N72-24318

measurement systems laboratory

massachusetts institute of technology, cambridge, massachusetts 02139

TE-49

APPLICATION OF CONTRACTION MAPPINGS TO THE CONTROL OF NONLINEAR SYSTEMS

BY

William Robert Killingsworth, Jr.

CASE FILE COPY

TE-49

APPLICATION OF CONTRACTION MAPPINGS TO THE CONTROL OF NONLINEAR SYSTEMS

by

William Robert Killingsworth, Jr.

January, 1972

Measurement Systems Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

APPROVED: W. Markey
Director
Measurement Systems Laboratory

APPLICATION OF CONTRACTION MAPPINGS TO THE CONTROL OF NONLINEAR SYSTEMS

by

WILLIAM ROBERT KILLINGSWORTH, JR.

B.S., Auburn University, 1966

M.S., Massachusetts Institute of Technology, 1968

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January 1972

Signature of Author William R Killingsworth Jr
Department of Aeronautics and Astronautics
January 1972

Certified by Peter Falb
Thesis Supervisor

Certified by Edward B. Roberts
Thesis Supervisor

Certified by John J Deyst Jr.
Thesis Supervisor

Accepted by Sidney R. Barry
Chairman, Departmental Committee on Graduate Students

APPLICATION OF CONTRACTION MAPPINGS TO

THE CONTROL OF NONLINEAR SYSTEMS

by

William Robert Killingsworth, Jr.

Submitted to the Department of Aeronautics and Astronautics on January 14, 1972 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

This research considers the theoretical and applied aspects of successive approximation techniques for the determination of controls for nonlinear dynamical systems. Particular emphasis is placed upon the methods of contraction mappings and modified contraction mappings. It is shown that application of the Pontryagin principle to the optimal nonlinear regulator problem results in necessary conditions for optimality in the form of a two point boundary value problem (TPBVP). The TPBVP is represented by an operator equation and functional analytic results on the iterative solution of operator equations are applied. The general convergence theorems are translated and applied to those operators arising from the optimal regulation of nonlinear systems. It is shown that simply structured matrices and similarity transformations may be used to facilitate the calculation of the matrix Green's functions and the evaluation of the convergence criteria. A controllability theory based on the integral representation of TPBVP's, the implicit function theorem, and contraction mappings is developed for nonlinear dynamical systems. Contraction mappings is theoretically and practically applied to a nonlinear control problem with bounded input control, and the Lipschitz norm is used to prove convergence for the nondifferentiable operator. A dynamic model representing community drug usage is developed and the contraction mappings method is used to study the optimal regulation of the nonlinear system.

Thesis Supervisors:

Professor John J. Deyst, Jr.

Title: Associate Professor of Aeronautics and Astronautics, MIT

Professor Peter L. Falb

Title: Professor, Division of Applied Mathematics, Brown University

Professor Edward B. Roberts

Title: Professor of Management, MIT

Page intentionally left blank

ACKNOWLEDGEMENTS

The author wishes to thank his thesis committee: Professor John J. Deyst, committee chairman, whose guidance and encouragement were invaluable; Professor Peter L. Falb who generously provided many hours of fruitful discussions and whose work provided the basis and motivation for the thesis; and Professor Edward B. Roberts whose incisive comments and thought provoking discussions were highly valued. The unique perspective and penetrating insight of each of these gentlemen has made association with this committee a particularly rewarding experience for the author.

Special thanks are due to Professor Walter Wrigley who provided guidance and encouragement throughout the author's doctoral program.

The author is indebted to Dr. Robert Stern and the staff of the Measurement Systems Laboratory who gave freely of their time and energy.

Thanks are also due to Miss Marjorie Goldstein whose conscientious efforts in typing the thesis are gratefully acknowledged, and to Mrs. Ann Preston for preparation of the figures and attending to the many details of publication.

Finally, the author wishes to extend special thanks to his dear wife Joyce for her unfailing support, encouragement, and patience throughout this endeavor.

This research was supported by a grant from the National Aeronautics and Space Administration, NGR 22-009-010 and NsG 22-009-270.

The publication of this thesis does not constitute approval by the National Aeronautics and Space Administration or by the MIT Measurement Systems Laboratory of the findings or the conclusions contained therein. It is published only for the exchange and stimulation of ideas.

Page intentionally left blank

TABLE OF CONTENTS

Chapter		Page
1	INTRODUCTION	1
	1.1 Background	1
	1.2 Description of the Problem	3
	1.3 Synopsis	4
2	OPTIMAL REGULATION OF NONLINEAR SYSTEMS	7
	2.1 Introduction	7
	2.2 Optimal Linear Regulator	7
	2.3 Optimal Regulation of Nonlinear Systems	11
3	METHODS OF SOLVING TPBVP's	19
	3.1 Introduction	19
	3.2 Representation of TPBVP's	19
	3.3 Frechet Derivatives and Lipschitz Norms	23
	3.4 Contraction Mappings Method	26
	3.5 Modified Contraction Mappings	31
	3.6 Applications of Contraction Mappings	36
4	CALCULATION OF CONVERGENCE CRITERIA	43
	4.1 Introduction	43
	4.2 Evaluation of Convergence Criteria	44
	4.3 Boundary Value Sets of Interest	46
	4.4 Boundary Set for Regulation wiht Terminal Cost	47

Page

4.5 Boundary Set for Regulation with No Terminal Cost 54

4.6 Application of Similarity Transformation 56

4.7 Approximate Technique 66

5 CONTROLLABILITY FOR NONLINEAR SYSTEMS 71

5.1 Introduction 71

5.2 Controllability for Linear Systems 71

5.3 Nonlinear Controllability 75

5.4 Evaluation of Controllability Convergence Parameters 84

6 NUMERICAL EXAMPLES 91

6.1 Introduction 91

6.2 Van der Pol's Equation 92

6.3 Null Controllability with Bounded Control 102

6.4 Controllability of Satellite Pitch Motion 109

7 PRELIMINARY STUDY ON THE DYNAMICS OF DRUG USAGE WITHIN A COMMUNITY 119

7.1 Introduction 119

7.2 Development of a Dynamic Model 120

7.3 Optimal Regulation of the Nonlinear System 130

8 SUMMARY, CONTRIBUTIONS AND RECOMMENDATIONS 141

8.1 Summary 141

8.2 Contributions 143

8.3 Recommendations 143

Appendix 145

A Description of Contraction Mappings Computer Algorithm 145

Bibliography 173

LIST OF FIGURES

Figure		Page
6.1	The sphere $\bar{S}(y_0,r)$	96
6.2	The sphere $\bar{S}(y_0,r)$	98
6.3	Control iterations	101
6.4	Comparison of performance for contraction mappings and modified contraction mappings	102
6.5	The sphere $\bar{S}(y_0,r)$	106
6.6	The sphere $\bar{S}(y_0,r)$ for $T = \pi$	112
6.7	The sphere $\bar{S}(y_0,r)$ for $T = \pi/2$	115
6.8	State and control history	117
7.1	Levels of drug usage	121
7.2	Feedback structure of drug usage	121
7.3	Availability of Potential Users multiplier	124
7.4	Availability of Drug Users multiplier	125
7.5	Effect of education on addiction growth rate	126
7.6	Police effectiveness	127
7.7	Availability of Addicts multiplier	128
7.8	Police effectiveness	134
7.9	Effect of education on addiction growth rate	134
7.10	The function $y_0(t)$ for T = 12 months	135
7.11	Addicts, Police, and Education for T = 12 months	137
7.12	The function $y_0(t)$ for T = 48 months	138
7.13	Addicts, Police, and Education for T = 48 months	139

CHAPTER I

INTRODUCTION

1.1. Background

Optimal control theory has experienced an increasing growth of interest in the past two decades. Initially motivated by the aerospace effort, optimal control theory is now involved in many aspects of general systems engineering. Applications range from chemical process control to attempts at managerial and economic planning.

