlinear partial differential equations, we have solved moving boundary problems by using a discretization of the system and consider it as a finite-dimensional problem. Boundary controllability has been shown using standard techniques from Sobolev theory. Clearly the method can be applied to a wide variety of such problems.

References

1. Alexiades, V., Solomon, A.D.: Mathematical Modeling of Melting and Freezing Processes. Taylor and Francis, London (1993)
2. Tomás-Rodríguez, M., Banks, S.P.: Linear Approximations to Nonlinear Dynamical Systems with Applications to Stability and Spectral Theory. IMA Journal of Math. Control and Inf. 20, 89–103 (2003)
3. Curtain, R.F., Pritchard, A.J.: Infinite-Dimensional Linear Systems Theory. Springer, London (1978)
4. Kato, T.: On the Cauchy Problem for the (Generalised) Korteweg-de Vries Equations. In: Advances in Mathematics Supplementary Studies. Stud. Appl. Math., vol. 8, pp. 93–128. Academic Press, New York (1983)
5. Zhang, B.-Y.: Exact Boundary Controllability of the Korteweg-de Vries Equation. SIAM J. Control 37(2), 543–565 (1999)
6. Banks, S.P.: Exact boundary controllability and optimal control for a generalised Korteweg-de Vries equation. J. Nonlinear Anal. - Apps and Methods 47, 5537–5546 (2001)

Chapter 10
Lie Algebraic Methods

10.1 Introduction

In this chapter we shall consider systems of the form

$$\dot{x} = A(x)x \tag{10.1}$$

where $A : \mathbb{R}^n \to \mathfrak{g}$ and $\mathfrak{g}$ is the Lie algebra of a Lie group G. The classical structure theory of Lie groups and Lie algebras (see Appendix B and [1,2]) will be used to decompose the system (10.1) into simpler subsystems in a way which generalises the classical Jordan decomposition of single matrices. In the latter case, of course, if we have a linear system

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

then we may use the generalised eigenspaces of A to write the system in the form

$$\dot{y} = Jy$$

where J is a Jordan matrix of the block-diagonal form

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix}$$

and each J_i is of the form

$$J_i = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 169–194.
springerlink.com

where λ is some eigenvalue of A. Here,

$$y = P^{-1}x$$

where

$$P^{-1}AP = J$$

and P is the matrix of generalised eigenvectors of A. (One drawback here is that P is a complex matrix if the eigenvalues are complex, but one can always use the real normal form, if necessary.)

In the case of the nonlinear system (10.1), we first decompose $\mathfrak{g}$ into the Levi form

$$\mathfrak{g} = \mathfrak{s} + \mathfrak{g}_1 \tag{10.2}$$

where $\mathfrak{s}$ is a solvable Lie algebra and $\mathfrak{g}_1$ is semi-simple. Note that the sum in (10.2) is not direct. Since $\mathfrak{g}_1$ is semi-simple it has a direct sum decomposition

$$\mathfrak{g}_1 = \mathfrak{h} \oplus \sum_{\alpha \in \Delta} \mathfrak{g}_\alpha$$

where $\mathfrak{h}$ is a Cartan subalgebra and $\mathfrak{g}_\alpha$ is a one-dimensional root space. It follows that any system of the form (10.1) may be written as

$$\dot{x} = S(x)x + H(x)x + \sum_{\alpha \in \Delta} e_\alpha(x) E_\alpha x$$

where $S(x) \in \mathfrak{s}$, $H(x) \in \mathfrak{h}$ and $E_\alpha \in \mathfrak{g}_\alpha$ for each $x \in \mathbb{R}^n$ and $\alpha \in \Delta$. Note that all the matrices in $\mathfrak{h}$ are simultaneously diagonalisable.

10.2 The Lie Algebra of a Differential Equation

In this section we consider the basic properties of a Lie algebra associated with a nonlinear system. Thus, consider the nonlinear system

$$\dot{x} = A(x)x \tag{10.3}$$

where A is continuous and let

$$\mathscr{L}_{\{A(x)\}} = \text{ Lie subalgebra of } \mathfrak{g}\ell(n, \mathbb{C}) \text{ generated by } A(x),\ \ x \in \mathbb{R}^n.$$

If A is also (real) analytic, we may expand it in a Taylor series:

$$A(x) = \sum_{|\mathbf{i}| \geq \mathbf{0}} A_{\mathbf{i}} x^{\mathbf{i}}$$

where $\mathbf{i} = (i_1, \cdots, i_n)$, $x^{\mathbf{i}} = x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}$ and $A_{\mathbf{i}}$ is a constant matrix. Then we also consider the Lie algebra $\mathscr{L}_{A_{\mathbf{i}}}$ defined by

$$\mathscr{L}_{A_{\mathbf{i}}} = \text{ Lie subalgebra of } \mathfrak{g}\ell(n,\mathbb{C}) \text{ generated by } A_{\mathbf{i}}, \ \ |\mathbf{i}| \geq \mathbf{0}.$$

Lemma 10.1. *If $A(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is analytic, then $\mathscr{L}_{\{A(x)\}} = \mathscr{L}_{A_i}$.*

Proof. If $\{B_i\}_{1\leq i\leq K}$ is a basis of $\mathscr{L}_{A_{\mathbf{i}}}$, then

$$A(x) = \sum_{i=1}^{K} p_i(x) B_i,$$

where $p_i(x)$ exists for all $x \in \mathbb{R}^n$ by the analyticity of $A(\cdot)$. Hence we have $A(x) \in \mathscr{L}_{\{A_{\mathbf{i}}\}}$ and so $\mathscr{L}_{\{A(x)\}} \subseteq \mathscr{L}_{\{A_{\mathbf{i}}\}}$.

For the converse, we show that $A_{\mathbf{i}} \in \mathscr{L}_{\{A(x)\}}$ for each $\mathbf{i}$. Clearly, $A_0 = A(0)$ so it is true for $\mathbf{i} = 0$. We prove that $A_{1,0,\cdots,0} \in \mathscr{L}_{\{A(x)\}}$, the others being similar. Now,

$$A_{1,0,\cdots,0} = \frac{\partial}{\partial x_1} A(x)\bigg|_{x=0} = \lim_{h\to 0} \frac{A(he_i) - A(0)}{h}.$$

If $\{B_i'\}_{1\leq i\leq M}$ is a basis of $\mathscr{L}_{\{A(x)\}}$, then

$$A_{1,0,\cdots,0} = \lim_{h\to 0} \sum_{i=1}^{M} q_i(h) B_i'$$

for some functions $q_i(h)$. Since the B_i''s are linearly independent, each limit $\lim_{h\to 0} q_i(h)$ must exist, so that $A_{1,0,\cdots,0} \in \mathscr{L}_{\{A(x)\}}$. □

Remark 10.1. *(a) The Lie algebra $\mathscr{L}_{\{A(x)\}}$ is defined even if $A(\cdot)$ is not analytic, so this is the more general case.*

(b) $\mathscr{L}_{\{A(x)\}}$ depends on the representation (10.3); however, different representations are equivalent in the sense that they operate on $\mathbb{R}^n$ to give the same solutions. Hence we denote the Lie algebra of (10.1) by $\mathscr{L}_A$.

Theorem 10.1. *Any nonlinear system of the form (10.1) can be written (in a suitable basis) as*

$$\dot{x} = S(x)x + \begin{pmatrix} \Gamma_1(x) & & & \\ & \Gamma_2(x) & & \\ & & \ddots & \\ & & & \Gamma_r(x) \end{pmatrix} x$$

where $(\widetilde{S}(x), \widetilde{S}(x)) = 0$, $\widetilde{S}(x) = [S(x), S(x)]$ and Γ_i belongs to one of the simple Lie algebras $A_n, B_n, C_n, D_n, G_2, F_4, E_6, E_7$ or E_8 where $(\cdot,\cdot)$ denotes the Killing form of $\mathscr{L}_A$.

Proof. Let $\mathfrak{r}$ be the radical of $\mathscr{L}_A$ and let

$$\mathscr{L}_A = \mathfrak{r} + \mathfrak{m}$$

be a Levi decomposition of $\mathscr{L}_A$, where $\mathfrak{m}$ is a semi-simple subalgebra. (Note that the sum is not direct, so the decomposition is not unique.) The semi-simple part $\mathfrak{m}$ may be written

$$\mathfrak{m} = \mathfrak{m}_1 \oplus \cdots \oplus \mathfrak{m}_r$$

as a direct sum of simple Lie algebras $\mathfrak{m}_i$ (which are ideals in $\mathscr{L}_A$). The condition $(\widetilde{S}(x), \widetilde{S}(x)) = 0$ follows from Cartan's criterion for semi-simplicity. □

Theorem 10.2. *Any system*

$$\dot{x} = A(x)x$$

where $\mathscr{L}_A$ is generates a simple Lie algebra can be written in one of the following forms:
Type A:

$$\dot{x} = A(x)x, \ \ tr\ (A(x)) = 0.$$

Type B:

$$\dot{x} = \begin{pmatrix} 0 & u(x) & v(x) \\ -v^T(x) & A_{11}(x) & A_{12}(x) \\ -u^T(x) & A_{21}(x) & -A_{11}^T(x) \end{pmatrix} x,$$
$$A_{12}^T(x) = -A_{12}(x), \ \ A_{21}^T(x) = -A_{21}(x).$$

Type C:

$$\dot{x} = \begin{pmatrix} A_{11}(x) & A_{12}(x) \\ A_{21}(x) & -A_{11}^T(x) \end{pmatrix} x,$$
$$A_{12}^T(x) = A_{12}(x), \ \ A_{21}^T(x) = A_{21}(x).$$

Type D:

$$\dot{x} = \begin{pmatrix} A_{11}(x) & A_{12}(x) \\ A_{21}(x) & -A_{11}^T(x) \end{pmatrix} x,$$
$$A_{12}^T(x) = -A_{12}(x), \ \ A_{21}^T(x) = -A_{21}(x).$$

Type G_2, where $\mu = \lambda_1(x) + \lambda_2(x), r = \sqrt{2}$:

$$\dot{x} = \begin{pmatrix} 0 & -rb_1(x) & -rb_2(x) & -rb_3(x) & ra_1(x) & ra_2(x) & ra_3(x) \\ -ra_1(x) & \lambda_1(x) & c_1(x) & c_3(x) & 0 & b_3(x) & -b_2(x) \\ -ra_2(x) & c_2(x) & \lambda_2(x) & c_5(x) & -b_3(x) & 0 & b_1(x) \\ -ra_3(x) & c_4(x) & c_6(x) & -\mu & b_2(x) & -b_1(x) & 0 \\ rb_1(x) & 0 & -a_3(x) & a_2(x) & -\lambda_1(x) & -c_2(x) & -c_4(x) \\ rb_2(x) & a_3(x) & 0 & -a_1(x) & -c_1(x) & -\lambda_2(x) & -c_6(x) \\ rb_3(x) & -a_2(x) & a_1(x) & 0 & -c_3(x) & -c_5(x) & \mu \end{pmatrix} x.$$

Types F_4, E_6, E_7, E_8:

$$\dot{x} = \left(\sum_{i=1}^{r} a_i(x) X_i + \sum_{i=1}^{s} b_i(x) Y_i \right) x$$

where X_i, Y_j satisfy

$$[X_i, Y_j] = \sum_{p=1}^{s} x^i_{pj} Y_p, \ \ 1 \le i \le r, \ \ 1 \le j \le s$$

$$[Y_\alpha, Y_\beta] = \sum_{i=1}^{r} x^i_{\alpha\beta} X_i, \ \ 1 \le \alpha, \beta \le s$$

where $X_i = (x^i_{\alpha\beta})$ and X_i, Y_j can be realised on a 16-dimensional space for type F_4 (with $r = 36, s = 16$) or a 2^7-dimensional space for types E_6, E_7, E_8. For $E_8, r = 120, s = 128$ and E_6, E_7 are subalgebras of dimensions 78 and 133, respectively.

(See Appendix B.)

Example 10.1. The system

$$\begin{aligned} \dot{x}_1 &= 4x_1 - x_2 x_3 + x_2^4 x_3 \\ \dot{x}_2 &= -x_1 + x_1 x_3 - x_1^2 x_3 \\ \dot{x}_3 &= x_1 + 4x_3 - x_1 x_2^4 + x_1^2 x_2 \end{aligned}$$

can be written

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 & -x_3 & x_2^4 \\ x_3 & 0 & -x_1^2 \\ -x_2^4 & x_1^2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The matrix $\begin{pmatrix} 4 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 4 \end{pmatrix}$ belongs to a trivial solvable Lie algebra and $\begin{pmatrix} 0 & -x_3 & x_2^4 \\ x_3 & 0 & -x_1^2 \\ -x_2^4 & x_1^2 & 0 \end{pmatrix}$ belongs to the simple Lie algebra $\mathfrak{g}_3$ generated by the matrices

$$M_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \ M_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \ M_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Note that

$$[M_i, M_j] = M_k$$

where (i, j, k) is an even permutation of $1, 2, 3$.

10.3 Lie Groups and the Solution of the System

In this section we consider solution of the equation

$$\dot{x} = A(x)x, \ \ x(0) = x_0$$

and show that it is given by

$$x(t;x_0) = \exp[A(t;x_0)]x_0,$$

where $A(t;x_0) \in \mathscr{L}_A$ for each t, x_0; *i.e.* the solution can be regarded as an operation of the Lie group of the Lie algebra $\mathscr{L}_A$ as a transformation group on $\mathbb{R}^n$, since we have

$$\exp[A(t;x_0)]x_0 = \exp\{A(t_2; \exp[A(t_1;x_0)]\} \exp[A(t_1;x_0)]x_0, \text{ for } t = t_1 + t_2.$$

Theorem 10.3. *Consider the nonlinear system*

$$\dot{x} = A(x)x, \ \ x(0) = x_0 \in \mathbb{R}^n$$

where $A : \mathbb{R}^n \to \mathbb{R}^{n^2}$ is locally Lipschitz. Then the solution for each t (for which the solution exists) can be written in the form

$$x(t;x_0) = \exp[A(t;x_0)]x_0$$

where $A(t;x_0) \in \mathscr{L}_A$ for each t, x_0. Moreover we have

$$\exp[A(t;x_0)]x_0 = \exp\{A(t_2; \exp[A(t_1;x_0)]\} \exp[A(t_1;x_0)]x_0, \textit{ for } t = t_1 + t_2.$$

Proof. Since the last equation is obvious from the group property of the solutions of differential equations, we need only prove the first part. As we know from Chapter 2, we can replace the system (on any compact time interval on which the solutions exist) by a sequence of linear, time-varying approximations

$$\begin{aligned} \dot{x}^{[1]}(t) &= A(x_0)x^{[1]}(t) \\ \dot{x}^{[i]}(t) &= A(x^{[i-1]}(t))x^{[i]}(t), \ \ x^{[i]}(0) = x_0, \ \ i \geq 2 \end{aligned}$$

the solutions of which converge uniformly on compact time intervals on which the nonlinear system has a solution. Now recall that for any time-varying system

$$\dot{x} = B(t)x, \;\; x(0) = x_0,$$

where $B(\cdot) : \mathbb{R} \to \mathbb{R}^{n^2}$ is continuous, we have

$$\begin{aligned} x(t) \; = \; \lim_{n\to\infty} \Big\{ & \exp\left[\frac{t}{n}B\left((n-1)\frac{t}{n}\right)\right] \exp\left[\frac{t}{n}B\left((n-2)\frac{t}{n}\right)\right] \cdots \\ & \exp\left[\frac{t}{n}B\left(\frac{t}{n}\right)\right] \exp\left(B(0)\right) x_0 \Big\} \end{aligned}$$

(see Chapter 3). Applying this to each term in the sequence $x^{[i]}$ above and taking a diagonal subsequence, the result follows from the Campbell-Hausdorff formula, which we recall says that if A and B are sufficiently close to 0, then $C = \ln(\exp A \exp B)$ is given by

$$C = B + \int_1^0 g[\exp(tA\, dA)\exp(A\, dB)](A)dt$$

where

$$\begin{aligned} g(z) \; &= \; \frac{\ln z}{z-1} \\ &= 1 + \frac{1}{2}(1-z) + \frac{1}{3}(1-z)^2 + \cdots \\ &= \sum_{\ell=0}^{\infty} \frac{1}{\ell+1}(-1)^{\ell}(z-1)^{\ell}. \end{aligned}$$

□

Of course, this result states that the solution of an equation of the form (10.3) is given by

$$x(t; x_0) \; = \; \gamma(t; x_0)x_0$$

where $\gamma(t; x_0) = \exp[A(t; x_0)]$ is a smooth curve in the Lie group G_A of $\mathscr{L}_A$.

Example 10.2. Any system of the form

$$\dot{x} = \begin{pmatrix} a_{11}(x) & \cdots & a_{1n}(x) \\ \vdots & \ddots & \vdots \\ a_{n1}(x) & \cdots & a_{nn}(x) \end{pmatrix} x, \;\; x(0) = x_0$$

where $a_{ij}(x) = -a_{ji}(x)$. that is A is skew-symmetric, has a solution of the form

$$x(t; x_0) = O(t; x_0)x_0$$

where $O(t;x_0)$ is an orthogonal matrix for each t. Thus, all systems of this form generate rotations of x_0 for each t. Of course, in this case, this result follows also from the elementary fact that the norm $\|x(t;x_0)\|$ is invariant; for

$$\frac{d}{dt}\|x(t;x_0)\|^2 = \sum_{i=1}^{n} x_i\dot{x}_i = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij}(x)x_j = 0.$$

Example 10.3. For any system which satisfies $\sum_{i=1}^{n} a_{ii} = 0$, *i.e.* A has trace 0, the solution is of the form

$$x(t;x_0) = D(t;x_0)x_0$$

where $\det[D(t;x_0)] = 1$ for each t, since in this case the Lie algebra L_A is $\mathfrak{sl}(n)$ and G_A is $SL(n)$. For example, the equation

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},\ x(0) = \begin{pmatrix} x_{10} \\ x_{20} \end{pmatrix}$$

has solution

$$\begin{aligned} x_1(t) &= \exp(t)\exp\{[1-\exp(-t)]x_{20}x_{10}\} \\ x_2(t) &= \exp(-t)x_{20} \end{aligned}$$

so that

$$x(t) = \begin{pmatrix} \exp(t) & \exp(t)(\exp\{[1-\exp(-t)]x_{20}x_{10}\})\frac{x_{10}}{x_{20}} \\ 0 & \exp(-t) \end{pmatrix}\begin{pmatrix} x_{10} \\ x_{20} \end{pmatrix}.$$

Note that we have

$$\exp\{[1-\exp(-t)]x_{20}\} - 1)/x_{20} \to 1 - \exp(-t)$$

as $x_{20} \to 0$, and so this function is well-defined. Note however that, just as with $A(x)$, this representation of the solution is not unique.

Example 10.4. Let J denote the $2n \times 2n$ matrix

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

A *symplectic matrix* A is one which satisfies the relation

$$A^T JA = J.$$

Differentiating this gives the Lie algebra $\mathfrak{sp}(n)$ of infinitesimal symplectic matrices B which satisfy the equation

$$JB + B^T J = 0.$$

Thus any differential equation of the form

$$\dot{x} = \begin{pmatrix} A_{11}(x) & A_{12}(x) \\ A_{21}(x) & -A_{11}^T(x) \end{pmatrix}, \ x(0) = x_0, \ x_0 \in \mathbb{R}^{2n},$$

where $A_{12}(x), A_{21}(x) \in \mathbb{R}^{n^2}$ are symmetric, has solutions of the form

$$x(t;x_0) = \begin{pmatrix} B_{11}(t;x_0) & B_{12}(t;x_0) \\ B_{21}(t;x_0) & B_{22}(t;x_0) \end{pmatrix} x_0,$$

where

$$\begin{aligned} B_{11}^T B_{21} - B_{21}^T B_{11} &= 0, \\ B_{11}^T B_{22} - B_{21}^T B_{12} &= I \\ -B_{22}^T B_{21} + B_{12}^T B_{22} &= 0. \end{aligned}$$

10.4 Solvable Systems

In this section we shall generalise some results of Chapter 3 on solvable systems. Thus, consider again a system of the form (10.3) where the Lie algebra $\mathscr{L}_A$ of the system is solvable. Then all the matrices $A(x)$ can be put simultaneously in triangular form. Therefore we can write the system as

$$\dot{x} = \begin{pmatrix} a_{11}(x) & a_{12}(x) & \cdots & a_{1n}(x) \\ & a_{21}(x) & \cdots & a_{2n}(x) \\ & & \ddots & \vdots \\ & & & a_{nn}(x) \end{pmatrix} x.$$

From the results of Chapter 2 we know that the solution of this system through the initial point x_0 is given by the limit of the sequence of systems

$$\dot{x}^{[i]}(t) = \begin{pmatrix} a_{11}(x^{[i-1]}(t)) & a_{12}(x^{[i-1]}(t)) & \cdots & a_{1n}(x^{[i-1]}(t)) \\ & a_{21}(x^{[i-1]}(t)) & \cdots & a_{2n}(x^{[i-1]}(t)) \\ & & \ddots & \vdots \\ & & & a_{nn}(x^{[i-1]}(t)) \end{pmatrix} x^{[i]}(t), \ x^{[i]}(0) = x_0.$$

We can solve each of these upper triangular time-varying systems explicitly to obtain

$$x^{[i]}(t) = S(x^{[i-1]}(\cdot))(t), \ x^{[0]}(t) = x_0,$$

where

$$S(\xi(t)) = \begin{pmatrix} \sigma_1(\xi(t)) \\ \sigma_2(\xi(t)) \\ \vdots \\ \sigma_n(\xi(t)) \end{pmatrix}$$

and

$$\begin{aligned}
\sigma_n(\xi(t)) &= \exp\left(\int_0^t a_{nn}(\xi(s))ds\right)x_{0n}\\
\sigma_k(\xi(t)) &= \exp\left(\int_0^t a_{kk}(\xi(s))ds\right)x_{0k} + \int_0^t \sum_{\ell=k+1}^{n} a_{k\ell}(\xi(s))\sigma_\ell(\xi(s))\\
&\quad \times \exp\left(\exp\left(\int_s^t a_{kk}(\xi(\tau))d\tau\right)\right)ds,\\
&\quad n-1 \geq k \geq 1.
\end{aligned}$$

Hence, iterating this sequence gives the solution

$$x^{[i]}(t) = S^i(x^{[0]})(t).$$

We can obtain an explicit expression for $A(t;x_0)$ in this case:

$$\sigma_n(\xi(t)) = \exp\left(\int_0^t a_{nn}(\xi(s))ds\right)x_0 = \alpha_{nn}(\xi;0,t)x_{0n}, \text{ say,}$$

$$\begin{aligned}
\sigma_{n-1}(\xi(t)) &= \alpha_{n-1,n-1}(\xi;0,t)x_{0,n-1} +\\
&\qquad \int_0^t a_{n-1,n}(\xi(s))\alpha_{n-1,n-1}(\xi;s,t)x_{0n}ds\\
&= \alpha_{n-1,n-1}(\xi;0,t)x_{0,n-1} + \alpha_{n-1,n}(\xi;0,t)x_{0,n}, \text{ say,}\\
&\ \vdots
\end{aligned}$$

and so we can write

$$S(\xi(\cdot))(t) = \overline{A}(\xi;t)x_0,$$

where

$$\overline{A}(\xi;t) = \begin{pmatrix} \alpha_{1,1}(\xi;0,t) & \alpha_{1,2}(\xi;0,t) & \cdots & \alpha_{1,n}(\xi;0,t)\\ & \ddots & \vdots & \vdots\\ & & \alpha_{n-1,n-1}(\xi;0,t) & \alpha_{n-1,n}(\xi;0,t)\\ & & & \alpha_{n,n}(\xi;0,t) \end{pmatrix}$$

and therefore

$$A(t;x_0) = \lim_{i\to\infty}[S^i(x^{[1]}(\cdot))\overline{A}(x^{[0]};t)x_0].$$

This gives the following stability result, for example.

Theorem 10.4. *Let $K > 0, M > 0$ and suppose that*

$$a_{ii}(x) \leq -\varepsilon_i < 0,\ \ 1 \leq i \leq n$$

and put $\delta = \min_i(\varepsilon_i/2),\ \ \alpha = \min_i(\varepsilon_i - \delta)$. *Moreover, suppose that*

$$|a_{k\ell}| \le L,\ \ k \ne \ell,\ \ \|x\| \le K.$$

Then if
(a) $nL/\alpha < 1$
(b) $|x_{0k}| \le (1 - nL/\alpha)M,\ \ 1 \le k \le n$
(c) $M \le K/n^{1/2}$
the system is asymptotically stable in the ball $\{x : \|x\| \le K\}$.

Proof. We assume that $|\sigma_\ell(t)| \le N\exp(-\delta t)$, $k+1 \le \ell \le n$. This is certainly true for $\ell = n$ by the above assumptions. Also we have

$$\begin{aligned}
|\sigma_k(t)| &\le \exp(-\varepsilon_k t)|x_{0k}| + \int_0^t \sum_{\ell=k+1}^{n} |a_{k\ell}(\xi)||\sigma_\ell(s)| \exp[-(t-s)\varepsilon_k] ds \\
&\le \exp(-\varepsilon_k t)|x_{0k}| + LM(n-k)\exp(-\varepsilon_k t)\int_0^t \exp(-\delta s)\exp(s\varepsilon_k) ds \\
&\le LMn\exp(-\delta t)\exp[-t(\varepsilon_k - \delta)]\int_0^t \exp[(-\delta + \varepsilon_k)s] ds \\
&\le \exp(-\varepsilon_k t)|x_{0k}| + LMn\frac{\exp(-\delta t)}{\alpha} \\
&\le M\exp(-\delta t)
\end{aligned}$$

by the assumptions of the theorem. It follows that if we have

$$\|x^{[i-1]}(t)\| \le K$$

then we have

$$\|x^{[i]}(t)\| \le K$$

and in fact

$$\|x^{[i]}(t)\| \le K\exp(-\delta t).$$

However, the same argument shows that $\|x^{[0]}(t)\| \le K$ and so the result follows by induction and the convergence of the sequence $x^{[i]}(t)$. □

10.5 The Killing Form and Invariant Spaces

The Killing form of a Lie algebra is given by (see Appendix B) the symmetric bilinear form

$$(X, Y) = \mathrm{Tr}(\mathrm{ad}\, X, \mathrm{ad}\, Y).$$

Since the Killing form is important in determining the structure of semi-simple Lie algebras we consider next the determination of the Killing form of the Lie algebra $\mathscr{L}_a$ generated by the differential equation

$$\dot{x} = \left(\sum_{|\mathbf{i}|=0}^{\infty} A_{\mathbf{i}} x^{\mathbf{i}} \right) x.$$

Recall that this is the same as the Lie algebra $\mathscr{L}_{A_{\mathbf{i}}}$. We suppose that, as a vector space, dim $\mathscr{L}_A = M$. Since the dimension of the vector space spanned by $\{A_{\mathbf{i}}\}$ is no larger than M we suppose that its dimension is $K \le M$. Thus, there are K linearly independent matrices in the set $\{A_{\mathbf{i}}\}$ - denote them by $C_1, \cdots, C_K$. If $K = M$ then we have a basis of $\mathscr{L}_A$; if not, we can extend them to a basis of $\mathscr{L}_A$ by adding the matrices $C_{K+1}, \cdots, C_M$, where

$$C_j = [C_{j_1}, C_{j_2}], \quad K+1 \le j \le M,$$

for some $j_1, j_2 < j$. Let

$$[C_\alpha, C_\beta] = \sum_\gamma d^\gamma_{\alpha\beta} C_\gamma;$$

i.e. $d^\gamma_{\alpha\beta}$ are the structure constants of $\mathscr{L}_A$ with respect to the basis $C_1, \cdots, C_M$.

Then we have:

Lemma 10.2. *The Killing form of $\mathscr{L}_A$ is given by*

$$(X,Y) = \sum_i \sum_\ell \sum_k \sum_\gamma y_\ell x_k d^\gamma_{\ell i} d^i_{k\gamma},$$

where $X = \sum x_k C_k$, $Y = \sum y_\ell C_\ell$.

Proof. Suppose that

$$\left[\sum x_k C_k, \left(\sum y_\ell C_\ell, C_i \right) \right] = \sum_j a_{ij} C_j.$$

Then the left hand side equals

$$\begin{aligned} \sum_\ell \sum_k y_\ell x_k [C_k, [C_\ell, C_i]] &= \sum_\ell \sum_k y_\ell x_k \sum_\gamma d^\gamma_{\ell i} [C_k, C_\gamma] \\ &= \sum_\ell \sum_k \sum_\gamma y_\ell x_k d^\gamma_{\ell i} \sum_j d^j_{k\gamma} C_j \end{aligned}$$

so that

$$a_{ji} = \sum_\ell \sum_k \sum_\gamma y_\ell x_k d^\gamma_{\ell i} d^j_{k\gamma}$$

and the result follows. □

Corollary 10.1. *$\mathscr{L}_A$ is semi-simple if and only if the form*

$$((x_1,\cdots,x_M),(y_1,\cdots,y_M)=\sum_i\sum_\ell\sum_k\sum_\gamma y_\ell x_k d^\gamma_{\ell i}d^j_{k\gamma}$$

is non-degenerate.

Proof. This follow from the lemma and Cartan's criterion for semi-simplicity (see Appendix B). □

Example 10.5. Define the matrices

$$M_1=\begin{pmatrix}0&0&0\\0&0&-1\\0&1&0\end{pmatrix}$$
$$M_2=\begin{pmatrix}0&0&1\\0&0&0\\-1&0&0\end{pmatrix}$$
$$M_3=\begin{pmatrix}0&-1&0\\1&0&0\\0&0&0\end{pmatrix}.$$

The matrices are linearly independent and involutive in the sense that

$$[M_i,M_j]=M_k,$$

where $\{i,j,k\}$ is an even permutation of $\{1,2,3\}$. Hence we have dim $\mathscr{L}_{\{M_1,M_2,M_3\}}=3$ and the structure constants are non-zero only if i,j,k are distinct and hence a permutation of $\{1,2,3\}$. Clearly,

$$(X,Y)=-2(x_1y_1+x_2y_2+x_3y_3)$$

where

$$X=x_1M_1+x_2M_2+x_3M_3,\ \ Y=y_1M_1+y_2M_2+y_3M_3$$

and so $\mathscr{L}_{\{M_1,M_2,M_3\}}$ is semisimple (in fact, it is simple). Consider the system

$$\dot{x}=[f_1M_1+f_2M_2+f_3M_3]x.$$

By Theorem 10.3, this has a solution of the form

$$x(t)=\exp[A(t;x_0)]x_0,$$

where

$$A(t;x_0)=\sum_{i=1}^{3}\alpha_i(t,x_0)M_i\in\mathscr{L}_{\{M_1,M_2,M_3\}}$$

for some functions α_1, α_2, ,α_3. In fact, the system generates rotations as can be seen directly:

$$\begin{aligned}\frac{d}{dt}\|x(t)\|^2 &= 2\sum_{i=1}^{3} x_i\dot{x}_i \\ &= 2x_1(f_2x_3 - f_3x_2) + 2x_2(-f_1x_3 + f_3x_1) + 2x_3(f_1x_2 - f_2x_1) \\ &= 0.\end{aligned}$$

Hence, spheres are invariant for this dynamical system. Now suppose that $f = (f_1, f_2, f_3)$ is of the form

$$f = \left(\frac{\partial V}{\partial x_1}, \frac{\partial V}{\partial x_2}, \frac{\partial V}{\partial x_3}\right) = \mathrm{grad}V$$

for some function V. Then the equation becomes

$$\dot{x} = \sum_{i=1}^{3} \frac{\partial V}{\partial x_i} M_i x$$

and the level curves of V are also invariant, *i.e.* $dV/dt = 0$, as can easily be checked. Hence for this system, the trajectory starting at x_0 remains in the set

$$\{x : \|x\| = \|x_0\|, V(x) = V(x_0)\}.$$

Consider now the more general system of the form

$$\dot{y} = \sum_{i=1}^{3} h_i(y) E_i y,$$

where $[E_i, E_j] = E_k$ for an even permutation $\{i, j, k\}$ of $\{1, 2, 3\}$. Then

$$\mathscr{L}_{\{E_1,E_2,E_3\}} \approx \mathscr{L}_{\{M_1,M_2,M_3\}}$$

and so there exists P such that

$$E_i = P^{-1} M_i P, \;\; 1 \le i \le 3.$$

Hence if we define the new coordinates $x = Py$, then we have

$$\dot{x} = \sum_{i=1}^{3} h_i(P^{-1}x) M_i x$$

and if there exists a function $V(x)$ such that

$$(h_1(P^{-1}x), h_2(P^{-1}x), h_3(P^{-1}x)) = \left(\frac{\partial V}{\partial x_1}, \frac{\partial V}{\partial x_2}, \frac{\partial V}{\partial x_3}\right),$$

then the system will be invariant on the level curves $V(x) = V(x_0)$. For this we must have

$$\frac{\partial h_i(P^{-1}x)}{\partial x_j} = \frac{\partial h_j(P^{-1}x)}{\partial x_i}, \ \ i \neq j,$$

i.e.

$$\frac{\partial h_i(y)}{\partial y} P_j^{'} = \frac{\partial h_j(y)}{\partial y} P_i^{'}, \ \ i \neq j,$$

where $P_i^{'}$ is the i^{th} column of P^{-1}. Hence we have proved:

Theorem 10.5. *Given a system of the above form where $[E_i, E_j] = E_k$ for an even permutation $\{i, j, k\}$ of $\{1, 2, 3\}$, if the functions h_i satisfy the condition*

$$\frac{\partial h_i(y)}{\partial y} P_j^{'} = \frac{\partial h_j(y)}{\partial y} P_i^{'}, \ \ i \neq j, \tag{10.4}$$

where P is given by the condition

$$E_i = P^{-1} M_i P, \ \ 1 \leq i \leq 3,$$

then there exists a function $W(y)$ such that the trajectories of the system with initial state y_0 lie in the set

$$\{y : \|Py\| = \|Py_0\|, \ \ W(y) = W(y_0)\}.$$

Example 10.6. Consider the system

$$\begin{aligned}\dot{y} = &\left[h_1(y) \begin{pmatrix} 0 & -2/3 & 1/3 \\ 0 & -1 & 1 \\ 0 & -2 & 1 \end{pmatrix} + h_2(y) \begin{pmatrix} -1 & 0 & 2/3 \\ -3/2 & 0 & 1/2 \\ -3 & 0 & 1 \end{pmatrix} + \right. \\ &\left. h_3(y) \begin{pmatrix} 0 & 2/3 & -1/3 \\ -3/2 & 0 & 1/2 \\ 0 & 0 & 0 \end{pmatrix} \right] y \\ \doteq &\ (h_1(y)E_1 + h_2(y)E_2 + h_3(y)E_3]y.\end{aligned}$$

The matrix P given by

$$P = \begin{pmatrix} 3 & 0 & -1 \\ 0 & -2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

satisfies the condition

$$PE_iP^{-1} = M_i, \ \ \ 1 \leq i \leq 3$$

and the conditions in (10.4) become

$$\begin{aligned}
-\frac{1}{2}\frac{\partial h_1}{\partial y_2} &= \frac{1}{3}\frac{\partial h_2}{\partial y_1} \\
\frac{1}{3}\frac{\partial h_1}{\partial y_1}+\frac{1}{2}\frac{\partial h_1}{\partial y_2}+\frac{\partial h_1}{\partial y_3} &= \frac{1}{3}\frac{\partial h_3}{\partial y_1} \\
\frac{1}{3}\frac{\partial h_2}{\partial y_1}+\frac{1}{2}\frac{\partial h_2}{\partial y_2}+\frac{\partial h_2}{\partial y_3} &= -\frac{1}{2}\frac{\partial h_3}{\partial y_2}.
\end{aligned}$$

These equations are satisfied by the functions

$$\begin{aligned}
h_1(y) &= 6y_1-2y_3, \\
h_2(y) &= -12y_2+6y_3, \\
h_3(y) &= 6y_3^2,
\end{aligned}$$

as can be easily checked. Substituting $x=Py$, we obtain

$$\begin{aligned}
\frac{\partial V}{\partial x_1} &= 2x_1 \\
\frac{\partial V}{\partial x_2} &= 6x_2 \\
\frac{\partial V}{\partial x_3} &= 6x_3^2,
\end{aligned}$$

from which we see that V is given by

$$V(x)=x_1^2+3x_2^2+2x_3^3,$$

i.e.

$$\begin{aligned}
W(y) &= V(x) \\
&= V(Py) \\
&= (3y_1-y_3)^2+2y_3^3+3(-2y_2+y_3)^2,
\end{aligned}$$

and so for the system

$$\dot{y}=\left[(6y_1-2y_3)\begin{pmatrix} 0 & -2/3 & 1/3 \\ 0 & -1 & 1 \\ 0 & -2 & 1 \end{pmatrix}+(-12y_2+6y_3)\begin{pmatrix} -1 & 0 & 2/3 \\ -3/2 & 0 & 1/2 \\ -3 & 0 & 1 \end{pmatrix}+\right.$$
$$\left. 6y_3^2\begin{pmatrix} 0 & 2/3 & -1/3 \\ -3/2 & 0 & 1/2 \\ 0 & 0 & 0 \end{pmatrix}\right]y$$

i.e.

$$\begin{aligned}
\dot{y}_1 &= 8y_1y_2 - 4y_1y_3 - \left(\frac{20}{3}\right) y_2y_3 + 4y_2y_3^2 + \left(\frac{10}{3}\right) y_3^2 - 2y_3^3 \\
\dot{y}_2 &= 12y_1y_2 - 3y_1y_3 - 9y_1y_3^2 - 4y_2y_3 + y_3^2 + 3y_3^3 \\
\dot{y}_3 &= 24y_1y_2 - 12y_1y_3 - 8y_2y_3 + 4y_3^2,
\end{aligned}$$

the sets

$$\{y : (3y_1 - y_3)^2 + y_3^2 + (-4 + y_3)^2 = \text{constant}\}$$

and

$$\{y : (3y_1 - y_3)^2 + 2y_3^2 + 3(-2y_2 + y_3)^2 = \text{constant}\}$$

are invariant.

10.6 Compact Lie Algebras

We next consider systems which generate compact Lie algebras. For the general theory of compact Lie algebras, see Appendix B. As one would expect, compactness of a Lie algebra has consequences for stability and invariance. First we have

Lemma 10.3. *If the system*

$$\dot{x} = A(x)x \tag{10.5}$$

generates a compact Lie algebra $\mathscr{L}_A$ then it is stable. Moreover, the system

$$\dot{y} = -\alpha y + A(y)y \tag{10.6}$$

is asymptotically stable.

Proof. The Lie group G_A generated by $\mathscr{L}_A$ is compact and the solution of (10.5) is of the form

$$x(t) = \exp[A(t;x_0)]x_0$$

by Theorem 10.3, where $\exp[A(t;x_0)] \in G_A$ (and $A(t;x_0) \in \mathscr{L}_A$). Since G_A is compact,

$$\|x(t)\| \leq K\|x_0\|$$

for some K independent of x_0 and hence we have stability.

In (10.6), put

$$z = \exp(\alpha t)y.$$

Then

$$\dot{z} = A(\exp(-\alpha t)z)z$$

and A generates a compact Lie algebra, so the system is stable. Hence, $y = \exp(-\alpha t)z$ is asymptotically stable. □

Lemma 10.4. *Let A and B be two square matrices and let $K = ker\, B$. Suppose that $K_1 \subseteq K$ is the largest invariant subspace of K under A, i.e. $AK_1 \subseteq K_1$. Then*

$$\exp(A+B)x = (\exp A)x,$$

for al $x \in K_1$.

Proof. Use the power series expansion of $\exp(A+B)$ and note that, for any term in the expansion which contains at least one B, it must be zero when operating on $x \in K_1$ since K_1 is invariant under A. □

Theorem 10.6. *Consider the nonlinear differential equation*

$$\dot{x} = A(x)x \tag{10.7}$$

and suppose that $\{A(x)\}$ generates a semi-simple Lie algebra $\mathscr{L}_A$ which has a Cartan decomposition

$$\mathscr{L}_A = \mathfrak{t}_0 + \mathfrak{p}_0.$$

Let ker $\mathfrak{p}_0$ denote the set

$$ker\, \mathfrak{p}_0 = \cap\{ker\, B : B \in \mathfrak{p}_0\}$$

and let K be the largest invariant subspace of ker $\mathfrak{p}_0$ under $\mathfrak{t}_0$, i.e.

$$AK \subseteq K$$

for all $A \in \mathfrak{t}_0$. Then the solutions of (10.7) are stable in K and we can choose appropriate coordinates so that the equation has the form

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} A_1(y) & A_2(y) \\ & A_3(y) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

where $\{A_1(y)\}$ generates a compact Lie algebra and

$$\begin{pmatrix} y_1 \\ 0 \end{pmatrix}$$

parameterises K.

Proof. By Theorem 10.3 and the decomposition of $\mathscr{L}_A$, we may write the solution in the form

$$x(t) = \exp[A_1(t;x_0) + A_2(t;x_0)]x_0$$

where $A_1 \in \mathfrak{t}_0$ and $A_2 \in \mathfrak{p}_0$. If $x_0 \in K$ then, by Lemma 10.6, we have

$$x(t) = \exp[A_1(t;x_0)]x_0$$

and so stability follows from lemma 10.6, and the decomposition of the system follows by standard linear algebra. □

Note that we can obtain a similar result in the general case where $\mathscr{L}_A$ is not necessarily semi-simple. We simply write it in the form

$$\mathscr{L}_A = \mathfrak{g} + \mathfrak{s}$$

where $\mathfrak{g}$ is semisimple and $\mathfrak{s}$ is solvable. Then if $\mathfrak{g} = \mathfrak{k}_0 + \mathfrak{p}_0$ is a Cartan decomposition of $\mathfrak{g}$ we may replace $\mathfrak{p}_0$ by $\mathfrak{p}_0 + \mathfrak{s}$ in the theorem.

Example 10.7. Consider the system

$$\dot{x} = [x_1A_1 + x_1^2x_4A_2 + x_2^3A_3 + x_2x_3A_4 + x_1x_3x_4A_5]x$$

where

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 \end{pmatrix}$$

and

$$A_5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then,

$$\mathscr{L}_A = \mathscr{L}\{A_1, A_2, A_3, A_4, A_5\}$$

which has basis

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 1 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -2 & 2 & 0 & 0 & -2 \\ 0 & 0 & 0 & 2 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The Killing form is

$$K(\xi,\eta) = -6\xi_1\eta_1 - 6\xi_2\eta_2 - 6\xi_6\eta_6 - 56\xi_{10}\eta_{10} + 14\xi_4\eta_4 + 14\xi_5\eta_5,$$

as can be easily checked by using a computer algebra package. Since this is degenerate, the Lie algebra $\mathscr{L}_A$ of the system is not semi-simple so we cannot use Theorem 10.6 directly - we must use the remark following that theorem; *i.e.* we find a semi-simple subalgebra. By examining the structure constants we see that the subalgebra generated by the basis elements

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

is compact and is isomorphic to the Lie algebra $\{M_1, M_2, M_3\}$ above. By a change of coordinates, therefore, we can write these matrices in the form

$$\begin{pmatrix} 0 & 1 & 0 & \\ -1 & 0 & 0 & * \\ 0 & 0 & 0 & \\ & & * & * \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 & \\ 0 & 0 & 0 & * \\ -1 & 0 & 0 & \\ & & * & * \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & \\ 0 & 0 & 1 & * \\ 0 & -1 & 0 & \\ & & * & * \end{pmatrix}.$$

Such a map is given by $y = P^{-1}x$, where

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

The system then becomes

$$\dot{y}=[x_1B_1+x_1^2x_4B_2+x_2^3B_3+x_2x_3B_4+x_1x_3x_4B_5]y$$

where

$$B_1=\begin{pmatrix}0&1&0&0&0\\-1&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix},\quad B_2=\begin{pmatrix}0&0&0&0&0\\0&0&1&0&0\\0&-1&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix},$$

$$B_3=\begin{pmatrix}0&0&0&1&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix},\quad B_4=\begin{pmatrix}0&0&0&0&0\\0&0&0&0&1\\0&0&0&0&0\\0&0&0&-1&0\\0&0&0&0&1\end{pmatrix}$$

and

$$B_5=\begin{pmatrix}0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&1\\0&0&0&1&0\end{pmatrix},$$

i.e.

$$\dot{y}=\begin{pmatrix}0&y_3&0&(y_1+y_3)^3&0\\-y_3&0&y_3^2y_4&0&(y_1+y_3)(y_1+y_5)\\0&-y_3^2y_4&0&0&0\\0&0&0&-(y_1+y_3)(y_1+y_5)&y_2y_3y_4\\0&0&0&y_2y_3y_4&(y_1+y_3)(y_1+y_5)\end{pmatrix}y.$$

It follows that the sphere $y_1^2+y_2^2+y_3^2=$const., $y_4=y_5=0$ is invariant under the dynamics of this equation and so the sets

$$\{(x_1,x_2,x_3,x_4,x_5):\ 2x_1^2-2x_1x_2+x_2^2+x_3^2=\text{ constant},\\ x_4=0,\ x_1-x_2+x_5=0\}$$

are invariant for the original system, *i.e.* the system

$$
\begin{aligned}
\dot{x}_1 &= -x_1^2 x_3 x_4\\
\dot{x}_2 &= x_1 x_3 - x_1^2 x_3 x_4 + x_2^3 x_4\\
\dot{x}_3 &= x_1^2 - x_1 x_2 + x_1^3 x_4 + -x_2^2 x_5 + x_2 x_5^2\\
\dot{x}_4 &= -x_2 x_4 x_5 + x_1^2 x_3 x_4 - x_1 x_2 x_3 x_4 + x_1 x_3 x_4 x_5\\
\dot{x}_5 &= x_1 x_3 + x_2^3 x_4 + x_1 x_2 x_5 - x_2^2 x_5 + x_1 x_3 x_4^2.
\end{aligned}
$$

We see, therefore, that a careful study of the Lie algebra generated by a differential equation, involving a Cartan decomposition of the algebra can give some insight into the invariant sets of the system.

10.7 Modal Control

In the classical theory of control and servomechanisms, an effective technique for linear systems

$$\dot{x} = Ax + bu$$

is to diagonalise A (or reduce it to Jordan form) by changing the state variables to

$$y = P^{-1}x$$

so that, in the y-coordinates, we have

$$\dot{y} = \Lambda y + (P^{-1}b)u,$$

where

$$\Lambda = P^{-1}AP.$$

We can then choose the control u (if possible) in a simple way to stabilise the system. In this section, we generalise this approach to nonlinear systems of the form

$$\dot{x} = A(x)x + b(x)u$$

by using the above Lie algebraic methods. Thus, let $\mathscr{L}_A$ denote the Lie algebra generated by the system and let

$$\mathscr{L}_A = \mathfrak{s} + \mathfrak{g}$$

be a Levi decomposition of $\mathscr{L}_A$. Here, $\mathfrak{s}$ is the solvable part and $\mathfrak{g}$ is the semi-simple part of $\mathscr{L}_A$. (This is not unique, of course.) Now choose a Cartan decomposition

$$\mathfrak{g} = \mathfrak{h} + \sum_{\alpha\in\Sigma} \mathfrak{g}^{\alpha}$$

of $\mathscr{L}_A$, where $\mathfrak{h}$ is a Cartan subalgebra and the root spaces $\mathfrak{g}^{\alpha}$ are one-dimensional, where Σ is the set of non-zero roots. Thus, we can write the system in the form

$$\dot{x} = S(x)x + H(x)x + \sum_{\alpha \in \Sigma} e_\alpha(x) E_\alpha x + b(x)u$$

where $S(x)$ is upper triangularisable and $H(x)$ is diagonalisable (simultaneously, independent of x). Let P be an invertible matrix which diagonalises $H(x)$, *i.e.* if $y = P^{-1}x$, then

$$P^{-1}H(x)P = \Lambda(x) = \text{diag}\,(\lambda_1(x), \cdots, \lambda_n(x))$$

and the system becomes

$$\dot{y} = \Lambda(Py)y + R(y)y + P^{-1}b(Py)u, \tag{10.8}$$

where

$$R(y) = P^{-1}\left(S(Py) + \sum_{\alpha \in \Sigma} e_\alpha(y) E_\alpha\right)P.$$

Suppose that there exists a control $u = u(y)$ such that

$$\sum_{i=1}^{n} \lambda_i(Py)y_i^2 + y^T P^{-1} b(Py)u(y) \leq -\mu \|y\|^2$$

for some $\mu > 0$. Then we clearly have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|y\|^2 &= y^T\dot{y} \\ &= -\mu\|y\|^2 + y^T R(y)y, \end{aligned}$$

and so

$$\begin{aligned} \|y\|^2 &= \exp(-2\mu t)\|y_0\|^2 + \int_0^t 2\exp[-2\mu(t-s)]y^T R(y)y ds \\ &\leq \exp(-2\mu t)\|y_0\|^2 + \int_0^t 2\exp[-2\mu(t-s)]\|y(s)\|^2\|R(y(s))\|ds. \end{aligned}$$

If we assume that

$$\|R(y)\| \leq \lambda$$

for $y \in B_{0,\Delta} = \{x : \|x\| \leq \Delta\}$ then, by Gronwall's inequality, we have

$$\|y\|^2 \leq \exp[-2(\mu - \lambda)t]\|y_0\|^2$$

and so we have stability if $\lambda < \mu$. The original system will therefore be stable in the set

$$\{x : x^T(P^T)^{-1}P^{-1}x \leq \Delta\}.$$

and we have proved:

Theorem 10.7. *Consider the control system*

$$\dot{x} = A(x)x + b(x)u$$

and suppose that $\mathscr{L}_A = \mathfrak{s} + \mathfrak{g} = \mathfrak{s} + \mathfrak{h} + \sum_{\alpha \in \Sigma} \mathfrak{g}^{\alpha}$. *Then we can write the equation in the form*

$$\dot{x} = H(x)x + R(x)x + b(x)u$$

where $H(x) \in \mathfrak{h}$, $R(x) \in \mathfrak{s} + \sum_{\alpha \in \Sigma} \mathfrak{g}^{\alpha}$. *Suppose that* P *is a non-singular matrix which diagonalises* $H(x)$ *(independently of* x*) and that the pair* $(P^{-1}H(x)P, P^{-1}b(x))$ *is exponentially stabilisable in the sense that we can choose a control* $u = u(y)$ *so that (10.8) holds for some* $\mu > 0$. *Moreover, if*

$$\|P^{-1}R(Py)P\| \le \lambda$$

for $y \in B_{0,\Delta}$ *and* $\lambda < \mu$, *then the system is exponentially stabilisable in* $\{x : x^T(P^T)^{-1}P^{-1}x \le \Delta\}$ *and*

$$\|x\|^2 \le \exp[-2(\mu - \lambda)t]\|P\|^2\|x_0\|^2,$$

where $\|Px\|^2 \ge \rho^2\|x\|^2$.

Example 10.8. Consider the system

$$\begin{aligned}
\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} &= \begin{pmatrix} 3 - x_1^2 & -4 + x_2^2 & x_2^2 \\ x_2^2 & -1 - x_1^2 & x_2^2 \\ -5 - x_2^2 & 4 - x_2^2 & -2 - x_1^2 - x_2^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\
&\quad + (1 + x_1^2 \sin^2 x_3) \begin{pmatrix} 1/3 \\ -1/6 \\ -2/3 \end{pmatrix} u \\
&= \begin{pmatrix} 3 - x_1^2 & -4 & 0 \\ 0 & -1 - x_1^2 & 0 \\ -5 & 4 & -2 - x_1^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\
&\quad + x_2^3 \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\
&\quad + (1 + x_1^2 \sin^2 x_3) \begin{pmatrix} 1/3 \\ -1/6 \\ -2/3 \end{pmatrix} u.
\end{aligned}$$

The matrix

$$P = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ -1 & 1 & -1 \end{pmatrix}$$

diagonalises the first matrix on the right hand side of the equation, so that if $y = P^{-1}x$ then we have

$$\dot{y} = \begin{pmatrix} -1-x_1^2 & 0 & 0 \\ 0 & -2-x_1^2 & 0 \\ 0 & 0 & 3-x_1^2 \end{pmatrix} y + x_2^2 \begin{pmatrix} 0\ 1\ 0 \\ 0\ 0\ 0 \\ 1\ 0\ 0 \end{pmatrix} y$$
$$+(1+x_1^2\sin^2 x_3)P^{-1}\begin{pmatrix} 1/3 \\ -1/6 \\ -2/3 \end{pmatrix} u.$$

The Lie algebra of this system is $\mathscr{L}_A = \mathfrak{s} + \mathfrak{g}$ where $\mathfrak{g} = A_2$ and $\mathfrak{s}$ is generated by

$$\begin{pmatrix} 0\ 0\ 0 \\ 0\ 0\ 0 \\ 0\ 0\ 3 \end{pmatrix} x_1^2.$$

However, it is better to combine this part into the Cartan subalgebra generated by

$$\begin{pmatrix} -1-x_1^2 & 0 & 0 \\ 0 & -2-x_1^2 & 0 \\ 0 & 0 & 3+2x_1^2 \end{pmatrix}$$

since the combined system is stable everywhere. If we choose the control

$$\begin{aligned} u &= \frac{1}{1+x_1^2\sin^2 x_3}\begin{pmatrix} -12\ 12\ 0 \end{pmatrix} x \\ &= \frac{12}{1+x_1^2\sin^2 x_3}\begin{pmatrix} -1\ 1\ 0 \end{pmatrix} Py, \end{aligned}$$

then we obtain the equation

$$\dot{y} = \begin{pmatrix} -1-x_1^2 & 0 & 0 \\ 0 & -2-x_1^2 & 0 \\ 0 & 0 & -3-x_1^2 \end{pmatrix} y + y.$$

In this case we have $\mu = 1$ and

$$\left\| x_2^2 \begin{pmatrix} 0\ 1\ 0 \\ 0\ 0\ 0 \\ 1\ 0\ 0 \end{pmatrix} \right\| = x_2^2$$

and the system is stable in the region $\|x\|^2 \le \delta < 1$ for some δ. By choosing the control to move the poles of the Cartan subalgebra further into the left-half plane, we can obtain a larger region.

10.8 Conclusions

In this chapter we have studied systems by using the theory of Lie algebras and their decompositions. We have seen that the Lie algebra of a system generates a Lie group which operates as a transformation group on the state-space, so that the solutions are represented by continuous curves in the Lie group operating on the initial condition. Thus the system solutions have the properties of the Lie group so that studying the Lie algebra of the system can give some insight into the invariant sets of solutions. Compactness of the Lie algebra (or group) is strongly connected with the stability of the system and we can use Cartan decompositions of the system Lie algebra to obtain a generalization of modal control to nonlinear systems. We can also use the theory to study chaos in high-dimensional systems (see [3]).

References

1. Carter, R.: Lie Algebras of Finite and Affine Type. Cam. Univ.Press, Cambridge (2005)
2. Jacobson, N.: Lie Algebras. Interscience Tracts in Pure and Applied Mathematics, vol. 10. Wiley, Chichester (1962)
3. Banks, S.P., McCaffrey, D.: Lie Algebras, Structure of Nonlinear Systems and Chaotic Motion. Int. J. Bif. and Chaos 32, 157–174 (2001)

Chapter 11
Global Analysis on Manifolds

11.1 Introduction

In this chapter we shall consider the non-local theory of systems – *i.e.* the theory of sections of the tangent bundle of a differentiable manifold which are called *vector fields*. Most control systems are described in terms of local operating points, *i.e.* they are linearised about some equilibrium point and then local feedback control is applied to hold this system 'near' this point. (Such is the case, for example, with aircraft systems, where the operating point is called a 'trim condition'.) Thus, suppose that

$$\dot{x} = f(x,u) \tag{11.1}$$

is some local representation of some system on a manifold and suppose that (x_d, u_d) is an equilibrium point, *i.e.*

$$f(x_d, u_d) = 0.$$

Then the control u_d keeps the system at the point x_d. Let

$$y = x - x_d, \;\; v = u - u_d.$$

Then, by Taylor's theorem, we have

$$\dot{x} \cong f(x_d, u_d) + \frac{\partial}{\partial x} f(x_d, u_d)(x - x_d) + \frac{\partial}{\partial u} f(x_d, u_d)(u - u_d) + \cdots,$$

i.e.

$$\dot{y} = Ay + Bv \tag{11.2}$$

where

$$A = \frac{\partial}{\partial x} f(x_d, u_d), \;\; \frac{\partial}{\partial u} f(x_d, u_d).$$

This is a local linear approximation to (11.2) 'near' the operating point (x_d, u_d). The system (11.2) is then controlled by standard linear techniques.

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 195–217.
springerlink.com

There are various questions which naturally arise from the above considerations. Firstly, given a number of local versions of the system (*i.e.* operating points), how do we switch smoothly between these operating values? Secondly, given a number of local representations of a system on a manifold, how do we reconstruct the manifold and the global vector field from the local models? We shall tackle these problems in this chapter and show how to obtain information about the underlying state-space (*e.g.* its topology and differentiable structure) from the local models.

11.2 Dynamical Systems on Manifolds

Let X be an n-dimensional differentiable manifold and let U be an m-dimensional vector space, regarded as a differentiable manifold in the obvious way. Then we form the product manifold $X \oplus U$ and let $T(X \oplus U \cong T(X) \oplus T(U) \cong T(X) \oplus U$ be its tangent bundle. Let

$$P : T(X) \oplus U \to T(X)$$

be the projection. Then a *control system* (with controls U) on X is a section of the bundle $T(X \oplus U)$ followed by P. (See Figure 11.1.)

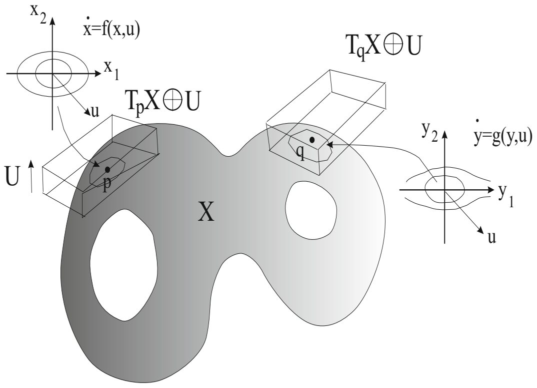


Fig. 11.1 Global control systems

Thus, in local coordinates x, a control system has the form

$$\dot{x} = f(x, u). \tag{11.3}$$

We shall assume that the system has a local operating point (x_d, u_d) at $x_d \in X$, *i.e.* for certain $x_d \in X$, there exists a control u_d such that

$$f(x_d, u_d) = 0.$$

Then, as seen above, at $x_d \in X$, we have a local linear approximation

$$\dot{x} = A_{(x_d,u_d)}x + B_{(x_d,u_d)}u \tag{11.4}$$

in terms of the local coordinate.

11.3 Local Reconstruction of Systems

In this section we shall consider the problem of the local reconstruction of a system of the form (11.3) from a number of local systems (11.4). Suppose we have some open subset $\mathscr{O}$ of $\mathbb{R}^{n+m}$ together with a finite open covering $\{O_i\}$ of $\mathscr{O}$ such that a linear system of the form (11.4) is defined on each O_i. We write the i^{th} linear system as

$$\dot{x}_{(i)} = A_{(x_{d_i},u_{d_i})}x_{(i)} + B_{(x_{d_i},u_{d_i})}u_{(i)}, \tag{11.5}$$

where $(x_{d_i}, u_{d_i}) \in O_i,\ \ 1 \le i \le K$. We shall assume that the function $f(x,u)$ has a polynomial approximation:

$$f(x,u) = \sum_{|\mathbf{i}|=0,|\mathbf{j}|=0}^{K} \alpha_{\mathbf{ij}} x^{\mathbf{i}} u^{\mathbf{j}}$$

for some K, where

$$x^{\mathbf{i}} = x_1^{i_1} \cdots x_n^{i_n},\ \ u^{\mathbf{j}} = u_1^{j_1} \cdots u_m^{j_m}$$

and $\alpha_{\mathbf{ij}}$ is an n-vector for each index $\mathbf{i}, \mathbf{j}$. From the ‘trim’ conditions for each point (x_{d_ℓ}, u_{d_ℓ}), we have

$$0 = f(x_{d_\ell}, u_{d_\ell}) = \sum_{|\mathbf{i}|=0,|\mathbf{j}|=0}^{K} \alpha_{\mathbf{ij}} x_{d_\ell}^{\mathbf{i}} u_{d_\ell}^{\mathbf{j}},\ \ 1 \le \ell \le K. \tag{11.6}$$

Also, we know that

$$\begin{aligned} A_{(x_{d_\ell},u_{d_\ell})} &= \frac{\partial f}{\partial x}(x_{d_\ell}, u_{d_\ell}) \\ B_{(x_{d_\ell},u_{d_\ell})} &= \frac{\partial f}{\partial u}(x_{d_\ell}, u_{d_\ell}) \end{aligned}$$

and so if we write

$$\begin{aligned} A_{(x_{d_\ell},u_{d_\ell})} &= (a^{\ell}_{\mu\nu}),\ \ B_{(x_{d_\ell},u_{d_\ell})} = (b^{\ell}_{\mu'\nu'}) \\ &\qquad (1 \le \mu, \nu \le n,\ \ 1 \le \mu' \le n,\ 1 \le \nu' \le m) \end{aligned}$$

then we have

$$a^{\ell}_{\mu\nu} = \frac{\partial f_\mu}{\partial x_\nu}(x_{d_\ell}, u_{d_\ell}) = \sum_{k=1}^{n} \sum_{|\mathbf{i}|=0,|\mathbf{j}|=0}^{K} i_k \alpha^{\mu}_{\mathbf{ij}} x_{\ell}^{\mathbf{i}-\mathbf{1}_k} u^{\mathbf{j}}_{d_\ell} \tag{11.7}$$

$$b^{\ell}_{\mu'\nu'} = \frac{\partial f_{\mu'}}{\partial x_{\nu'}}(x_{d_\ell}, u_{d_\ell}) = \sum_{k=1}^{n} \sum_{|\mathbf{i}|=0,|\mathbf{j}|=0}^{K} j_k \alpha^{\mu'}_{\mathbf{ij}} x^{\mathbf{i}}_{\ell} u^{\mathbf{j}-\mathbf{1}_k}_{d_\ell} \tag{11.8}$$

where

$$\mathbf{i}-\mathbf{1}_k = (i_1, \cdots, i_k - 1, \cdots, i_n), \ \ \mathbf{j}-\mathbf{1}_k = (j_1, \cdots, j_k - 1, \cdots, j_m).$$

Writing the unknowns $\alpha^{\mu}_{\mathbf{ij}}$ $(1 \le \mu \le n, 0 \le |\mathbf{i}|) \le K, \ 0 \le |\mathbf{j}|) \le K$ as a column vector Ξ, the Equations 11.6 – 11.8 give a linear equation of the form

$$M\Xi = N.$$

If M is invertible then the solution is

$$\Xi = M^{-1}N.$$

If M is not square then we can use the generalized inverse $(M^T M)^{-1} M^T$ if $M^T M$ is invertible. In general, it will be necessary to have the information contained in the local linear systems at an appropriate set of points, so that the above system is soluble.

Example 11.1. As a trivial example, suppose that we have just one linear approximation to a two-dimensional nonlinear system at the point $(x_{1d}, x_{2d}, u_d) = (2, 6, 2)$, where the linear approximation is

$$\dot{x} = \begin{pmatrix} -11 & 1 \\ -1 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u$$

and suppose we assume that the nonlinear system is of the form

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2) = a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_1^3 + a_5 u \\ \dot{x}_2 &= f_2(x_1, x_2) = a_6 x_1 + a_7 u. \end{aligned}$$

Then

$$\begin{aligned} A_{(x_d, u_d} &= \left.\begin{pmatrix} a_1 + 2a_3 x_1 + 3a_4 x_1^2 & a_2 \\ a_6 & a_7 \end{pmatrix}\right|_{(x_1,x_2)=(2,6)} = \begin{pmatrix} -11 & 1 \\ 1 & 0 \end{pmatrix} \\ B_{(x_d, u_d} &= \left.\begin{pmatrix} a_5 \\ a_7 \end{pmatrix}\right|_{(x_1,x_2)=(2,6)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} 0 &= a_1 x_{1d} + a_2 x_{2d} + a_3 x_{1d}^2 + a_4 x_{1d}^3 + a_5 u_d = 2a_1 + 6a_2 + 36a_3 + 8a_4 + 2a_5 \\ 0 &= a_6 x_{1d} + a_7 u_d = 2a_6 + 2a_7. \end{aligned}$$

We see that

$$a_1 = 1,\ a_2 = 1,\ a_3 = 0,\ a_4 = -1,\ a_5 = 0,\ a_6 = -1,\ a_7 = 1$$

so that the system is

$$\begin{aligned}\dot{x}_1 &= x_2 - x_1^3 + x_1\\ \dot{x}_2 &= -x_1 + u\end{aligned}$$

i.e. a controlled Van der Pol oscillator.

11.4 Smooth Transition Between Operating Conditions

Now that we have a method of joining together distinct operating conditions to obtain a nonlinear system

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

we can consider the question of smooth transition between such operating points. Thus, suppose $(x(1),u(1))$ and $(x(2),u(2))$ are operating points and let $t \to (x_d(t),u_d(t))$ be a differentiable function such that

$$(x_d(t_i),u_d(t_i)) = (x(i),u(i)),\ \ i = 1,2$$

and which satisfies the equation

$$\dot{x}_d(t) = f(x_d(t),u_d(t)),\ \ x_d(0) = x(1),\ x_d(T) = x(2), \tag{11.9}$$

i.e. there exists a control $u_d(\cdot)$ such that we can track the desired trajectory $x_d(\cdot)$. Hence u_d is an open-loop control and so we will require to control the system around $x_d(\cdot)$ to ensure good tracking. First consider the existence problem for $u_d(\cdot)$. The system (11.9) can be written in the integral form

$$x_d(t) = x(1) + \int_0^t f(x_d(s),u_d(s))ds.$$

For a fixed desired trajectory $t \to x_d(t)$, we consider the map $u \in C^0[0,T;\mathbb{R}^m] \to F(u) \in C^1[0,T;\mathbb{R}^n]$ given by

$$F(u)(t) = x_d(t) - x(1) - \int_0^t f(x_d(s),u(s))ds.$$

where we assume that $x_d, f(x_d(\cdot),u(\cdot)) \in C^1[0,T;\mathbb{R}^n]$ for all $u \in C^0[0,T;\mathbb{R}^m]$. The Fréchet derivative of F is given by

$$F'(0)(t) = -\int_0^t \frac{\partial f}{\partial u}(x_d(s),0)ds$$

and it is a linear operator acting as

$$F'(0)(u)(t) = -\int_0^t \frac{\partial f}{\partial u}(x_d(s),0)u(s)ds.$$

If $\frac{\partial f}{\partial u}(x_d(s),0)$ is non-singular along the trajectory x_d, then the operator $F'(0)$ is bounded and invertible, with inverse

$$\left[\frac{\partial f}{\partial u}(x_d(s),0)\right]^{-1}\frac{d}{dt}.$$

Note that the above reasoning is a special case of the implicit function theorem in a Hilbert (or Banach) space (see Appendix D). If the condition for the existence of a local trim controller is satisfied along the desired trajectory joining operating points, then we expand the system around this trajectory $(x_d(t),u_d(t))$. Thus we put

$$y(t) = x(t) - x_d(t), \;\; v(t) = u(t) - u_d(t).$$

Then,

$$\begin{aligned}\dot{y}(t) &= \dot{x}(t) - \dot{x}_d(t)\\ &= f(x(t),u(t)) - f(x_d(t),u_d(t))\\ &= g(y(t),v(t)),\end{aligned}$$

say, where

$$g(0,0) = 0.$$

Writing g in the form

$$g(y(t),v(t)) = A(y(t),v(t))y(t) + B(y(t),v(t))v(t)$$

we have the system

$$\dot{y}(t) = A(y(t),v(t))y(t) + B(y(t),v(t))v(t)$$

and so we can introduce the iteration scheme

$$\dot{y}^{[i]}(t) = A(y^{[i-1]}(t),v^{[i-1]}(t))y^{[i]}(t) + B(y^{[i-1]}(t),v^{[i-1]}(t))v^{[i]}(t).$$

These systems can be controlled by any of the methods studied earlier in the book an we will have smooth transfer from one operating condition to another.

The above theory has been presented for ‘local systems’ – *i.e.* ones of the form

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

written in terms of a local coordinate. Now suppose that we have a global system of the form defined in Section 11.2. If we two operating points as in Figure 11.2, then (assuming the manifold is totally geodesic for simplicity), we can choose a geodesic connecting P_1 to P_2 and cover it with a finite set of local coordinates. (See Figure 11.2.) In each local coordinate system, the system may be written in the form

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

and we may apply the above theory, provided the control keeps the trajectory in the same coordinate neighbourhood (See Figure 11.3.)

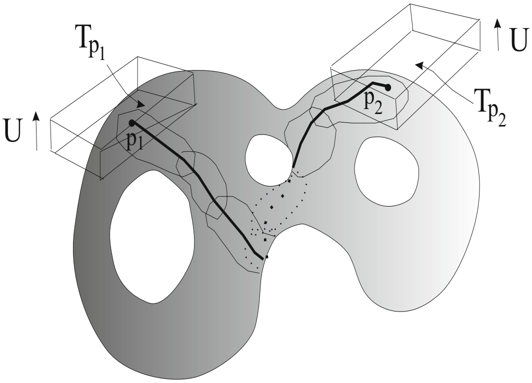


Fig. 11.2 Controlling along a geodesic

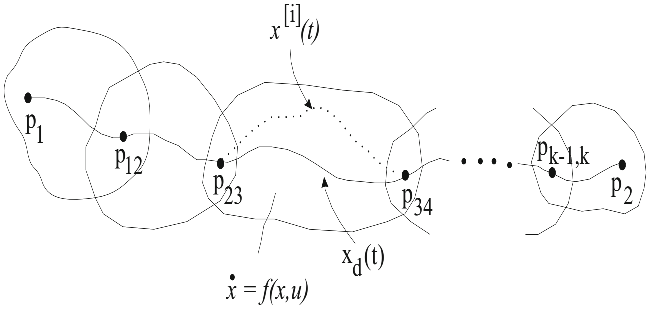


Fig. 11.3 Controlling to a desired trajectory

11.5 From Local to Global

We now ask the converse question to the ones above; namely, if we are given a number of local systems

$$\dot{x}_{(k)} = f_{(k)}x_{(k)}, u_{(k)}), \quad 1 \le k \le K \tag{11.10}$$

defined in local coordinate systems (on $\mathbb{R}^n \times \mathbb{R}^m$), how do we piece them together to form a (parameterised) vector field on a (compact) manifold M, and how do we get topological (and possibly differential) information about M? For this we shall need some algebraic topology and differential geometry, the background for which is given in Appendix C.

We shall assume that the local systems (11.10) are *complete* in the sense that there exists some differentiable (C^∞) compact manifold M covered by K coordinate patches (U_k, φ_k) with U_k open in M and $\varphi_k : U_k \to \mathbb{R}^n$ together with a parameterised vector field X_u, such that in U_k, X_u defines the local system in (11.10). To simplify matters we shall assume that the topology and differentiable structure of M can be determined entirely from the unforced systems. Hence we shall consider the local systems

$$\dot{x}_{(k)} = f_{(k)} x_{(k)}, 0), \quad 1 \le k \le K. \tag{11.11}$$

In many cases, the topology of M will be reflected in these unforced systems. However, if the control spaces are very 'twisted' then characteristic classes of the more general bundle $T(X \oplus U)$ may have to be considered.

In the first instance, let us consider the case of two-dimensional local systems. The two invariants of a compact surface are its genus and its orientability. In the case of orientable 2-manifolds S we have the Poincaré index theorem:

$$I_X = \chi_S \tag{11.12}$$

where X is a vector field on S, I_X is its index and χ_S is the Euler characteristic of the surface. Suppose each of the local systems (11.11) has only one fixed point of index I_k, $1 \le k \le K$. (Of course, some local systems may have no equilibria, in which case we take $I_k = 0$.) If we know that the local systems are such that their equilibria (if they exist) are not shared by any other local system, then from the index theorem (11.12) we can immediately say that the global system must be defined on a surface, which if it is orientable, has genus g given by

$$2(1-g) = \chi_S = I_X = \sum\{\text{indices of local systems}\},$$

i.e.

$$g = 1 - \frac{1}{2} \sum\{\text{indices of local systems}\}.$$

Since g is an integer, for orientable surfaces, we see immediately that

$$\sum\{\text{indices of local systems}\}$$

is an even integer. Hence if the sum of all the indices of our local systems is odd, we know that it cannot fit onto an orientable surface.

If the manifold is non-orientable, then (11.10) is still valid, but the Euler characteristic χ_S may not be even. For example, for the projective plane $S = \mathbb{P}^1(\mathbb{R})$ we have $\chi_S = 1$ and $g = 1/2$. (For example the system shown in Figure 11.4 on the unit disc has one saddle and two nodes, one stable and one unstable. By identifying antipodal points on the unit circle, we obtain a system on the projective plane with total index 1.)

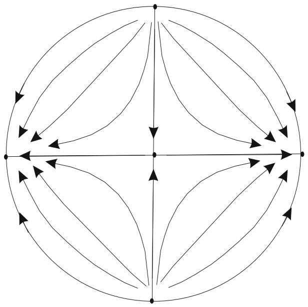


Fig. 11.4 A system on the projective plane

Of course, several local coordinate neighbourhoods may share some equilibria and so the best we can say is

$$2(1-g) = \chi_S = I_X = \sum\{\text{indices of non-equivalent local systems}\},$$

where we define local systems to be *equivalent* if they share an equilibrium point and can be pieced together with a local change of coordinates. In the next few sections we shall make some remarks about the possible structures of 2, 3 and 4-dimensional systems on compact manifolds which in many ways generalise the notion of index. We first need to recall some basic facts about Smale systems.