One of the most important and most widely treated problems to date in optimal control theory is the so-called "Linear Regulator Problem". Historically, this problem arose in Wiener's work concerning stationary time series and linear filtering and prediction [W1]. Under the name "Minimum Integral Squared Error", development of this problem was continued through the 1950's by Newton [N1], Booten [B3], and Zadeh [Z1]. Finally utilizing the techniques of modern control theory, Kalman [K1] presented important new aspects of the problem.

The prominence of this problem is due to two primary factors. First, the problem provides a strong link between the classical methods of analytic feedback system design via frequency domain methods and the more recent variational approach favoring analysis in the time domain [K2], [W2]. Secondly, the problem allows the determination of optimal controls in closed form with mathematical ease. (For general development and presentation of the problem, see Athans and Falb [A1] and Lee and Markus [L1]). Finally, a pragmatic motivation for considering the problem is the ease with which the quadratic cost criteria can be interpreted

physically. Consequently, optimal linear regulation has been extensively applied to various systems. For example, the theory has found widespread applications in the area of automatic flight control systems. Much of this work is based on the significant efforts of Rynaski [R4], [R5]. Other examples of optimal linear regulation are contained in Dyer and McReynolds [D2].

However, few systems can adequately be described by a linear dynamic model. In particular, increasing effort is now being devoted to the development of models representing systems as varied and as complex as urban areas, natural resource depletion, management of R and D efforts, and drug usage within a community. These models are primarily due to the efforts of Forrester [F3, F4,F5,F6] and Roberts [R2]. Along with many engineering systems, these systems contain inherent nonlinearities which must be included in any meaningful study.

In contrast to linear systems, the regulation of nonlinear dynamical systems has received limited attention, most of a specialized nature. The primary reason for this seems to lie in the fact that nonlinear optimal control problems can rarely be solved analytically or, more specifically, in feedback form as for linear regulators. As a result, one must often resort to iterative numerical techniques for the determination of the optimizing solutions. Consequently, much of the analysis regarding regulation of nonlinear systems concerns techniques for determining suboptimal feedback controllers. (See for example [D1], [G2], [L3], [P1], [S2], [J1], [F8], and [B5]). Most of these approaches involve the modeling of the nonlinear system as a linear system in some manner. A somewhat different approach, not suboptimal, is taken by Brunovsky [B4] and Lukes [L4]. Both of these treatments are closely related to the basic hypothesis that the system be stabilizable [L1]. Under the assumption of complete controllability, Brunovsky approached the problem via Lyapunov functions. Lukes requires the system be

stabilizable and then uses Lyapunov-like theory to obtain results for feedback controllers.

The direction of these various approaches is primarily generated by the desire for a feedback controller. However, there is a second, more esoteric reason, and that is the desire for general results. Unfortunately, the undiscerning application of an algorithm often limits insight into the underlying structure of the problem being considered. This loss of general information is often due to the fact that practical convergence criteria are few for most of the iterative methods used in the solution of optimal control problems. Theoretical aspects of these criteria have been investigated by numerous applied mathematicians (see Kantorovich [K4] and Collatz [C2]). The Russian Kantorovich [K4] was one of the first to develop and unify the mathematical theory of iterative methods. Using the power of functional analysis methods, he presented convergence results for such basic iterative schemes as contraction mappings and Newton's method. These basic results have been considerably broadened, modernized, and made practical by the efforts of Falb and de Jong [F1]. In their book, they present the derivation of general convergence criteria for the application of various successive approximation methods to the solution of optimal control problems.

1.2. Description of the Problem

The primary goal of this research is the consideration of the theoretical and applied aspects of successive approximation techniques for the solution of optimal nonlinear regulator problems. Application of the Pontryagin principle to the posed optimization problem results in necessary conditions for optimality in the form of a two point boundary value problem (TPBVP). Hence, the central

theme of this study shall be the application of successive approximation methods to the solution of nonlinear TPBVP's which arise from optimal nonlinear regulation. The basic approach to be used is to represent the TPBVP by an operator equation and then apply functional analytic results in the iterative solution of operator equations.

In particular, we shall investigate the contraction mappings method and the modified contraction mappings method. We have as our first objective the translation and application of the general convergence theorems to those operators originating in the optimal regulation of a nonlinear system. A second objective is the development of techniques to facilitate the evaluation of the convergence criteria. Finally, example problems will be solved to demonstrate the usefulness of the theory.

1.3. Synopsis

A brief summary of the dissertation is as follows: In Chapter 2, the optimal regulation of dynamical systems is introduced. In particular, we discuss the reduction of optimization problems to two point boundary value problems by means of Pontryagin's principle. Results are derived for optimal regulation of linear dynamical systems (Section 2.2) and several classes of nonlinear systems (Section 2.3). Optimal system regulation is considered for both unconstrained and bounded controls. In Chapter 3, methods of solving two point boundary value problems are presented. In particular, the integral equation representation of two point boundary value problems is introduced (Section 3.2). The book by Falb and de Jong [F1] was used as the main reference for this chapter. The integral representation makes it possible to consider the solution of a two point boundary value problem as the solution of a corresponding operator equation.

A review of Lipschitz norms and derivative norms for the integral operator is presented (Section 3.3) and the methods of contraction mappings (Section 3.4) and modified contraction mappings (Section 3.5) are introduced. Convergence theorems for both methods are presented. Chapter 3 concludes with the application of contraction mappings to the solution of two point boundary value problems arising in Chapter 2 and the derivation of translated convergence theorems. Chapter 4 is devoted to a rather detailed investigation into the calculation of the theoretical convergence criteria. Upper bounds are presented for the Lipschitz norm and derivative norm (Section 4.2) and various techniques for evaluating these bounds are introduced. In particular, the use of simply structured matrices (Sections 4.4, 4.5) and similarity transformations (Section 4.6) are considered. The use of partitioned matrices in these developments provides considerable insight into the generic structure of the Green's matrices contained within the integral representation. In Chapter 5 the issue of controllability for nonlinear systems is considered. Specifically, it is shown that controllability for linear systems (Section 5.2) and nonlinear systems (Section 5.3) may be studied via the integral representation and contraction mappings. In Chapter 6 we present numerical examples to illustrate the theoretical and practical application of contraction mappings to the regulation and control of nonlinear systems. In Chapter 7, a dynamic model is developed for a socio-economic system and contraction mappings is used to investigate the optimal regulation of this nonlinear system. Finally, in Chapter 8, we summarize our results and indicate directions in which future research may be done. We conclude with an appendix which gives the actual computer program (written in the FORTRAN language) which was used in the application of contraction mappings to the problem discussed in Chapter 7.

Page intentionally left blank

CHAPTER 2

OPTIMAL REGULATION OF DYNAMICAL SYSTEMS

2.1. Introduction

An optimal control problem is a composite concept consisting of four basic elements: (1) a dynamical system, (2) a set of initial states and a set of final states, (3) a set of admissible controls, and (4) a cost functional to be minimized. The problem consists of finding the admissible control which transfers the state of the dynamical system from the set of initial states to the set of final states and, in so doing, minimizes the cost functional. In this chapter we discuss the optimal regulation of nonlinear systems and the reduction of the optimization problem to a TPBVP by means of Pontryagin's principle.

2.2. Optimal Linear Regulator

As a preface to the nonlinear system analysis, we shall present the basic results for the optimal linear regulator. (For a very thorough treatment of this problem see Kleinman [K4]).

Definition 2.2.1. Linear Dynamical System

A linear dynamical system is characterized by the following elements:

(1) A state vector x of dimension n

(2) A control input vector u of dimension r

(3) A linear differential equation which describes the evolution of the system in time, i.e.,

$$\dot{x}(t) = A(t)\, x(t) + B(t)\, u(t) \qquad 2.2.2$$

where A(t) is an n×n matrix and B(t) is an n×r matrix.

Now given an initial state, $x(t_0) = x_0$, and assuming the control $u(t)$ is not constrained, the optimal linear regulator problem is then to determine the control $u(t)$ which minimizes the quadratic cost function

$$J(u) = \frac{1}{2}\langle x(T), Kx(T)\rangle + \frac{1}{2}\int_{t_0}^{T} [\langle x(t), Q(t)\, x(t)\rangle + \langle u(t), R(t)\, u(t)\rangle]dt \qquad 2.2.3$$

where

The terminal time T is specified 2.2.4

K is a constant n×n positive semidefinite matrix

Q(t) is an n×n positive semidefinite matrix

R(t) is an r×r positive definite matrix

and K and Q(t) are not both identically zero.

Application of the minimum principle to the optimization problem posed above yields necessary conditions for optimality in the form of the 2n × 2n canonical system of equations

$$\begin{bmatrix} \dot{x}(t) \\ \dot{p}(t) \end{bmatrix} = \begin{bmatrix} A(t) & -B(t)R^{-1}(t)B'(t) \\ -Q(t) & -A'(t) \end{bmatrix} \begin{bmatrix} x(t) \\ p(t) \end{bmatrix} \qquad 2.2.5$$

subject to the boundary conditions

$$x(t_0) = x_0$$
$$p(T) = K\, x(T). \qquad 2.2.6$$

The H-minimal control for $t \in [t_0, T]$ is then given by

$$u(t) = -R^{-1}(t)\, B'(t)\, p(t). \qquad 2.2.7$$

The boundary conditions specified by eq(2.2.6) may be expressed more compactly as

$$\begin{bmatrix} I & 0 \\ \theta & 0 \end{bmatrix} \begin{bmatrix} x(t_0) \\ p(t_0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \begin{bmatrix} x(T) \\ p(T) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.2.8$$

where I is the n×n identity matrix and 0 is the n×n zero matrix. This form of expressing boundary conditions will become important in the sequel. The TPBVP arising from the linear optimal regulator problem may then be put into the form

$$\dot{y}(t) = S(t)\ y(t) \qquad 2.2.9$$

$$My(t_0) + Ny(T) = c$$

where y is the 2n composite vector

$$y(t) = \begin{bmatrix} x(t) \\ p(t) \end{bmatrix} \qquad 2.2.10$$

S(t) is the 2n × 2n matrix

$$S(t) = \begin{bmatrix} A(t) & -B(t)R^{-1}(t)B'(t) \\ -Q(t) & -A'(t) \end{bmatrix} \qquad 2.2.11$$

and M and N are the 2n × 2n boundary value matrices

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \qquad 2.2.12$$

and c is the 2n constant matrix

$$c = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.2.13$$

In many physical situations, the input control u(t) may not take on all values. As an introduction to systems with bounded control, let us now suppose the input control to the linear dynamical system is constrained in magnitude by the relation

$$|u_j(\cdot)| \leq 1 \qquad j = 1,\ldots,r\ . \qquad 2.2.14$$

Then given an initial state for the linear dynamical system, the optimal linear regulator problem is to determine an admissible control $u(t) \in \Omega$ which minimizes the quadratic cost functional given in (2.2.3).

It is shown in [A1] that the necessary conditions for optimality reduce to the 2n × 2n canonical system of equations

$$\dot{x}(t) = A(t)\ x(t) - B(t)\ \mathrm{SAT}\ \{R^{-1}(t)\ B'(t)\ p(t)\} \qquad 2.2.15$$

$$\dot{p}(t) = Q(t)\ x(t) - A'(t)\ p(t)$$

subject to the boundary conditions

$$x(t_0) = x_0 \qquad 2.2.16$$

$$p(T) = K\ x(T),$$

where the SAT function is defined as

$$\mathrm{SAT}\{y\} = \begin{cases} 1 \ , & y > 1 \\ y \ , & |y| \leq 1 \\ -1 \ , & y < -1 \end{cases} \qquad 2.2.17$$

It is seen this system of 2n differential equations is not linear. The necessary conditions thus reduce to a nonlinear TPBVP of the form

$$\dot{y}(t) = S(t)\ y(t) + f(y(t)) \qquad 2.2.18$$

$$My(t_0) + Ny(T) = c$$

where y is the composite 2n vector

$$y(t) = \begin{bmatrix} x(t) \\ p(t) \end{bmatrix} \qquad 2.2.19$$

S(t) is the 2n × 2n matrix

$$S(t) = \begin{bmatrix} A(t) & 0 \\ -Q(t) & -A'(t) \end{bmatrix} \qquad 2.2.20$$

f(y(t)) is the 2n vector function

$$f(y(t)) = \begin{bmatrix} -B(t)\ \mathrm{SAT}\{R^{-1}(t)\ B'(t)p(t)\} \\ 0 \end{bmatrix} \qquad 2.2.21$$

M and N are 2n × 2n matrices and c is the 2n vector

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad N = \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \quad c = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.2.22$$

This example illustrates a nonlinear TPBVP arising from the optimization of a linear system. We shall now consider the optimization of nonlinear systems and the forms of the resultant TPBVP's.

2.3. Optimal Regulation of Nonlinear Systems

In this section we shall consider the control of several classes of nonlinear systems subject to the quadratic cost functional given in (2.2.3). Our aim in this section is to reduce the necessary conditions for optimality to two point boundary value problems.

Example 2.3.1.