11.6 Smale Theory

In this section we outline Smale's theory of dynamical systems (see [35]) on manifolds and the notion of basic set. A continuous dynamical system is defined by an $\mathbb{R}$-action, on a compact manifold, called a flow, *i.e.*, a map $\Phi : \mathbb{R} \times M \to M$ such that:

(a) $\Phi(t,\cdot) : M \to M$ is a homeomorphism of M for all t

(b) $\Phi(0,\cdot) : M \to M$ is the identity on M

(c) $\Phi\big(t, \Phi(s,x)\big) = \Phi(t+s,x)$ for all $s,t \in \mathbb{R}$, $x \in M$.

We usually write $\phi_t(\cdot) = \Phi(t,\cdot)$.

A subset $M_1 \subseteq M$ is said to be *invariant* for the flow Φ if

$$\phi_t(M_1) \subseteq M_1 \quad \text{for all } t.$$

An invariant set M_1 is *hyperbolic* if there is a continuous ϕ_t invariant splitting of $TM_1 (= TM|_{M_1}$, the tangent bundle of M restricted to $M_1)$ given by

$$TM_1 = E^s_{M_1} \oplus E^u_{M_1} \oplus E^c_{M_1}$$

where

$$\begin{aligned}
&\|D\phi_t(v)\| \le Ce^{-t\lambda}\|v\| && \forall\, v \in E^s_{M_1}, t > 0 \\
&\|D\phi_{-t}(v)\| \le Ce^{-t\lambda}\|v\| && \forall\, v \in E^u_{M_1}, t > 0 \\
&\frac{d\phi_t^{(x)}}{dt}|_{t=0} \text{ spans } E^c_z \text{ for all } x \in M_1.
\end{aligned}$$

$E^c_{M_1}$ is just the space of all the orbits in the invariant set. Let $\widetilde{M} \subseteq M_1$ be a subset of a hyperbolic invariant set of a flow Φ on M. The local stable and unstable manifold of $\widetilde{M}$ are defined by

$$\begin{aligned}
W^s_{loc}(\widetilde{M}) = \{y \in M :\ & \lim_{t\to\infty} \|\phi_t(\widetilde{M}) - \phi_t(y)\| = 0, \\
& \text{and } \exists\, \varepsilon > 0 \text{ such that } \|\phi_t(\widetilde{M}) - \phi_t(y)\| < \varepsilon,\ \forall\, t \ge 0\}, \\
W^u_{loc}(\widetilde{M}) = \{y \in M :\ & \lim_{t\to-\infty} \|\phi_t(\widetilde{M}) - \phi_t(y)\| = 0, \\
& \text{and } \exists\, \varepsilon > 0 \text{ such that } \|\phi_t(\widetilde{M}) - \phi_t(y)\| < \varepsilon,\ \forall\, t \le 0\}.
\end{aligned}$$

The classical stable (unstable) manifold theorem (see [1]) then says that these local manifolds have global extensions. Given a flow ϕ_t on M we define a kind of recurrence in terms of the *chain-recurrent* set. Thus $x \in M$ is a chain-recurrent if for any $\varepsilon > 0\ \exists$ points $\{x_1, x_2, \cdots, x_{n-1}, x_n\}$ where $x_1 = x_n = x$ and positive real numbers $t_1, \cdots, t_{n-1}$ such that $\|\phi_{t_i}(x_i) - x_{i+1}\| < \varepsilon$, for $1 \le i \le n-1$. Then we have

Theorem 11.1. *If M is a compact orientable manifold and ϕ_t is a dynamical systems defined on M. Then if the chain recurrent set is hyperbolic, then it is the union of a finite number of disjoint basic sets, each of which is closed, invariant and contains a dense orbit. Moreover, the periodic points in the basic sets are dense in each such set.*

Similar results hold for discrete dynamical systems (*i.e.*, homeomorphisms of M). A variety of dynamical behaviours have been defined, largely with a view to studying structural stability (*i.e.*, the rigidity of the topology of the dynamics under 'small' perturbations in the vector field (or homeomorphism). Thus we define:

(a) An *Anosov system* on M is one which is hyperbolic everywhere on M.

(b) A flow ϕ_t is *Morse-Smale* if:

(i) the chain recurrent set is hyperbolic,
(ii) the stable and unstable manifolds of basic sets meet transversely,
(iii) each basic set is a single closed orbit or a field.

(c) a *Smale flow* ϕ_t on M is one for which:

(i) the chain recurrent set R is hyperbolic,

(ii) the basic subsets of the chain recurrent set are zero or one-dimensional,
(iii) the stable manifold of any orbit in R and the unstable manifold of any other orbit in R have transverse intersection.

The importance of Smale flows on compact manifolds is that they are structurally stable under C^1 perturbations. They are not dense, however, in the space of C^1 flows. The suspension of s Smale horseshoe at a saddle point is a Smale flow.

11.7 Two-dimensional Manifolds

We can find dynamical systems on surfaces of genus p by using the hyperbolic plane to ‘unfold’ the surface. Thus we have the following results (see [2]).

Theorem 11.2. *Suppose that we choose distinct points $\{p_1, \cdots, p_k\}$ on a torus such that there exists a set of points $\{q_1, \cdots, q_k\}$ in a fundamental parallelogram in $\mathbb{C}$ for which*

$$\left(\sum_{i=1}^{k_1} q_i + \sum_{i=k_1+1}^{k_2} \left(\frac{e_i}{2}+1\right) q_i - \sum_{i=k_2+1}^{k} \left(\frac{h_i}{2}\right) q_i\right) \qquad \mathit{mod}\ \Omega$$

where

$$k + \frac{1}{2} \sum_{i=k_1+1}^{k_2} e_i - \frac{1}{2} \sum_{i=k_2+1}^{k} h_i = 0$$

and Ω is the lattice

$$\Omega = \{k + li : k, l \in \mathbb{Z}\},$$

then there exists a dynamical system on the torus such that $p_1, \cdots, p_k$ are stable or unstable points, p_i has e_i elliptic sectors, $k_1 + 1 \le i \le k_2$ and p_i has h_i hyperbolic sectors for $k_2 + 1 \le i \le k$. Then the equation is given by

$$\dot{z} = E(z),$$

where E is an elliptic function.

To generalise this result to higher genus surfaces, we must use automorphic functions. let Γ be a Fuchsian group, *i.e.*, a discrete subgroup of $PSL(2, \mathbb{R})$ given by

$$\Gamma = \{T_0 = I, T_1, T_2, \cdots\},$$

and let

$$\theta_1(z) = \sum_{i=0}^{\infty} (c_i z + d_i)^{-2m} H_i\big(T_i(z)\big)$$

be a ‘generalised’ θ-series. Then we have

Theorem 11.3. *The function*

$$F(z) = \frac{\tilde{\theta}_2(z)}{\theta_1(z)}$$

satisfies

$$F(T_i z) = \frac{a_i d_i - b_i c_i}{(c_i z + d_i)^2} F(z)$$

for each i and defines a Γ-invariant vector field if $m \geq 3$. Functions of this type give rise to dynamical systems on the Riemann surface of the Fuchsian group Γ in the form

$$\dot{z} = F(z).$$

The existence of periodic cycles in a dynamical systems cannot be obtained from the *Poincaré-Hopf* theorem, but there are more general approaches using Morse theory and Floer homology which give a more general index theorem (see [3]). Moreover, any surface of genus $p > 0$ can carry an infinite number of knotted trajectories. However, we have the following result (see [4]).

Theorem 11.4. *A dynamical system on a orientable surface of genus p can carry at most p topologically distinct knot types as periodic solutions.*

Another simple way of generating systems on 2-manifolds is by the operation of *connected sum*. Given any two 2-manifolds S_1 and S_2, we define their connected sum $S_1 \# S_2$ as the 2-manifold obtained by cutting out disks from S_1 and S_2 and sewing together their boundaries (see Figure 11.5). Recall (see [1]) that for any two-dimensional polynomial vector field the equilibria have the general local form consisting of e elliptic sectors, h hyperbolic sectors and p parabolic sectors and the index of such a point is given by

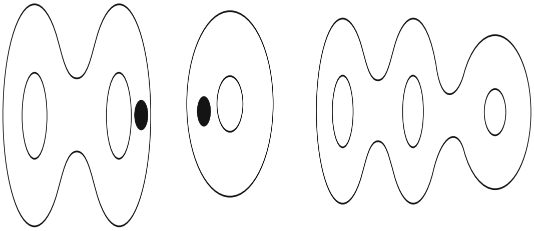


Fig. 11.5 Connected sum of S_1 and S_2

$$I = 1 + \left(\frac{e-h}{2}\right).$$

Suppose we have two 2-manifolds S_1 and S_2 on which there are defined dynamical systems and suppose one has an equilibrium point with the local structure consisting of e elliptic sectors, h hyperbolic sectors and p parabolic sectors and the other has n equilibrium point with the 'dual' structure. By this we mean that the point has h elliptic sectors, e hyperbolic sectors and p parabolic sectors arranged with the opposite orientation of the first point. Thus if the sectors of the first point are $\sigma_1, \sigma_2, \cdots, \sigma_k$ arranged anti-clockwise, then the sectors of the second point are $\bar{\sigma}_1, \bar{\sigma}_2, \cdots, \bar{\sigma}_k$ arranged clockwise, where

$$\bar{\sigma}_i = \begin{cases} \text{dual hyperbolic sector if } \sigma_i \text{ is elliptic} \\ \text{dual elliptic sector if } \sigma_i \text{ is hyperbolic} \\ \text{dual parabolic sector if } \sigma_i \text{ is parabolic} \end{cases}.$$

The dual sectors are shown in Figure 11.6 Then we have

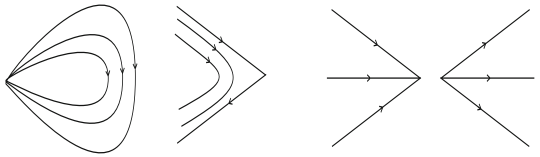


Fig. 11.6 Dual sectors: elliptic-hyperbolic and parabolic-parabolic

Theorem 1. *With the above notation we can form a dynamical system on $S_1\#S_2$ by introducing $e+h$ equilibrium points at the sew boundaries of the excised disks each of the form shown in Figure 11.7.*

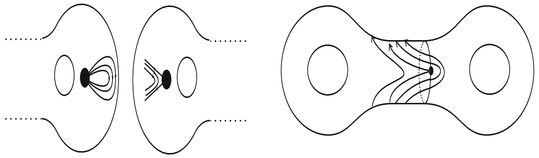


Fig. 11.7 Matching of the system dynamics after performing connected sum

Note that we could also match elliptic sectors to elliptic sectors and hyperbolic sectors to hyperbolic sectors by introducing saddle points for the hyperbolic points and the centres for the elliptic points. The topology of invariant sets on general manifolds is clearly important and the nature of nonlinear oscillations provides an insight into the global structure of invariant basic sets. In fact, in [5] the system

$$\ddot{x} + h(x)\dot{x} + g(t,x) = 0$$

is considered and it is shown that there exists an invariant set A which is not homeomorphic to a circle if the system contains an inversely unstable periodic solution.

(This means that the linearised Poincaré map near the solution has one unstable eigenvalue.) In [6] the following generalised version of this result is proved.

Theorem 11.5. *Given a system defined on a genus-p surface which is dissipative with respect to a knot K on this surface and suppose that there exists an inversely unstable solution within a knotted attractor A_I, then A_I is not homeomorphic to the circle.*

11.8 Three-dimensional Manifolds

The theory of dynamical systems on 3-manifolds is much more complicated than that for 2-manifolds, since there is no complete set of topological invariants for 3-manifolds (see [7]): although a great deal of invariants have been found, the most interesting being related to quantum groups and braided tensor categories (see [8,9]). Moreover, the *Euler* characteristic of a 3-manifold is 0 so the index theorem does not give too much information in this case. However, there are useful results in three-manifold theory which can be used to obtain a kind of decomposition of 3-dimensional dynamical systems in terms of simpler ones. These results are related to Dehn surgery, Heegaard splittings and branched covering manifolds, each of which we shall discuss below. We begin with the important result of [10] and [11].

Theorem 11.6. *Every closed, orientable, connected 3-manifold can be obtained by surgery on a link in S^3 (the 3-sphere).*

By a surgery on a knot K in S^3 we mean the following – cut out a tubular neighbourhood of K in S^3 (which is topologically a torus) and then glue it back in by some homeomorphism from the boundary of the excised torus to the boundary of the toroidal 'hole' in S^3. This leads to the possibility of defining dynamical systems on three-manifolds by first choosing one on S^3 which has periodic solutions defining some link in S^3. We then perform Dehn surgery on the link to obtain a dynamical system on the 3-manifold.

Example 11.2. From [12], we know that $3/4$-surgery on a trivial knot in S^3 yields the lens space $L(3,4)$ $(= L(3,1))$. Consider the non-singular Morse-Smale flow on S^3 with a Hopf link as periodic solutions shown in Figure 11.8.

Now do $3/4$-surgery on the trivial knot K_1 and we obtain a system on $L(3,4)$ with two periodic solutions such that the stable periodic solution is surrounded by stable solutions which wind around it three times, as shown in the right half of Figure 11.8.

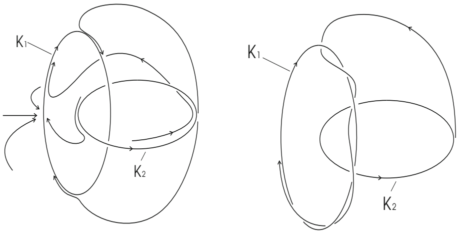


Fig. 11.8 3/4-surgery on a Morse-Smale flow

We therefore have:

Theorem 11.7. *There is a non-singular Morse-Smale system on any lens space.*

Morse-Smale diffeomorphisms of certain types on 3-manifolds are classified in [13] and Franks [14] shows how to generate non-singular Smale flows on S^3.

Next we consider the concept of Heegaard splittings (see [15]). An genus-p surface bounds a 3-manifold which is a ball with handles attached. A *Heegaard splitting* of a genus-p 3-manifold M is a pair (H_1, H_2) of such 'handle-bodies' of the same genus (and orientation) such that $M = H_1 \cup H_2$ and $H_1 \cap H_2 = \partial H_1 = \partial H_2$. It can be shown that every closed, connected 3-manifold has a Heegaard splitting. The *Heegaard diagram* of a Heegaard splitting $M = H_1 \cup H_2$ of genus-p is the surface ∂H_1 on which p distinct closed curves are drawn to which the fundamental meridians of ∂H_2 are attached. For example, the trefoil knot on the torus is a Heegaard diagram for the lens space $L(2,3)$, as in Figure 11.9 Clearly any dynamical system on a closed, compact 3-manifold M which has an invariant genus p-surface S that gives rise to a Heegaard splitting $M = H_1 \cup H_2$ where $H_1 \cap H_2 = S$ and to dynamical systems on H_1 and H_2. Conversely any two dynamical systems on solid genus-p handlebodies H_1, H_2 such that the induced dynamics on $\partial H_1 = \partial H_2$ are related by $\phi_t^{H_1} = \psi\left(\phi_t^{H_2}\right)$ for some diffoemorphism ψ define a dynamical system on $M = H_1 \cup H_2$. (see [16,17].) Handlebody decompositions are also important for

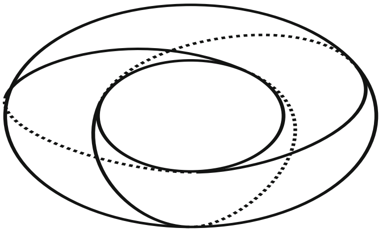


Fig. 11.9 Heegaard diagram for $L(2,3)$

systems containing basic sets which are solenoids. To define a solenoid, let M be a 3-manifold and let $T \subseteq M$ be a solid torus and let K be a knot (non-toroidal) in T which has a tubular neighbourhood K_ε (topologically a torus) also contained in T. Suppose that $f : M \to M$ is a diffeomorphism which maps T into K_ε, then the set

$$S = \bigcap_{k=1}^{\infty} f^k(T)$$

is called a contracting (Smale) solenoid. Similarly, if f^{-1} maps T onto K_ε then we get an expanding solenoid

$$S = \bigcap_{k=1}^{\infty} f^{-k}(T).$$

It can then be shown (see [18]) that the following statement are equivalent for a closed orientable 3-manifold:

(a) There exists a diffeomorphism $f : M \to M$ such that the non-wandering set $\Omega(f)$ contains a Smale solenoid.

(b) M has a lens space $L(p,q)$, $(p \neq 0, \pm 1)$ as a prime factor.

Moreover, the following two statements are also equivalent.

(a) There exists a diffeomorphism $f : M \to M$ whose non-wandering set is the union of finitely many solenoids (in fact, the only 2 solenoids).

(b) M is a lens space $L(p,q)$, $(p \neq 0)$.

By using higher genus Heegaard splittings and branched covering manifolds, the following generalisation of these statements can be proved (see [19]):

Theorem 2. *Any 3-manifold can carry a pair of generalized Smale solenoids of arbitrary genus.*

The above result is based on the covering theorem of Alexander and Montesinos [20]:

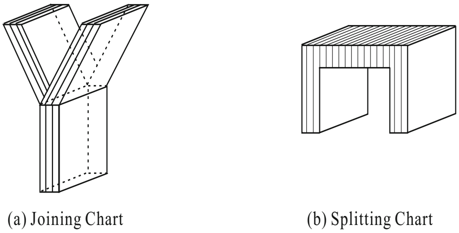
Theorem 3. *Any three manifold is a branched covering of S^3 branched over a ‘universal knot’.*

Applying this result to the theory of surface homeomorphisms of Thurston ([21]) gives the following result:

Theorem 11.8. *Every close, orientable 3-manifold M has a singular Anosov flow.*

Moreover, the singularities can be made to occur on the inverse of a figure-8 knot in the covering projection of M over S^3.

Another important aspect of global dynamical systems theory is that of branch manifolds and templates (Figure 11.10) on which knots ‘live’ (see [22]). A *branched*


(a) Joining Chart

(b) Splitting Chart

Fig. 11.10 Two types of chart

manifold in n-dimensions is a topological space for which every point has a neighbourhood which is homeomorphic to $\mathbb{R}^n$ or to a branched chart. A *three-dimensional template* is a compact branched 3-manifold with boundary and a smooth foliation made up of two types of charts: joining and splitting, as shown in the figure. A similar definition can be given for a two-dimensional template, this time the charts are oriented by the dynamics. An important type of template is the Lorenz template which 'sits' in the Lorenz system (Figure 11.11). It is well-known that this template carries any torus knot and that any link is a positive fibred braid (see [23]). Moreover, there are 'universal templates' which contain any knot or link. This follows from a study of the symbolic dynamics of the strips making up the template. A crucial part of this requires a proof of the fact that any non-trivial template contains a non-trivial knot. The proof in [22] requires inequality and this is difficult to generalise to higher-dimensional templates. However, in [24] a simple proof of this result given by showing that a very simple template consisting of just two neighbourhoods of the form in Figure 11.11 (plus 'twists') is a subtemplate of every template and these simple templates carry non-trivial knots. In [24] the theory is generalised to spun knots in $\mathbb{R}^4$.

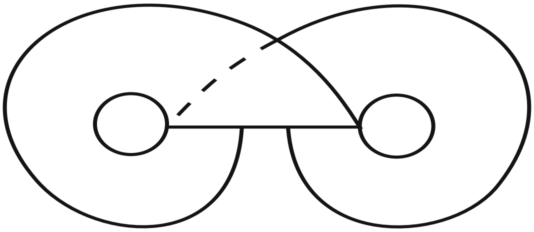


Fig. 11.11 The Lorenz template

11.9 Four-dimensional Manifolds

The topology and differentiable structure of 4-manifolds is probably the most difficult of any dimension and they occupy a unique position in that the topology of two-dimensional manifolds is well understood, that of three-dimensional manifolds is now well developed and high-dimensional manifolds can be effectively studied by the h-cobordism theorem (see [25]). 4-manifolds, on the other hand, exhibit somewhat more strange behaviour. In fact, $\mathbb{R}^4$ is the only Euclidean space to carry different differentiable structures ([26]). The existence of Smale, More-Smale and (pseudo) Anosov diffeomorphisms in the case of 4-manifolds then becomes more difficult because of the complex interaction of the topology and differentiable structures. The differentiable structure can be, to some extent, measured by some recent invariants discovered by Donaldson [27,28] and simplified in terms of Seiberg-Witten invariants. We first give an outline of this theory and then apply it to questions about dynamical systems on 4-manifolds (see [29]). Thus let M be a smooth 4-manifold and consider a real vector bundle E on M given by the transition function

$$g_{\alpha\beta} : U_\alpha \cap U_\beta \to GL(m, \mathbb{R}),$$

where $\{U_\alpha : \alpha \in A\}$ is an open covering of M. These functions satisfy the condition

$$g_{\alpha\beta} \cdot g_{\beta\gamma} = g_{\alpha\gamma} \qquad \text{on } U_\alpha \cap U_\beta \cap U_\gamma,$$

Let $\pi : E \to M$ be the projection map. (If we take $GL(m, \mathbb{C})$ or $GL(m, \mathbb{H})$ instead, we get complex or quaternionic vector bundles, if we take a Lie subgroup G of $GL(m, \cdot)$ we get a G-vector bundle, if $G = O(m)$ we can define a fibre metric $<,>_p$: $E_p \times E_p \to \mathbb{R}(\text{ or } \mathbb{C})$ where $E_p = \pi^{-1}(p)$, $p \in M$.) A section of a bundle (E, π) is a smooth map $\sigma : M \to E$ such that $\pi \circ \sigma = id$. The transition expressions for local sections is

$$\sigma_\alpha = g_{\alpha\beta}\sigma_\beta \quad \text{on } U_\alpha \cap U_\beta.$$

The topology of 4-manifolds is strongly connected with the space of connections in the tangent bundle. A *connection* on the vector bundles E is an map

$$d_A : \Gamma(E) \to \Gamma(T^*M \otimes E),$$

which satisfies

$$d_A(f\sigma + \tau) = (df) \otimes \sigma + f d_A\sigma + d_A\tau,$$

where σ, τ are sections of E and f is a function on M. Any connection d_A on a trivial bundle is of the form

$$d_A\sigma = d\sigma + \omega\sigma = (d + \omega)\sigma, \tag{11.13}$$

where d is the exterior derivative and ω is a matrix of one-forms. Thus we can think of a connection as a collection of differential operator $d+\omega\alpha$ which ω_α transforms according to

$$\omega_\alpha = g_{\alpha\beta} d g_{\alpha\beta}^{-1} + g_{\alpha\beta}\omega_\beta g_{\alpha\beta}^{-1} \qquad \text{on } U_\alpha \cap U_\beta .$$

Moreover, a connection on E is a linear map from the space of E-valued zero-forms to one-forms:

$$d_A : \Omega^0(E) \to \Omega^1(E)$$

and can be extended to $\Omega^p(M)$ by

$$d_A(\omega\otimes\sigma) = d\omega\otimes\sigma + (-1)^p \omega\wedge d_A\sigma, \qquad \text{for } \omega\in\Omega^p(M),\ \ \sigma\in\Gamma(E).$$

Then $(d_A)^2$ is a tensor field and is called the *curvature* of d_A. In the case of a trivial bundle we have

$$d_A^2(\sigma) = \Omega\sigma$$

where Ω is a matrix of 2-forms. Now

$$\begin{aligned} d_A^2(\sigma) &= (d+\omega_\alpha)(d\sigma+\omega\sigma) \\ &= (d\omega+\omega\wedge\omega)\sigma \end{aligned}$$

by (11.13), in any local coordinate system, so $\Omega = d\omega+\omega\wedge\omega$, which gives the Bianchi identity $d\Omega = [\Omega,\omega]$. If $\{U_\alpha : \alpha\in A\}$ is an open covering of M, then the local $\Omega' s$ transform as

$$\Omega_\alpha = g_{\alpha\beta}\Omega_\beta g_{\alpha\beta}^{-1}.$$

The local diffeomorphism forms $Tr\big[(\frac{i}{2\pi}\Omega_\alpha)^k\big]$, which gives rise to a global closed form and defines an element of $H^{2k}(M;\mathbb{R})$ which is independent of the (unitary) connection and Hermitian metric on E. It is called the characteristic class of E and is denoted by $\tau_k(E)$.

The Chern classes c_1, c_2 of E are defined by

$$c_1(E) = \tau_1(E) \ \text{ and } \ c_2(E) = \frac{1}{2}[\tau_1(E)^2 - \tau_2(E)].$$

By the Gauss-Bonnet theorem:

$$< c_1(TM), [M] >= \chi(M).$$

The Seiberg-Witten invariants are based on the Dirac operators given by

$$D_A(\psi) = \sum_{i=1}^{\infty} e_i \cdot d_A\psi(e_i)$$

where the e_i's are the standard Dirac matrices forming a basis of a Clifford algebra. The Weitzenbock formula

$$D_A^{\alpha}(\psi) = \Delta\psi + \frac{s}{4}\psi - \sum_{i<j} F_A(e_i,e_j)(ie_i \cdot e_j \cdot \psi)$$

where s is the scalar curvature of M and F_A is the curvature of the connection of a line bundle L. The important point is that D_A splits into two parts

$$D_A^{\pm} : \Gamma(W_{\pm} \otimes L) \to \Gamma(W_{\mp} \otimes L)$$

where $W_{\pm}$ are $U(2)$ bundles. Then the Seiberg-Witten equations are

$$D_A^{+}\psi = 0, \;\; F_A^{+} = \sigma(\psi) + \phi = -\frac{i}{2} < \psi, e_i \cdot e_j \cdot \psi > +\phi$$

where F_A^{+} is the self-dual part of F_A, σ is a quadratic form, and ϕ is a given self-dual 2-form. The solutions of these equations are pairs (d_A, ψ) consisting of a connection d_A on a line bundle L and a section ψ of $W_{+} \otimes L$.

We form the moduli space U_{ϕ} of all gauge-equivalent solutions of these equations – it turns out to have a compact closure which is a manifold. Then the Seiberg-Witten invariants of a line bundle L over a manifold M are defined by

$$SW(L) = < c_1^d, U_{\phi} >, \qquad d = \frac{1}{2}\dim U_{\phi}$$

and c_1 is the first Chern class of a line bundle over the space of all gauge equivalent pairs (d_A, ψ).

Using these invariants, it can be shown that the compact manifold

$$P^2\mathbb{C}\#_q\overline{P^2\mathbb{C}}$$

has infinitely many distinct smooth structures. Dynamical systems on 4-manifolds can be approached in similar ways to those for two- and three dimensions. Thus we can start with simple systems on S^4 and use connected sums, covering manifolds or blowing up techniques to generate more complex systems. For example, we have the following result.

Theorem 11.9. *Given a Morse-Smale system on a smooth simply connected 4-manifold X with $b^+ > 1$, SW invariant SW_X and an embedded periodic solution K with A-polynomial $P(t)$, then we can obtain a dynamical system on a 4-manifold X_P with SW-invariant SW_{X_P} where*

$$SW_{X_P} = SW_X \cdot P(t).$$

Proof. This follows from Fintushel and Stern [30] and Etgu [31], by doing Dehn surgery on the knot K. □

Note that in Gompf [32], the following definition of a generalised connected sum is given:

Let M and N be smooth, closed, oriented manifolds of dimension n and $n-2$, respectively and let $j_{1,2} : N \to M$ be disjoint embeddings with normal bundle ν_i over N and normal Euler classes $e(\nu_i) \in H^2(N;\mathbb{Z})$ which are opposite: $e(\nu_2) = -e(\nu_1)$. Then there is an orientation preserving diffeomorphism $\phi : V_1 - j_1(N) \to V_2 - j_2(N)$ for tubular neighbourhoods V_i of $j_i(N)$. Then we denote by $\#_\phi M$ to denote the manifold obtained by taking $M - \big(j_1(N) \cup j_2(N)\big)$ and identifying $V_1 - j_1(N)$ with $V_2 - j_2(N)$ by ϕ. If $M = M_1 \cup M_2$, then M_ϕ is the connected sum of M_1, M_2 along N via ϕ and is denoted by $M_1 \#_\phi M_2$.

Thus if $M = M_1 \cup M_2$ is a 4-manifold and M_1, M_2 carry dynamical systems with an invariant surface N such that the dynamics on $j_1(N)$ and $j_2(N)$ 'match up' by $\phi_* : T\big(j_1(N)\big) \to T\big(j_2(N)\big)$, then we obtain a well-defined system on $M_1 \#_\phi M_2$.

Finally, we can use the theory of covering surfaces to obtain dynamical systems on any 4-manifold. The theory of branched coverings of S^4 can be found in [33]. In particular, Piergallini [34] shows that every closed oriented PL 4-manifold is a single 4-fold covering of S^4 branched over a transversally immersed PL surface. The question of whether there exist 'universal surfaces' over which all coverings of S^4 branch seems to be open. We have:

Theorem 11.10. *A 4-manifold which is a simple 4-fold covering of S^4 over an immersed surface S carries a dynamical system which is the lift of a system of S^4, which has S as an invariant surface – so that the immersed double and treble points of S in S^4 are invariant sets of the flow.*

11.10 Conclusions

In this chapter we have discussed the nature of systems defined globally on manifolds and their relationship to local representations in terms of nonlinear differential equations. The main difficulty is in reconstructing a system from its local representatives and determining the topology and differentiable structure of the resulting manifold. In the case of surfaces, orientable or not, we can say a considerable amount from the indices of the local systems. However, when we come to higher-dimensional manifolds, we are faced with much greater difficulties. To emphasise the problems we have given an overview of the theory of dynamical systems on 2, 3 and 4-manifolds. The two- and three-dimensional cases are fundamentally different from the four-dimensional case since in the former two cases, the topological and differentiable structures are the same. However in the four-dimensional case, the fact that many different differentiable structures exists on the same topological manifold, leads to many different dynamical systems of various kinds. The classification of the structure of all four-dimensional dynamical systems appears to be very challenging. Putting together local systems in higher dimensions can only be approached at present by a consideration of the characteristic classes of the tangent bundle or some related bundle.

References

1. Perko, L.: Differential equations and dynamical systems. Springer, New York (1991)
2. Banks, S.P., Song, Y.: Elliptic and automorphic dynamical systems on surfaces. Int. J. of Bifurcation and Chaos 16(4), 911–923 (2006)
3. Banks, S.P.: Three-dimensional stratifications, knots and bifurcations of two-dimensional dynamical systems. Int. J. of Bifurcation and Chaos 12(1), 1–21 (2002)
4. Jost, J.: Dynamical systems - examples of conplex behavior. Springer, Berlin (2005)
5. Martins, R.: The effect of inversely unstable solutions on the attractor of the forced pendulum equation with friction. J. of Differential Equations 212(2), 351–365 (2005)
6. Song, Y., Banks, S.P.: Inversely Unstable Solutions of Two-Dimensional Systems on Genus-*p* Surfaces and the Topology of Knotted Attractors. Int. J. of Bifurcation and Chaos (in print)
7. Markov, A.A.: Insolubility of the problem of homeomorphy. Proc. Inc. Cong. Math., pp. 300–306. Cambridge University Press, Cambridge (1958)
8. Chari, V., Pressley, A.: A guide to quantum groups. CUP, Cambridge (1994)
9. Kassel, C.: Quantum groups. Springer, Heidelberg (1995)
10. Lickorish, W.B.R.: A representation of orientable combinatorial 3-manifolds. Ann. of Math. 76, 531–540 (1962)
11. Wallace, A.D.: Modifications and cobounding manifolds. Can. J. Math. 12, 503–528 (1960)
12. Rolfsen, D.: Knots and links. Publish or Perish, Berkeley (1976)
13. Wade, M.: Closed orbits of non-singular Morse-Smale flows on S^3'. J. Math. Soc. Jap. 41(3), 405–413 (1989)
14. Franks, J.: Non-singular Smale flows on S^3. Topology 24, 265–282 (1985)
15. Hempel, J.: 3-manifolds. Ann. of Math. Studies, vol. 86. Princeton University Press, Princeton (1976)
16. Song, Y., Banks, S.P., Diaz, D.: Dynamical Systems on Three Manifolds–Part I: Knots, Links and Chaos. Int. J. of Bifurcation and Chaos 17(6), 2073–2084 (2007)
17. Song, Y., Banks, S.P.: Dynamical Systems On Three Manifolds–Part II: 3-Manifolds, Heegaard Splittings and Three-Dimensional Systems. Int. J. of Bifurcation and Chaos 17(6), 2085–2095 (2007)
18. Jiang, B., Ni, Y., Wang, S.: 3-manifolds that admit knotted solenoids as attractors. Trans. AMS 356(11), 4371–4382 (2004)
19. Song, Y., Banks, S.P.: Generalized Smale solenoids on 3-manifolds (2009) (preprint)
20. Montesinos, J.: A representation of closed, orientable 3-manifolds as 3-fold branched coverings of S^3. Bull. Amer. Math. Soc. 80, 845–846 (1974)
21. Thurston, W.: On the geometry and dynamics of diffeomrphisms of surfaces. Bull. AMS 19, 417–431 (1988)
22. Ghrist, R.W., Holmes, P.J., Sullivan, M.C.: Knots and links in three-dimensional flows. LNM, vol. 1654. Springer, Heidelberg (1997)
23. Birman, J., Williams, R.: Knotted periodic orbits in dynamical systems - I: Lorenz's equations. Topology 22(1), 47–82 (1983)
24. Chen, W., Banks, S.P.: Branched manifolds, knotted surfaces and dynamical systems. Int. J. Bir. and Chaos (2008) (to appear)
25. Milnor, J.: Lectures on the h-cobordism theorem. Princeton Univ. Press, Princeton (1965)
26. Gompf, R.: Three exotic $\mathbb{R}^4$'s and other anomalies. J. Diff. Geom. 18, 317–328 (1983)
27. Donaldson, S.: An application of gauge theory to four-dimensional topology. J. Diff. Geom. 18, 279–315 (1988)

28. Donaldson, S.: Polynomial invariants for smooth 4-manifolds. Topology 29, 257–315 (1990)
29. Moore, J.D.: Lectures on Seiberg-Witten invariants. Springer, New York (1996)
30. Fintushel, R., Stern, R.J.: Knots, links and 4-manifolds. Invent. Math. 134, 363–400 (1998)
31. Etgu, T., Doug Park, B.: Non-isotopic symplectic tori in the same homology class. Trans. AMS 359(9), 3739–3750 (2003)
32. Gompf, R.: A new construction of symplectic manifolds. Ann. Math. 142(3), 527–595 (1995)
33. Montesinos, J.: A note on moves and irregular coverings of S^4. Contemp. Math. 44, 345–349 (1985)
34. Piergallini, R.: Four-manifolds as 4-fold branched covers of S^4. Topology 34(3), 497–508 (1995)
35. Smale, S.: Differentiable dynamical systems. Bull. AMS 73, 747–817 (1967)

Chapter 12
Summary, Conclusions and Prospects for Development

12.1 Introduction

In this book we have presented a theory which provides a general approach to nonlinear problems in systems theory. The method consists of writing a nonlinear system as the limit of a sequence of an approximating sequence of linear, time-varying ones and applying linear theory to each of the approximating systems. We have proved general convergence theorems and given applications to frequency-domain theory of nonlinear systems, optimal control, nonlinear sliding control and to nonlinear partial differential equations.

In this final chapter we shall show that there are many more potential applications of the method by using two illustrative examples which are now in the process of development. One is the application of the method to the problem of travelling waves in nonlinear partial differential equations and the other is to the separation theorem for nonlinear stochastic systems.

12.2 Travelling Wave Solutions in Nonlinear Lattice Differential Equations

We consider the existence of travelling wave solutions in lattice differential equations of the form

$$\dot{u}_{ij} = -\triangle u_{ij} + f(u_{ij})$$

where $\triangle$ is a two-dimensional difference operator. These solutions correspond to waves moving through the material to yield different structures on each side of the wave. These ideas can be applied to discrete Cahn-Hilliard and Cahn-Allen equations for the dynamical behaviour of binary alloys (see [1], [2]). A two-dimensional lattice differential equation on the lattice $\mathbb{Z}^2 \subseteq \mathbb{R}^2$ is an equation of the form

$$\dot{u}_{ij} = F(u_{i+k,j+k}, 0 \leq |K| \leq \ell)$$

M. Tomas-Rodriguez and S.P. Banks: Linear, Time-varying Approximations, LNCIS 400, pp. 219–228.
springerlink.com

for some finite natural number ℓ. We shall be interested in equations of the form

$$\dot{u}_{ij} = \lambda \triangle u_{ij} - f(u_{ij}), \ \ (i,j) \in \mathbb{Z}^2$$

where

$$\triangle u_{ij} = u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}$$

and λ is a real parameter.