Many nonlinear systems contain nonlinearities involving only the state variables. Hence, rather than initially considering the most general formulation, we shall first consider the class described by the differential equation

$$\dot{x}(t) = A(t)\ x(t) + B(t)\ u(t) + \psi(x(t)) \qquad 2.3.2$$

where we assume $\psi(x(\cdot))$ and $(\partial\psi/\partial x)(x(\cdot))$ are continuous on R^n. We shall initially consider the control to be unconstrained, i.e., $u \in \Omega = R^n$. Again we shall consider the quadratic cost functional

$$J(u) = \frac{1}{2}\langle x(T), Kx(T)\rangle + \frac{1}{2}\int_{t_0}^{T} \left[\langle x(t),\ Q(t)\ x(t)\rangle + \langle u(t),\ R(t)\ u(t)\rangle\right]dt \qquad 2.3.3$$

subject to the assumptions of (2.2.4). Application of the minimum principle yields necessary conditions for optimality in the form of the $2n \times 2n$ canonical system of equations

$$\begin{bmatrix} \dot{x}(t) \\ \dot{p}(t) \end{bmatrix} = \begin{bmatrix} A(t) & -B(t)R^{-1}(t)B'(t) \\ Q(t) & -A'(t) \end{bmatrix} \begin{bmatrix} x(t) \\ p(t) \end{bmatrix} + \begin{bmatrix} \psi(x(t)) \\ -\ (\partial\psi/\partial x)'(x(t))\ p(t) \end{bmatrix} \qquad 2.3.4$$

subject to the boundary conditions

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(t_0) \\ p(t_0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \begin{bmatrix} x(T) \\ p(T) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.3.5$$

The H-minimal control for $t \in [t_0, T]$ is then given by

$$u(t) = -R^{-1}(t)\ B'(t)\ p(t). \qquad 2.3.6$$

It is often advantageous to standardize the time interval over which the TPBVP is defined. This standardization is accomplished by the introduction of a new variable. Let (see Long [L2])

$$t = t_0 + (T-t_0)s = b + as. \qquad 2.3.7$$

Here s is the new variable which varies between 0 and 1. In most cases we may take $t_0 = b = 0$. In terms of s and a, the TPBVP then becomes

$$\begin{bmatrix} \dot{x}(s) \\ \dot{p}(s) \end{bmatrix} = a \begin{bmatrix} A(as) & -B(as)R^{-1}(as)B'(as) \\ -Q(as) & -A'(as) \end{bmatrix} \begin{bmatrix} x(s) \\ p(s) \end{bmatrix} + a \begin{bmatrix} \psi(x(s)) \\ -(\partial\psi/\partial x)'(x(s))p(s) \end{bmatrix} \qquad 2.3.8$$

with

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(0) \\ p(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \begin{bmatrix} x(1) \\ p(1) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.3.9$$

where $(\dot{\ })$ indicates differentiation with respect to s. In the sequel, the TPBVP's which shall be considered will generally be normalized in this fashion.

Example 2.3.10.

As an illustration of the ideas presented in Example (2.3.1), let us consider the driven, second order nonlinear oscillator studied by Van der Pol. We have the system given as

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -x_1(t) + \varepsilon(1-x_1^2(t))x_2(t) + u(t), \end{aligned} \qquad 2.3.11$$

or in vector-matrix form as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0 \\ \varepsilon(1-x_1^2)x_2 \end{bmatrix} \qquad 2.3.12$$

The optimization problem to be considered is that of minimizing the cost functional

$$J = \frac{1}{2} \int_0^T (x_1^2(t) + x_2^2(t) + u^2(t))dt \qquad 2.3.13$$

subject to the boundary conditions

$$x_1(0) = x_0 \ , \ x_1(T) \ \text{ unspecified} \qquad 2.3.14$$
$$x_2(0) = 0 \ , \ x_2(T) \ \text{ unspecified.}$$

From eq (2.3.4), we have the 2n x 2n canonical system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{p}_1 \\ \dot{p}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ p_1 \\ p_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \varepsilon(1-x_1^2)x_2 \\ 2\varepsilon x_1 x_2 p_2 \\ -\varepsilon(1-x_1^2)p_2 \end{bmatrix} \qquad 2.3.15$$

subject to the boundary conditions

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \\ p_1(0) \\ p_2(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(1) \\ x_2(1) \\ p_1(1) \\ p_2(1) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad 2.3.17$$

or, in the more compact form,

$$\dot{y}(t) = Sy(t) + f(y(t)) \qquad 2.3.18$$
$$My(0) + Ny(1) = c.$$

In the sequel, this example will reappear as we consider the iterative solution of TPBVP's of the form (2.3.18).

The class of systems studied in Example 2.3.1 will now be reconsidered with a magnitude constraint upon the control.

Example 2.3.19.

Let us now consider the regulation of the previous system

$$\dot{x}(t) = A(t)\ x(t) + B(t)\ u(t) + \psi(x(t)) \qquad 2.3.20$$

where the input control vector is constrained in magnitude by

$$|u_j(\cdot)| \leq 1 \ , \ j=1,\ldots,r. \qquad 2.3.21$$

The cost functional is again given by (2.3.3). Application of the minimum principle yields the 2n × 2n system of canonical equations as

$$\begin{bmatrix} \dot{x}(t) \\ \dot{p}(t) \end{bmatrix} = \begin{bmatrix} A(t) & 0 \\ -Q(t) & -A'(t) \end{bmatrix} \begin{bmatrix} x(t) \\ p(t) \end{bmatrix} + \begin{bmatrix} \psi(x(t)) - B(t)SAT\{R^{-1}(t)B'(t)p(t)\} \\ - (\partial\psi/\partial x)'(x(t))\ p(t) \end{bmatrix} \qquad 2.3.22$$

subject to the boundary conditions

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(t_0) \\ p(t_0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \begin{bmatrix} x(T) \\ p(T) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.3.23$$

where the SAT function is specified in (2.2.16). For this system, the H-minimal control is given as

$$u(t) = -SAT\{R^{-1}(t)\ B'(t)\ p(t)\} \quad , \quad t \in [t_0,T]. \qquad 2.3.24$$

Example 2.3.25

In the previous example, we discussed the large class of nonlinear systems in which the nonlinearity is a function of only the state variable. Let us now consider the more general system described by the differential equation

$$\dot{x} = A(t)x + B(t)u + \psi(x,u) \qquad 2.3.26$$

where $A(t)$ is an $n \times n$ matrix, $B(t)$ is an $n \times r$ matrix, u is an unconstrained r-vector, and $\psi(x,u)$ and $(\partial\psi/\partial x)(x(\cdot),u(\cdot))$ are continuous in $R^n \times R^n$. The system is subject to the quadratic cost criteria given as

$$J = \frac{1}{2}\langle x(T),Kx(T)\rangle + \frac{1}{2} \int_{t_0}^{T} [\langle x(t),\ Q(t)x(t)\rangle + \langle u(t),\ R(t)u(t)\rangle]dt \qquad 2.3.27$$

under the assumptions of (2.2.4).

Following the Pontryagin minimum principle, the Hamiltonian for the optimization problem posed above is given as

$$H = \frac{1}{2}\langle x(t), Q(t)\ x(t)\rangle + \frac{1}{2}\langle u(t),\ R(t)u(t)\rangle + \langle A(t)\ x(t),\ p(t)\rangle + \langle B(t)\ u(t),\ p(t)\rangle + \langle \psi(x(t),\ u(t)),\ p(t)\rangle . \qquad 2.3.28$$

Formally applying the Pontryagin principle, the costate vector is then described by the differential equation

$$\dot{p}(t) = -\ Q(t)\ x(t) - A'(t)\ p(t) - (\partial\psi/\partial x)'(x(t),u(t))p(t). \qquad 2.3.29$$

Along the optimal trajectory we must have $u^*(t)$ minimizing the Hamiltonian, i.e.,

$$H(x^*(t),\ p^*(t),\ u^*(t),t) \leq H(x^*(t),p^*(t),\omega,t) \qquad 2.3.30$$

for all admissible ω and where $(\cdot)^*$ denotes optimal trajectories. If the Hamiltonian is normal [A1], the minimization equation (2.3.30) may be solved for the H-minimal u in terms of x,p, and t, i.e.,

$$u = \xi(x,p,t). \qquad 2.3.31$$

Now using (2.3.31) we define

$$\psi(x,\xi(x,p)) = \phi(x,p) \qquad 2.3.32$$

and

$$(\partial\psi/\partial x)(x,\xi(x,p)) = Z(x,p) \qquad 2.3.33$$

where ϕ is an n vector function and Z is an n x n matrix function.

$$\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} A(t) & 0 \\ -Q(t) & -A'(t) \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix} + \begin{bmatrix} B(t)\xi(x,p) + \phi(x,p) \\ -\ Z'(x,p)p \end{bmatrix} \qquad 2.3.34$$

subject to the boundary conditions

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x(t_0) \\ p(t_0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix}\begin{bmatrix} x(T) \\ p(T) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.3.35$$

These results may be applied to various forms of $\psi(x,p)$. In the following example we consider one such form.

Example 2.3.36

Consider the system described by

$$\dot{x} = A(t)x + B(t)u + D(x)u \qquad 2.3.36$$

where $A(t)$ is an $n \times n$ matrix, $B(t)$ is an $n \times r$ matrix, $D(x)$ is an $n \times r$ matrix, $D_{ij}(x)$ and $(\partial D_{ij}/\partial x_k)(x(\cdot))$ are continuous in R^n, and $u(\cdot)$ is an unconstrained r vector. Consider system (2.3.36) subject to the cost functional (2.3.27) and the initial condition $x(t_0) = x_0$.

Define the vector function $\xi(x,p,t)$ to contain the elements

$$\xi_i(x,p,t) = \left\langle p, -(\partial D/\partial x_i)(x)\, R^{-1}(t)[B'(t) + D'(x)]p \right\rangle \qquad 2.3.37$$

and define the matrix $c(x,t)$ as

$$C(x,t) = -B(t)\, R^{-1}(t)\, D'(t) - D(x)\, R^{-1}(t)[B'(t) + D'(x)]. \qquad 2.3.38$$

Using the results of Example 2.3.25 and (2.3.37), (2.3.38), the $2n \times 2n$ canonical system of equations is given as

$$\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} A(t) & -B(t)R^{-1}(t)B'(t) \\ -Q(t) & -A(t) \end{bmatrix}\begin{bmatrix} x \\ p \end{bmatrix} + \begin{bmatrix} C(x,t)p \\ \xi(x,p,t) \end{bmatrix} \qquad 2.3.39$$

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x(t_0) \\ p(t_0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix}\begin{bmatrix} x(T) \\ p(T) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 2.3.40$$

The various two point boundary value problems presented in the previous

examples can very rarely be solved analytically. Thus we shall investigate successive approximation techniques for the solution of TPBVP's in the next chapter.

Page intentionally left blank

CHAPTER 3

METHODS OF SOLVING TPBVP's

3.1. Introduction

In the analysis of optimal control problems, the necessary conditions for optimality are often in a form which may be reduced to a TPBVP of the form

$$\dot{y}(t) = F(y,t), \; g(y(0)) + h(y(1)) = c. \qquad 3.1.1$$

In particular, we presented in Chapter 2 various TPBVP's which originate in the optimal regulation of certain classes of nonlinear systems. We shall now illustrate that under certain conditions, such TPBVP's may be represented by operator equations of the form

$$y = T(y). \qquad 3.1.2$$

Then, following the lead of Falb and deJong [F1], we shall investigate the application of successive approximation techniques to the iterative solution of these operator equations.

3.2 Representation of TPBVP's

In this section we consider the (normalized) two point boundary value problem

$$\dot{y}(t) = F(y,t), \; g(y(0)) + h(y(1)) = c \qquad 3.2.1$$

where G, g, and h are vector valued functions and c is an element of R^P. We shall first review some results relating to the development of equivalent integral equation representations of the TPBVP(3.2.1). Most results in this section

come from Falb and de Jong [F1]. Since linear TPBVP's will play an important role in the integral equation representations, we begin our discussion with a consideration of linear TPBVP's.

Consider the linear TPBVP

$$\dot{y}(t) = V(t)y(t) + f(t) \ , \ My(0) + Ny(1) = c \qquad 3.2.2$$

where $V(t)$, M, and N are $p \times p$ matrices, and $f(t)$ and d are p vectors. We present the following theorem on the existence of a solution of equation (3.2.2).

Theorem 3.2.3

Suppose that the functions $V(t)$ and $f(t)$ satisfy appropriate smoothness and boundedness conditions and $\det[M + N\Phi^V(1,0)] \neq 0$ where $\Phi^V(t,s)$ is the fundamental matrix of $\dot{y} = Vy$. Then (3.2.2) has a unique solution $y(t)$ on $[0,1]$ which can be written in the form

$$y(t) = H(t)\ c + \int_0^1 G^J(t,s)f(s)ds \qquad 3.2.4$$

where the Green's matrices H and G are given by

$$H(t) = \Phi^V(t,0)[M+N\Phi^V(1,0)]^{-1} \qquad 3.2.5$$

and

$$G(t,s) = \begin{cases} \Phi^V(t,0)\ [M+N\Phi^V(1,0)]^{-1}M\Phi^V(0,s) \ , & 0 \leq s \leq t \\ -\Phi^V(t,0)[M+N\Phi^V(1,0)]^{-1}N\Phi^V(1,s) \ , & t < s \leq 1 \end{cases} \qquad 3.2.6$$

for all t,s in $[0,1]$.

Proof: (See [F1] for proof of theorem and technical conditions specified for $V(t)$ and $f(t)$).

The requirement in Theorem 3.2.3 that $\det[M+N\Phi^V(1,0)] \neq 0$ is crucial to the integral representation of TPBVP's. We therefore make the following definition.

Definition 3.2.7

Let V, M, N be $p \times p$ matrices. Then $J = \{V(t),M,N\}$ is called a boundary compatible set if and only if $V(t)$ satisfies certain technical conditions and $\det[M+N\Phi^V(1,0)] \neq 0$ where $\Phi^V(t,s)$ is the fundamental matrix solution of $\dot{y} = V(t)\, y$.

In the sequel we shall often be given two boundary related matrices M and N and will be required to determine a matrix $V(t)$ so that the set $J = \{V(t),M,N\}$ is boundary compatible. In the next lemma we give necessary and sufficient conditions for the existence of a matrix $V(t)$ which is boundary compatible with the prescribed matrices M and N.

Lemma 3.2.8

Let M and N be $p \times p$ matrices. A necessary and sufficient condition that there be a $V(t)$ with $J = \{V(t),M,N\}$ boundary compatible is that the $p \times 2p$ matrix $[M\ N]$ have full rank p.