12.3 Travelling Waves

A travelling wave is a solution of the form

$$u_{ij}(t) = \tau(iv_1 + jv_2 - ct),$$

where $\|(v_1,v_2)\| = 1$ (*i.e.* $(v_1,v_2) \in S$ is a unit vector). Thus τ satisfies the functional differential equation

$$\begin{aligned} -c\frac{d\tau}{dx} = \lambda(\tau(x+v_1) + \tau(x-v_1) + \\ \tau(x+v_2) + \tau(x-v_2) - 4\tau(x) - f(\tau(x)), \end{aligned}$$

where x is the wave parameter. This is a nonlinear functional differential equation of mixed type and as such is extremely difficult to solve. To simplify matters, we shall assume that v_1 and v_2 are commensurate, so that

$$v_1 = mh, \ \ v_2 = nh$$

for some integers m,n (where $m \geq n$) and real number h. We shall also assume that m,n are large, so that h is small relative to v_1 and v_2. Thus, we can approximate $d\tau/dx$ by $(\tau(x+h) - \tau(x))/h$. Thus, if $x = k - mh$, we have

$$\begin{aligned} -\frac{c}{h}(\tau(k-(m-1)h) - \tau(k-mh)) = \lambda(\tau(k) + \tau(k-2mh) + \\ \tau(k-(m-n)h) + \tau(k-(m+n)h) - 4\tau(k-mh) - f(\tau(k-mh)), \end{aligned}$$

i.e.

$$\begin{aligned} \tau(k) = & -\frac{c}{h\lambda}\tau(k-(m-1)h) + \frac{c}{h\lambda}\tau(k-mh) \\ & -\tau(k-2mh) - \tau(k-(m-n)h) - \tau(k-(m+n)h) + 4\tau(k-mh) \\ & +\frac{1}{\lambda}f(\tau(k-mh)) \end{aligned}$$

and so, if we define the state

$$\Gamma(k) = (\tau(k-(2m-1)h), \tau(k-(2m-2)h), \cdots, \tau(k))^T$$

we obtain the nonlinear difference equation

$$\Gamma(k) = A\Gamma(k-1) + F(\Gamma(k-1))$$

where

$$A = \begin{array}{c}\left(\begin{array}{ccccccccccccccccccc}
0 & 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\
0 & 0 & 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\
-1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & \frac{c}{h\lambda} & -\frac{c}{h\lambda} & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & &
\end{array}\right) \\
\begin{array}{ccccccccc}
\uparrow & & \uparrow & & \uparrow & \uparrow & & \uparrow & \\
1 & & 1+m-n & & 1+m & 2+m & & m+n-1 &
\end{array}\end{array}$$

and

$$F(\Gamma(k-1)) = \frac{1}{\lambda}\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ f(\Gamma_{m+1}(k-1)) \end{pmatrix}.$$

12.4 An Approach to the Solution

Consider the linear system

$$\widetilde{\Gamma}(k) = A\widetilde{\Gamma}(k-1), \;\; \widetilde{\Gamma}(0) = \Gamma_0.$$

The solution is

$$\widetilde{\Gamma}(k) = A^k\Gamma_0.$$

For the nonlinear problem we shall assume that F can be written in the form

$$F(\Gamma) = G(\Gamma)\Gamma.$$

Then we have

$$\begin{aligned}\Gamma(k) &= A(\Gamma(k-1)) + G(\Gamma(k-1))\Gamma(k-1) \\ &= (A + G\Gamma(k-1))\Gamma(k-1).\end{aligned}$$

In order to solve this equation we introduce the system of approximations $\Gamma^{[i]}(k)$ given by

$$\Gamma^{[i]}(k) = (A + G\Gamma^{[i-1]}(k-1))\Gamma^{[i]}(k-1)$$

with initial guess

$$\Gamma^{[0]}(k) = \Gamma_0.$$

These are linear, non-shift invariant systems which can be shown to converge to a solution of the nonlinear problem in the same way as in Chapter 2, provided that G is (locally) Lipschitz. This clearly gives a simple way to solve difficult mixed-type functional differential equations and other kinds of complex nonlinear problems.

12.5 A Separation Theorem for Nonlinear Systems

The theory of nonlinear optimal control was presented in Chapter 6. In this last section we shall show how to apply similar methods to the nonlinear separation problem. First recall that the linear separation theorem says that if we control a system with additive noise of the form

$$dX_t = (A_t X_t + B_t u_t)dt + \sigma_t d\mathscr{B}_t \ , \ \ t \geq s \ ; \ X_s = x$$

together with a cost functional

$$J^u(s,x) = E^{s,x}\left[\int_s^{t_1} \left\{ X_t^T Q_t X_t + u_t^T R_t u_t \right\} dt + X_{t_1}^T F X_{t_1} \right] \ , \ \ s \leq t_1,$$

where the matrices $A_t \in R^{n\times n}$, $B_t \in R^{n\times m}$, $\sigma_t \in R^{n\times k}$, $Q_t \in R^{n\times n}$, $R_t \in R^{m\times m}$, $F \in R^{n\times n}$ are continuous, and $\mathscr{B}_t$ is a standard Brownian motion, then (see [3]) we find that the optimal control is given by

$$u^*(t,X_t) = -R_t^{-1} B_t^T P_t X_t \ ,$$

where P_t satisfies the Riccati equation

$$\begin{aligned} &\frac{dP_t}{dt} = -A_t^T P_t - P_t A_t - Q_t + P_t B_t R_t^{-1} B_t^T P_t \ , \\ &P_{t_1} = F. \end{aligned}$$

If we do not have complete knowledge of X_t, but only a noisy observation

$$dZ_t = C_t X_t dt + \gamma_t d\widetilde{\mathscr{B}}_t$$

then the optimal control is given by

$$u^*(t,X_t) = -R_t^{-1} B_t^T P_t \widehat{X}_t(\omega)$$

where $\widehat{X}_t$ is the filtered estimate of X_t given by the Kalman-Bucy filter

$$\begin{aligned} d\widehat{X}_t &= \left(A_t - \widetilde{P}_t C_t^T (\gamma_t \gamma_t^T)^{-1} C_t\right) \widehat{X}_t dt + B_t u_t dt \\ &\quad + \widetilde{P}_t C_t^T (\gamma_t \gamma_t^T)^{-1} dZ_t \ ; \ \ \widehat{X}_0 = E[X_0] \end{aligned}$$

where $\widetilde{P}_t = E\left[(X_t - \widehat{X}_t)(X_t - \widehat{X}_t)^T\right]$ satisfies the Riccati equation

$$\frac{d\widetilde{P}_t}{dt} = A_t\widetilde{P}_t + \widetilde{P}_t A_t^T - \widetilde{P}_t C_t^T(\gamma_t\gamma_t^T)^{-1}C_t\widetilde{P}_t + \sigma_t\sigma_t^T \ ;$$
$$\widetilde{P}_0 = E\Big[\big(X_0 - E[X_0]\big)\big(X_0 - E[X_0]\big)^T\Big].$$

This is the content of the (linear) separation theorem, *i.e.* if we have a noisy system and noisy measurements, then we may separate out the filtering and the control, so that we may filter the output to obtain an estimate $\widehat{X}_t$ of the state and use that, instead of the unknown state $x(t)$ to obtain an optimal control.

There have been many attempts to solve the nonlinear separation problem (see, for example, [4,5,6,7,8]) using a variety of methods. Here we shall show that we can use our methods to give an effective solution to the problem (see [9] for more details). The problem we consider is given by the nonlinear stochastic equation

$$dX_t = \big(A_t(X_t)X_t + B_t(X_t,u_t)u_t\big)dt + \sigma_t d\mathscr{B}_t$$

together with the nonlinear measurement equation

$$dZ_t = C_t(X_t)X_t dt + \gamma_t d\widetilde{\mathscr{B}}_t$$

and the non-quadratic cost functional

$$J^u(s,x) = E^{s,x}\left[\int_s^{t_1}\Big\{X_t^T Q_t(X_t)X_t + u_t^T R_t(X_t)u_t\Big\}dt + X_{t_1}^T F X_{t_1}\right].$$

First we shall give the formal solution and then discuss the convergence of the approximations. The optimality of the solution is discussed in [9]. As in the case of linear, quadratic problems, we are lead to a system of approximations of the following form:

$$\begin{aligned} d\widehat{X}_t^{[i]} &= \Big(A_t\big(\widehat{X}_t^{[i-1]}\big)\widehat{X}_t^{[i]} + B_t\big(\widehat{X}_t^{[i-1]}, u_t^{[i-1]}\big)u_t^{[i]}\Big)dt \\ &\quad + \widetilde{P}_t^{[i]} C_t^T\big(\widehat{X}_t^{[i-1]}\big)\widetilde{R}_t^{-1}\Big[dZ_t - C_t\big(\widehat{X}_t^{[i-1]}\big)\widehat{X}_t^{[i]}\Big], \\ x^{[i]}(0) &= x_0 \end{aligned}$$

together with the Riccati filtering equation

$$\begin{aligned} \dot{\widetilde{P}}_t^{[i]} &= A_t\big(\widehat{X}_t^{[i-1]}\big)\widetilde{P}_t^{[i]} + \widetilde{P}_t^{[i]} A_t^T\big(\widehat{X}_t^{[i-1]}\big) + \widetilde{Q}_t - \widetilde{P}_t^{[i]} C_t^T \times \\ &\qquad \big(\widehat{X}_t^{[i-1]}\big)\widetilde{R}_t^{-1} C_t\big(\widehat{X}_t^{[i-1]}\big)\widetilde{P}_t^{[i]}, \\ \widetilde{P}^{[i]}(t_0) &= \widetilde{P}_0, \end{aligned}$$

the control sequence

$$u_t^{[i]} = -R_t^{-1}\big(\widehat{X}_t^{[i-1]}\big)B_t^T\big(\widehat{X}_t^{[i-1]}, u_t^{[i-1]}\big)P_t^{[i]}\widehat{X}_t^{[i]},$$

and the control Riccati equation

$$\begin{aligned}\dot{P}_t^{[i]} &= -A_t^T\big(\widehat{X}_t^{[i-1]}\big)P_t^{[i]} - P_t^{[i]}A_t\big(\widehat{X}_t^{[i-1]}\big)\\ &\quad -Q_t\big(\widehat{X}_t^{[i-1]}\big) + P_t^{[i]}B_t\big(\widehat{X}_t^{[i-1]}, u_t^{[i-1]}\big)\\ &\quad \times R_t^{-1}\big(\widehat{X}_t^{[i-1]}\big)B_t^T\big(\widehat{X}_t^{[i-1]}, u_t^{[i-1]}\big)P_t^{[i]}, P^{[i]}(t_f) = F\end{aligned}$$

where

$$\widetilde{R}_t = \gamma_t\gamma_t^T, \;\; \widetilde{Q}_t = \sigma_t\sigma_t^T .$$

If the sequence of functions $\left\{\widehat{X}_t^{[i]}, \widetilde{P}_t^{[i]}, u_t^{[i]}, P_t^{[i]}\right\}_{i\geq 1}$ converges in some sense, we denote the limit functions by $\left\{\widehat{X}_t^{[\infty]}, \widetilde{P}_t^{[\infty]}, u_t^{[\infty]}, P_t^{[\infty]}\right\}_{i\geq 1}$. The controlled dynamics then becomes

$$dX_t = \Big(A_t(X_t)X_t + B_t\big(X_t, u_t^{[\infty]}\big)u_t^{[\infty]}\Big)dt + \sigma_t d\mathscr{B}_t$$

and so we must decide in what sense does the sequence of systems

$$dX_t^{[1]} = \big(A_t(x_0)X_t^{[1]} + B_t(x_0, 0)u_t^{[1]}\big)dt + \sigma_t d\mathscr{B}_t^{[1]}$$

$$dX_t^{[i]} = \Big(A_t\big(X_t^{[i-1]}\big)X_t^{[i]} + B_t\big(X_t^{[i-1]}, u_t^{[i-1]}\big)u_t^{[i]}\Big)dt + \sigma_t d\mathscr{B}_t^{[i]}, \;\; i \geq 2$$

converge to the solution of the nonlinear problem. (Here $u_t^{[i]}$ can be chosen to be the optimal control of a standard linear regulator, and we assume that $\mathscr{B}_t^{[i]}, \; i \geq 1$ are independent Ito processes.) The sequence of functions The sequence of functions $\left\{\widehat{X}_t^{[i]}, \widetilde{P}_t^{[i]}, u_t^{[i]}, P_t^{[i]}\right\}_{i\geq 1}$ converges uniformly on $[0, t_f]$ by the standard theory of Chapter 2 and so we need only to consider the last two equations above, which we can write in the form

$$dX_t^{[1]} = \Big(A_t(x_0)X_t^{[1]} + V_t^{[1]}(x_0)\Big)dt + \sigma_t d\mathscr{B}_t^{[1]}$$

and

$$dX_t^{[i]} = \Big(A_t\big(X_t^{[i-1]}\big)X_t^{[i]} + V_t^{[i]}\big(X_t^{[i-1]}\big)\Big)dt + \sigma_t d\mathscr{B}_t^{[i]}, \;\; i \geq 2, \;\; X_0^{[i]} = X_0$$

where each $V_t^{[i]}(\cdot)$ is a (local) Lipschitz continuous function. From the standard theory of Ito stochastic differential equations we see that each of these equations has a unique solution ([10]). The (unique) solutions of these equations can be shown to be given by

$$\begin{aligned}X_t^{[1]} &= \Phi\Big(A_t(x_0), t\Big)\Big[X_0 + \Phi\big(-A_t(x_0), t\big)\sigma_t\mathscr{B}_t^{[1]}\Big] + \int_0^t \Phi\big(A_t(x_0), t-s\big)\\ &\quad \times\Big\{V_s^{[1]}(x_0) + A_s(x_0)\sigma_s\mathscr{B}_s^{[1]}\Big\}ds\end{aligned}$$

and

$$\begin{aligned}X_t^{[i]} &= \Phi\Big(A_t\big(X_t^{[i-1]}\big),t\Big)\Big[X_0+\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)\sigma_t\mathscr{B}_t^{[i]}\Big]\\&\quad+\int_0^t\Phi\Big(A_t\big(X_t^{[i-1]}\big),t-s\Big)\Big\{V_s^{[i]}\big(X_s^{[i-1]}\big)+A_s\big(X_s^{[i-1]}\big)\sigma_s\mathscr{B}_s^{[i]}\Big\}ds,\end{aligned}$$

(for $i\geq 2$), where $\Phi(A_t,t)$ is the fundamental matrix of A_t. To prove the convergence of the method we note that we have

$$\begin{aligned}\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)X_t^{[i]} &= \Big[X_0+\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)\sigma_t\mathscr{B}_t^{[i]}\Big]\\&\quad+\int_0^t\Phi\Big(-A_t\big(X_t^{[i-1]}\big),s\Big)\\&\quad\times\Big\{V_s^{[i]}\big(X_s^{[i-1]}\big)+A_s\big(X_s^{[i-1]}\big)\sigma_s\mathscr{B}_s^{[i]}\Big\}ds\end{aligned}$$

and

$$\begin{aligned}\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)X_t^{[i-1]} &= \Big[X_0+\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)\sigma_t\mathscr{B}_t^{[i-1]}\Big]\\&\quad+\int_0^t\Phi\Big(-A_t\big(X_t^{[i-2]}\big),s\Big)\\&\quad\times\Big\{V_s^{[i-1]}\big(X_s^{[i-2]}\big)+A_s\big(X_s^{[i-2]}\big)\sigma_s\mathscr{B}_s^{[i-1]}\Big\}ds\end{aligned}$$

and so

$$\begin{aligned}\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)X_t^{[i]} &-\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)X_t^{[i-1]}\\&=\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)\sigma_t\mathscr{B}_t^{[i]}\\&-\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)\sigma_t\mathscr{B}_t^{[i-1]}\\&+\int_0^t\Big\{\Phi\Big(-A_t\big(X_t^{[i-1]}\big),s\Big)\\&\times\Big\{V_s^{[i]}\big(X_s^{[i-1]}\big)+A_s\big(X_s^{[i-1]}\big)\sigma_s\mathscr{B}_s^{[i]}\Big\}\\&-\Phi\Big(-A_t\big(X_t^{[i-2]}\big),s\Big)\\&\times\Big\{V_s^{[i-1]}\big(X_s^{[i-2]}\big)+A_s\big(X_s^{[i-2]}\big)\sigma_s\mathscr{B}_s^{[i-1]}\Big\}\Big\}ds.\end{aligned}$$

If we denote the right hand side by Ψ, then we have

$$\begin{aligned}\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)X_t^{[i]} - \Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)X_t^{[i-1]} &= \Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)X_t^{[i-1]} \\ &\quad -\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)X_t^{[i-1]}+\Psi\end{aligned}$$

and so

$$\begin{aligned}X_t^{[i]}-X_t^{[i-1]} = \Phi\Big(A_t\big(X_t^{[i-1]}\big),t\Big)\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)X_t^{[i-1]} \\ -X_t^{[i-1]}+\Phi\Big(A_t\big(X_t^{[i-1]}\big),t\Big)\Psi.\end{aligned}$$

It follows that

$$\begin{aligned}E\Big(\|X_t^{[i]}-X_t^{[i-1]}\|^2\Big) \le E\Big(2\|\Phi\Big(A_t\big(X_t^{[i-1]}\big),t\Big) \\ \times\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)X_t^{[i-1]}-X_t^{[i-1]}\|^2\Big) \\ +2E\Big(\|\Phi\Big(A_t\big(X_t^{[i-1]}\big),t\Big)\Psi\|^2\Big)\end{aligned}$$

To estimate the first term on the right note that

$$\begin{aligned}\Phi\Big(A_t\big(X_t^{[i-1]}\big),t\Big)\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)X_t^{[i-1]}-X_t^{[i-1]} &= \Phi\Big(A_t\big(X_t^{[i-1]}\big),t\Big) \\ &\quad\times\Big[\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big) \\ &\quad-\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)\Big]X_t^{[i-1]}\end{aligned}$$

and this can be bounded by $\|X_t^{[i-1]}-X_t^{[i-2]}\|$ as in [11]. The second term in the inequality is bounded by

$$\begin{aligned}E(\|\Psi\|^2) \le 3E\Big(\|\Phi\Big(-A_t\big(X_t^{[i-1]}\big),t\Big)\sigma_t\mathscr{B}_t^{[i]}-\Phi\Big(-A_t\big(X_t^{[i-2]}\big),t\Big)\sigma_t\mathscr{B}_t^{[i-1]}\|^2\Big) \\ +3\int_0^t E\Big(\|\Phi\Big(-A_t\big(X_t^{[i-1]}\big),s\Big) \\ \times\Big\{V_s^{[i]}\big(X_s^{[i-1]}\big)+A_s\big(X_s^{[i-1]}\big)\sigma_s\mathscr{B}_s^{[i]}\Big\}\|^2\Big)ds \\ +3\int_0^t E\Big(\|\Phi\Big(A_t\big(X_t^{[i-2]}\big),s\Big)\Big\{V_s^{[i-1]}\big(X_s^{[i-2]}\big) \\ +A_s\big(X_s^{[i-2]}\big)\sigma_s\mathscr{B}_s^{[i-1]}\Big\}\|^2\Big)ds.\end{aligned}$$

Applying Ito's isometry (see [10]) to the second two terms and using the independence of the $\mathscr{B}^{[i]}$'s and $\mathscr{B}^{[i-1]}$'s, the proof of local and global convergence then follows as in [11]. For a discussion of the optimality of the process and a numerical example, see [9].

12.6 Conclusions

The two problems considered in this chapter show that the iteration method does have a wide variety of applications, particularly because the conditions for the applicability of the technique are very mild (local Lipschitz continuity of the system vector field), unlike many other nonlinear methods. The ideas also have applications to the control of chaos for secure communication (see [12]) and for the stabilisation of semiconductor lasers ([13]). By choosing the matrices Q and R in a cost functional to be dependent on the state x and the control u, it may also be possible to study hard constraint problems by this method. (This is the case, for example, when the control function u satisfies a pointwise bound of the form $|u| \leq u_{max}$, rather than an average energy bound as in the standard quadratic regulator problem.) Of course, this is the more practically useful form of control, since in most systems the control action is limited by some maximum value, such as the aileron deflection in an aircraft. The development of the control in this case is much more difficult than the linear, quadratic regulator case, since we must use the full Hamilton-Jacobi equations to solve the problem. If we can solve the problem by a sequence of linear, quadratic regulators then we could provide an easy classical solution to the hard constraint problem.

References

1. Allen, S.M., Cahn, J.W.: A Microscopic Theory for Antiphase Boundary Motion and its Application to Antiphase Domain Coursening. Acta. Met. 27, 1085–1095 (1979)
2. Cahn, J.W., Hilliard, J.E.: Free Energy of a Non-uniform Systems: I: Interfacial Free Energy. J. Chem. Phys. 28, 258–267 (1958)
3. Fleming, W.H., Rishel, R.W.: Deterministic and Stochastic Optimal Control. Springer, New York (1975)
4. Arslan, G., Basar, T.: Decentralized Risk-Sensitive Controller Design for Strict-Feedback Systems. Systems and Control Letters 50(5), 383–393 (2003)
5. Deng, H., Krstic, M.: Output-Feedback Stochastic Nonlinear Stabilisation. IEEE Transactions on Automatic Control 44(2), 328–333 (1999)
6. Germani, A., Manes, C., Palumbo, P.: Polynomial Extended Kalman Filter. IEEE Transactions on Automatic Control 50(12), 2059–2064 (2005)
7. Germani, A., Manes, C., Palumbo, P.: Filtering of Stochastic Nonlinear Differential Systems via a Carleman Approximation Approach. IEEE Transactions on Automatic Control 52(11), 2166–2172 (2007)
8. Kushner, H.J., Budhiraja, A.S.: A Nonlinear Filtering Algorithm Based on an Approximation of the Conditional Distribution. IEEE Transactions on Automatic Control 45(3), 580–585 (2000)

9. Kilicaslan, S., Banks, S.P.: A Separation Theorem for Nonlinear Systems. Automatica (to appear)
10. Oksendal, B.: Stochastic differential equations: An introduction with applications, (6th edn., Corrected 4th printing). Springer, New York (2007)
11. Tomás-Rodríguez, M., Banks, S.P.: Linear Approximations to Nonlinear Dynamical Systems with Applications to Stability and Spectral Theory. IMA Journal of Math. Control and Inf. 20, 89–103 (2003)
12. Hugues-Salas, O., Banks, S.P.: Control of Chaos for Secure Communication. Int. J. Bifur. and Chaos 18(11), 3355–3374 (2008)
13. Hugues-Salas, O., Shore, K.A., Banks, S.P.: Stabilisation of Chaotic Dynamics in Semiconductor Lasers with Optical Feedback using Optimal Control. IET Optoelectronics 2, 231–240 (2008)

Appendix A
Linear Algebra

A.1 Vector Spaces

We outline here briefly the basic ideas of linear algebra and spectral theory used in the book – more details can be found in many standard texts on vector spaces (see, for example, [1,2]). We begin with the definition of a vector space:

Definition A.1. A *vector space* $(V,+,\cdot)$ over a field $\mathbb{F}$ (here, as usual, $\mathbb{F}=\mathbb{R}$ or $\mathbb{C}$) is a set on which there are defined two operations of addition and scalar multiplication by elements of $\mathbb{F}$ which satisfy the following axioms:

(a) $(v+w)+x=v+(w+x)$, for all $v,w,x\in V$.
(b) $v+w=w+v$, for all $v,w\in V$.
(c) There exists an element 0 such that $0+w=w+0=w$ for all $w\in V$.
(d) For all $v\in V$, there exists an element $-v$ such that

$$v+(-v)=(-v)+v=0.$$

(e) $\lambda(v+w)=\lambda v+\lambda w$ for all $\lambda\in\mathbb{F}$ and $v,w\in V$.
(f) $(\lambda+\mu)v=\lambda v+\mu v$ for all $\lambda,\mu\in\mathbb{F}$ and $v\in V$.
(g) $(\lambda\mu)v=\lambda(\mu v)$ for all $\lambda,\mu\in\mathbb{F}$ and $v\in V$.
(h) $1\cdot v=v$ for all $v\in V$.

Axioms (a)–(d) say that $(V,+)$ is an abelian group, (e) and (f) are distributive laws, (g) is the commutativity of scalar multiplication and (h) says that $1\in\mathbb{F}$ behaves as an identity.

Example A.1. $\mathbb{R}^n$ and $\mathbb{C}^n$ with their usual structures.

Example A.2. $C[0,1]$ = set of functions $f:[0,1]\to\mathbb{R}$ which are continuous. The vector space operations are

$$\begin{aligned}(f+g)(t) &= f(t)+g(t), \text{for all } f,g \in C[0,1]\\ (\lambda f)(t) &= \lambda(f(t)), \text{for all } f \in C[0,1], \lambda \in \mathbb{R}\end{aligned}$$

Definition A.2. A *metric space* (M,δ) is a set M together with a metric (or distance function) δ which satisfies the axioms:
(a) $\delta(x,y) \geq 0$ for all $x,y \in M$.
(b) $\delta(x,y) = 0$ if and only if $x = y$.
(c) $\delta(x,y) = \delta(y,x)$ (symmetry).
(d) $\delta(x,y) \leq \delta(x,z)+\delta(z,y)$ (the triangle inequality).

Definition A.3. A *norm* on a vector space V is a function $f : \|\cdot\| : V \to \mathbb{R}^+$ such that
(N1) $\|v\| = 0$ if and only if $v = 0$.
(N2) $\|\alpha v\| = |\alpha| \, \|v\|$ for all $v \in V$ and $\alpha \in \mathbb{F}$.
(N3) $\|v+w\| \leq \|v\| + \|w\|$, for all $v,w \in V$.

Example A.3. $\mathbb{R}^n$ with the usual Euclidean norm

$$\|(v_1,\cdots,v_n)\| = \left(\sum_{i=1}^{n} v_i^2\right)^{1/2},$$

or the norm

$$\|(v_1,\cdots,v_n)\| = \max_i |v_i|.$$

Note that all norms on $\mathbb{R}^n$ are equivalent in the sense that they define the same topology (which has a neighbourhood basis consisting of open balls $B_{v_0}(\delta) = \{v : \|v - v_0\| < \delta\}$).

Example A.4. $\mathbb{C}^n$ with the norm

$$\|(z_1,\cdots,z_n)\| = \left(\sum_{i=1}^{n} |z_i|^2\right)^{1/2}.$$

Example A.5. $C[0,1]$ with the sup norm:

$$\|f\| = \max_{t\in[0,1]} |f(t)|, \; f \in C[0,1].$$

Definition A.4. An inner product space $(V,(\cdot,\cdot))$ is a vector space V together with a map $(\cdot,\cdot) : V \times V \to \mathbb{F}$ such that:
(a) $(v,v) \geq 0$ for all $v \in V$ (in particular, $(v,v) \in \mathbb{R}$.
(b) $(v,v) = 0$ if and only if $v = 0$.
(c) $(v,w) = \overline{(w,v)}$ for all $v,w \in V$.
(d) $(\lambda v + \mu w, x) = \lambda(v,x) + \mu(w,x)$, for all $v,w,x \in V$ and $\lambda,\mu \in \mathbb{F}$.

An inner product is also called a scalar product. A vector space with an inner product is a normed space if we define

$$\|v\| = \sqrt{(v,v)}.$$

Example A.6. $\mathbb{R}^n$ and $\mathbb{C}^n$ with their usual structures.

Example A.7. $C[0,1]$ with the inner product

$$(f,g) = \int_0^1 f(t)g(t)dt$$

(or $(f,g) = \int_0^1 f(t)\overline{g(t)}dt$ if we allow complex-valued functions). Note that this inner product gives a different norm on $C[0,1]$ from the one above.

A.2 Linear Dependence and Bases

If V is a vector space over $\mathbb{F}$, a *linear combination* of the vectors $v_1,\cdots,v_k \in V$ is an element of V of the form

$$w = \alpha_1 v_1 + \cdots + \alpha_k v_k, \ \ \alpha \in \mathbb{F}.$$

We say that the vectors $v_1,\cdots,v_k$ are *linearly dependent* if there exist scalars $\alpha_1,\cdots,\alpha_k \in \mathbb{F}$, not all zero, such that

$$\alpha_1 v_1 + \cdots + \alpha_k v_k = 0. \tag{A.1}$$

If, on the other hand, expressions of the form (A.1) always imply that $\alpha_1 = \cdots = \alpha_k = 0$, we say that $v_1,\cdots,v_k$ are *linearly independent*. A maximal linearly independent set in V is called a *basis* of V, and the number of elements in such a set is called the dimension of V (this is clearly well-defined and can be infinite).

Definition A.5. The standard basis of $\mathbb{R}^n$ (or $\mathbb{C}^n$) is the set $\{e_1,\cdots,e_n\}$, where e_i is the n-vector with a single '1' as the i^{th} component and zeros otherwise.

Definition A.6. A *linear operator* (or *homomorphism*) A from a vector space V into a vector space W is a function $A: V \to W$ such that

$$A(\alpha v_1 + \beta v_2) = \alpha A(v_1) + \beta A(v_2), \text{ for all } \alpha,\beta \in \mathbb{F}, v_1, v_2 \in V.$$

If A maps V onto W, it is called an *epimorphism*, if it is one-to-one it is a *monomorphism* and if it is both it is an *isomorphism* . For finite-dimensional vector spaces, A is an isomorphism if and only if dim V = dim W.

Theorem A.1. *Every finite-dimensional vector space over $\mathbb{R}$ is isomorphic to $\mathbb{R}^n$ for some n and every finite-dimensional vector space over $\mathbb{C}$ is isomorphic to $\mathbb{C}^n$ for some n.*

Because of this theorem, we can think of any finite-dimensional vector space over $\mathbb{R}$ (or $\mathbb{C}$) as a set of n-tuples $(x_1, \cdots, x_n)$ (or $(z_1, \cdots, z_n)$) with their usual operations. Thus any basis can be thought of as the columns of a matrix E which is invertible. To find the components $(\overline{v}_1, \cdots, \overline{v}_n)$ of any vector v in terms of the basis determined by E is equivalent to solving the linear equation

$$E\begin{pmatrix} \overline{v}_1 \\ \vdots \\ \overline{v}_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

where $v_1, \cdots, v_n$ are the components of v in the standard basis.

If $A : \mathbb{R}^m \to \mathbb{R}^n$ is a linear operator (similar remarks apply, of course, if $A : \mathbb{C}^m \to \mathbb{C}^n$), we define the *matrix* of A with respect to the bases $b_1^1, \cdots, b_1^m$ of $\mathbb{R}^m$ and $b_2^1, \cdots, b_2^n$ of $\mathbb{R}^n$ by

$$Ab_1^i = \sum_{j=1}^{n} a_{ji} b_2^j, \quad 1 \le i \le m.$$

If the bases are clear from the context, we usually denote the operator and its matrix representation by the same letter A. Hence, if we write an equation of the form

$$Av = w,$$

we shall mean an operator equation in basis-free form or a given matrix representation. Since different matrix representations A and B are related by a similarity transformation of the form

$$A = P^{-1}BP$$

for some invertible matrix P, we can define invariant properties of A as in the following definition.

Definition A.7. Let $A : \mathbb{R}^n \to \mathbb{R}^n$ (so that A is represented by a square matrix). The *kernel* of A (Ker A) is the set of vectors $v \in \mathbb{R}^n$ such that $Av = 0$. The range of A is called the *image* of A (Im A) and the *determinant* of A (det A) is the determinant of any matrix representation of A.

Note that the equation

$$Av = w \tag{A.2}$$

has a unique solution if and only if $\det A \neq 0$, *i.e.* if and only if $\operatorname{Ker} A = \{0\}$. In this case, $\operatorname{Im} A = \mathbb{R}^n$. If $\operatorname{Ker} A \neq 0$, then A is not one-to-one and the Equation A.2 has a solution v if and only if $w \in \operatorname{Im} A$, in which case $v + \alpha x$ is also a solution for any $x \in \operatorname{Ker} A$ and all $\alpha \in \mathbb{R}$.

A.3 Subspaces and Quotient Spaces

Definition A.8. If a subset W of a vector space V is closed under the operations of addition and scalar multiplication, it is called a (*linear*) *subspace* of V.

Definition A.9. If $W \subseteq V$ is a subspace, we define the *quotient space* V/W to be the set of affine subsets of V 'parallel' to W, *i.e.* the sets of the form

$$v+W=\{v+w : w\in W\}.$$

We can make V/W into a vector space by defining

$$\begin{aligned}(v_1+W)+(v_2+W) &= (v_1+v_2)+W\\ \lambda(v+W) &= \lambda v+W.\end{aligned}$$

This is clearly well-defined. If $b_1,\cdots,b_n$ is a basis of V such that $b_1,\cdots,b_m$ is a basis of W (*i.e.* dim $W=m$), then

$$b_{m+1}+W,\cdots,b_n+W$$

is a basis of V/W and we have

$$\dim V=\dim W+\dim V/W.$$

We usually write $\overline{v}=v+W$. Note that $\overline{v_1}=\overline{v_2}$ if and only if $v_1-v_2\in W$ and if we write $v_1\sim v_2$ if and only if $v_1-v_2\in W$, then $\sim$ is an equivalence relation, so the sets $\overline{v}$, $v\in V$ partition V.

Now let $f : V\to X$ be a linear map between vector spaces V and X and suppose that $f(U)\subseteq Y$ where U is a subspace of V and Y is a subspace of X. Then we define a linear map $\overline{f} : V/U\to X/Y$ by

$$\overline{f}(\overline{v})=\overline{f(v)}.$$

In particular, if $V=X$ and $U=Y$, we say that U is an invariant subspace if $f(U)\subseteq U$. In this case, f induces a map

$$\overline{f} : V/U\to V/U.$$

If $V=\{0\}$, then $X/Y\cong X$ and so if $f : V\to X$ has kernel Ker f, then

$$\overline{f} : V/\text{Ker}\, f\to \text{Im}\, f$$

is an isomorphism.

A.4 Eigenspaces and the Jordan Form

Definition A.10. An *eigenvector* of a linear operator (or matrix) A is a non-zero vector v which satisfies the equation

$$Av = \lambda v$$

for some (complex) number λ, called an *eigenvalue*.

Thus an eigenvector spans an invariant subspace of A. The eigenvector equation can be written

$$(A - \lambda I)v = 0.$$

This equation can have a non-zero solution v if and only if

$$\det\,(A - \lambda I) = 0.$$

This is called the *characteristic equation* of A and has n solutions (counting multiplicity).

Suppose first that A has n linearly independent eigenvectors $v_1, \cdots, v_n$ with corresponding eigenvalues $\lambda_1, \cdots, \lambda_n$. Then

$$(A - \lambda_i)v_i = 0, \;\; 1 \le i \le n$$

and the $n \times n$ matrix

$$P = [v_1, v_2, \cdots, v_n]$$

is invertible. Hence,

$$\begin{aligned} AP &= [Av_1, Av_2, \cdots, Av_n] \\ &= [\lambda_1 v_1, \lambda_2 v_2, \cdots, \lambda_n v_n] \\ &= P\Lambda \end{aligned}$$

where $\Lambda = \text{diag}\,[\lambda_1, \lambda_2, \cdots, \lambda_n]$, and so

$$P^{-1}AP = \Lambda$$

and therefore a similarity transformation diagonalises A.

If A does not have a linearly independent set of n eigenvectors, we consider the *generalised eigenspace* of λ, *i.e.*

$$V_A(\lambda) = \{v \in \mathbb{C}^n : (A - \lambda I)^k v = 0 \text{ for some integer } k\}.$$

(We must consider $\mathbb{C}^n$ because the eigenvalues may be complex.) We have:

Theorem A.2. *If $A : \mathbb{C}^n \to \mathbb{C}^n$ is any linear operator and $V_A(\lambda)$ are the corresponding generalised eigenspaces, then:*

(a) $V_A(\lambda)$ is an invariant subspace of A.

(b) $V_A(\lambda) \ne \{0\}$ if and only if λ is an eigenvalue of A.

(c) $A|_{V_A(\lambda)}$ *has all eigenvalues equal to* λ *if* $V_A(\lambda) \neq \{0\}$.

(d) If $V_A(\lambda) \neq \{0\}$ *and* $m = dim\, V_A(\lambda)$*, then*

$$(A - \lambda I)^m v = 0 \textit{ for all } v \in V_A(\lambda).$$

(e) The multiplicity of the eigenvalue λ *of* A *is* $m = dim\, V_A(\lambda)$.

(f) $\mathbb{C}^n = \bigoplus_{\lambda \in \Delta} V_A(\lambda)$*, where* Δ *is the set of distinct eigenvalues of* A.

(Here $\bigoplus$ *means the* direct sum *of subspaces, so that if* $v \in V_1 \bigoplus V_2$*, then* v *has a unique representation in the form* $v = v_1 + v_2$ *where* $v_1 \in V_1, v_2 \in V_2$.*)*

Proof.

(a) If $v \in V_A(\lambda)$, then $(A - \lambda I)^k Av = A(A - \lambda I)^k v = 0$, so $Av \in V_A$.

(b) If λ is an eigenvalue, then $(A - \lambda I)v = 0$, so $v \in V_A$. Conversely, if $V_A \neq \{0\}$, then $(A - \lambda I)^k v = 0$, for some minimal $k > 0$ and $v \neq 0$. Then $(A - \lambda I)(A - \lambda I)^{k-1} v = 0$, so λ is an eigenvalue with eigenvector $(A - \lambda I)^{k-1} v \neq 0$, by minimality of k.

(c) If $\lambda \neq \mu$ is an eigenvalue of $A|_{V_A(\lambda)}$, then $Av = \mu v$ for some $v \neq 0$ in $V_A(\lambda)$. Then $(A - \lambda I)^k v = 0$ for some $k > 0$. Substituting $Av = \mu v$ into this gives $(\mu - \lambda)^k v = 0$, so $v = 0$, which is a contradiction, so $\lambda = \mu$.