Proof: (See [F1].)

Theorem 3.2.3 and Lemma 3.2.8 form the basis for the integral equation representation of nonlinear TPBVP's of the form (3.2.1). We now have the following.

Theorem 3.2.9

Suppose that $F(y,t)$ satisfies certain technical conditions and $J = \{V(t),M,N\}$ is a boundary compatible set of dimension p. Then the boundary value problem

$$\dot{y} = F(y,t), \quad g(y(0)) + h(y(1)) = c \qquad 3.2.10$$

has the equivalent representation

$$y(t) = H^J(t)\{c - g(y(0) - h(y(1)) + My(0) + Ny(1)\} + \int_0^1 G^J(t,s)\{F(y(s),s) - V(s)y(s)\}ds \qquad 3.2.11$$

where the Green's functions $H^J(t)$ and $G^J(t,s)$ are given by

$$H^J(t) = \Phi^V(t,0)[M + N\Phi^V(1,0)]^{-1} \qquad 3.2.12$$

and

$$G^J(t,s) = \begin{cases} \Phi^V(t,s)[M + N\Phi^V(1,0)]^{-1}M\Phi^V(0,s) \ , \ 0 \leq s \leq t \\ -\Phi^V(t,s)[M + N\Phi^V(1,0)]^{-1}N\Phi^V(1,s) \ , \ t < s \leq 1 \end{cases} \qquad 3.2.13$$

where $\Phi^V(t,s)$ is the fundamental matrix of the linear system $\dot{y} = V(t)y$.

Proof: (See [F1] for complete conditions assumed for $F(y,t)$ and a proof of the theorem.)

Theorem 3.2.9 presents an integral equation representation for TPBVP's of the form (3.2.1). It is now a simple matter to demonstrate that solving (3.2.1) is equivalent to solving a certain fixed point problem in an appropriate Banach space. In particular, assuming that the conditions of the previous theorem are satisfied, we can define a mapping T^J of the Banach space $Y = \mathcal{C}([0,1],R^p)$ into itself by setting

$$T^J(y) = H^J(t)\{c-g(y(0)) - h(y(1)) + My(0) + Ny(1)\} + \int_0^1 G^J(t,s)\{F(y(s),s) - V(s)y(s)\}ds. \qquad 3.2.14$$

Then, (3.2.11) is equivalent to the fixed point problem

$$y = T^J(y) \qquad 3.2.15$$

on $\mathscr{C}([0,1],R^p)$. The operator equation (3.2.14) can now be solved by successive approximation iterative techniques as presented by Kantorovich [K4] and particularly Falb and de Jong [F1].

3.3. Frechet Derivatives and Lipschitz Norms

In the discussion of successive approximation iterative techniques, we shall require an expression for the Frechet derivative or Lipschitz norm of the operator T^J. In this section we shall present a brief treatment of these concepts. (Again, many of these basic results are from Falb [F1].) Let us begin with the following definition.

Definition 3.3.1.

Let Y be a Banach space with $\|\cdot\|$ as norm. Let Ω be a closed subset of Y and let T map Y into Y. The Lipschitz norm of T on Ω, in symbols: $\|T\|_\Omega$, is given by

$$\|T\|_\Omega = \sup_{u,v\in\Omega} \{\|T(u) - T(v)\|/\|u-v\|\}. \qquad 3.3.2$$

If T is Frechet differentiable on Ω, then derivative norm of T on Ω, in symbols: $\|T\|'_\Omega$ is given by

$$\|T\|'_\Omega = \sup_{y\in\Omega} \|(T_y)'\| . \qquad 3.3.3$$

We shall now compute expressions for $(T_y{}^J)'$ and $(T_y{}^J)''$. We have

$$(T_y{}^J)'(u) = H^J(t)\{[M-(\partial g/\partial y)(y(0))]u(0) + [N-(\partial h/\partial y)(y(1))]u(1)\} + \int_0^1 G^J(t,s)\{(\partial F/\partial y)(y(s)) - V(s)\}u(s)ds \qquad 3.3.4$$

and

$$(T_y^J)''(u,v) = H^J(t)\left\{\sum_{i=1}^{p}[(\partial/\partial y_i)(-\partial g/\partial y)](y(0))u_i(0)v(0)\right.$$

$$\left. + \sum_{i=1}^{p} [(\partial/\partial y_i)(-\partial h/\partial y)](y(1))u_i(1)v(1)\right\}$$

$$+ \int_0^1 G^J(t,s)\left\{\sum_{i=1}^{p}[(\partial/\partial y_i)(\partial F/\partial y)](y(s))u_i(s)v(s)\right\}ds \qquad 3.3.5$$

provided the indicated partial derivatives exist. When evaluating convergence criteria, we shall require estimates, say for example of the norm of the operator $(T_y{}^J)'$. There are of course several expressions for calculating or estimating $\|(T_y{}^J)'\|$. Since the more accurate expressions are difficult to evaluate in practice, we shall present a coarse estimate that is more amenable to future applications. We recall first of all that if $v(\cdot)\in \mathscr{C}([0,1], R^p)$, then

$$\|v(\cdot)\| = \sup_{i\in P}\ \sup_{t\in[0,1]} |v_i(t)| \qquad 3.3.6$$

is the norm of $v(\cdot)$ where $P = \{1,\ldots,p\}$ and $v_i(\cdot)$ is the ith component of $v(\cdot)$. Noting that $\|(T_y{}^J)'\| = \sup_{\|u\|\le 1}\left\{\|(T_u{}^J)'u\|\right\}$ and letting $H^J(t) = [H^J_{ij}(t)]$, $G^J(t,s) = [G^J_{ij}(t,s)]$, $M = [m_{jk}]$, $N = [n_{jk}]$, $V(s) = [v_{jk}(s)]$, we have as a coarse estimate

$$\|(T_j{}^J)'\| = \sup_{\|u\|\le 1}\left\{\|(T_y{}^J)'u\|\right\}$$

$$\le \sup_{i\in P}\ \sup_{t}\left\{\sum_{j=1}^{P}(|H^J_{ij}(t)|)\cdot\left(\sum_{k=1}^{P}\{|m_{jk} - (\partial g_j/\partial y_k)(y(0))|\right.\right.$$

$$\left.\left. + |n_{jk} - (\partial h_j/\partial y_k)(y(1))|\}\right)\right. \qquad 3.3.7$$

$$+ \sum_{j=1}^{P} \left(\int_0^1 |G_{ij}^J (t,s)| ds \right) \cdot \left(\sup_s \left\{ \sum_{k=1}^{p} |(\partial F_j/\partial y_k)(y(s),s) - v_{jk}(s)| \right\} \right) \Bigg\}$$

Expression (3.3.7) will become quite important in the sequel. One of our primary objectives shall be determining techniques for easily estimating this expression.

In some cases, the smoothness conditions required to obtain Frechet derivatives are too strong. As an example, we have the nonlinearity containing the SAT function in equation (2.2.15). This fact does not imply that successive approximating techniques may not be applied to the iterative solution of the operator equation. It simply means we have lost one method of evaluating convergence criteria. Hence, under somewhat weaker smoothness conditions, we shall compute the Lipschitz norm of the operator $T^J(y)$.