(d) If $v \in V_A(\lambda)$ and $v \neq 0$, let

$$v_0 = v,\ v_1 = (A - \lambda I)^i v,\ i \geq 1.$$

If $v_p = (A - \lambda I)^p v \neq 0$ for $p \geq m$ and $v_{p+1} = 0$, then $v_0, v_1, \cdots, v_p$ are linearly independent, since if not,

$$\alpha_0 v_0 + \alpha_1 v_1 + \cdots + \alpha_p v_p = 0$$

for some α's, not all zero. Applying $(A - \lambda I)^p$ to this equation gives $\alpha_0 = 0$, and then applying $(A - \lambda I)^{p-1}$ gives $\alpha_1 = 0$, etc. Thus,$v_0, \cdots, v_p$ are linearly independent for $p \geq m$, which contradicts the definition of m.

(e) We must show that $\overline{A} : \mathbb{C}^n / V_A(\lambda) \to \mathbb{C}^n / V_A(\lambda)$ does not have λ as an eigenvalue. If it does, then there exists $0 \neq \overline{v} \in \mathbb{C}^n / V_A(\lambda)$ such that

$$(\overline{A} - \lambda I)\overline{v} = 0$$

i.e.

$$(A - \lambda I)v \in V_A(\lambda)$$

so that

$$(A - \lambda I)^{m+1} v = 0$$

i.e. $v \in V_A(\lambda)$, which contradicts $\overline{v} \neq 0$.

(f) Let $\lambda_1, \cdots, \lambda_r$ be the distinct eigenvalues of A. We show that the sum $\sum_{i=1}^r V_A(\lambda_i)$ is direct. Suppose that

$$v_1 + \cdots + v_r = 0,\ v_i \in V_A(\lambda_i).$$

If $m_i = \dim V_A(\lambda_i)$, we define the polynomials

$$f_i(\lambda) = \prod_{\substack{j=1 \\ j \neq i}}^{r} (\lambda - \lambda_j)^{m_j}, \; g_i(\lambda) = (\lambda - \lambda_i)^{m_i}.$$

Then f_i and g_i are relatively prime (by the distinctness of the λ_i's) and so, by classical number theory, there exists polynomials p, q such that

$$p(\lambda) f_i(\lambda) + q(\lambda) g_i(\lambda) = 1.$$

Substituting A into this equation gives

$$p(A) f_i(A) + q(A) g_i(A) = I$$

i.e.

$$p(A) \prod_{\substack{j=1 \\ j \neq i}}^{r} (A - \lambda_j I)^{m_j} + q(A)(A - \lambda_i I)^{m_i} = I.$$

Now apply this to v_i $(= -\sum_{\substack{j=1 \\ j \neq i}}^{r} v_j$, by assumption). This gives $v_i = 0$, so the sum is direct. Since m_i is the multiplicity of λ_i and $\sum m_i = \dim \mathbb{C}^n = n$, the result follows. □

This theorem says that we can choose a basis of $\mathbb{C}^n$ consisting of generalised eigenvectors of A so that A takes the form

$$\begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_r \end{bmatrix}$$

where r is the number of distinct eigenvalues of A. To determine the structure of each Λ_i, let λ_i be the corresponding eigenvalue. For each element v of $V_A(\lambda_i)$, we have

$$(A - \lambda_i I)^p v = 0 \text{ and } (A - \lambda_i I)^{p-1} v \neq 0$$

for some integer $p \geq 1$ (depending on v). Let v_1 be an element of $V_A(\lambda_i)$ for which p is maximal. Then we have the linearly independent vectors

$$w_1 = v_1, w_2 = (A - \lambda_i I) v_1, \cdots, w_p = (A - \lambda_i I)^{p-1} v_1$$

which span a subspace V_1 of $V_A(\lambda_i)$. The matrix of A restricted to this subspace is

$$\begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{bmatrix}$$

(a $p \times p$ matrix). If $V_1 \neq V_A(\lambda_i)$, we consider $V_A(\lambda_i)/V_1$ and repeat the process. Using the resulting basis for $\mathbb{C}^n$ gives the *Jordan form* of A.

References

1. Mirsky, L.: An Introduction to Linear Algebra. OUP, Oxford (1955)
2. Halmos, P.R.: Finite Dimensional Vector Spaces. Princeton (1958)

Appendix B
Lie Algebras

B.1 Elementary Theory

In this appendix we give a brief outline of the theory of Lie algebras. Most of the proofs are omitted and can be found in a number of excellent monographs on the subject (for example, [1,2]). Lie algebras abstract the notion of non-commutation of matrices; they are vector spaces with an additional (non-commutative) product structure:

Definition B.1. A Lie algebra $\mathfrak{g}$ is a vector space over $\mathbb{F}$, together with a binary operation $[\cdot,\cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ which satisfies the axioms:

(a) $[\lambda_1 X_1 + \lambda_2 X_2, Y] = \lambda_1 [X_1, Y] + \lambda_2 [X_2, Y]$ for all $X_1, X_2, Y \in \mathfrak{g}$.
(b) $[X,Y] = -[Y,X]$, for all $X,Y \in \mathfrak{g}$.
(c) $[X,[Y,Z]] + [Y,[Z,X]] + [Z,[X,Y]] = 0$, for all $X,Y,Z \in \mathfrak{g}$.

The *dimension* of $\mathfrak{g}$ is its dimension as a vector space. Axiom (c) of Definition B.1 is called the *Jacobi identity*. Note that

$$[X,X] = 0 \text{ for all } X \in \mathfrak{g}.$$

Example B.1. Any vector space V with bracket $[X,Y] = 0$ for all $X,Y \in V$. These are the *abelian* Lie algebras.

Example B.2. $\mathbb{R}^3$ with the usual vector space structure and the bracket

$$[x,y] = x \times y, \ \ x,y \in \mathbb{R}^3,$$

where × denotes the vector cross product.

Example B.3. The set of all $n \times n$ complex matrices $\mathfrak{g}\ell(n;\mathbb{C})$ with the obvious bracket

$$[X,Y] = XY - YX.$$

More generally, if V is any given vector space, we write $\mathfrak{gl}(V)$ for the set of all linear operators $L : V \to V$.)

Example B.4. Any associative algebra A (*i.e.* vector space with an associative product), together with the bracket

$$[a,b] = a \cdot b - b \cdot a, \text{ for all } a,b \in A,$$

where $\cdot$ is the algebra product. (This clearly generalises Example B.3.)

If $\mathfrak{g}$ is a finite-dimensional Lie algebra, let $\{b_i\}_{1 \le i \le n}$ be a basis of $\mathfrak{g}$. Then we have

$$[b_i, b_j] = \sum_{k=1}^{n} c_{ij}^k b_k,$$

for some n^3 constants c_{ij}^k called the *structure constants* of $\mathfrak{g}$. Using the axioms (L1)–(L3), it is easy to prove the identities

$$c_{ij}^k = -c_{ji}^k, \;\; 1 \le i,j,k \le n$$

$$\sum_{\ell=1}^{n} \sum_{m=1}^{n} \left(c_{ik}^{\ell} c_{i\ell}^{m} + c_{ki}^{\ell} c_{j\ell}^{m} + c_{ij}^{\ell} c_{j\ell}^{m} \right) = 0, \;\; 1 \le i,j,k \le n.$$

The structure constants clearly determine the Lie algebra, for any given fixed basis. A linear map (operator) $A : \mathfrak{g}_1 \to \mathfrak{g}_2$, from a Lie algebra $\mathfrak{g}_1$ to another Lie algebra $\mathfrak{g}_2$ is a *homomorphism* (of Lie algebras) if it preserves the bracket, *i.e.*

$$A[X,Y] = [AX, AY], \text{ for all } X,Y \in \mathfrak{g}_1.$$

Monomorphisms and isomorphisms are defined as in the vector space case.

Definition B.2. Let $\mathfrak{g}$ be a Lie algebra and $\mathfrak{h}$ a subspace of $\mathfrak{g}$. Then $\mathfrak{h}$ is a *subalgebra* of $\mathfrak{g}$ if

$$[\mathfrak{h}, \mathfrak{h}] \subseteq \mathfrak{h}$$

and an *ideal* if

$$[\mathfrak{h}, \mathfrak{g}] \subseteq \mathfrak{h}.$$

A Lie algebra $\mathfrak{g}$ with no non-trivial ideals (*i.e.* $\neq \mathfrak{g}$ or $\{0\}$) is called *simple*. The (unique) 1-dimensional Lie algebra is the *trivial simple Lie algebra*.

Clearly, if $\mathfrak{h}_1, \mathfrak{h}_2$ are ideals, then so are $\mathfrak{h}_1 + \mathfrak{h}_2$ and $\mathfrak{h}_1 \cap \mathfrak{h}_2$. If $\mathfrak{h} \subseteq \mathfrak{g}$ is an ideal, then we can define the *quotient space* $\mathfrak{g}/\mathfrak{h}$ and make it into a Lie algebra by defining

$$[\overline{X}, \overline{Y}] = \overline{[X,Y]}, \text{ for all } X,Y \in \mathfrak{g}.$$

Induced maps then behave in the same way as induced linear maps of vector spaces.

Definition B.3. If $\mathfrak{g}$ is a Lie algebra, then $\mathfrak{D}\mathfrak{g} = [\mathfrak{g}, \mathfrak{g}]$ is called the *derived algebra* of $\mathfrak{g}$. The *derived series* of $\mathfrak{g}$ is the sequence of ideals

$$\mathfrak{g} = \mathfrak{D}^{(0)}\mathfrak{g} \supseteq \mathfrak{D}^{(1)}\mathfrak{g} \supseteq \cdots \supseteq \mathfrak{D}^{(n)}\mathfrak{g} \supseteq \cdots,$$

where

$$\mathfrak{D}^{(n)}\mathfrak{g} = \mathfrak{D}(\mathfrak{D}^{(n-1)}\mathfrak{g}).$$

The sequence of ideals

$$\mathfrak{g} = \mathfrak{C}^{(0)}\mathfrak{g} \supseteq \mathfrak{C}^{(1)}\mathfrak{g} \supseteq \cdots \supseteq \mathfrak{C}^{(n)}\mathfrak{g} \supseteq \cdots,$$

where $\mathfrak{C}^{(n)}\mathfrak{g} = [\mathfrak{g}, \mathfrak{C}^{(n-1)}\mathfrak{g}]$, is the *descending central series* of $\mathfrak{g}$.

If $\mathfrak{D}^{(n)}\mathfrak{g} = \{0\}$ for some finite n, then $\mathfrak{g}$ is called a *solvable* Lie algebra. If $\mathfrak{C}^{(n)}\mathfrak{g} = \{0\}$ for some finite n, then $\mathfrak{g}$ is called *nilpotent*. Note that

$$\mathfrak{D}^{(n)}\mathfrak{g} \subseteq \mathfrak{C}^{(n)}\mathfrak{g}$$

so that if $\mathfrak{g}$ is nilpotent then it is solvable. Subalgebras, homomorphic images and direct sums of solvable (nilpotent) Lie algebras are solvable (nilpotent). It follows that any Lie algebra $\mathfrak{g}$ has a unique maximal solvable ideal $\mathfrak{r}$ called the *radical* of $\mathfrak{g}$. If the only solvable ideal in $\mathfrak{g}$ is $\{0\}$ then $\mathfrak{g}$ is called *semi-simple*. Thus, if $\mathfrak{g}$ is not solvable, then $\mathfrak{g}/\mathfrak{r}$ is semi-simple. (Any decomposition of the form

$$\mathfrak{g} = \mathfrak{r} + \mathfrak{m}$$

is called a *Levi decomposition*. It is not a direct sum and so the decomposition is not unique.) Also, $\mathfrak{g}$ is solvable if and only if $\mathfrak{D}\mathfrak{g}$ is nilpotent. If $\mathfrak{g}$ is a solvable subalgebra of $\mathfrak{g}\ell(V)$, then there exists a basis of V such that the matrix representations of each operator in $\mathfrak{g}$ in this basis are all upper triangular.

Definition B.4. If $\mathfrak{g}$ is a Lie algebra then, for any $X \in \mathfrak{g}$, we define the map $\operatorname{ad} X : \mathfrak{g} \to \mathfrak{g}$ by

$$(\operatorname{ad} X)(Y) = [X, Y].$$

Then $\mathfrak{g}$ is nilpotent if and only if for any $X \in \mathfrak{g}$, $\operatorname{ad} X$ is nilpotent. Moreover, if $\mathfrak{g}$ is a subalgebra of $\mathfrak{g}\ell(V)$, then if every element X of $\mathfrak{g}$ is nilpotent (*i.e.* $X^k = 0$ for some k), then $\mathfrak{g}$ is a nilpotent Lie algebra.

B.2 Cartan Decompositions of Semi-simple Lie Algebras

Definition B.5. Let $\mathfrak{h}$ be a subalgebra of $\mathfrak{g}\ell(V)$. A function $\lambda : \mathfrak{h} \to \mathbb{C}$ is called a *weight* of $\mathfrak{h}$ if there exists $0 \neq v \in V$ such that

$$Hv = \lambda(H)v, \text{ for all } H \in \mathfrak{h}.$$

The vector v is called a *weight vector*.

Nilpotent Lie subalgebras of $\mathfrak{g}\ell(V)$ have a decomposition similar to the Jordan decomposition of a single $n \times n$ matrix. In fact, we have:

Theorem B.1. *Let $\mathfrak{h} \subseteq \mathfrak{g}\ell(V)$ be a nilpotent Lie algebra and let $\lambda : \mathfrak{h} \to \mathbb{C}$ be a linear function. Define the* weight subspace

$$V_{\mathfrak{h}}(\lambda) = \{v \in V : (H - \lambda(H)I)^k v = 0 \textit{ for some } k > 0 \textit{ and all } H \in \mathfrak{h}\}.$$

Then:

(a) $V_{\mathfrak{h}}(\lambda)$ is an invariant subspace and $V_{\mathfrak{h}}(\lambda) = \bigcap_{H \in \mathfrak{h}} V_H(\lambda)$.

(b) $V_{\mathfrak{h}}(\lambda) \neq 0$ if and only if λ is a weight of $\mathfrak{h}$, and λ is the only weight of $\mathfrak{h}$ in $V_{\mathfrak{h}}(\lambda)$.

(c) If $V_{\mathfrak{h}}(\lambda) \neq 0$, then $(H - \lambda(H)I)^{dim\, V_{\mathfrak{h}}(\lambda)} v = 0$, for all $H \in \mathfrak{h}$, $v \in V_{\mathfrak{h}}(\lambda)$.

(d) $V = \bigoplus_{\lambda \in \Delta} V_{\mathfrak{h}}(\lambda)$, where Δ is the set of weights of $\mathfrak{h}$.

This gives an immediate generalisation of Jordan's theorem:

Corollary B.1. *If $\mathfrak{h} \subseteq \mathfrak{g}\ell(V)$ is nilpotent, where $V \cong \mathbb{C}^n$, then $\mathfrak{h}$ is isomorphic to a subalgebra of $\mathfrak{n}(n_1, \mathbb{C}) \bigoplus \cdots \mathfrak{n}(n_r, \mathbb{C})$ where $n_1 + \cdots + n_r = n$ and r is the number of weights. Here, $\mathfrak{n}(m, \mathbb{C})$ is the set of upper triangular matrices with equal diagonal elements.*

We now apply this to a general Lie algebra $\mathfrak{g}$ with a nilpotent subalgebra $\mathfrak{h}$. Then the set

$$\mathrm{ad}_{\mathfrak{g}} \mathfrak{h} = \{\mathrm{ad}_{\mathfrak{g}} H : H \in \mathfrak{h}\}$$

is a nilpotent Lie subalgebra of $\mathfrak{g}\ell(\mathfrak{g}) \cong \mathfrak{g}\ell(\mathbb{C}^n)$, if $\dim \mathfrak{g} = n$. By the above theorem we have the decomposition

$$\mathfrak{g} = \bigoplus_{\lambda \in \Delta} \mathfrak{g}_{\mathrm{ad}\, \mathfrak{h}}(\lambda), \tag{B.1}$$

where Δ is the set of weights of ad $\mathfrak{h}$. (Note that Δ is also referred to as the set of weights of $\mathfrak{h}$.) To study this decomposition in more detail we introduce the *Killing form* of $\mathfrak{g}$ to be the symmetric, bi-linear function $(\cdot,\cdot) : \mathfrak{g} \times \mathfrak{g} \to \mathbb{C}$ given by

$$(X,Y) = \mathrm{tr}\ [(\mathrm{ad}\, X)(\mathrm{ad}\, Y)], \ \ X,Y \in \mathfrak{g}.$$

Note that

$$((\mathrm{ad}\ )X,Y) + (X,(\mathrm{ad})Y) = 0.$$

For each $X \in \mathfrak{g}$, the characteristic polynomial $K_X(\lambda)$ of ad X is called the *Killing polynomial* of X. Since

$$(\mathrm{ad}\, X)X = [X,X] = 0,$$

$K_X(\lambda)$ has at least one zero root (*i.e.* ad X has a zero eigenvalue), so that $K_X(\lambda)$ has the form

$$K_X(\lambda) = \lambda^r + \alpha_1(X)\lambda^{r-1} + \cdots + \alpha_{r-k(X)}(X)\lambda^{k(X)}$$

where $k(X) \geq 1$. Set

$$\kappa = \min_{X\in\mathfrak{g}} k(X).$$

κ is called the *rank* of $\mathfrak{g}$. If $X \in \mathfrak{g}$ is such that $k(X) = \kappa$ then X is called *regular*; otherwise it is *singular*.

Returning to the decomposition (B.1), the following properties of the weight (root) spaces $\mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\lambda)$ are simple consequences of the definitions:

(a) $[\mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\alpha), \mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\beta)] \subseteq \mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\alpha+\beta)$, for all $\alpha, \beta \in \Delta$, so that $\mathfrak{g}_{\text{ad}\,\mathfrak{h}}(0)$ is a subalgebra of $\mathfrak{g}$.

(b) If $\alpha+\beta$ is not a root, then $[\mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\alpha), \mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\beta)] = 0$.

(c) $(\mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\alpha), \mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\beta)) = 0$, if $\alpha+\beta \neq 0$, where $(\cdot,\cdot)$ is the Killing form.

(d) $\mathfrak{h} \subseteq \mathfrak{g}_{\text{ad}\,\mathfrak{h}}(0)$.

Definition B.6. If $\mathfrak{h} = \mathfrak{g}_{\text{ad}\,\mathfrak{h}}(0)$, then $\mathfrak{h}$ is called a *Cartan subalgebra* of $\mathfrak{g}$ and (B.1) is the *Cartan decomposition* of $\mathfrak{g}$.

Theorem B.2. *(a) A Cartan subalgebra of $\mathfrak{g}$ is a maximal nilpotent subalgebra.*

(b) Every Lie algebra has a Cartan subalgebra.

(c) If X_0 is a regular element of $\mathfrak{g}$, then $\mathfrak{g}_{ad\,X_0}(0)$ is a Cartan subalgebra.

(d) If $\mathfrak{h}_i$ is a Cartan subalgebra of $\mathfrak{g}_i$, ($i = 1,2$), then $\mathfrak{h}_1 \bigoplus \mathfrak{h}_2$ is a Cartan subalgebra of $\mathfrak{g}_1 \bigoplus \mathfrak{g}_2$.

(e) If $\mathfrak{h}$ is a nilpotent subalgebra of $\mathfrak{g}$ and

$$\mathfrak{n}(\mathfrak{h}) = \{X \in \mathfrak{g} : [X, \mathfrak{h}] \subseteq \mathfrak{h}\}$$

is the normaliser *of $\mathfrak{h}$ in $\mathfrak{g}$, then $\mathfrak{h}$ is a Cartan subalgebra if and only if $\mathfrak{h} = \mathfrak{n}\mathfrak{h}$.*

(f) Cartan subalgebras of a given Lie algebra $\mathfrak{g}$ are conjugate under inner automorphisms (i.e. automorphisms of the form exp$(ad\,X)$ where $ad\,X$ is a nilpotent linear operator).

The decomposition (B.1) can be written in the form

$$\mathfrak{g} = \mathfrak{h} \oplus \bigoplus_{\lambda\in\Sigma} \mathfrak{g}(\lambda)$$

where $\mathfrak{h}$ is a given Cartan subalgebra, $\mathfrak{g}(\lambda) = \mathfrak{g}_{\text{ad}\,\mathfrak{h}}(\lambda)$ and $\sum$ is the set of non-zero roots of $\mathfrak{g}$. If $n_\lambda = \dim\,\mathfrak{g}(\lambda)$, then the Killing form is

$$(X,Y) = \sum_{\lambda\in\Sigma} n_\lambda \lambda(X)\lambda(Y).$$

Theorem B.3. *(Cartan)*

(a) A Lie algebra $\mathfrak{g}$ is solvable if and only if $(X,X) = 0$ for all $X \in \mathfrak{D}\mathfrak{g}$.

(b) $\mathfrak{g}$ is semi-simple if and only if the Killing form $(\cdot,\cdot)$ is non-degenerate.

It follows that a semisimple Lie algebra is the direct sum of all its minimal ideals, which are simple Lie algebras which are orthogonal with respect to $(\cdot,\cdot)$.

B.3 Root Systems and Classification of Simple Lie Algebras

Now consider the Cartan decomposition of a semisimple Lie algebra $\mathfrak{g}$ with Cartan subalgebra $\mathfrak{h}$. If dim $\mathfrak{g} = n$ and dim $\mathfrak{h} = m$, then the Cartan decomposition of $\mathfrak{g}$ is

$$\mathfrak{g} = \mathfrak{h} \oplus \bigoplus_{\lambda \in \Sigma} \mathfrak{g}(\lambda)$$

where $\mathfrak{h}$ is an abelian subalgebra and $(\cdot,\cdot)|_{\mathfrak{h}}$ is non-degenerate. Moreover, there are m linearly independent root subspaces $\mathfrak{g}_\lambda$ $(\lambda \neq 0)$ which are one-dimensional and if $\lambda \in \Sigma$ then so is $-\lambda$ and $k\lambda \notin \Sigma$ for $k \neq \pm 1$. Note that if $\lambda + \mu$ is a root, then

$$[\mathfrak{g}_\lambda, \mathfrak{g}_\mu] = \mathfrak{g}_{\lambda+\mu}. \tag{B.2}$$

Since the restriction of the Killing form $(\cdot,\cdot)$ to $\mathfrak{h}$ is non-degenerate and any root $\lambda : \mathfrak{h} \to \mathbb{C}$ is a linear functional on $\mathfrak{h}$, *i.e.* $\lambda \in \mathfrak{h}^*$ (the dual space of $\mathfrak{h}$), there exists a unique $H_\lambda \in \mathfrak{h}$ such that

$$(H, H_\lambda) = \lambda(H), \text{ for all } \lambda \in \Sigma.$$

Clearly,

$$\lambda(H_\lambda) \neq 0 \text{ for all } \lambda \in \Sigma.$$

Also, for any $\lambda \in \Sigma$ and any $E_\lambda \in \mathfrak{g}_\lambda$, there exists a unique $E_{-\lambda} \in \mathfrak{g}_{-\lambda}$ such that

$$(E_\lambda, E_{-\lambda}) = 1 \text{ and } [E_\lambda, E_{-\lambda}] = H_\lambda. \tag{B.3}$$

Now let λ, μ be roots such that $\lambda \neq \pm\mu$. Then there exist non-negative integers p, q such that

$$-p\lambda + \mu, -(p-1)\lambda + \mu, \cdots, -\lambda + \mu, \mu, \lambda + \mu, \cdots, q\lambda + \mu$$

are roots, but $-(p+1)\lambda + \mu$ and $(q+1)\lambda + \mu$ are not. Then we have

$$2\frac{(H_\lambda, H_\mu)}{(H_\lambda, H_\lambda)} = p - q.$$

Note that the elements H_λ span $\mathfrak{h}$ and since dim $\mathfrak{h} = m$, there exist $H_{\lambda_1}, H_{\lambda_2}, \cdots, H_{\lambda_m}$ which form a basis of $\mathfrak{h}$. Moreover it follows that we can write any H_λ in the form

$$H_\lambda = \sum_{i=1}^{m} \alpha_i H_{\lambda_i}$$

where α_i is a rational number. If $\mathfrak{h}_{\mathbb{R}}$ denotes the real vector space corresponding to $\mathfrak{h}$ (*i.e.* the space generated by the H_λ over $\mathbb{R}$) then the Killing form on $\mathfrak{h}_{\mathbb{R}}$ makes it into a Euclidean space of dimension m, *i.e.* the Killing form $(\cdot,\cdot)$ is a standard Euclidean inner product. Note that, by duality, this gives a metric on $\mathfrak{h}_{\mathbb{R}}^*$ by defining

$$(\lambda,\mu) = (H_\lambda, H_\mu).$$

We define a total order $<$ on $\mathfrak{h}_{\mathbb{R}}^*$ (this is a standard set-theoretic total order which satisfies $\lambda < \mu \Rightarrow \lambda + \nu < \mu + \nu$, for all λ, μ, ν and if $\lambda < \mu$ and $0 < r \in \mathbb{R}$, then $r\lambda < r\mu$). Every real vector space has a total order defined lexicographically. For a given total order on $\mathfrak{h}_{\mathbb{R}}^*$, we call Σ^+ the set of *positive roots* (*i.e.* $\lambda \in \Sigma$ such that $\lambda > 0$). A *fundamental system* of roots $\Pi \subseteq \Sigma^+$ is a subset of positive roots λ such that

$$\lambda \neq \mu + \nu$$

where $\mu, \nu \in \Sigma^+$ $(\mu \neq \lambda,\ \nu \neq \lambda)$. We have:

Theorem B.4. *(a) Every root in Σ^+ is a sum of roots in Π.*
(b) If $\lambda, \mu \in \Pi$ and $\lambda \neq \mu$, then $(\lambda, \mu) \leq 0$.
(c) Any fundamental system of roots Π is a basis of $\mathfrak{h}_{\mathbb{R}}^*$.

If $\Pi = \{\lambda_1, \cdots, \lambda_m\}$ is a fundamental system of roots of a semisimple Lie algebra $\mathfrak{g}$, then the matrix $A = (A_{ij})$ defined by

$$A_{ij} = 2\frac{(\lambda_i, \lambda_j)}{(\lambda_i, \lambda_i)}, \ \ 1 \leq i, j \leq m$$

is called the *Cartan matrix* of $\mathfrak{g}$.

From (B.2) it follows that, if E_λ is chosen as in (B.3), then we have

$$[E_\lambda, E_\mu] = N_{\lambda\mu} E_{\lambda+\mu}$$

for some constants $N_{\lambda\mu} \in \mathbb{C}$. (These numbers are sometimes also called the *structure constants* of $\mathfrak{g}$.) They satisfy the properties:

(a) $N_{\lambda\mu} = -N_{\mu\lambda}$, for all $\lambda, \mu \in \Sigma$ with $\lambda + \mu \neq 0$.

(b) If $\lambda, \mu, \nu \in \Sigma$ and $\lambda + \mu + \nu = 0$, then

$$N_{\lambda\mu} = N_{\mu\nu} = N_{\nu\lambda}.$$

(c) If $\alpha, \beta, \gamma, \delta \in \Sigma$ and $\alpha + \beta + \gamma + \delta = 0$, and the sum of any two of $\alpha, \beta, \gamma, \delta$ is not zero, then

$$N_{\alpha\beta}N_{\gamma\delta} + N_{\alpha\gamma}N_{\delta\beta} + N_{\alpha\delta}N_{\beta\gamma} = 0.$$

(d) If $\lambda, \mu, \nu \in \Sigma$ and $\lambda + \mu \neq 0$, and $\mu - p\lambda, \mu - (p-1)\lambda, \cdots, \mu + q\lambda$ is a maximal chain of roots, then

$$N_{\lambda\mu}N_{-\lambda,-\mu} = -\frac{q(p+1)}{2}(\lambda, \lambda).$$

It follows that for any semi-simple Lie algebra $\mathfrak{g}$, there exist root vectors E_λ ($\lambda \in \Sigma$) such that $(E_\lambda, E_{-\lambda}) = 1$ and all the structure constants $N_{\lambda\mu}$ ($\lambda, \mu, \lambda + \mu \in \Sigma$) are non-zero real numbers such that

$$N_{\lambda\mu} = -N_{-\lambda,-\mu}$$

and

$$N_{\lambda\mu}^2 = \frac{q(p+1)}{2}(\lambda, \lambda) > 0.$$

The basis $\{H_1, H_2, \cdots, H_m\} \cup \{E_\lambda : \lambda \in \Sigma\}$ of $\mathfrak{g}$, where the E_λ's satisfy the above conditions and $\{H_i\}$ is any basis of the Cartan subalgebra $\mathfrak{h}$ of $\mathfrak{g}$, is called a *Weyl basis* of $\mathfrak{g}$.

Recall now that a fundamental system of roots Π is a linearly independent set of vectors such that if $\lambda, \mu \in \Pi$ and $\lambda \neq \mu$, then $2(\lambda, \mu)/(\lambda, \lambda)$ is zero or a negative integer. Since

$$4\cos^2(\angle(\lambda, \mu)) = 4\frac{(\lambda, \mu)^2}{(\lambda, \lambda)(\mu, \mu)} = 2\frac{(\lambda, \mu)}{(\lambda, \lambda)} \cdot 2\frac{(\lambda, \mu)}{(\mu, \mu)},$$

where $\angle(\lambda, \mu)$ is the angle between λ and μ, we have

$$\cos(\angle(\lambda, \mu)) = -\frac{1}{2}\sqrt{r}$$

where $r = 0, 1, 2$ or 3. If $2(\lambda, \mu)/(\lambda, \lambda) \neq 0$ for $\lambda, \mu \in \Pi$, we say that Π is a *simple root system*. Then the possible angles between roots in a simple root system are $120°, 135°$ and $150°$. We construct a graph called a *Dynkin diagram* of Π with one vertex for each root in Π and 1,2 or 3 lines joining pairs of roots if the angle between them is respectively $120°, 135°$ or $150°$. Note that if $\angle(\lambda, \mu) = 120°$ then $(\mu, \mu) = (\lambda, \lambda)$ so the roots have the same length, while if $\angle(\lambda, \mu) = 135°$ then $(\mu, \mu) = 2(\lambda, \lambda)$ and if $\angle(\lambda, \mu) = 150°$ then $(\mu, \mu) = 3(\lambda, \lambda)$. It can be shown that, for a simple Π system of roots, the possible Dynkin diagrams are as shown in Figure B.1. (The shorter roots are denoted by a circle $\circ$ and the larger ones by $\bullet$.)

It can be shown that a Lie algebra is simple if and only if it has a simple Π system of roots, so the above Dynkin diagrams characterise all (finite-dimensional) simple Lie algebras. The fact that each one of these diagrams represents a valid simple Lie algebra can be shown by using representation theory. We shall simply give a realisation of each one – the details can be found in the references.

Type $\mathbf{A_n}$

Consider the subspace L_A of $\mathfrak{g}\ell(n+1)$ consisting of matrices with trace zero. Then L_A is a simple Lie algebra and has the decomposition

$$L_A = H \bigoplus \sum_{i \neq j} \mathbb{C} E_{ij}$$

where H is the set of diagonal matrices of trace zero and E_{ij} is the $(n+1) \times (n+1)$ matrix with zeros everywhere except for a 1 in the ij^{th} place. If diag $(\lambda_1, \cdots, \lambda_{n+1}) \in H$, then the roots are the functions

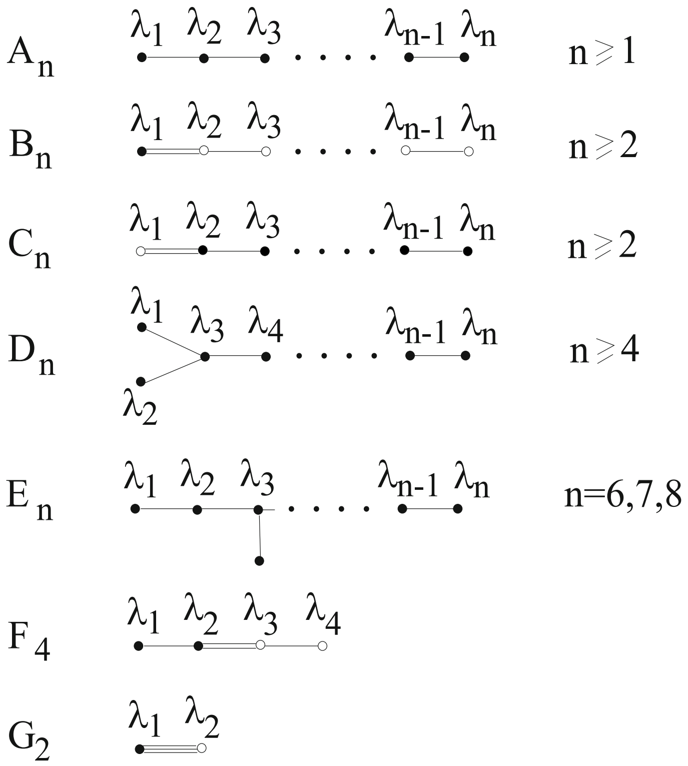


Fig. B.1 Dynkin diagrams of the simple Lie algebras

$$\text{diag }(\lambda_1,\cdots,\lambda_{n+1}) \to \lambda_i - \lambda_j,\ \ i \neq j.$$

A fundamental set of roots is given by

$$\text{diag }(\lambda_1,\cdots,\lambda_{n+1}) \to \lambda_i - \lambda_j,\ \ 1 \leq i \leq n.$$

The dimension of L_A is $n(n+2)$. Note that Lie algebras of type A_n are also written $\mathfrak{s}\ell(n+1,\mathbb{C})$ since they are the Lie algebras of $SL(n,\mathbb{C}) = \{g \in GL(n,\mathbb{C}) : \det g = 1\}$.

Type $\mathbf{B_n}$

The Lie algebras are the subspaces L_B of $\mathfrak{g}\ell(2n+1)$ consisting of matrices X which satisfy $X^T M + MX = 0$ where M is the matrix

$$M = \begin{pmatrix} 2 & 0_{1\times n} & 0_{1\times n} \\ 0_{n\times 1} & 0_{n\times n} & I_n \\ 0_{n\times 1} & I_n & 0_{n\times n} \end{pmatrix}$$

where the zeros represent zero matrices of appropriate dimensions. We have the decomposition

$$L_B = H \bigoplus \sum_{\lambda} \mathbb{C} E_{\lambda}$$

where

$$H = \{X : X = \text{diag}\ (0, \lambda_1, \cdots, \lambda_n, -\lambda_1, \cdots, -\lambda_n)\}$$

and the (non-zero) roots are $\Sigma = \{\pm\lambda_i \pm \lambda_j, i \neq j; \lambda_i\}$. The matrices E_{λ} for each root type are

$$E_{\lambda_i-\lambda_j} = \begin{pmatrix} 0 & & \\ & E_{ij} & \\ & & -E_{ji} \end{pmatrix}, E_{-\lambda_i+\lambda_j} = \begin{pmatrix} 0 & & \\ & E_{ji} & \\ & & -E_{ij} \end{pmatrix},\ i<j,$$

$$E_{\lambda_i+\lambda_j} = \begin{pmatrix} 0 & & \\ & 0 & E_{ij}-E_{ji} \\ & & 0 \end{pmatrix},\ E_{-\lambda_i-\lambda_j} = \begin{pmatrix} 0 & & \\ & 0 & \\ & -E_{ij}+E_{ji} & 0 \end{pmatrix},\ i<j,$$

$$E_{\lambda_i} = \begin{pmatrix} 0 & 0 & e_i \\ -e_i^T & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ E_{-\lambda_i} = \begin{pmatrix} 0 & -e_i & 0 \\ 0 & 0 & 0 \\ e_i^T & 0 & 0 \end{pmatrix},$$

where e_i is the standard i^{th} unit basis vector (of dimension n). The generators of H are

$$H_{\lambda_i-\lambda_j} = \begin{pmatrix} 0 & & \\ & E_{ii}-E_{jj} & \\ & & -E_{ii}+E_{jj} \end{pmatrix},\ i<j$$

$$H_{\lambda_i+\lambda_j} = \begin{pmatrix} 0 & & \\ & E_{ii}+E_{jj} & \\ & & -E_{ii}-E_{jj} \end{pmatrix},\ i<j$$

$$H_{\lambda_i} = \begin{pmatrix} 0 & & \\ & E_{ii} & \\ & & -E_{ii} \end{pmatrix}.$$

Lie algebras of type B_n are also written $\mathfrak{so}(2n+1,\mathbb{C})$ since they are the Lie algebras of the Lie groups $SO(2n+1,\mathbb{C})$ - the special orthogonal group.

Type $\mathbf{C_n}$

In this case we consider the Lie algebra of matrices X which satisfy $X^T M + MX = 0$ where $M = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}$. Then

$$L_C = H \bigoplus \sum_{\lambda} \mathbb{C} E_{\lambda}$$

where

$$H = \{X : X = \text{diag}\ (\lambda_1, \cdots, \lambda_n, -\lambda_1, \cdots, -\lambda_n)\}$$

and the roots are $\{\pm\lambda_i \pm \lambda_j, i \neq j; \pm 2\lambda_i\}$. The matrices E_λ are

$$E_{\lambda_i-\lambda_j} = \begin{pmatrix} E_{ij} & 0 \\ 0 & -E_{ji} \end{pmatrix},\ E_{-\lambda_i+\lambda_j} = \begin{pmatrix} E_{ji} & 0 \\ 0 & -E_{ij} \end{pmatrix},\ i<j,$$

$$E_{\lambda_i+\lambda_j} = \begin{pmatrix} 0 & E_{ij}+E_{ji} \\ 0 & 0 \end{pmatrix},\ E_{-\lambda_i-\lambda_j} = \begin{pmatrix} 0 & 0 \\ E_{ij}+E_{ji} & 0 \end{pmatrix},\ i<j,$$

$$E_{2\lambda_i} = \begin{pmatrix} 0 & E_{ii} \\ 0 & 0 \end{pmatrix},\ E_{-2\lambda_i} = \begin{pmatrix} 0 & 0 \\ E_{ii} & 0 \end{pmatrix},$$

and the generators of H are

$$H_{\lambda_i-\lambda_j} = \begin{pmatrix} E_{ii}-E_{jj} & 0 \\ 0 & -E_{ii}+E_{jj} \end{pmatrix},\ i<j,$$

$$H_{\lambda_i+\lambda_j} = \begin{pmatrix} E_{ii}+E_{jj} & 0 \\ 0 & -E_{ii}-E_{jj} \end{pmatrix},\ i<j,$$

$$H_{2\lambda_i} = \begin{pmatrix} E_{ii} & 0 \\ 0 & -E_{ii} \end{pmatrix}.$$

Lie algebras of type C_n are also written $\mathfrak{sp}(n,\mathbb{C})$ since they are the Lie algebras of the Lie groups $Sp(n,\mathbb{C}) = \{g \in SL(2n,\mathbb{C}) : g^T J g = J,$ where $J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}\}$ - the *symplectic* groups.

Type $\mathbf{D_n}$

Here we consider the Lie algebra of matrices X which satisfy $X^T M + MX = 0$ where $M = \begin{pmatrix} 0 & I_n \\ I_n & 0 \end{pmatrix}$. Then

$$L_D = H \bigoplus \sum_\lambda \mathbb{C} E_\lambda$$

where

$$H = \{X : X = \text{diag}\,(\lambda_1, \cdots, \lambda_n, -\lambda_1, \cdots, -\lambda_n)\}$$

and the roots are $\{\pm\lambda_i \pm \lambda_j, i \neq j\}$. In this case,

$$E_{\lambda_i-\lambda_j} = \begin{pmatrix} E_{ij} & 0 \\ 0 & -E_{ji} \end{pmatrix},\ E_{-\lambda_i+\lambda_j} = \begin{pmatrix} E_{ji} & 0 \\ 0 & -E_{ij} \end{pmatrix},\ i<j,$$

$$E_{\lambda_i+\lambda_j} = \begin{pmatrix} 0 & E_{ij}-E_{ji} \\ 0 & 0 \end{pmatrix},\ E_{-\lambda_i-\lambda_j} = \begin{pmatrix} 0 & 0 \\ -E_{ij}+E_{ji} & 0 \end{pmatrix},\ i<j.$$

and the generators of H are

$$H_{\lambda_i-\lambda_j} = \begin{pmatrix} E_{ii}-E_{jj} & 0 \\ 0 & -E_{ii}+E_{jj} \end{pmatrix},\ i<j,$$

$$H_{\lambda_i+\lambda_j} = \begin{pmatrix} E_{ii}+E_{jj} & 0 \\ 0 & -E_{ii}-E_{jj} \end{pmatrix}, \ i<j.$$

Lie algebras of type D_n are also written $\mathfrak{so}(2n,\mathbb{C})$.

The Lie algebras A_n, B_n, C_n and D_n are called the *classical* Lie algebras. The remaining simple Lie algebras E_6, E_7, E_8, F_4 and G_2 are the *exceptional* (or sporadic) Lie algebras. We begin by giving a realisation of G_2.

Type $\mathbf{G_2}$

The Lie algebra G_2 has a realisation as a subalgebra of B_3. We know that

$$H = \{X : X = \text{ diag } (0, \lambda_1, \lambda_2, \lambda_3, -\lambda_1, -\lambda_2, -\lambda_3\}$$

is a Cartan subalgebra of B_3. Let $\widetilde{H}$ be the subalgebra

$$\widetilde{H} = \{X \in H : \lambda_1 + \lambda_2 + \lambda_3 = 0\}.$$

Then G_2 can be seen to be the 14-dimensional Lie algebra generated by the matrices

$$\left.\begin{aligned} G(i) &= \sqrt{2}E_{\lambda_i} + E_{-\lambda_j-\lambda_k} \\ G(i+3) &= \sqrt{2}E_{-\lambda_i} + E_{\lambda_j+\lambda_k} \end{aligned}\right\} \ (i,j,k) = \begin{cases} (1,2,3) \\ (2,3,1) \\ (3,1,2) \end{cases}$$

$$\begin{aligned} G(7) &= E_{\lambda_1-\lambda_2} \\ G(8) &= E_{\lambda_2-\lambda_1} \\ G(9) &= E_{\lambda_1-\lambda_3} \\ G(10) &= E_{\lambda_3-\lambda_1} \\ G(11) &= E_{\lambda_2-\lambda_3} \\ G(12) &= E_{\lambda_3-\lambda_2} \end{aligned}$$

and $\widetilde{H}$. If λ_1, λ_2 are fundamental roots of G_2, then all the roots of G_2 are given by

$$\begin{aligned} &\{\lambda_1, \lambda_2, \lambda_1+\lambda_2, \lambda_1+2\lambda_2, \lambda_1+3\lambda_2, 2\lambda_1+3\lambda_2, \\ &-\lambda_1, -\lambda_2, -\lambda_1-\lambda_2, -\lambda_1-2\lambda_2, -\lambda_1-3\lambda_2, -2\lambda_1-3\lambda_2\}. \end{aligned}$$

In order to find a realisation of the Lie algebra F_4, we need the following definition

Definition B.7. A *representation* of a Lie algebra $\mathfrak{g}$ on a vector space V is a homomorphism $\rho : \mathfrak{g} \to \mathfrak{gl}(V)$.
Thus, in particular, ρ preserves the barcket:

$$\rho[X,Y] = [\rho(X), \rho(Y)].$$

Definition B.8. A representation ρ of $\mathfrak{g}$ is *reducible* if there exists a proper invariant subspace of V under ρ, *i.e.* there exists a subspace V_1 of V such that $V_1 \neq \{0\}$ and $V_1 \neq V$ and

$$\rho(X)V_1 \subseteq V_1 \text{ for all } X \in \mathfrak{g}.$$

Then, $\overline{\rho}(X)\overline{v} = \overline{\rho(X)v}$ is a representation of $\mathfrak{g}$ on V/V_1.

Definition B.9. A representation ρ of $\mathfrak{g}$ is *irreducible* if it is not reducible.

The representation $\rho : X \to \operatorname{ad} X$ $(X \in \mathfrak{g})$ of $\mathfrak{g}$ on itself is called the *regular* (or adjoint) representation. The representation $\rho : X \to 0$ is called the *trivial representation*. Then we have the following theorem of Witt [3]:

Theorem B.5. *Let $\mathfrak{g}$ be an n-dimensional simple Lie algebra with basis $\{X_1,\cdots,X_n\}$. Let V be a vector space of dimension m with basis $\{e_1,\cdots,e_m\}$ and let ρ be an irreducible representation of $\mathfrak{g}$ on V, which is not the trivial or regular representation. Suppose that $Y_i = (y^i_{jk}|_{1\le j,k\le m}$ is the matrix of $\rho(X_i)$ with respect to the basis $\{e_1,\cdots,e_m\}$ of V, for $1 \le i \le n$, and assume that it satisfies the properties*

(a) Y_i is real and skew-symmetric for $1 \le i \le n$.

(b) Tr $(Y_iY_j) = -m\delta_{ij}$.

(c) Tr $(\Sigma_{i,j}(Y_iY_j)^2) = \frac{1}{2}nm^2$.

Then $\mathfrak{g}' = \mathfrak{g} \oplus V$ is a simple Lie algebra with bracket defined by

$$\begin{aligned}
[X_i,X_j] &= \sum_{k=1}^{n} c^k_{ij}X_k, \;\; i,j = 1,\cdots,n \\
-[e_i,X_j] &= [X_j,e_i] = \sum_{k=1}^{m} y^j_{ki}e_k, \;\; 1 \le i \le m, 1 \le j \le n \\
[e_i,e_j] &= \sum_{k=1}^{n} y^k_{ij}X_k, \;\; 1 \le i,j \le m.
\end{aligned}$$

Type $\mathbf{F_4}$

Define the following 16×16 matrices:

$$M_1 = \begin{pmatrix} I_8 & 0 \\ 0 & -I_8 \end{pmatrix}, \; M_2 = \begin{pmatrix} 0 & I_8 \\ I_8 & 0 \end{pmatrix},$$

$$M_3 = \begin{pmatrix} & & 0 & I_4 \\ & & -I_4 & 0 \\ 0 & -I_4 & & \\ I_4 & 0 & & \end{pmatrix},$$

$$M_4 = \begin{pmatrix} 0 & N_2 \\ N_2^T & 0 \end{pmatrix} \text{ where } N_2 = \begin{pmatrix} 0 & I_2 & & \\ -I_2 & 0 & & \\ & & 0 & -I_2 \\ & & I_2 & 0 \end{pmatrix},$$

$$M_5 = \begin{pmatrix} & & N_1 & 0 \\ & & 0 & N_1^T \\ N_1^T & 0 & & \\ 0 & N_1 & & \end{pmatrix}, \text{ where } N_1 = \begin{pmatrix} 0 & 1 & & \\ -1 & 0 & & \\ & & 0 & -1 \\ & & 1 & 0 \end{pmatrix},$$

$$M_6 = \begin{pmatrix} & & N_1^{'} & 0 \\ & & 0 & N_1^{'T} \\ N_1^{'T} & 0 & & \\ 0 & N_1^{'} & & \end{pmatrix}, \text{ where } N_1^{\prime} = \begin{pmatrix} 0 & 1 & & \\ -1 & 0 & & \\ & & 0 & 1 \\ & & -1 & 0 \end{pmatrix},$$

$$M_7 = \begin{pmatrix} 0 & N_3 \\ N_3^T & 0 \end{pmatrix} \text{ where } N_3 = \begin{pmatrix} N_4 & 0 \\ 0 & N_4^T \end{pmatrix} \text{ and } N_4 = \begin{pmatrix} & & & 1 \\ & & -1 & \\ & 1 & & \\ -1 & & & \end{pmatrix}$$

$$M_8 = \begin{pmatrix} 0 & N_5 \\ N_5^T & 0 \end{pmatrix} \text{ where } N_5 = \begin{pmatrix} 0 & N_6 \\ N_6 & 0 \end{pmatrix} \text{ and } N_6 = \begin{pmatrix} & & & 1 \\ & & 1 & \\ & -1 & & \\ -1 & & & \end{pmatrix}$$

and

$$M_0 = M_1 M_2 \cdots M_8.$$

If we define

$$X_i = M_j M_k, \ \ 0 \le j < k \le 8, \ \ 1 \le i \le 36$$

where $i = 1$ corresponds to $j = 0, k = 1$; $i = 2$ to $j = 0, k = 2$, etc., we obtain a set of matrices which satisfies the conditions of Theorem B.5 with $n = 36$. Since Tr $(X_j^2) = -16$, $1 \le i \le 36$, we must choose $m = 16$. Hence, F_4 has dimension 36+26=52. The roots of F_4 are (for basic roots $\lambda_1, \cdots, \lambda_4$),

$$\Sigma = \{\pm\lambda_i, \pm\lambda_i \pm \lambda_j, i \ne j; \frac{1}{2}(\pm\lambda_1 \pm \lambda_2 \pm \lambda_3 \pm \lambda_4)\}.$$

A matrix realization of F_4 is easy to produce in $\mathbb{R}^{52\times 52}$; namely, we can take the regular representation using the expressions at the end of Theorem B.5. Thus the matrices for X_i are

$$\begin{pmatrix} (c_{ij}^k)_{1\le j,k\le 36} & 0 \\ 0 & 0 \end{pmatrix}, \ \ 1 \le i \le 36$$

and for e_i they are

$$\begin{pmatrix} 0 & -(y_{ij}^k)_{1\le j\le 16, 1\le k\le 36} \\ (y_{ij}^k)_{1\le j\le 16, 1\le k\le 36} & 0 \end{pmatrix}, \ \ 1 \le i \le 16.$$

Type $\mathbf{E_8}$

In this case we choose the 128×128 matrices

$$K_1 = \begin{pmatrix} 0 & I_{64} \\ -I_{64} & 0 \end{pmatrix}, \quad K_2 = \begin{pmatrix} & I_{32} & & \\ -I_{32} & & & \\ & & & -I_{32} \\ & & I_{32} & \end{pmatrix},$$

$$K_3 = \begin{pmatrix} L_1 & 0 \\ 0 & L_1^T \end{pmatrix}, \text{ where } L_1 = \begin{pmatrix} & I_{16} & & \\ -I_{16} & & & \\ & & & -I_{16} \\ & & I_{16} & \end{pmatrix},$$

$$K_4 = \begin{pmatrix} L_2 & 0 \\ 0 & L_2^T \end{pmatrix}, \text{ where } L_2 = \begin{pmatrix} & & & I_{16} \\ & & -I_{16} & \\ & I_{16} & & \\ -I_{16} & & & \end{pmatrix},$$

$$K_5 = \begin{pmatrix} 0 & L_3 \\ L_3 & 0 \end{pmatrix}, \text{ where } L_3 = \begin{pmatrix} & & I_{16} & \\ & & & -I_{16} \\ -I_{16} & & & \\ & I_{16} & & \end{pmatrix},$$

$$K_6 = \begin{pmatrix} 0 & L_4 \\ L_4 & 0 \end{pmatrix}, \text{ where } L_4 = \begin{pmatrix} & I_{16} & & \\ -I_{16} & & & \\ & & & I_{16} \\ & & -I_{16} & \end{pmatrix},$$

$$K_7 = \begin{pmatrix} 0 & L_5 \\ L_5 & 0 \end{pmatrix}, \text{ where } L_5 = \begin{pmatrix} & & M_8 & \\ & & & -M_8 \\ -M_8 & & & \\ & M_8 & & \end{pmatrix},$$

$$K_{i+7} = \begin{pmatrix} 0 & L_{i+7} \\ L_{i+7} & 0 \end{pmatrix}, \text{ where } L_{i+8} = \begin{pmatrix} & & M_i & \\ & & & -M_i \\ -M_i & & & \\ & M_i & & \end{pmatrix},$$

for $i = 1, \cdots, 7$ where M_i is as for type F_4 and

$$K_8 = K_1 K_2 \cdots K_7 K_9 K_{10} \cdots K_{15}.$$

Then we define

$$X_i = K_j K_k, \;\; 1 \leq j < k \leq 15, \;\; 1 \leq i \leq 120$$

where $i = 1$ corresponds to $j = 1, k = 2$, etc. Since Tr $(X_j^2) = -128$ for all j, we have dim E_8 = 248. The roots of E_8 are

$$\{\pm\lambda_i \pm \lambda_j, i \neq j, i, j = 1, \cdots, 8; \frac{1}{2}(\pm\lambda_1 \pm \lambda_2 \cdots \pm \lambda_8),$$
$$\text{with an even number of negative signs}\}$$

Type $\mathbf{E_7}$

This is a subalgebra of E_8 with roots

$$\{\pm\lambda_i \pm \lambda_j, i \neq j, i, j = 2, \cdots, 7; \pm(\lambda_1 + \lambda_8); \frac{1}{2}(\varepsilon_1\lambda_1 + \cdots + \varepsilon_8\lambda_8),$$
$$\varepsilon_i = \pm 1, \prod \varepsilon_i = 1, \varepsilon_1 = \varepsilon_8\}$$

and dimension 133. Type $\mathbf{E_6}$ This is a subalgebra of E_8 with roots

$$\{\pm\lambda_i \pm \lambda_j, i \neq j, i, j = 3, \cdots, 7; \pm(\lambda_1 + \lambda_8); \frac{1}{2}(\varepsilon_1\lambda_1 + \cdots + \varepsilon_8\lambda_8),$$
$$\varepsilon_i = \pm 1, \prod \varepsilon_i = 1, \varepsilon_1 = \varepsilon_2 = \varepsilon_8\}$$

and dimension 78.

B.4 Compact Lie Algebras

We shall outline the theory of compact Lie algebras – more details can be found in [4].

Definition B.10. A real Lie algebra $\mathfrak{g}_0$ is called *compact* if one can define a symmetric, negative-definite bi-linear form $B(X,Y)$ on it for which

$$B((\operatorname{ad} A)X, Y) + B(X, (\operatorname{ad} A)Y) = 0, \text{ for all } X \in \mathfrak{g}_0.$$

The Lie group of a compact Lie algebra is compact in the topological sense.

If $\mathfrak{g}$ is a complex Lie algebra, we denote by $\mathfrak{g}^{\mathbb{R}}$ the real Lie algebra obtained from the vector space $\mathfrak{g}$ by restricting to real scalars with the same bracket (which satisfies $[X, iY] = i[X,Y]$. Note that on $\mathfrak{g}^{\mathbb{R}}$, $J = i$ is a isomorphism for which $J^2 = i^2 = -I$. Such a map is called a *complex structure*. Then $\mathfrak{g}$ is said to have a *real form* $\mathfrak{g}_0$ if

$$\mathfrak{g}^{\mathbb{R}} = \mathfrak{g}_0 \oplus J\mathfrak{g}_0$$

(vector space direct sum). Every element $Z \in \mathfrak{g}$ can be written

$$Z = X + JY = X + iY, \;\; X, Y \in \mathfrak{g}_0$$

and so $\mathfrak{g}$ is isomorphic to the complexification of $\mathfrak{g}_0$. The map

$$\sigma : X + iY \to X - iY, \;\; X, Y \in \mathfrak{g}_0$$

is called the *conjugation* of $\mathfrak{g}$ with respect to $\mathfrak{g}_0$. A direct sum decomposition

$$\mathfrak{g}_0 = \mathfrak{k}_0 + \mathfrak{p}_0$$

where $\mathfrak{t}_0$ is a subalgebra and $\mathfrak{p}_0$ is a vector subspace is called a *Cartan decomposition* if the complexification $\mathfrak{g}$ of $\mathfrak{g}_0$ has a compact real form $\mathfrak{g}_k$ such that

$$\sigma\mathfrak{g}_k \subseteq \mathfrak{g}_k, \;\; \mathfrak{t}_0 = \mathfrak{g}_0 \cap \mathfrak{g}_k, \;\; \mathfrak{p}_0 = \mathfrak{g}_0 \cap (i\mathfrak{g}_k).$$

Lemma B.1. *(a) A real Lie algebra is compact if and only if its Killing form is strictly negative definite (and hence the Lie algebra is necessarily semi-simple).*

(b) Every compact Lie algebra $\mathfrak{g}$ is a direct sum

$$\mathfrak{g} = \mathfrak{z} + [\mathfrak{g}, \mathfrak{g}],$$

where $\mathfrak{z}$ is the centre of $\mathfrak{g}$ and the ideal $[\mathfrak{g}, \mathfrak{g}]$ is compact (and semi-simple).

Lemma B.2. *Let $\mathfrak{g}_0$ be a real semi-simple Lie algebra which is a direct sum $\mathfrak{t}_0 + \mathfrak{p}_0$ where $\mathfrak{t}_0$ is a subalgebra and $\mathfrak{p}_0$ is a vector subspace. Then the following statements are equivalent:*

(a) $\mathfrak{g}_0 = \mathfrak{t}_0 + \mathfrak{p}_0$ is a Cartan decomposition of $\mathfrak{g}_0$.

(b) $B(T,T) < 0$ for all $T \neq 0$ in $\mathfrak{t}_0$, $B(X,X) < 0$ for all $X \neq 0$ in $\mathfrak{p}_0$ and the mapping $s : T + X \to T - X$, $T \in \mathfrak{t}_0, X in \mathfrak{p}_0$ is an automorphism.

References

1. Carter, R.: Lie Algebras of Finite and Affine Type. Cam. Univ.Press, Cambridge (2005)
2. Jacobson, N.: Lie Algebras. Interscience Tracts in Pure and Applied Mathematics, vol. 10. Wiley, Chichester (1962)
3. Witt, E.: Spiegelungsgruppen und Aufzahlung halbeinfacher Liescher Ringe. Abh. Math. Sem. Univ. Hamburg 14, 289–337 (1941)
4. Helgason, S.: Differential Geometry and Symmetric Spaces. Academic Press, New York (1962)

Appendix C
Differential Geometry

C.1 Differentiable Manifolds

There are many excellent books on differential geometry and vector bundles and so we give only a brief outline of the ideas we use in the book (see [1,2,3] for further details).

Definition C.1. An *n-dimensional (topological) manifold M* is a Hausdorff space which is locally homeomorphic to $\mathbb{R}^n$.

This means that, for each $x \in M$, there exists a neighbourhood U of x and a homeomorphism $\varphi : U \to \mathbb{R}^n$ called a *local coordinate map*. If U, V are open sets in M such that $U \cap V \neq \emptyset$ and $(U, \varphi), (V, \psi)$ are local coordinate maps then $\psi \circ \varphi^{-1}$ and $\varphi \circ \psi^{-1}$ are homeomorphisms of $\mathbb{R}^n$.

Definition C.2. If $\mathfrak{U} = \{(U_i, \varphi_i)\}_{i \in I}$ is an open covering of M by local coordinate charts such that the maps $\varphi_i \circ \varphi_j^{-1}$ are C^r on the intersections of their domains, then M is called a *C^r-differentiable manifold* and $\mathfrak{U}$ is called a C^r *differentiable structure* for M.

(If $r = \infty$ we simply call M a *differentiable manifold* and if $r = \omega$ we call M an *analytic manifold*. Similar definitions apply if we use $\mathbb{C}^n$ instead of $\mathbb{R}^n$, in which case we obtain complex manifolds.) To emphasise the dimension of a manifold M we write M_n.

Example C.1. Let

$$\mathbb{P}^n = \mathbb{P}(\mathbb{R}^n) = \{[x] : \ x \in \mathbb{R}^{n+1},\ x \neq 0 \text{ and } [x] \text{ is the line through } x\}.$$

Define the open sets

$$U_i = \{[x_1, \cdots, x_{n+1}] : \ x_i \neq 0\}$$

and the maps

$$\varphi_i([x]) = \left(\frac{x_1}{x_i}, \frac{x_2}{x_i}, \cdots, \widehat{x_i}, \cdots, \frac{x_{n+1}}{x_i} \right).$$

Then, on $U_i \cap U_j$, we have

$$\varphi_i \circ \varphi_j^{-1}(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$$

where

$$y_k = \begin{cases} \frac{x_k}{x_i}, & k \neq j \\ \frac{1}{x_i}, & k = j \end{cases}$$

(provided $i \neq j$). The manifold $\mathbb{P}^n$ is called the n^{th} *real projective space*.

Definition C.3. If $f : M \to N$ is a continuous map between smooth manifolds M_m and N_n such that if $(U, \varphi), (V, \psi)$ are coordinate neighbourhoods of some point $p \in M$ and $f(p) \in N$, then f is said to be C^∞ at p if

$$\psi \circ f^{-1} \circ \varphi : \mathbb{R}^m \to \mathbb{R}^n$$

is C^∞ at the point $\varphi(p)$. If f is C^∞ at all points of M, then we say that f is C^∞ or *smooth*.

Definition C.4. If $f : M \to M$ is smooth and $f^{-1} : M \to M$ exists and is smooth, then f is called a *diffeomorphism*.

C.2 Tangent Spaces

Let M be an n-dimensional manifold and let (U, φ) be a local coordinate system at $p \in M$. If $I \subseteq \mathbb{R}$ denotes an interval containing 0, then a *smooth curve* on M (at p) is a map $\alpha : I \to M$ with $\alpha(0) = p$. Two curves α_1 and α_2 are *equivalent* if $(\varphi \circ \alpha_1)'(0) = (\varphi \circ \alpha_2)'(0)$. This is clearly an equivalence relation and is independent of the chart φ. The set of all equivalence classes of smooth curves at p is called the *tangent space* of M at p and is denoted by T_pM. It is clearly isomorphic (as an n-dimensional vector space) to $\mathbb{R}^n$.

Let $f : M_m \to N_n$ be a smooth function. We define the *differential* df_p of f at $p \in M$ by

$$df_p([\alpha]) = [f \circ \alpha]$$

where [ ] denotes an equivalence class of curves. Clearly, df_p is represented in local coordinates (φ, ψ) by the Jacobian matrix of $\psi \circ f \circ \varphi^{-1}$. If (U, φ) is a chart at $p \in M$, then we denote by $\left(\frac{\partial}{\partial x_i} \right)_p$ the image of $e_i = (0, 0, \cdots, 1, 0, \cdots, 0) \in \mathbb{R}^n$ (*i.e.* the standard i^{th} basis vector) under the map

$$d\varphi^{-1}_{\varphi(p)} : \mathbb{R}^n \to T_pM.$$

Then any tangent vector $X_p \in T_pM$ can be written as

$$X_p = \sum_{i=1}^{n} a_i \left(\frac{\partial}{\partial x_i}\right)_p$$

where $a = (a_1, \cdots, a_n)$ is given by

$$a = (\varphi \circ \alpha)'(0),$$

where $X_p = [\alpha]$.

The *cotangent space* of M at p is the dual space $(T_pM)^*$ of $T_p(M)$. Note that $(T_pM)^*$ is usually written as (T_p^*M).

C.3 Vector Bundles

A (smooth) *real vector bundle E* of rank m on a differentiable manifold M_n is a differentiable manifold together with a smooth projection $\pi : E \to M$ such that for each $p \in M$ the set $\pi^{-1}(p)$ has the structure of a (real) m-dimensional vector space and there exists a neighbourhood U of p in M and a homeomorphism $\varphi : U \times \mathbb{R}^m \to \pi^{-1}(U)$ such that the map $v \to \varphi(q, v)$ is a vector space isomorphism for each $q \in U$. (The latter condition says that the bundle is *locally trivial*.)

Example C.2. The *tangent bundle TM* of a differentiable manifold M is the (disjoint) union of all tangent spaces to M:

$$TM = \cup_{p \in M} T_p(M)$$

with the obvious differentiable structure and projection.

Example C.3. The cotangent bundle is the disjoint union of the dual spaces $\cup_{p \in M}(T_p^*M)$ again with the obvious differentiable structure and projection.

Example C.4. The normal bundle to $M \subseteq \mathbb{R}^k$, $(k \geq n)$ is the union of all spaces

$$N_pM = \{v \in \mathbb{R}^k : \ v \perp T_pM\}$$

with the structure induced from the tangent bundle.

Definition C.5. A *section* of a vector bundle $\pi : E \to M$ is a smooth map $s : M \to E$ such that

$$\pi \circ s = \mathrm{id}_M,$$

i.e. such that $s(p) \in \pi^{-1}(p)$ for each $p \in M$. The set of all sections of a vector bundle $\pi : E \to M$ is usually denoted by $\Gamma(E)$.

C.4 Exterior Algebra and de Rham Cohomology

Definition C.6. Let V be an n-dimensional vector space. The vector space $\bigwedge^r V$, $(0 \le r \le n)$ is defined to be the collection of all linear combinations of products of the form $v_1 \wedge v_2 \wedge \cdots \wedge v_r$ $(v_i \in V)$ subject to the relations

$$v \wedge w = -w \wedge v \text{ for any } v, w \in V,$$
$$(v \wedge w) \wedge x = v \wedge (w \wedge x), \text{ for any } v, w, x \in V$$

with the obvious linear structure.

If $\{e_i\}_{1 \le i \le n}$ is a basis of V, then any element $\xi \in \bigwedge^r V$ can be written

$$\xi = \sum_{i_1 < \cdots < i_r} \xi^{i_1 \cdots i_r} e_{i_1} \wedge \cdots \wedge e_{i_r}.$$

Note that $\dim(\bigwedge^r V) = \begin{pmatrix} n \\ r \end{pmatrix}$.

Definition C.7. If M is an n-dimensional smooth manifold, the bundle of exterior r-forms on M is defined as

$$\bigwedge^r(M) = \bigcup_{p \in M} \bigwedge^r(T_p^* M).$$

The space of smooth sections of $\bigwedge^r(M)$ is denoted by

$$A^r(M) = \Gamma(\bigwedge^r(M))$$

and is called the space of *(exterior) r-forms* on M. The space of *exterior differential forms* on M is the set

$$A(M) = \sum_{r=0}^{n} A^r(M).$$

Clearly, we have a map

$$\bigwedge : A^r(M) \times A^s(M) \to A^{r+s}(M)$$

given by

$$\bigwedge(\omega_1, \omega_2)(p) = \omega_1(p) \wedge \omega_2(p).$$

In local coordinates (U, φ), $x = \varphi(p)$, we have

$$\omega = \sum_{\mathbf{i}} \alpha_{i_1 \cdots i_r} dx_{i_1} \wedge \cdots \wedge dx_{i_r}$$

for any r-form ω.

The most important operation on $A(M)$ is the (unique) differential operator d called the *exterior derivative* - it has the properties:

(a) $d(\omega_1 + \omega_2) = d\omega_1 + d\omega_2,\ \ \omega_1, \omega_2 \in A(M)$.

(b) If ω_1 is an r-form, then

$$d(\omega_1 \wedge \omega_2) = d\omega_1 \wedge \omega_2 + (-1)^r \omega_1 \wedge d\omega_2.$$

(c) If f is a smooth function on M, then df is the differential of f.

(d) If f is a smooth function, then $d(df) = 0$.

In fact, these properties uniquely specify d.

Therefore we have the differential complex

$$\cdots \xrightarrow{d} A^{k-1}(m) \xrightarrow{d} A^k(m) \xrightarrow{d} A^{k+1}(m) \xrightarrow{d} \cdots$$

for an n-dimensional manifold M, where $A^k(M) = 0$ if $k > n$. Clearly, $A^0(M) =$ set of smooth functions on M.

Definition C.8. The r^{th} *de Rham cohomology group* of M is defined as

$$H^r(M) = \text{Ker}\ (d : A^r(M) \to A^{r+1}(M))/\text{Im}\ (d : A^{r-1}(M) \to A^r(M)),$$

It is a topological invariant and, in fact, also a homotopy invariant.

C.5 Degree and Index

Let $f : M_n \to N_n$ be a smooth map between n-manifolds M and N. We define the *degree* of f to be the unique number $\deg(f)$ which satisfies the equation

$$\int_M f^*(\omega) = \deg(f) \int_N \omega,\ \ \omega \in \bigwedge^n (N),$$

where f^* denotes the pull-back of f (this is essentially the change of coordinates in an integral). Note that $\deg(f)$ is a homotopy invariant and satisfies the functorial condition

$$\deg(fg) = \deg(f)\deg(g).$$

Now let $f : M_n \to N_n$ be a smooth map. We say that $p \in N$ is a *regular value* of f if $d_q f : T_q M \to T_p N$ is surjective for all $q \in f^{-1}(p)$. Regular values of f are dense in N by Sard's theorem [3].

Now, if $f : M_n \to N_n$ is a smooth map and $p \in N$ is regular, then if $q \in f^{-1}(p)$, then index of f at q is given by

$$\text{ind}\ (f)_q = \begin{cases} 1 \text{ if } \det(d_q f) = +1 \\ -1 \text{ if } \det(d_q f) = -1 \end{cases}.$$

Then we have

$$\deg(f) = \sum_{q\in f^{-1}(p)} \operatorname{ind}(f)_q.$$

Next consider a vector field $f \in C^\infty(U;\mathbb{R}^n)$ on an open subset of $\mathbb{R}^n$), such that $0 \in U,\ \ f(0) = 0$ and $f(x) \neq 0$ (*i.e.* f has an isolated singularity at 0). Then we define a map $f_\rho : S^{n-1} \to S^{n-1}$ by

$$f_\rho(x) = \frac{f(\rho x)}{\|f(\rho x)\|}$$

for $\rho > 0$ sufficiently small. Then the homotopy class of f_ρ is independent of ρ and we define

$$\operatorname{ind}(f)_0 = \deg(f_\rho).$$

Note that deg is functorial in the sense that

$$\operatorname{ind}(\psi_* f)_0 = \operatorname{ind}(f)_0$$

for any diffeomorphism ψ. (Here $(\psi_* f)(q) = d_p\psi(f(p)),\ \ \psi(p) = q$, where d is the Jacobian derivative.) Hence, for any vector field X on a manifold M, we can define the index of a singularity of the vector field at $p \in M$ by

$$\operatorname{ind}(X)_p = \operatorname{ind}(\,\varphi_* X|_U)_0$$

where φ is a local coordinate system at p where $p \in U \subseteq M$.

The *index* of a vector field on a compact subset $K \subseteq M$ is just the sum of all the indices at all the singularities in K (assuming these are finite in number).

Lemma C.1. *Let $f : M_{n+1} \to N_n$ be a smooth map, where M and N are oriented, compact and connected manifolds, and let K be a compact subset of M with smooth (n-dimensional) boundary $B = \partial K$ such that*

$$B = B^1 \cup B^2 \cup \cdots \cup B^k$$

is a disjoint union of submanifolds of M, then

$$\sum_{i=1}^{k} deg(f_i) = 0$$

where $f_i = f|_{B^i}$.

Proof. We have

$$\deg(\,f|_B) = \sum_{i=1}^{k} deg(f_i),$$

and if $\omega \in \bigwedge^n(N)$ with $\int_N \omega = 1$, then

$$\deg(\,f|_B) = \int_B f^*(\omega) = \int_K df^*(\omega) = \int_K f^*(d\omega) = 0,$$

by Stoke's theorem. □

It follows from the lemma that if M_n is a compact manifold and X is a smooth vector field on M_n containing a discrete (finite) set of singularities at $p_i \in M, \; 1 \le i \le k$, then the index of X is given by

$$\text{ind}\, X = \sum_{i=1}^{k} \text{ind}\, (X)_{p_i}.$$

(To prove this, just remove 'small' balls around each p_i and apply the theorem.)

If M_n is a smooth manifold embedded in $\mathbb{R}^{n+k}$, then we define a *tubular neighbourhood* T_ε of M_n in $\mathbb{R}^{n+k}$ to be an open set containing M in its interior, has a smooth boundary and such that, for $x \in T_\varepsilon$, there exists $y \in M$ such that $\|x - y\| < \varepsilon$.

Definition C.9. The *Gauss map* $G : \partial T_\varepsilon \to S^{n+k-1}$ is the map which takes a point on the boundary of T_ε to its outward pointing unit normal.

Then by the above lemma we can see that, for any vector field X on M_n, we have

$$\text{ind}\, X = \deg G,$$

so that the index is independent of the vector field and is just a property of the topology of M.

In order to prove the Poincaré-Hopf theorem, therefore, we can consider any vector field X on a compact manifold M. This theorem states that

$$ind(X) = \chi(M) = \text{ Euler characteristic of } M.$$

We shall choose a gradient vector field on M. To illustrate the idea consider the oriented, two-dimensional case. Then M is a surface of some genus, say 2. A gradient vector field is shown in Figure C.1. Note that it has four saddle points and two nodes, one stable and one unstable. The gradient field is so-called because it is essentially

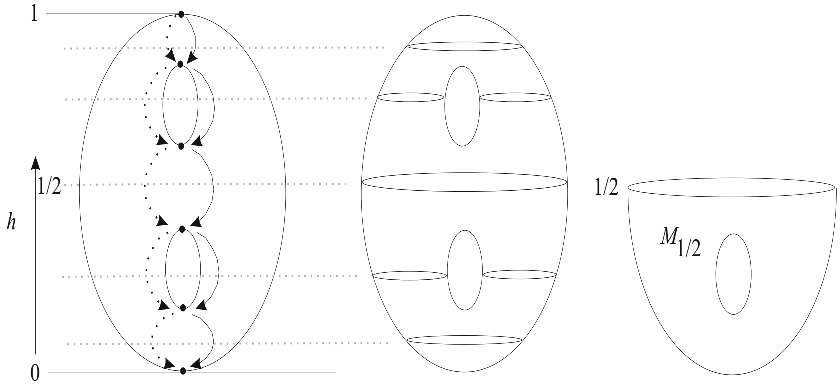


Fig. C.1 A gradient vector field on a 2-manifold

the gradient of the height function h as in Figure C.1. At the singular points, h is locally a quadratic form which is non-degenerate and has signature (*i.e.* the number of negative eigenvalues of the Hessian matrix) equal to the index. The theorem then follows by some elementary algebraic topology. The general case is an extension of these ideas (called *Morse theory*). The important point is how the topology of the level set of h changes as we pass a singular point. In the two-dimensional case, we see that between level sets which contain a single singularity, we either get a disk or a 'pair of pants'.