We have the following result from Falb [F1].

Lemma 3.3.8

Let S be a bounded open set in $\mathcal{C}([0,1],R^p)$ and let D be an open set in R^p containing the range of S. Suppose that (i) $K(t,y,s)$ is a map of $[0,1] \times D \times [0,1]$ into D which satisfies certain technical conditions, and (ii) there is an integrable function $m(t,s)$ of s with $\sup_t \int_0^1 m(t,s)ds = \mu < \infty$ such that $\| K(t,y,s) \| \leq m(t,s)$ and $\|K(t,y_1,s) - K(t,y_2,s)\| \leq m(t,s) \|y_1-y_2\|$ on $[0,1] \times D \times [0,1]$. Then the mapping T given by

$$T(u)(t) = \int_0^1 K(t,u(s),s)ds$$

maps $\mathcal{C}([0,1],R^p)$ into $\mathcal{C}([0,1],R^p)$ and the Lipschitz norm, $\| T \|_S$, satisfies

$$\| T \|_S \leq \mu. \qquad 3.3.9$$

Proof: (See [F1] for proof of theorem and specific conditions on K.)

Corrollary 3.3.10

Suppose that the function

$$K(t,y,s) = G^J(t,s)\{F(y,s) - V(s)y\}$$

satisfies the conditions of Lemma 3.3.8 and that

$$\| g(y_1) - g(y_2) \| \leq \mu_1 \| y_1 - y_2 \| \text{ and } \| h(y_1) - h(y_2) \| \leq \mu_2 \| y_1 - y_2 \| \quad 3.3.11$$

Let

$$\alpha = \max \{\mu, \| H^J(\cdot) \| \mu_1, \| H^J(\cdot) \| \mu_2, \| H^J(\cdot)M \|, \| H^J(\cdot)N \| \}. \quad 3.3.12$$

Then

$$\| T^J \|_{\bar{S}} \leq \alpha \quad 3.3.13$$

This result will prove useful in particular when investigating regulators with bounded input controls.

3.4. Contraction Mappings Method

Contraction mappings (or Picard's method, [P2]) is well known in the mathematical literature and has long been a standard approach for proving existence and uniqueness properties for ordinary differential equations. (See for example Coddington and Levinson [C1], specifically Section 1.3 entitled "The Method of Successive Approximations.") To formalize our discussion of this technique, let us begin with the following definition.

Definition 3.4.1

Let Y be a topological space and let T map Y into itself. Let y_0 be an element of Y. The sequence $\{y_n(\cdot)\}$ generated by the algorithm

$$y_{n+1} = T(y_n) \quad n = 0,1,2,\ldots \quad 3.4.2$$

is called a contraction mapping or CM sequence for T based on y_0.

The following theorem is central to our future discussions concerning the contraction mappings method.

Theorem 3.4.3

Let Y be a Banach space and let $\bar{S} = \bar{S}(y_0,r)$ be the closed sphere in Y with center y_0 and radius r. Let T map Y into Y and suppose that (i) T is defined on $\bar{S}(y_0,r)$, and (ii) there are real numbers η and α with $\eta \geq 0$ and $0 \leq \alpha < 1$ such that

$$\| y_1 - y_0 \| \leq \eta \tag{3.4.4}$$

$$\| T \|_{\bar{S}} \leq \alpha < 1 \quad \text{or} \quad \| T \|'_{\bar{S}} \leq \alpha < 1 \tag{3.4.5}$$

$$\frac{1}{1-\alpha} \eta \leq r \tag{3.4.6}$$

where $y_1 = T(y_0)$. Then the CM sequence $\{y_n\}$ for T based on y_0 converges to the unique fixed point y^* of T in $\bar{S}$ and the rate of convergence is given by

$$\| y^* - y_n \| \leq \frac{\alpha}{1-\alpha} \| y_n - y_{n-1} \| \leq \frac{\alpha^n}{1-\alpha} \| y_1 - y_0 \| . \tag{3.4.7}$$

Proof: (See [F]).

Let us now consider the application of this theorem to operator equations of the form

$$\begin{aligned} y(t) = T^J(y)(t) &= H^J(t)\{c-g(y(0))- h(y(1) + My(0) + Ny(1)\} \\ &+ \int_0^1 G^J(t,s)\{F(y(s),s) - V(s)y(s)\}ds \end{aligned} \tag{3.4.8}$$

where $J = \{V(t),M,N\}$ is a boundary compatible set. Following the contraction mapping prescription, we select an initial element $y_0(\cdot)$ in $\mathscr{C}([0,1], R_p)$ and successively generate a CM sequence $\{y_n(\cdot)\}$ for T^J based on $y_0(\cdot)$ by means of the algorithm

$$y_{n+1} = T^J(y_n) \tag{3.4.9}$$

or equivalently, by

$$y_{n+1}(t) = H^J(t)\{c-g(y_n(0)) - h(y_n(0)) + My_n(0) + Ny_n(1)\} + \int G^J(t,s)\{F(y_n(s),s) - V(s)y_n(s)\}ds. \qquad 3.4.10$$

Since we know $y_n(\cdot)$ at each successive step, we can write (3.4.10) in the form

$$y_{n+1}(t) = H^J(t)c_n + \int G^J(t,s)f_n(s)ds \qquad 3.4.11$$

where

$$c_n = c - g(y_n(0)) - h(y_n(1)) + My_n(0) + Ny_n(1) \qquad 3.4.12$$

and

$$f_n(s) = F(y_n(s)) - V(s)\ y_n(s). \qquad 3.4.13$$

Hence, it is seen from (3.4.11) and our results on linear TPBVP (eq. 3.2.4) that the method of contraction mappings when applied to (3.4.8) essentially amounts to the successive solution of the linear TPBVP's (3.4.11).

If the partial derivatives of (3.3.) exist, we then have the following.

Theorem 3.4.14.

Let $y_0(\cdot)$ be an element of $\mathcal{C}([0,1],R^p)$ and let $\bar{S} = \bar{S}(y_0,r)$. Suppose that (i) $J = \{V(t),M,n\}$ is a boundary compatible set for which

$$\dot{y} = F(y(t),t) \quad g(y(0)) + h(y(1)) = c \qquad 3.4.15$$

is differentiable on $\bar{S}$, and (ii) there are real numbers η and α with $\eta \geq 0$ and $0 \leq \alpha < 1$ such that

$$\| T^J(y_0) - y_0 \| = \sup_i \sup_{t \in [0,1]} \{ |T^J(y_0)_i(t) - y_{0,i}(t)| \} \leq \eta \qquad 3.4.16$$

$$\sup_{y \in \bar{S}} \{ \| (T_y^J)' \| \} \leq \alpha \qquad 3.4.17$$

$$\frac{1}{1-\alpha} \eta \leq r \qquad 3.4.18$$

Then the CM sequence $\{y_n(\cdot)\}$ for the TPBVP based on y_0 and J converges uniformly to the unique solution $y^*(\cdot)$ of (3.4.15) in $\bar{S}$ and the rate of convergence is given by

$$\| y^*(\cdot) - y_n(\cdot) \| \leq \frac{\alpha^n}{1-\alpha} \| y_1(\cdot) - y_0(\cdot) \| . \qquad 3.4.19$$

Proof: Simply apply Theorem 3.4.3.