In the general case, some standard analysis shows that, for a compact manifold M, we can find a *Morse function* h which is similar to the two-dimensional case, in that it is a 'height function' and has only a finite number of non-degenerate critical values (*i.e.* the Hessian matrix is non-singular). Let

$$M_a = \{x \in M : \ f(x) < a\},$$

i.e. the manifold with boundary consisting of the subset of M which is 'below' a. (For example, $M_{1/2}$ is shown in Figure C.1.) If there are no critical values of H in the interval $[a_1, a_2]$ then clearly M_{a_1} and M_{a_2} are diffeomorphic. Suppose that in $[a_1, a_2]$ there is a single critical values with one critical point p. Then we can find a small neighbourhood U of p in M such that U is diffeomorphic to an open contractible set in $\mathbb{R}^n$ and

$$U \cap M_{a_1} \cong S^{\ell-1} \times V$$

where $V \subseteq \mathbb{R}^{n-\ell+1}$ is open and contractible and

$$M_{a_2} \text{ is diffeomorphic to } U \cup M_{a_2}.$$

Then, elementary algebraic topology shows that

$$\chi(M_{a_2}) = \chi(M_{a_1}) + (-1)^{\ell},$$

and the index theorem follows from this.

C.6 Connections and Curvature

Sections of vector bundles on manifolds, such as the tangent bundle, take values in different spaces (*i.e.* the fibres of the bundles) and so it is not possible to differentiate such a section s directly. This is because a standard difference quotient $(s(t+\delta t) - s(t))/\delta t$ is not defined since the values $s(t+\delta t)$ and $s(t)$ are in different vector spaces. In order to overcome this difficulty we need a connection on the manifold, which relates different fibres of the bundle.

Definition C.10. A *connection* D on a vector bundle $\pi : E \to M$ is a map

$$D : \Gamma(E) \to \Gamma(T^*(M) \otimes E)$$

such that

(1) $D(s_1+s_2)=D(s_1)+D(s_2)$, for all $s_1,s_2\in\Gamma(E)$,
(2) $D(fs)=df\otimes s+fDs$, for all $s\in\Gamma(E)$, $f\in C^\infty(M)$.

Let $e_1,\cdots,e_k\in\Gamma(E)$ be sections such that $e_1(p),\cdots,e_k(p)$ is a basis of E for each $p\in U$ (= a trivialising neighbourhood). Then the elements of $T^*(M)\otimes E$ can be written as $\sum f_i\otimes e_i$ for some $f_i\in T^*(U)$ and so

$$D(e_i)=\sum_{j=1}^{k}\omega_{ij}\otimes e_j \tag{C.1}$$

where $\omega_{ij}\in T^*(U)$ is a matrix of one-forms, called the connection form with respect to $\{e_1,\cdots,e_k\}$. We write $\omega=(\omega_{ij})$. If $\langle\cdot,\cdot\rangle$ is the pairing between $T(M)$ and $T^*(M)$, then we define D_Xs for any $X\in T(M)$ and any section $s\in\Gamma(E)$ by setting

$$D_Xe_i=\langle X,\sum_{j=1}^{k}\omega_{ij}\otimes e_j\rangle=\sum_{j=1}^{k}\langle X,\omega_{ij}\rangle e_j$$

and extending by linearity. D_Xs is called the *covariant derivative of s along X*. Note that D_X satisfies the properties:

(a) $D_{fX}s=fD_Xs$, $f\in C^\infty(M)$.
(b) $D_{X+Y}s=D_Xs+D_Ys$.
(c) $D_X(s_1+s_2)=D_Xs_1+D_Xs_2$.
(d) $D_X(fs)=(Xf)s+fD_Xs$, $f\in C^\infty(M)$.

Definition C.11. Let D be a connection on a differentiable bundle $\pi:E\to M_n$ of rank m and let (U,φ) be a local coordinate system on M, with coordinates x^i, $1\le i\le m$. Let s_j, $1\le j\le m$ be m smooth sections which are linearly independent on U. Then $T_p^*\otimes E_p$ has a basis $dx^i\otimes s_j$ at each point $p\in U$. The set of sections $\{s_j\}_{1\le j\le m}$ is called a *local frame* for E on U.

We have

$$Ds_j=\sum_{i,k}\Gamma_{ji}^{k}dx^i\otimes s_k,$$

since Ds_j is a local section of $T^*(M)\otimes E$, for some constants Γ_{ji}^k. From (C.1) we have

$$Ds_j=\sum\omega_{ji}\otimes s_i \tag{C.2}$$

so that, locally,

$$\omega_{ij}=\sum\Gamma_{jk}^{i}dx^k.$$

We can write (C.2) in the matrix form

$$DS=\omega\otimes S;$$

then ω is called the *connection matrix*. Since it depends on the local frame, we can check what happens if we change the frame: in that case, the new frame is

$$S' = PS$$

for some invertible matrix P, whose elements are smooth functions on U. By the definition of a connection, we have

$$\begin{aligned} DS' &= dP \otimes S + P \cdot DS \\ &= (dP + P \cdot \omega) \otimes S \\ &= (dP \cdot P^{-1} + P \cdot \omega \cdot P^{-1}) \otimes S' \end{aligned}$$

and so the connection matrices are related by

$$\omega' = dP \cdot P^{-1} + P\omega P^{-1}. \tag{C.3}$$

From this it can be seen that any vector bundle has a connection. If we (exterior) differentiate (C.3), we obtain

$$d\omega' \cdot P - \omega' \wedge dP = dP \wedge \omega + P \cdot d\omega$$

and since

$$dP = \omega' \cdot P - P \cdot \omega,$$

we have

$$(d\omega' - \omega' \wedge \omega') \cdot P = P \cdot (d\omega - \omega \wedge \omega)$$

i.e.

$$\Omega' = P\Omega P^{-1}$$

where

$$\Omega = d\omega - \omega \wedge \omega$$

is the *connection matrix* of D on U. We can show that Ω satisfies the *Bianchi identity*

$$d\Omega = \omega \wedge \Omega - \Omega \wedge \omega.$$

If we think of $\Gamma(T^*(M) \otimes E)$ as the space of sections of vector-valued one-forms, we will write it as $\Gamma^1(E)$. Similarly, the space of sections of vector-valued r-forms will be denoted by $\Gamma^r(E)$. Then it can be shown that we can extend the definition of a connection to a map

$$\Gamma^{r-1}(E) \xrightarrow{D} \Gamma^r(E),$$

and so we get a differential complex

$$0 \to \Gamma^0(E) \xrightarrow{D} \Gamma^1(E) \xrightarrow{D} \Gamma^2(E) \xrightarrow{D} \cdots .$$

This complex is not exact, in general, so that

$$D \circ D \neq 0.$$

We write

$$F = D \circ D.$$

Then F is called the (global) *curvature form*; it is the global form of Ω. The above sequence is exact precisely when the connection is *flat*, *i.e.* $F = 0$. Not every vector bundle has a flat connection. Note that, in the case of complex line bundles, the connection matrix is one-dimensional, so the wedge product commutes and the Bianchi identity becomes

$$dF = 0.$$

C.7 Characteristic Classes

In this section we shall outline the theory of characteristic classes which relates the topology of a bundle to the cohomology groups of the base manifold. If G is a Lie group with Lie algebra $\mathfrak{g}$, let $S(\mathfrak{g}) = \oplus_{r \geq 0} S^r(\mathfrak{g})$ denote its symmetric algebra, where

$$S^r(\mathfrak{g}) = T^r(\mathfrak{g})/I$$

and

$$T^r(\mathfrak{g}) = \mathfrak{g} \otimes \cdots \otimes \mathfrak{g}$$

is the r^{th} tensor algebra of $\mathfrak{g}$ and I is the ideal generated by elements of the form

$$A_1 \otimes \cdots \otimes A_r - A_{\sigma(1)} \otimes \cdots \otimes A_{\sigma(r)},$$

for all permutations σ. An element $P \in S^r(\mathfrak{g})$ is *G-invariant* if it is invariant under the adjoint action of G, *i.e.*

$$P(\mathrm{Ad}_g A_1, \cdots, \mathrm{Ad}_g A_r) = P(A_1, \cdots, A_r)$$

where $\mathrm{Ad}_g A_i = g^{-1} A_i g$, for $g \in G$. Let $\mathscr{I}^r(G)$ denote the G-invariant elements of $S^r(\mathfrak{g})$. We can define a multiplication map

$$\mu : \mathscr{I}^{r_1}(G) \otimes \mathscr{I}^{r_2}(G) \to \mathscr{I}^{r_1+r_2}(G)$$

by

$$\mu(PQ)(A_1, \cdots, A_{r_1+r_2}) = \frac{1}{(r_1+r_2)!} \sum_{\sigma} P(A_{\sigma(1)}, \cdots, A_{\sigma(r_1)}) Q(A_{\sigma(r_1+1)}, \cdots, A_{\sigma(r_1+r_2)})$$

making $\mathscr{I}(G) = \oplus_{r \geq 0} \mathscr{I}^r(G)$ an algebra.

Now, if $\pi : E \to M$ is a smooth (complex) vector bundle, we can extend the above ideas to $\mathfrak{g}$-valued p-forms on M as follows. If $\mu_i \in A^{p_i}(M)$ (*i.e.* μ_i are p_i-forms on M) and $A_i \in \mathfrak{g}$, $1 \leq i \leq r$, then we define

$$P(A_1\mu_1,\cdots,A_r\mu_r)=\mu_1\wedge\cdots\wedge\mu_r P(A_1,\cdots,A_r).$$

The most useful case is when $A_1=\cdots=A_r$. Then we have invariant polynomials

$$\mathscr{P}(A)=P(A,A,\cdots,A).$$

If D, D_1 and D_2 are connections on the bundle E with corresponding curvatures F, F_1 and F_2, then it can be shown that

$$d\mathscr{P}(F)=0$$

and that

$$\mathscr{P}(F_1)-\mathscr{P}(F_2)$$

is exact (*i.e.* it is equal to $d\psi$ for some 1-form ψ), for any invariant polynomial $\mathscr{P}$. Since $d\mathscr{P}(F)=0$ for any connection D with corresponding curvature F, it follows that $\mathscr{P}(F)$ is closed and so defines a cohomology class in $H^*(M;\mathbb{C})$ (the de Rham cohomology algebra). Moreover, by the above remarks, it is independent of the connection.

In particular, if we define

$$p(F)=\det\left(1+\frac{i}{2\pi}F\right)$$

(where the det is expanded as normal, except that we obtain sums of even forms), then

$$p(F)=1+p_1(F)+p_2(F)+\cdots$$

and we write

$$c_i(F)=[p_i(F)]\in H^{2i}(M;\mathbb{C})$$

and

$$c(F)=1+[p_1(F)]+[p_2(F)]+\cdots=1+c_1+c_2+\cdots.$$

Then we call $c(F)$ the *total Chern class* and $c_i(F)$ the i^{th} *Chern class* of the bundle E. If E is a rank k vector bundle, then

$$\begin{aligned}
c_0(F)&=1\\
c_1(F)&=\frac{i}{2\pi}\text{tr}\,F\\
c_2(F)&=-\frac{1}{8\pi^2}[\text{tr}\,F\wedge\text{tr}\,F-\text{tr}\,(F\wedge F)]\\
&\vdots\\
c_k(F)&=\left(\frac{i}{2\pi}\right)^k\det F.
\end{aligned}$$

Note that the Chern classes are natural in the sense that

$$c(f^*E) = f^*c(E)$$

for any smooth map $f : N \to M$ and if $E \oplus F$ is a Whitney sum bundle, then

$$c(E \oplus F) = c(E) \wedge c(F).$$

For a sum of n complex lines bundles

$$E = L_1 \oplus \cdots \oplus L_n$$

we have

$$c(E) = c(L_1) \cdots c(L_n)$$

so the Chern class cannot distinguish between an n-dimensional vector bundle and a sum of line bundles (this is the *splitting principle*).

Many other types of characteristic classes exist such as Pontryagin classes (for detemining if 2-manifolds bound another), Steifel-Whitney classes which are useful for determining embeddings of projective and similar spaces, etc. For more details see [4,5].

References

1. Helgason, S.: Differential Geometry and Symmetric Spaces. Academic Press, New York (1962)
2. Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry, vol. 1, 2. Wiley, New York (1963)
3. Spivak, M.: A Comprehensive Introduction to Differential Geometry, vol. 1-5. Publish or Perish, New York (1970)
4. Milnor, J.W., Stasheff, J.D.: Characteristic Classes. PUP, Princeton (1974)
5. Moore, J.D.: Lectures on Seiberg-Witten invariants. Springer, New York (1996)

Appendix D
Functional Analysis

D.1 Banach and Hilbert Spaces

In the first four sections we give an introduction to functional analysis. More details can be found in [1,2,3].

Definition D.1. A *real (complex) Banach space* X is a complete normed vector space over the field of scalars $\mathbb{R}$ (or $\mathbb{C}$) such that the following axioms for the norm are satisfied:

(a) $\|x\| = 0$ if and only if $x = 0$.
(b) $\|\alpha x\| = |\alpha|\|x\|$, for all $\alpha \in \mathbb{R}(\mathbb{C})$, $x \in X$.
(c) $\|x+y\| \leq \|x\| + \|y\|$, for all $x, y \in X$.

If X is a Banach space and M is a closed subspace of X we define the quotient space

$$\widetilde{X} \overset{\Delta}{=} X/M = \{\overline{x} : \overline{x} = x + M\}.$$

Thus, $\widetilde{X}$ is the set of all affine subspaces parallel to M. It is clearly a linear space, and is, in fact, a Banach space under the norm

$$\|x\| = \inf_{y \in \overline{x}} \|y\| = \inf_{m \in M} \|x - m\| = \text{ dist } (x, M).$$

Definition D.2. A *real (complex) Hilbert space* H is a Banach space the norm of which is defined by an inner product $\langle \cdot, \cdot \rangle$ which is linear in the first slot and conjugate linear in the second. In order to emphasise the space H we sometimes write the inner product as $\langle \cdot, \cdot \rangle_H$.

The structure theory of Hilbert spaces is based on the parallelogram law and the projection onto closed convex subspaces which are described by the following two lemmas.

Lemma D.1. *In any Hilbert space H we have*

$$\|x+y\|^2+\|x-y\|^2=2\|x\|^2+2\|y\|^2,$$

for all $x,y \in H$.

Lemma D.2. *If A is a closed and convex subset of a Hilbert space H, then there exists a unique vector $x \in A$ such that*

$$\|x-h\| \leq \|y-h\|, \quad \textit{for all } y \in A.$$

Definition D.3. Two subsets S_1, S_2 of H are *orthogonal* (written $S_1 \perp S_2$) if

$$\langle h_1, h_2 \rangle = 0$$

for all $h_i \in S_i, \ i=1,2$.

Lemma D.3. *If $E \subseteq H$ is closed, then every element $h \in H$ has a unique representation as $h=h_1+h_2$ where $h_1 \in E$ and $h_2 \perp E$.*

Definition D.4. A set of vectors $\{e_i\}_{i \in I} \subseteq H$ if *orthonormal* if

$$\langle e_i, e_j \rangle = \delta_{ij},$$

for all $i,j \in I$. The set $\{e_i\}_{i \in I}$ is a *maximal orthonormal family* if it is not a proper subset of another orthonormal set.

Definition D.5. Let S be any subset of a Hilbert space H. We define $\overline{\text{span}(S)}$ to be the smallest closed subspace of H containing S. It is clearly the closure of the set of all finite linear combinations of elements of S.

Every Hilbert space contains a maximal orthonormal family $\{e_i\}_{i \in I}$ such that

$$H = \overline{\text{span}(\{e_i\}_{i \in I})}.$$

Such a family is called an (orthonormal) basis of H and if the index set I is countable, we say that H is *separable*. Most Hilbert spaces which appear in applications are separable.

Theorem D.1. *If H is a separable Hilbert space and $\{e_i\}_{i \in I}$ is a basis,then we have*
(a) $\sum_{i=1}^{k} |\langle x, e_i \rangle|^2 \leq \|x\|^2$, for all $x \in H$ and $k=1, \ cdots, \infty$.
(b) $x=\sum_{i=1}^{\infty} |\langle x, e_i \rangle| e_i$ and

$$\|x\|^2 = \sum_{i=1}^{\infty} |\langle x, e_i \rangle|^2.$$

The inequality in (a) is called *Bessel's inequality* and the second equality in (b) is called *Parseval's relation*.

Definition D.6. Let X be a (real or complex) Banach space. The *dual space* X^* of X is the linear space of all continuous linear forms on X, *i.e.* linear maps $x^* : X \to \mathbb{R}$ (or $\mathbb{C}$). X^* is a Banach space under the norm

$$\|x^*\| = \sup_{\|x\|=1} |\langle x^*, x\rangle|.$$

Definition D.7. The Banach space X is *reflexive* if the map

$$J : X \to X^{**}$$

defined by

$$\langle J(x), x^*\rangle = \langle x^*, x\rangle$$

is onto X^{**}, where $\langle x^*, x\rangle = x^*(x)$.

Lemma D.4. *If H is a Hilbert space and $h^* \in H^*$, then there exists a unique element $h \in H$ such that*

$$\langle h^*, x\rangle = \langle x, h\rangle$$

for all $h \in H$, and $\|h^*\| = \|h\|$.

Two of the main classical results of functional analysis are the following.

Theorem D.2. *(Hahn-Banach) Let M be a proper closed subspace of a Banach space X. If $m^* \in M$ is a linear form defined on M then it can be extended to a linear form $x^* \in X^*$ such that $\|x^*\| = \|m^*\|$.*

Theorem D.3. *(Banach-Steinhaus) Suppose that $\{x_i^* : \ i \in I\}$ is an indexed collection of linear forms such that the set*

$$\{\langle x_i^*, x\rangle : \ i \in I\}$$

is bounded for each $x \in X$, then the set

$$\{\|x_i^*\| : \ i \in I\}$$

is bounded.

Another result useful in the theory of partial differential equations is

Lemma D.5. *Let B be a continuous function (with values in the field of scalars) on a Hilbert space $H \times H$ which is linear in the first variable and complex-linear in the second. Suppose that*

$$|B(x,y) \ \geq m\|x\|^2, \textit{ for all } x \in H$$

for some $m > 0$. Then if $f \in H^*$, there exists a unique $h \in H$ such that

$$f(x) = B(x,h), \textit{ for all } x \in H.$$

D.2 Examples

In this section we present three examples of Banach and Hilbert spaces which occur frequently in systems theory.

Example D.1. Let $\Omega \subseteq \mathbb{R}^n$ be an open and bounded set and let $C^k(\overline{\Omega})$ (k a non-negative integer) be the set of all real or complex-valued functions on Ω whose partial derivatives to order k exist and are uniformly continuous on $\overline{\Omega}$. We define a norm on $C^k(\overline{\Omega})$ by

$$\|f\|_k \overset{\Delta}{=} \sup_{|p|\le k}\left[\sup_{x\in\Omega}|(\partial/\partial x)^p f(x)|\right],\ f\in C^k(\overline{\Omega})$$

where $p \in \mathbb{N}^n$, $|p| = p_1 + \cdots + p_n$ and

$$(\partial/\partial x)^p = (\partial/\partial x_1)^{p_1}\cdots(\partial/\partial x_n)^{p_n}.$$

It can be shown that $C^k(\overline{\Omega})$ is a Banach space which is not a Hilbert space. We also write

$$C(\overline{\Omega}) = C^0(\overline{\Omega})$$

and

$$C^\infty(\overline{\Omega}) = \cap_{k=0}^{\infty} C^k(\overline{\Omega}).$$

Example D.2. Let Ω be an open set in $\mathbb{R}^n$ with smooth boundary (*i.e.* it is a smooth $(n-1)$- dimensional manifold. We define the spaces $L^p(\Omega)$ for all real numbers $p \ge 1$ to be the space of all Lebesgue measurable (real or complex valued) functions defined on Ω such that

$$\|f\|_{L^p(\Omega)} \overset{\Delta}{=} \left(\int |f(x)|^p dx\right)^{1/p} < \infty$$

if $0 < p < \infty$, and

$$\|f\|_{L^\infty(\Omega)} \overset{\Delta}{=} \text{ ess sup }_{x\in\Omega}|f(x)| < \infty.$$

It can be shown that each space $L^p(\Omega)$ is a Banach space for $1 \le p \le \infty$. Moreover, $L^2(\Omega)$ is a Hilbert space for the inner product

$$\langle f, g\rangle_{L^2(\Omega)} = \int_\Omega f(x)\overline{g}(x)dx.$$

Example D.3.

Let $\ell^p,\ 1 \le p \le \infty$ be the space of all (real or complex) sequences $x = (x_i)_{i=1,2,\cdots}$ such that the following norms are finite:

$$\|x\|_{\ell^p} \overset{\Delta}{=} \left(\sum_{i=1}^{\infty}|x_i|^p\right)^{1/p}$$

if $p < \infty$ and

$$\|x\|_{\ell^\infty} \overset{\Delta}{=} \sup_{1 \le i < \infty} |x_i|.$$

The linear spaces ℓ^p are again Banach spaces and ℓ^2 is a Hilbert space. In fact, any separable Hilbert space H is isomorphic to ℓ^2, so that in some sense this is a 'universal' separable Hilbert space. This can be seen easily be choosing an orthonormal basis for H.

We have seen the map J above which is an isometric isomorphism between a Banach space and its second dual. Lemma D.1.4 also shows that there is a map $j_H : H \to H^*$ which is also an isometric isomorphism between H and H^*. Note, however, that while J is functorial (*i.e.* is a natural map, independent of any specific coordinates), j_H is not. To see the effect of this, let H be a (separable) Hilbert space and let V be a dense subspace of H such that the injection

$$i_V : V \subseteq H$$

is continuous. This means that there is a constant c such that

$$\|h\|_H \le c\|h\|_V, \text{ for all } h \in V.$$

Thus we have a sequence of injections

$$V \overset{i_V}{\subseteq} H \overset{j_H}{\subseteq} H^* \overset{i_{H^*}}{\subseteq} V^*.$$

The map $i_{H^*} \circ j_H \circ i_V$ is different from the map j_V.

D.3 Theory of Operators

A *linear operator* A between Banach spaces X and Y with domain $\mathscr{D}(A)$ (which may be different from X) is a function for which

$$A(\alpha x + \beta y) = \alpha A(x) + \beta A(y)$$

for all real (or complex) α, β and all $x, y \in \mathscr{D}(A)$ (the latter being assumed to be a linear subspace of X). The *kernel* of A is defined as the linear space

$$\ker A = \{x \in \mathscr{D}(A) : \ Ax = 0\}.$$

Clearly the operator A is one-to-one if and only if ker $A = \{0\}$. The *range* $\mathscr{R}(A)$ of A is the linear subspace of Y consisting of all images of elements of $\mathscr{D}(A)$. If ker $A = \{0\}$ then we can define the inverse operator $A^{-1} : \mathscr{R}(A) \to \mathscr{D}(A)$. If B is another linear operator with domain $\mathscr{D}(B) \subseteq \mathscr{D}(A)$ then we say that A is an extension of B and write $B \subseteq A$, if

$$Bx = Ax, \text{ for all } x \in \mathscr{D}(B).$$

If $\mathscr{D}(A) = X$ and the norm

$$\|A\| \overset{\Delta}{=} \sup_{\|x\|=1} \|A\|x$$

is finite, then we say that A is a *bounded (linear) operator* from X to Y. The set of all bounded linear operators from X to Y is written $\mathscr{B}(X,Y)$ or $\mathscr{B}(X)$ in the case where $X = Y$. Note that $\mathscr{B}(X)$ is a Banach algebra since

$$\|A_1A_2\| \leq \|A_1\|\|A_2\|$$

for all bounded operators A_1, A_2.

An important result for solving many types of equations is the *Neumann series*:

Theorem D.4. *If X is a Banach space and $A \in \mathscr{B}(X)$ satisfies $\|A\| \leq 1$, then $(I-A)$ is invertible, $(I-A)^{-1} \in \mathscr{B}(X)$ and it is given by*

$$(I-A)^{-1} = I + A + A^2 + \cdots$$

where the convergence is in the uniform topology of $\mathscr{B}(X)$.

An important class of bounded operators is the class of projections. An operator $P \in \mathscr{B}(X)$ is a projection operator with range $\mathscr{R}(P)$ if it satisfies $P^2 = P$. In the Hilbert space case, projections onto a closed subspace are particularly important, since we have

$$H = M \oplus M^{\perp}$$

for any Hilbert space H and closed subspace M. Here $M^{\perp}$ is the *orthogonal complement* of M given by

$$M^{\perp} = \{h \in H : \ \langle h, m \rangle = 0 \text{ for all } m \in M\}.$$

Of course, the map

$$P_M(h) = m, \ \ h = m + m^{\perp}$$

where $m \in M$ and $m^{\perp} \in M^{\perp}$ is a projection operator with

$$\|P_M\| \leq 1.$$

The notion of transpose of a matrix can be generalised to operators in the following way. If $A \in \mathscr{B}(X,Y)$ for two Banach spaces X, Y, let $y^* \in Y^*$. Then we define the linear form x^* by

$$\langle x^*, x \rangle_{X^*,X} = \langle y^*, Ax \rangle_{Y^*,Y}$$

and if we put

$$x^* = A^* y^*$$

then $A^* \in \mathscr{B}(Y^*, X^*)$ and $\|A^*\| = \|A\|$. A^* is called the *transpose* or *dual* of A. In the case where X and Y are Hilbert spaces we can identify X^* with X and Y^* with Y, using the maps j_X, j_Y defined above. Then the dual operator is called the *adjoint* and. However, it is usually defined by $j_X^{-1} A^* j_Y$, *i.e.* by the identity

$$\langle A^* h_1, h_2 \rangle_X = \langle h_1, Ah_2 \rangle_Y, \ \ h_1 \in Y, \ h_2 \in X.$$

Bounded operators form the most convenient class of operators and of those, clearly (finite) matrices are the simplest, for which the Jordan decomposition gives a complete structure theory, as seen in Appendix A. The next most amenable class consists of compact operators which are bounded operators belonging to $\mathscr{B}(X,Y)$ which have the property that any bounded subset of X is mapped to a relatively compact subset of Y. The class of compact operators is denoted by $\mathscr{K}(X,Y)$. There is a complete spectral theory for these operators. If an operator is not bounded, such as the operator

$$A = \frac{d}{dx}$$

defined on a dense subspace of $C[0,1]$ then it may have the following useful property.

Definition D.8. A linear operator $A : \mathscr{D}(A) \subseteq (X) \to Y$ is said to be *closed* if its graph is closed, *i.e.* if $x_n \to x$ and $Ax_n \to y$ then $x \in \mathscr{D}(A)$ and $y = Ax$. It is *closable* if $x_n \in \mathscr{D}(A)$ and $x_n \to 0, \ Ax_n \to y$ imply that $y = 0$. We denote the class of closed operators by $\mathscr{C}(X,Y)$.

If an operator $A : \mathscr{D}(A) \subseteq X \to Y$ has dense domain, *i.e.* $\overline{\mathscr{D}(A)} = X$ then we can define the dual of A by

$$\langle y, Ax \rangle = \langle A^* y, x \rangle, \ \ x \in \mathscr{D}(A), \ \ y \in \mathscr{D}(A^*),$$

such that $\overline{\mathscr{D}(A^*)} = Y^*$. If A is closable and has dense domain, then $A^{**} = (A^*)^*$ is the closure of A.

D.4 Spectral Theory

Let X be a complex Banach space and suppose that A is an operator with domain and range in X. For a complex number $\lambda \in \mathbb{C}$, the operator $(\lambda I - A)$ may or may not be invertible.

Definition D.9. The *resolvent set* $\rho(A)$ of the operator A is the set of all $\lambda \in \mathbb{C}$ such that the inverse of $(\lambda I - A)$ exists and is bounded, *i.e.* $(\lambda I - A)^{-1} \in \mathscr{B}(X)$.

Definition D.10. The *spectrum* of A, denoted by $\sigma(A)$ is the complement of the resolvent set, *i.e.* $\sigma(A) = \mathbb{C} \setminus \rho(A)$.

The spectrum of A consists of three disjoint subsets:

$$\begin{aligned}
\sigma_P(A) &= \{\lambda \in \sigma(A) : \ (\lambda I - A) \text{ is not 1-1 } \}\\
\sigma_C(A) &= \{\lambda \in \sigma(A) : \ \overline{(\lambda I - A)^{-1}} = X\}\\
\sigma_R(A) &= \{\lambda \in \sigma(A) : \ \overline{(\lambda I - A)^{-1}} \neq X\}
\end{aligned}$$

called, respectively, the *point spectrum*, the *continuous spectrum* and the *residual spectrum*.

For $\lambda \in \sigma(A)$ we write

$$R(\lambda; A) = (\lambda I - A)^{-1}$$

and we call R the *resolvent operator* of A. It satisfies the *resolvent equation*

$$R(\lambda; A) - R(\mu; A) = (\mu - \lambda) R(\lambda; A) R(\mu; A)$$

for all $\lambda, \mu \in \rho(A)$. The resolvent operator $R(\lambda; A)$ is analytic on $\rho(A)$ and it can be used with a generalisation of the Cauchy integral to define a functional calculus for operators and to obtain spectral decompositions. Consider first the case of bounded operators $A \in \mathscr{B}(X)$. For such an operator, let f be any function which is analytic in a neighbourhood U of $\sigma(A)$ such that U is open and has boundary ∂U consisting of a finite number of oriented rectifiable Jordan curves. Then we define the operator $f(A)$ by

$$f(A) = \frac{1}{2\pi i} \int_{\partial U} f(\lambda) R(\lambda; A) d\lambda.$$

Note that $f(A)$ depends only on f and A and not on the choice of neighbourhood U. If f and g are analytic on a neighbourhood of $\sigma(A)$, then so is their product $f \cdot g$ and we have

$$(f \cdot g)(A) = f(A) g(A).$$

From this we can obtain the *spectral mapping theorem*:

$$f(\sigma(A)) = \sigma(f(A)),$$

for all functions analytic on a neighbourhood of $\sigma(A)$. We define a *spectral subset* of A to be an open and closed subset of $\sigma(A)$. Then the characteristic function of a spectral set σ defined by

$$e(\lambda) = \begin{cases} 1 & \text{if } \lambda \in \sigma \\ 0 & \text{if } \lambda \in \sigma(A) \setminus \sigma \end{cases}$$

can be extended to an analytic function on a neighbourhood of $\sigma(A)$ since a spectral set is open and closed in $\sigma(A)$. Hence we can define the operator

$$E(\sigma; A) = e(A).$$

Clearly, $E(\sigma;A)$ is a projection operator and if we define

$$\begin{aligned} X_\sigma &= E(\sigma;A)X \\ A_\sigma &= A|_{X_\sigma} \end{aligned}$$

then we have

$$\begin{aligned} AX_\sigma &\subseteq X_\sigma \\ \sigma(A_\sigma) &= \sigma \end{aligned}$$

and, for any function f analytic on a neighbourhood of $\sigma(A)$,

$$f(A)_\sigma = f(A_\sigma).$$

Hence, if $\sigma(A)$ consists of a finite number of spectral subsets $\sigma_1, \cdots, \sigma_n$, then we have

$$E(\sigma_i;A)^2 = E(\sigma_i;A), \quad E(\sigma_i;A)E(\sigma_j;A) = 0, \text{ if } i \neq j$$

and we obtain a decomposition of the identity

$$I = \sum_{i=1}^{n} E(\sigma_i;A), \quad X = \oplus_{i=1}^{n} X_{\sigma_i}.$$

Theorem D.5. *Suppose $A \in \mathscr{K}(X)$. Then the spectrum of A is countable and has no accumulation point apart from 0. Every point $\lambda \in \sigma(A)$ is in the point spectrum and the eigenspace*

$$E(\lambda;A) = \{x : \ (A - \lambda I)^k = 0, \textit{ for some } k \geq 0\}$$

is finite-dimensional.

To apply the theory in the more general case of closed operators, we take a closed operator A and consider the homeomorphism $\varphi : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ defined for any $\alpha \in \rho(A)$ by

$$\varphi(\lambda) = (\lambda - \alpha)^{-1}, \ \lambda \neq \alpha$$

and

$$\varphi(\infty) = 0, \quad \varphi(\alpha) = \infty.$$

Then the operator $A' = -R(\alpha;A)$ is bounded and

$$\varphi(\sigma(A) \cup \{\infty\}) = \sigma(A')$$

and φ maps functions analytic on the spectrum of A to functions analytic on the spectrum of A'. Hence, if f is analytic on the spectrum of A, we can define

$$f(A) = f(\varphi^{-1}(A')).$$

It follows that

$$f(A)=f(\infty)I+\frac{1}{2\pi i}\int_{\Gamma}f(\lambda)R(\lambda;A)d\lambda$$

where Γ consists of a finite number of Jordan curves containing $\sigma(A)$. Moreover, we have the spectral mapping theorem

$$\sigma(f(A))=f(\sigma(A)\cup\{\infty\}).$$

From this we can get the spectral theory of closed self-adjoint operators on Hilbert space with compact resolvent.

Theorem D.6. *If A is a closed self-adjoint operator (i.e. $A^*=A$) on a Hilbert space H and $R(\lambda;A)$ is compact (for some and hence all $\lambda\in\rho(A)$), then the spectrum of A consists of just eigenvalues λ_i with finite multiplicities such that $|\lambda_i|\to\infty$ as $i\to\infty$. Moreover, there is an orthonormal set of eigenvectors φ_i such that, for any $x\in H$,*

$$x=\sum_{i=1}^{\infty}\langle x,\varphi_i\rangle\varphi_i,\quad Ax=\sum_{i=1}^{\infty}\lambda_i\langle x,\varphi_i\rangle\varphi_i$$

and

$$R(\lambda;A)x=\sum_{i=1}^{\infty}\frac{1}{\lambda-\lambda_i}\langle x,\varphi_i\rangle\varphi_i$$

for $\lambda\in\rho(A)$. (Note that we count the eigenvalues with multiplicity.)

D.5 Distribution Theory

In many types of linear and nonlinear partial differential equations it is necessary to consider non-classical (*i.e.* non-differentiable) solutions, which requires the theory of distributions. Here we shall give a brief outline of the theory; more details can be found in [4], [5].

Definition D.11. A topological vector space E (over the field $F=\mathbb{R}$ or $\mathbb{C}$) is a vector space over F together with a topology which is compatible with the vector space structure, *i.e.* the maps

(a) $(x,y)\to x+y$ from $E\times E$ into E

(b) $(\lambda,x)\to\lambda x$ from $F\times E$ into E

are continuous.

Definition D.12. A *filter* $\mathfrak{F}$ on a set X is a collection of subsets of X such that:

(a) If $A\subseteq X$ and $A\supseteq B\in\mathfrak{F}$, then $A\in\mathfrak{F}$.

(b) $\emptyset$ does not belong to $\mathfrak{F}$.

(c) If $A,B\in\mathfrak{F}$, then $A\cap B\in\mathfrak{F}$.

A *basis* of a filter $\mathfrak{F}$ is a non-empty sub-collection $\mathfrak{B}\subseteq\mathfrak{F}$ such that if $A,B\in\mathfrak{B}$ then there exists $C\subseteq A\cap B$ with $C\in\mathfrak{B}$.

Definition D.13. A subset A of a topological vector space E is called *absorbing* if for any $x \in E$ there exists an $\alpha > 0$ such that

$$x \in \lambda A, \text{ for all } \lambda \in F \text{ such that } |\lambda| \geq \alpha.$$

Definition D.14. A subset A of a topological vector space E is called *balanced* if

$$\lambda A \subseteq A, \text{ for all } \lambda \in F \text{ such that } |\lambda| \leq 1.$$

Definition D.15. A basis for the filter of neighbourhoods of a point x in a topological vector space is called a *fundamental systems of neighbourhoods* of x.

Since the translation map $y \to y + x$ is a homeomorphism, the topology of a topological vector space is determined entirely by a fundamental systems of neighbourhoods of 0.

Definition D.16. A *locally convex topological vector space* is a topological vector space with a fundamental system of convex neighbourhoods.
The basic result of topological vector spaces is the following:

Theorem D.7. *Let E be a vector space and $\mathfrak{N}$ a collection of absorbing, balanced and convex subsets of E. If we define*

$$\mathfrak{N}' = \{\cap_{i=1}^{n}(\lambda_i V_i) : \ \lambda_i > 0, \ V_i \in \mathfrak{N}\}$$

then $\mathfrak{N}'$ is a fundamental systems of neighbourhoods of zero for a unique locally convex topology on E. Moreover, an equivalent fundamental system (in the sense that it generates the same topology) is given by

$$\mathfrak{N}'' = \{\lambda V : \ \lambda > 0, \ V = \cap_{i=1}^{n} U_i \, for\ some\ U_i \in \mathfrak{N}'\}.$$

We can also define locally convex spaces by means of seminorms.

Definition D.17. A *seminorm* on a vector space E is a map $q : E \to \mathbb{R}^+$ such that
(a) $q(\lambda x) = |\lambda| q(x)$, for all $\lambda \in F$ and $x \in E$.
(b) $q(x+y) \leq q(x) + q(y)$, for all $x, y \in E$.

If q is a seminorm, then the set

$$V = \{x : \ q(x) \leq 1\}$$

is balanced, absorbing and convex and so, from Theorem D.7, if $(q_i)_i \in I$ is a family of seminorms, then the sets

$$V_{i_1,\cdots,i_n,\varepsilon} = \{x : \ q_{i_k}(x) \leq \varepsilon \text{ for } 1 \leq k \leq n\}, \ i_k \in I$$

constitute a fundamental systems of neighbourhoods of 0 for a locally convex space. Conversely, any locally convex space can be defined by the set of all continuous seminorms. Hence, if E and F are locally convex spaces defined by families of seminorms $\{p_i\}_{i\in I}$ and $\{q_j\}_{j\in J}$, then a linear map $f : E \to F$ is continuous if and only if for any q_j, there exists a corresponding p_i and some M such that

$$q_j(f(x)) \leq Mp_i(x), \text{ for all } x \in E.$$

We next present some examples of locally convex spaces which are the most important in distribution theory.

Example D.4. Let Ω be an open subset of $\mathbb{R}^n$. If $K \subseteq \Omega$ is compact, let $\mathscr{D}(K)$ denote the space of functions in (a neighbourhood) of K and whose derivatives of all orders exists and are continuous and which vanish outside K. We topologise $\mathscr{D}(K)$ as the locally convex space defined by the seminorms

$$q_{j,K}(f) = \max_{x\in K} |(\partial/\partial x)^j f(x)|, \quad \text{for all } j \in \mathbb{N}^n.$$

Then define

$$\mathscr{D}(\Omega) = \cup_K \mathscr{D}(K)$$

where the union is over all compact subsets of Ω. We give $\mathscr{D}(\Omega)$ the finest locally convex topology for which the inclusion maps

$$\mathscr{D}(K) \to \mathscr{D}(\Omega)$$

are all continuous. We call $\mathscr{D}(\Omega)$ the space of all *infinitely differentiable functions with compact support* on Ω.

Example D.5. $\mathscr{E}(\Omega)$ denotes the space of functions which are infinitely differentiable, but with not restriction on their supports, together with the same seminorms as in Example D.4.

Example D.6. The space of *rapidly decreasing functions* is useful in defining the Fourier transforms of distributions. A function $f \in \mathscr{E}(\mathbb{R}^n)$ is rapidly decreasing if, for any $j \in \mathbb{N}^n$, $k \in \mathbb{Z}$ and $\varepsilon > 0$, there exists $\rho > 0$ such that

$$|(1 + \|x\|^2)^k (\partial/\partial x)^j f(x)| \leq \varepsilon.$$

Then $\mathscr{S}$ denotes the space of all rapidly decreasing functions with the system of seminorms

$$q_{k,j}(f) = \max_{x\in\mathbb{R}^n} |(1 + \|x\|^2)^k (\partial/\partial x)^j f(x)|.$$

Then $\mathscr{S}$ is a locally convex space and we have the continuous inclusions

$$\mathscr{D}(\Omega) \subseteq \mathscr{E}(\Omega), \quad \mathscr{D}(\mathbb{R}^n) \subseteq \mathscr{S} \subseteq \mathscr{E}(\mathbb{R}^n)$$

where each space is dense in the next.

There are a variety of different topologies possible on a given topological vector space E. The weak topologies have many useful properties which the stronger ones do not have. For example, bounded sets are often relatively compact in a weak topology. To define these topologies, consider the duality $\langle \cdot, \cdot \rangle_{E,E^*}$ between E and its dual space E^* and consider the two systems of seminorms

$$q_{y^*} : x \longrightarrow |\langle x, y^* \rangle|$$

and

$$q_x : y^* \longrightarrow |\langle x, y^* \rangle|.$$

Then q_{y^*} is a seminorm on E for each $y^* \in E^*$ and q_x is a seminorm on E^* for each $x \in E$. Then the systems of seminorms $\{q_{y^*} : y^* \in E^*\}$ and $\{q_x : x \in E\}$ define locally convex topologies on E and E^* called the weak and weak* topologies, respectively. They are often denoted by $\sigma(E, E^*)$ and $\sigma(E^*, E)$. Note that we cannot use sequences in these spaces to determine conditions such as compactness - we require directed sets. However, we can say that a sequence $\{x_n\} \subseteq E$ *converges weakly* to x if and only if

$$\lim_n \langle x_n, x^* \rangle = \langle x, x^* \rangle$$

for all $x^* \in E^*$ (with a similar condition for weak* convergence in E^*). Also, it can be shown that the unit sphere in a Hilbert space H is compact in the weak topology (not, of course, in the norm topology, unless H is finite-dimensional).

Definition D.18. The dual space of $\mathscr{D}(\Omega)$, denoted by $\mathscr{D}'(\Omega)$ is called the *space of distributions* on Ω. The space $\mathscr{E}'(\Omega)$ is the *space of distributions with compact support* in Ω and $\mathscr{S}$ is the space of *tempered distributions*.

Using the spaces of distributions, defined as dual spaces of function spaces, we can use the duality and transposition (essentially generalized integration by parts) to define operations such as differentiation, Fourier transform, etc. on distributions in terms of their counterparts on the spaces of well-behaved functions $\mathscr{D}(\Omega)$, $\mathscr{E}(\Omega)$ and $\mathscr{S}$. For example, to define differentiation, let $T \in \mathscr{D}'(\Omega)$ and let $p \in \mathbb{N}^n$ with $|p| = p_1 + \cdots + p_n$. Then we define the *derivative* $(\partial/\partial x)^p$ of T to be the distribution given by

$$\langle (\partial/\partial x)^p T, \varphi \rangle = (-1)^{|p|} \langle T, (\partial/\partial x)^p \varphi \rangle$$

for all $\varphi \in \mathscr{D}(\Omega)$. (The duality here is, of course, between $\mathscr{D}'(\Omega)$ and $\mathscr{D}(\Omega)$.) In a similar way, we can define the *Fourier transform* $\mathscr{F}T$ of the distribution T by

$$\langle \mathscr{F}T, \varphi \rangle = \langle T, \mathscr{F}\varphi \rangle$$

for all $\varphi \in \mathscr{S}$. This time, the duality is between $\mathscr{S}'$ and $\mathscr{S}$. The function φ used in both these definitions is often called a *test function*.

Example D.7. The *Dirac delta function* δ is actually a distribution defined by

$$\langle \delta, \varphi \rangle = \varphi(0),$$

for any $\varphi \in \mathscr{D}(\mathbb{R}^n)$. Its derivative is given by

$$\langle(\partial/\partial x)^p\delta, \varphi\rangle = (-1)^{|p|}((\partial/\partial x)^p\varphi)(0),$$

for all $\varphi \in \mathscr{D}(\mathbb{R}^n)$.

Example D.8. The Heaviside step function

$$U(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

is, of course, not differentiable in the normal sense at 0. However, the associated distribution (which we also denote U) given by

$$U(\varphi) = \int_{-\infty}^{\infty} U(x)\varphi(x)dx = \int_{0}^{\infty} \varphi(x)dx$$

belongs to $\mathscr{D}'(\mathbb{R})$. Then we have

$$\langle(\partial/\partial x)U, \varphi\rangle = -\langle U, \partial\varphi/\partial x\rangle = \varphi(0) = \langle\delta, \varphi\rangle$$

and so

$$(\partial/\partial x)U = \delta.$$

Example D.9. If $f : \Omega \to \mathbb{R}$ is any locally integrable function, then the associated linear functional $\overline{f} : \mathscr{D}(\Omega) \to \mathbb{R}$ given by

$$\overline{f}(\varphi) = \langle\overline{f}, \varphi\rangle = \int_{\Omega} f(x)\varphi(x)dx$$

is a distribution. In particular, if $1(x) = 1$, for all $x \in \mathbb{R}^n$, then

$$\langle\overline{1}, \varphi\rangle = \int_{\mathbb{R}^n} \varphi(x)dx.$$

Hence we can calculate the Fourier transform of the delta function by

$$\langle\mathscr{F}\delta, \varphi\rangle = \langle\delta, \mathscr{F}\varphi\rangle = \int_{\mathbb{R}^n} \varphi(x)exp(-2\pi i\langle x, \xi\rangle)dx\Big|_{\xi=0} = \int_{\mathbb{R}^n} \varphi(x)dx = \langle\overline{1}, \varphi\rangle$$

i.e.

$$\mathscr{F}\delta = \overline{1}.$$

The next result is very useful in writing certain linear operators on function spaces in terms of the kernels of integrals. It is known as Schwartz' kernel theorem.

Theorem D.8. *Let $X \subseteq \mathbb{R}^m$ and $Y \subseteq \mathbb{R}^n$ be open sets. Then the product $X \times Y \subseteq \mathbb{R}^{m+n}$ is also open and we have the isomorphism*

$$\mathscr{D}'(X \times Y) \cong \mathscr{L}(\mathscr{D}(Y); \mathscr{D}'(X))$$

i.e. the space of distributions on $X \times Y$ is isomorphic (as a topological vector space) to the space of continuous linear maps from $\mathscr{D}(Y)$ into $\mathscr{D}'(X)$.

Thus if we have a continuous linear map $L : \mathscr{D}(Y) \to \mathscr{D}'(X)$, then we can associate with it a distribution $K \in \mathscr{D}'(X \times Y)$, such that

$$\langle K, \varphi\psi \rangle = \langle L\psi, \varphi \rangle$$

for all test functions $\varphi \in \mathscr{D}(X)$, $\psi \in \mathscr{D}(Y)$, where

$$\varphi\psi : (x,y) \in \mathbb{R}^{m+n} \to \varphi(x)\psi(y).$$

This is often written in the form

$$(L\psi)(x) = \int K(x,y)\psi(y)dy.$$

Since we can regard $L^2(X \times Y)$ as a subspace of $\mathscr{D}'(X \times Y)$, we have

$$L^2(X \times Y) \cong \mathscr{K}_N(L^2(X), L^2(Y))$$

where $\mathscr{K}_N(L^2(X), L^2(Y))$ is the space of compact *nuclear* operators from $L^2(X)$ into $L^2(Y))$, *i.e.* operators of the form

$$Nf = \sum_k \lambda_k \langle f, f_k \rangle g_k$$

where $\{f_k\}, (\{g_k\})$ is a basis of $L^2(X)$ $(L^2(Y))$ and $\sum |\lambda_k| < \infty$. Then we can write

$$(Nf)(y) = \int_X N(x,y)f(x)dx$$

where $f \in L^2(X)$ and $N(x,y) \in L^2(X \times Y)$.

D.6 Sobolev Spaces

Sobolev spaces are important in the theory of partial differential equations and we shall give a brief outline of them here. For more details see [6]. Note that

$$\mathscr{D}(\Omega) \subseteq L^p(\Omega) \subseteq \mathscr{D}'(\Omega)$$

for $1 \leq p \leq \infty$ and so we can talk about distributions in $L^p(\Omega)$.

Definition D.19. Define the space

$$H^{p,m}(\Omega) = \{f \in \mathscr{D}'(\Omega) : (\partial/\partial x)^{\alpha} f \in L^p(\Omega),\ \ |\alpha| \le m\}$$

together with the norm

$$\|f\|_{p,m} = \left\{ \sum_{|\alpha|\le m} \int_{\Omega} |(\partial/\partial x)^{\alpha} f(x)|^p \right\}^{1/p}.$$

It can be shown that $H^{p,m}(\Omega)$ is a Banach space and that $H^{2,m}(\Omega)$ is a Hilbert space with the inner product

$$\langle f, g\rangle_{p,m} = \sum_{|\alpha|\le m} \int_{\Omega} (\partial/\partial x)^{\alpha} f(x) \overline{(\partial/\partial x)^{\alpha} g(x)} dx.$$

We denote by $H_0^{p,m}(\Omega)$ $(1 \le p, \le \infty, m \ge 1)$ the closure of $\mathscr{D}(\Omega)$ in $H^{p,m}(\Omega)$ and we put

$$H^{p',-m}(\Omega) = (H_0^{p,m}(\Omega))^*$$

where $p' = p/(p-1)$, for $1 \le p < \infty$. Then it can be shown that

$$H^{p',-m}(\Omega) = \{f \in \mathscr{D}'(\Omega) :\ \ f = \sum_{|\alpha|\le m} (\partial/\partial x)^{\alpha} g_{\alpha},\ \ g_{\alpha} \in L^p(\Omega)\}.$$

Note that if $\Omega = \mathbb{R}^n$ then $H_0^{p,m}(\Omega) = H^{p,m}(\Omega)$. We can then define H^s for arbitrary real s.

Definition D.20. For any $s \in \mathbb{R}$ we define the space

$$H^s = \{f \in \mathscr{S}'(\mathbb{R}^n) :\ \ (1 + |\xi|^2)^{s/2} \widehat{f} \in L^2(\mathbb{R}^n)\}$$

where $\widehat{f}$ denotes the Fourier transform of f.

Example D.10. The Dirac delta function δ is in $H^{-n/2-\varepsilon}$ for any $\varepsilon > 0$.

We mention, finally, the Sobolev embedding theorem which is useful for proving regularity theorems for partial differential equations.

Theorem D.9. *Suppose that $\Omega \subseteq \mathbb{R}^n$ is an open bounded set with sufficiently smooth boundary (although this condition can be relaxed). Then:*
(a) if $m < k - n/p$, we have $H^{k,p}(\Omega) \subseteq C^m(\overline{\Omega})$
(b) if $1/q > 1/p - k/n$, with $1 \le p \le \infty$, $1 \le q \le \infty$, $k \ge 1$
then we have the embedding

$$H^{k,p}(\Omega) \subseteq L^q(\Omega)$$

which is compact if $p, q < \infty$.

D.7 Partial Differential Equations

A linear partial differential operator of order m is an expression of the form

$$L = \sum_{|\alpha| \le m} a_\alpha(x) D^\alpha = \sum_{k=0}^{m} \left\{ \sum_{\alpha_1 + \cdots + \alpha_n = k} a_{\alpha_1 \cdots \alpha_n}(x) D_1^{\alpha_1} \cdots D_n^{\alpha_n} \right\}$$

where $D_i = \partial / \partial x_i$, $\alpha \in \mathbb{N}^n$ and $a_\alpha(x)$ is defined in a bounded open set $\Omega \in \mathbb{R}^n$. The highest order term

$$\sum_{|\alpha| = m} a_\alpha(x) D^\alpha$$

is called the *principal part* and in a sense dominates the operator. We associate with it the polynomial

$$p_m(\xi, x) = \sum_{|\alpha| = m} a_\alpha(x) \xi^\alpha.$$

If $p_m(\xi, x) \neq 0$ for all non-zero $\xi \in \mathbb{R}^n$ and each fixed x, then we say that the differential operator L is *elliptic* (at x). If the coefficients a_α are real then m must be even and so if we put $m = 2q$ we say that the operator is (uniformly) *strongly elliptic* in Ω if

$$(-1)^q \Re(p_{2q}(\xi, x)) \ge c|\xi|^{2q}, \text{ for all } x \in \Omega$$

for some constant $c > 0$.

If the coefficients a_α are sufficiently differentiable, we can write the operator L in 'divergence form'

$$Lu = \sum_{0 \le |\beta|, |\gamma| \le q} (-1)^{|\beta|} D^\beta (a^{\beta\gamma}(x) D^\gamma u)$$

and so, integrating by parts, we define the bi-linear form B by

$$\langle v, Lu \rangle_{L^2} = B(v, u) = \sum_{0 \le |\beta|, |\gamma| \le q} (-1)^{|\beta|} \langle D^\beta v, a^{\beta\gamma}(x) D^\gamma u \rangle_{L^2}$$

for all $u, v \in \mathscr{D}(\Omega)$.

In order to prove the existence of solutions to elliptic partial differential equations, we can use Lemma D.4 in conjunction with Garding's inequality given in the next result.

Lemma D.6. *If L is a strongly elliptic operator in Ω such that the coefficients $a^{\beta\gamma}$ are bounded in Ω and satisfy the inequality*

$$|a^{\beta\gamma}(x) - a^{\beta\gamma}(y)| \le f(\|x - y\|), \textit{ for } x, y \in \Omega \textit{ and } |\beta| = |\gamma| = q,$$

where $f(t) \to 0^+$ as $t \to 0^+$, then we have the inequality

$$\Re(B(u,u)) \ge c_1 \|u\|_{2,m}^2 - c_2 \|u\|_{2,0}^2, \textit{ for all } u \in H_0^{2,m}(\Omega)$$

for some constants $c_1 > 0$ and c_2.

Now consider the typical *Dirichlet problem* for an elliptic operator L: this is to solve the equations

$$\begin{cases} Lu = f & \text{in } \Omega \\ \frac{\partial^j u}{\partial \nu^j} = g_j & \text{on } \partial\Omega,\ 0 \le j \le m-1 \end{cases}$$

for some functions f, g_j, where ν is the (outward pointing) normal to $\partial\Omega$. By modifying u we can show that, if each $g_j \in C^{2m}(\partial\Omega)$, then it is equivalent to studying the problem

$$\begin{cases} Lu = f & f \in L^2(\Omega) \\ \frac{\partial^j u}{\partial \nu^j} = 0 & \text{on } \partial\Omega, \quad 0 \le j \le m-1 \end{cases}.$$

By a *solution* of this system we mean an element u of $H_0^{2,m}(\Omega)$ such that

$$B(\varphi, u) = \langle \varphi, f \rangle_{L^2}$$

for all $\varphi \in \mathscr{D}(\Omega)$. We have

Theorem D.10. *If L satisfies the conditions of Lemma D.6, then there exists a constant c_2 such that, for all $c \ge c_2$, the Dirichlet problem for the operator $L + c$ has a unique solution for any $f \in L^2(\Omega)$.*

Proof. Consider the bi-linear form

$$B_1(u,v) = B(u,v) + c\langle u, v \rangle_{L^2(\Omega)}$$

for all $u, v, \in H_0^{2,m}(\Omega)$. Since L is associated with B, we have, by Garding's inequality,

$$\Re B(u,u) \ge c_1 \|u\|_{2,m}^2 - c_2 \|u\|_{L^2(\Omega)}^2$$

for some constants $c_1 > 0, c_2$. Hence,

$$\Re B_1(u,u) \ge c_1 \|u\|_{2,m}^2$$

if $c \ge c_2$. Of course, B_1 is associated with $L + c$. Now the linear form

$$\varphi \to \langle \varphi, f \rangle_{L^2(\Omega)}, \quad \varphi \in H_0^{2,m}(\Omega)$$

is continuous on $H_0^{2,m}(\Omega)$ and so by Lemma D.4 there exists $u \in H_0^{2,m}(\Omega)$ such that

$$B(\varphi, u) = \langle \varphi, f \rangle_{L^2}$$

holds, for each $\varphi \in H_0^{2,m}(\Omega)$. □

The following abbreviations are usually used for the Sobolev spaces:

$$H_0^m(\Omega) = H_0^{2,m}(\Omega), \;\; H^m(\Omega) = H^{2,m}(\Omega).$$

Now define a linear operator A with domain $\mathscr{D}(A) = H^{2m}(\Omega) \cap H_0^{2,m}(\Omega)$ by

$$(Au)(x) = Lu(x), \;\; u \in \mathscr{D}(A)$$

where L is the above partial differential operator. Then we have

Theorem D.11. *Let Ω be a bounded domain in $\mathbb{R}^n$ with sufficiently smooth boundary and let L be strongly elliptic in Ω, with coefficients $a_\alpha \in C^j(\Omega)$, where $j = \max(0, |\alpha| - m)$. Then the operator A associated with L as defined above is a closed operator defined in $L^2(\Omega)$ with domain $\mathscr{D}(A) = H^{2m}(\Omega) \cap H_0^{2,m}(\Omega)$. Also, the resolvent $(\lambda I - A)^{-1} : L^2(\Omega) \to L^2(\Omega)$ exists for all $\lambda \in \mathbb{C}$ belonging to the sector*

$$\{\lambda : \; \frac{1}{2}\pi < \; arg \; (\lambda + k) < \frac{3}{2}\pi, \, for \; some \; k > 0\}$$

and we have

$$\|(\lambda I - A)^{-1}\|_{\mathscr{L}(L^2(\Omega))} \le \frac{C}{|\lambda| + 1}$$

for some $C > 0$.

D.8 Semigroup Theory

When we try to solve parabolic problems such as that defined by the heat conduction equation

$$\frac{\partial \varphi}{\partial t} = \kappa \frac{\partial^2 \varphi}{\partial x^2}$$

we can try to write the equation as an ordinary differential equation of the form

$$\dot{x} = Ax$$

where A is an operator defined on some Hilbert space. The main problem is that A is usually not bounded and so we cannot define the operator $\exp(At)$ as in the finite-dimensional case. However, we can define, in many cases, an operator which has similar properties to the exponential. It is called a semigroup of operators and we outline the main ideas next. We shall state the main results without proof since they can be found in many standard texts (*e.g.* [7]).

Definition D.21. A (*strongly continuous*) *semigroup of operators* on a Banach space X is an operator-valued function $T : \mathbb{R}^+ \to \mathscr{B}(X)$ such that:

(a) $T(0) = I$.

(b) $T(t_1 + t_2) = T(t_1) + T(t_2)$.

(c) $\lim_{t \to 0+} T(t)x = x$, for all $x \in X$, *i.e.* T is strongly continuous at $t = 0$.

Definition D.22. If $T(t)$ is a strongly continuous semigroup of operators on a Banach space X then the limit

$$Ax \overset{\Delta}{=} \lim_{t \to 0+} \left\{ \frac{T(t)x - x}{t} \right\}, \text{ for } x \in \mathscr{D}(A)$$

is called the *infinitesimal generator* of $T(t)$.

It can be shown that A is a closed operator with dense domain in X. We have

Theorem D.12. *If $T(t)$ is a strongly continuous semigroup of operators on a Banach space X with generator A, then*
(a) $T(t) : \mathscr{D}(A) \to \mathscr{D}(A)$
(b) $\frac{d^n}{dt^n}(T(t)x) = A^n T(t)x = T(t)A^n x$, for all $x \in \mathscr{D}(A^n)$ and $t > 0$.

Moreover we have similar boundedness properties to the finite-dimensional case.

Theorem D.13. *Let $T(t)$ be a strongly continuous semigroup of operators on a Banach space X. Then:*
(a) $\|T(t)\|$ is locally bounded on $[0,\infty)$.
(b) $T(t)$ is strongly continuous in t.
(c) $\omega_0 \overset{\Delta}{=} \inf_{t>0} \frac{1}{t} \log \|T(t)\| = \lim_{t\to\infty} \frac{1}{t} \log \|T(t)\| < \infty$.
(d) for all $\omega > \omega_0$ there exists $M(= M(\omega))$ such that $\|T(t)\| \le Me^{\omega t}, t \ge 0$.

The resolvent operator of the generator of a semigroup has a simple representation. In fact, if $\Re(s) > \omega$ where $\|T(t)\| \le Me^{\omega t}$ then $s \in \rho(A)$ and

$$R(s;A)x = \int_0^\infty e^{-st} T(t)x dt, \text{ for all } x \in X.$$

The next result (the Hille-Yosida theorem) characterises the generators of semigroups.

Theorem D.14. *In order that a closed linear operator A with dense domain in a Banach space X generates a strongly continuous semigroup, it is necessary and sufficient that there exist real numbers M, ω such that, for all real $\sigma > \omega$, $\sigma \in \rho(A)$ and*

$$\|R(\sigma;A)^m\| \le M(\sigma - \omega)^{-m}, \;\; m \ge 1.$$

Then we have

$$\|T(t)\| \le Me^{\omega t}.$$

We have seen above that, for any semigroup $T(t)$, we have $T(t) : \mathscr{D}(A) \to \mathscr{D}(A)$. This means that we only obtain a solution of the equation

$$\dot{x} = Ax, \;\; x(0) = x_0$$

for $x_0 \in \mathscr{D}(A)$. There is an important class of semigroups, namely the analytic ones, for which we have $T(t) : X \to \mathscr{D}(A)$, *i.e.* the semigroup has a smoothing property on the initial data. Many parabolic systems have this property.

Definition D.23. Let A be a closed operator defined on a Banach space with $\overline{\mathscr{D}(A)} = X$, such that for some $\varphi \in (0, \pi/2)$ we have:

(a) $S_{\varphi+\pi/2} \overset{\Delta}{=} \{\lambda \in \mathbb{C} : \ |\arg \lambda| < \varphi + \pi/2\} \subseteq \rho(A)$
(b) $\|R(\lambda; A)\| < \frac{C}{|\lambda|}$ if $\lambda \in S_\varphi,\ \lambda \neq 0$

where C is a constant independent of λ. Then A is called a *sectorial operator*.

Theorem D.15. *If A is a sectorial operator, then A generates a strongly continuous semigroup $T(t)$ such that*

(a) $T(t)$ can be analytically continued into the sector

$$S_\varphi = \{t \in \mathbb{C} : \ |arg\ t| < \varphi,\ t \neq 0\}$$

(b) $AT(t)$ and $dT(t)/dt$ are bounded for each $t \in S_\varphi$ and

$$\frac{dT(t)}{dt} x = AT(t)x, \text{ for all } x \in X$$

(c) for any $\varepsilon \in (0, \varphi)$, there exists C' (=$C'(\varepsilon)$) such that

$$\|T(t)\| \leq C', \ \ \|AT(t)\| < \frac{C'}{|t|}, \text{ for } t \in S_{\varphi-\varepsilon}.$$

In many parabolic systems, we require fractional powers of operators. These are defined in the following way:

Definition D.24. If A is a sectorial operator which generates a strongly continuous semigroup $T(t)$ and $\Re\sigma(A) < 0$, then for $\alpha > 0$ we define

$$A^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} T(t) dt$$

and if $\alpha > 0$ we put

$$A^\alpha = (A^{-\alpha})^{-1}.$$

Then A^α has the properties

(a) A^α is closed and densely defined if $\alpha > 0$.
(b) If $\alpha \geq \beta$, then $\mathscr{D}(A^\alpha) \subseteq \mathscr{D}(A^\beta)$.
(c) $A^\alpha A^\beta = A^\beta A^\alpha = A^{\alpha+\beta}$ on $\mathscr{D}(A^\gamma)$ where $\gamma = \max(\alpha, \beta, \alpha+\beta)$.

The importance of the semigroup approach is that, given an inhomogeneous system

$$\dot{x}(t) = Ax(t) + f(t), \ \ x(0) = x_0 \in X$$

then the solution is given by

$$x(t) = T(t)x_0 + \int_0^t T(t-s)f(s)ds$$

for many types of functions such as locally Lipschitz ones.

D.9 The Contraction Mapping and Implicit Function Theorems

We finish this appendix by stating the contraction mapping theorem and using it to give a simple proof of the implicit function theorem on Banach spaces.

Theorem D.16. *(Banach) If (M,d) is a complete metric space and $f : M \to M$ is a contraction mapping, so that there exists $\mu < 1$ such that*

$$d(f(x), f(y)) \leq \mu d(x,y), \textit{ for all } x,y \in M,$$

then there exists a unique fixed point a of f in M, i.e.

$$f(a) = a.$$

The implicit function theorem for Banach spaces can now be stated in the following way.

Theorem D.17. *Let X,Y and Z be Banach spaces and let $U \subseteq X$, $V \subseteq Y$be open subsets. Suppose that $F : U \times V \to Z$ is a continuously differentiable (in the sense of Fréchet) function on $U \times V$. Let $(x_0, y_0) \in U \times V$ and assume that $F(x_0, y_0) = 0$ and that the partial derivative $F_x(x_0, y_0) \in \mathscr{B}(X,Z)$ (it is essential that it is bounded) and has continuous inverse.*

Then there exists a neighbourhood $U' \times V' \subseteq U \times V$ of (x_0, y_0) and a function $h : V' \to U'$ with $h(y_0) = x_0$ such that

$$F(x,y) = 0 \textit{ if and only if } x = h(y), \textit{ for all } (x,y) \in U' \times V'.$$

Moreover, h has the same level of differentiability as F.

Proof. Let $A = (F_x(x_0, y_0))^{-1} \in \mathscr{B}(Z,X)$ and define the operator

$$K(x,y) = x - AF(x,y).$$

Then K is continuously differentiable (the same number of times as F) with

$$K(x_0, y_0) = x_0, \quad K_x(x_0, y_0) = 0$$

and

$$\|K_x(x,y)\| < 1$$

in a neighbourhood of (x_0, y_0), by continuity of K_x. Now use the contraction mapping theorem. □

References

1. Dunford, N., Schwartz, J.T.: Linear Operators. Wiley Interscience, New York (1959)
2. Yosida, K.: Functional Analysis. Springer, New York (1974)
3. Taylor, A.: Functional Analysis. Wiley Interscience, New York (1958)
4. Horvath, J.: Topological Vector Spaces and Distributions. Addison-Wesley, New York (1966)
5. Treves, F.: Topological Vector Spaces, Distributions and Kernels. Academic Press, New York (1967)
6. Adams, R.: Sobolev Spaces. Academic Press, New York (1975)
7. Hille, E., Phillips, R.S.: Functional Analysis and Semigroups, vol. 31. Amer. Math. Soc. Coll. Publ. (1957)

Index

Lecture Notes in Control and Information Sciences

Edited by M. Thoma, F. Allgöwer, M. Morari

Further volumes of this series can be found on our homepage:
springer.com

Vol. 400: Tomás-Rodríguez, M.;
Banks, S.P.:
Linear, Time-varying Approximations
to Nonlinear Dynamical Systems
298 p. 2010 [978-1-84996-100-4]

Vol. 399: Edwards, C.; Lombaerts, T.;
Smaili, H. (Eds.):
Fault Tolerant Flight Control
appro. 350 p. 2010 [978-3-642-11689-6]

Vol. 398: Hara, S.; Ohta, Y.;
Willems, J.C.; Hisaya, F. (Eds.):
Perspectives in Mathematical System
Theory, Control, and Signal Processing
appro. 370 p. 2010 [978-3-540-93917-7]

Vol. 397: Yang, H.; Jiang, B.;
Cocquempot, V.:
Fault Tolerant Control Design for
Hybrid Systems
191 p. 2010 [978-3-642-10680-4]

Vol. 396: Kozlowski, K. (Ed.):
Robot Motion and Control 2009
475 p. 2009 [978-1-84882-984-8]

Vol. 395: Talebi, H.A.; Abdollahi, F.;
Patel, R.V.; Khorasani, K.:
Neural Network-Based State
Estimation of Nonlinear Systems
appro. 175 p. 2010 [978-1-4419-1437-8]

Vol. 394: Pipeleers, G.; Demeulenaere, B.;
Swevers, J.:
Optimal Linear Controller Design for
Periodic Inputs
177 p. 2009 [978-1-84882-974-9]

Vol. 393: Ghosh, B.K.; Martin, C.F.;
Zhou, Y.:
Emergent Problems in Nonlinear
Systems and Control
285 p. 2009 [978-3-642-03626-2]

Vol. 392: Bandyopadhyay, B.; Deepak, F.;
Kim, K.-S.:
Sliding Mode Control Using Novel Sliding
Surfaces
137 p. 2009 [978-3-642-03447-3]

Vol. 391: Khaki-Sedigh, A.; Moaveni, B.:
Control Configuration Selection for
Multivariable Plants
232 p. 2009 [978-3-642-03192-2]

Vol. 390: Chesi, G.; Garulli, A.;
Tesi, A.; Vicino, A.:
Homogeneous Polynomial Forms for
Robustness Analysis of Uncertain Systems
197 p. 2009 [978-1-84882-780-6]

Vol. 389: Bru, R.; Romero-Vivó, S. (Eds.):
Positive Systems
398 p. 2009 [978-3-642-02893-9]

Vol. 388: Jacques Loiseau, J.; Michiels, W.;
Niculescu, S-I.; Sipahi, R. (Eds.):
Topics in Time Delay Systems
418 p. 2009 [978-3-642-02896-0]

Vol. 387: Xia, Y.;
Fu, M.; Shi, P.:
Analysis and Synthesis of
Dynamical Systems with Time-Delays
283 p. 2009 [978-3-642-02695-9]

Vol. 386: Huang, D.;
Nguang, S.K.:
Robust Control for Uncertain Networked Control
Systems with Random Delays
159 p. 2009 [978-1-84882-677-9]

Vol. 385: Jungers, R.:
The Joint Spectral Radius
144 p. 2009 [978-3-540-95979-3]

Vol. 384: Magni, L.; Raimondo, D.M.;
Allgöwer, F. (Eds.):
Nonlinear Model Predictive Control
572 p. 2009 [978-3-642-01093-4]

Vol. 383: Sobhani-Tehrani E.:
Khorasani K.;
Fault Diagnosis of Nonlinear Systems
Using a Hybrid Approach
360 p. 2009 [978-0-387-92906-4]

Vol. 382: Bartoszewicz A.;
Nowacka-Leverton A.:
Time-Varying Sliding Modes for Second
and Third Order Systems
192 p. 2009 [978-3-540-92216-2]

Vol. 381: Hirsch M.J.; Commander C.W.;
Pardalos P.M.; Murphey R. (Eds.):
Optimization and Cooperative
Control Strategies: Proceedings of the 8th
International Conference on Cooperative
Control and Optimization
459 p. 2009 [978-3-540-88062-2]

Vol. 380: Basin M.:
New Trends in Optimal Filtering and Control for
Polynomial and Time-Delay Systems
206 p. 2008 [978-3-540-70802-5]

Vol. 379: Mellodge P.; Kachroo P.:
Model Abstraction in Dynamical Systems:
Application to Mobile Robot Control
116 p. 2008 [978-3-540-70792-9]

Vol. 378: Femat R.; Solis-Perales G.:
Robust Synchronization of Chaotic Systems
Via Feedback
199 p. 2008 [978-3-540-69306-2]

Vol. 377: Patan K.:
Artificial Neural Networks for
the Modelling and Fault
Diagnosis of Technical Processes
206 p. 2008 [978-3-540-79871-2]

Vol. 376: Hasegawa Y.:
Approximate and Noisy Realization of
Discrete-Time Dynamical Systems
245 p. 2008 [978-3-540-79433-2]

Vol. 375: Bartolini G.; Fridman L.;
Pisano A.; Usai E. (Eds.):
Modern Sliding Mode Control Theory
465 p. 2008 [978-3-540-79015-0]

Vol. 374: Huang B.; Kadali R.:
Dynamic Modeling, Predictive Control
and Performance Monitoring
240 p. 2008 [978-1-84800-232-6]

Vol. 373: Wang Q.-G.; Ye Z.; Cai W.-J.;
Hang C.-C.:
PID Control for Multivariable Processes
264 p. 2008 [978-3-540-78481-4]

Vol. 372: Zhou J.; Wen C.:
Adaptive Backstepping Control of Uncertain
Systems
241 p. 2008 [978-3-540-77806-6]

Vol. 371: Blondel V.D.; Boyd S.P.;
Kimura H. (Eds.):
Recent Advances in Learning and Control
279 p. 2008 [978-1-84800-154-1]

Vol. 370: Lee S.; Suh I.H.;
Kim M.S. (Eds.):
Recent Progress in Robotics:
Viable Robotic Service to Human
410 p. 2008 [978-3-540-76728-2]

Vol. 369: Hirsch M.J.; Pardalos P.M.;
Murphey R.; Grundel D.:
Advances in Cooperative Control and
Optimization
423 p. 2007 [978-3-540-74354-5]

Vol. 368: Chee F.; Fernando T.
Closed-Loop Control of Blood Glucose
157 p. 2007 [978-3-540-74030-8]

Vol. 367: Turner M.C.; Bates D.G. (Eds.):
Mathematical Methods for Robust and
Nonlinear Control
444 p. 2007 [978-1-84800-024-7]

Vol. 366: Bullo F.; Fujimoto K. (Eds.):
Lagrangian and Hamiltonian Methods for
Nonlinear Control 2006
398 p. 2007 [978-3-540-73889-3]

Vol. 365: Bates D.; Hagström M. (Eds.):
Nonlinear Analysis and Synthesis
Techniques for Aircraft Control
360 p. 2007 [978-3-540-73718-6]

Vol. 364: Chiuso A.; Ferrante A.;
Pinzoni S. (Eds.):
Modeling, Estimation and Control
356 p. 2007 [978-3-540-73569-4]

Vol. 363: Besançon G. (Ed.):
Nonlinear Observers and Applications
224 p. 2007 [978-3-540-73502-1]

Vol. 362: Tarn T.-J.; Chen S.-B.;
Zhou C. (Eds.):
Robotic Welding, Intelligence and Automation
562 p. 2007 [978-3-540-73373-7]

Vol. 361: Méndez-Acosta H.O.; Femat R.;
González-Álvarez V. (Eds.):
Selected Topics in Dynamics and Control of
Chemical and Biological Processes
320 p. 2007 [978-3-540-73187-0]

Vol. 360: Kozlowski K. (Ed.):
Robot Motion and Control 2007
452 p. 2007 [978-1-84628-973-6]

Vol. 359: Christophersen F.J.:
Optimal Control of Constrained
Piecewise Affine Systems
190 p. 2007 [978-3-540-72700-2]

Vol. 358: Findeisen R.; Allgöwer
F.; Biegler L.T. (Eds.): Assessment and Future
Directions of Nonlinear
Model Predictive Control
642 p. 2007 [978-3-540-72698-2]