It should be noticed that if the TPBVP of interest is not differentiable, but a Lipschitz norm can be obtained, then (3.4.17) is simply replaced by

$$[\![T^J]\!]_{\bar{S}} \leq \alpha . \qquad 3.4.20$$

We shall use (3.4.20) in the investigation of optimal regulators with bounded control.

At this point we shall make a few general comments concerning our representation of TPBVP's and, in partiuclar, the role of the boundary compatible set $J = \{V,M,N\}$. From Theorem 3.4.14, we see that the convergence rate factor, α, is determined by the Frechet derivative of the operator $T^J(y)$. In particular, from equation 3.3.6 we have an estimate for this norm given as

$$\|(T_y^J)'\| = \sup_{\|u\|\leq 1}\left\{\|(T_y^J)'u\|\right\}$$

$$\leq \sup_{i\in P}\sup_t\left\{\sum_{j=1}^{P}(|H_{ij}^J(t))|)\cdot\left(\sum_{k=1}^{P}\left\{|m_{jk}-\left(\frac{\partial g_j}{\partial y_k}\right)(y(0))|+|n_{jk}-\left(\frac{\partial h_j}{\partial y_k}\right)(y(1))|\right\}\right)\right.$$

$$\left.+\sum_{j=1}^{P}\left(\int_0^1|G_{ij}^J(t,s)|ds\right)\cdot\left(\sup_s\left\{\sum_{k=1}^{P}|\left(\frac{\partial F_j}{\partial y_k}\right)(y(s),s)-v_{jk}(s)|\right\}\right)\right\} \qquad 3.4.21$$

For convergence purposes we wish to make this quantity as small as possible, and in this light, we shall discuss the choice of $J = \{V(t), M,N\}$. All of the TPBVP's obtained in Chapter 2 have linear boundary conditions of the form $Ky(0) + Ly(1) = c$. From this we shall clearly choose M and N to equal the linear boundary conditions of the TPBVP, thus eliminating the first terms in (3.4.21). We then have the simplified expression

$$\|(T_y^J)'\|\leq\sup_{i\in P}\sup_t\left\{\sum_{j=1}^{P}\left(\int_0^1|G_{ij}^J(t,s)|ds\right)\cdot\left(\sup_s\left\{\sum_{k=1}^{P}|\left(\frac{\partial F_j}{\partial y_k}\right)(y(s),s)-v_{jk}(s)|\right\}\right)\right\} \qquad 3.4.22$$

Consideration of this expression allows us to deduce that if y_0 is a good initial estimate of the solution, then it is often effective to choose V(s) close to $(\partial F/\partial y)(y_0(s),s)$. In fact, for $V(s) = (\partial F/\partial y)(y_0(s),s)$, the iterative method is known as the "modified Newton's method." However, a general choice such as this for the V matrix usually precludes any attempt at calculating or estimating the term $\int_0^1|G^J(t,s)|ds$, thus preventing an easy estimation of the convergence criteria. In the next section we shall consider a technique which is often useful for evaluating convergence criteria.

3.5. Modified Contration Mappings

In some situations, the direct application of the contraction mappings method does not lead to a convergent sequence of approximations. However, it is frequently possible to modify T in such a way as to lead to a convergent sequence of approximations. We consider the following.

Lemma 3.5.1.

Let T and U be maps of Y into Y. Suppose that I - U is invertible and let P be the map of Y into itself given by

$$P(y) = [I-U]^{-1}[T(y) - U(y)]. \qquad 3.5.2$$

Then $y^*(\cdot)$ is a fixed point of T if and only if $y^*(\cdot)$ is a fixed point of P.

Proof: (See [F1]).

We shall then consider the selection of an initial approximate solution y_0 and the generation of a sequence $\{y_n\}$ by the algorithm

$$y_{n+1} = P(y_n) = [I-U]^{-1}\ [T(y_n) - U(y_n)]. \qquad 3.5.3$$

We shall call this algorithm the modified contraction mappings method. It should be noted that the modified contraction mapping sequence for T based on y_0 and U coincides with the contraction mapping sequence for P based on y_0. Hence we may translate the results on contraction mappings into theorems for modified contraction mappings. The primary theorem is given as follows.

Theorem 3.5.4.

If U is a linear operator with I-U invertible, if T is differentiable on $\bar{S}$, and if there are real numbers η,α with $\eta \geq 0$ and $0 \leq \alpha < 1$ such that

$$\| y_1 - y_0 \| \leq \eta \qquad 3.5.5$$

$$\sup_{y \in \bar{S}} \{ \| [I-U]^{-1}[T'_{(y)} - U] \| \} \leq \alpha \qquad 3.5.6$$

$$\frac{1}{1-\alpha} \eta \leq r, \qquad 3.5.7$$

then the modified contraction mappings sequence $\{y_n\}$ converges to the unique fixed point y^* of T and $\bar{S}$ and the rate of convergence is given by

$$\| y^* - y_n \| \leq \frac{\alpha}{1-a} \| y_n - y_{n-1} \| \leq \frac{\alpha^n}{1-\alpha} \| y_1 - y_0 \| \ . \qquad 3.5.8$$

Proof: Apply Theorem 3.4.3.

The importance of these results lies in the fact that they extend the range of applicability of the contraction mapping method to fixed point problems for operators T that are not contraction mappings. In other words, the basic contraction mapping criteria

$$\sup_{y \in \bar{S}} \{ \| (T_y^J)' \| \} \leq \alpha < 1 \qquad 3.5.9$$

is replaced by the condition that the Frechet derivative satisfies

$$\sup_{y \in \bar{S}} \{ \| [I-U]^{-1}[T_y^J)'-U] \| \} \leq \alpha < 1. \qquad 3.5.10$$

A second possibility is to replace the single norm in (3.5.10) by a product of two norms so that

$$\sup_{y \in \bar{S}} \{ \| [I-U]^{-1} \| \cdot \| [(T_y^J)'-U] \| \} \leq \alpha < 1. \qquad 3.5.11$$

This formulation offers the possible advantage of easier evaluation, but also results in less sharp convergence conditions.

We shall now specify the form of linear operator U that will be used in the modified contraction mappings algorithm. The following lemma involves

the relation between the operators T^J and $T^{\tilde{J}}$ for different boundary compatible sets $J = \{V(t),M,N\}$ and $\tilde{J} = \{W(t),K,L\}$.

Lemma 3.5.12.

Let $J = \{V(t),M,N\}$ and $\tilde{J} = \{W(t),K,L\}$ be boundary compatible sets. Let $F(y,t)$ be continuous in y for each fixed t and measurable in t for each fixed y with $\|F(y,t)\| \leq m(t)$, $m(t)$ integrable. Let Γ be the linear manifold of absolutely continuous functions in $\mathcal{E}([0,1],R_p)$. Let U_{KL}^J be the operator given by

$$U_{KL}^J(y)(t) = H^J(t)\ \{-Ky(0) - Ly(1) + My(0) + Ny(1)\} + \int_0^1 G^J(t,s)\{W(s)y(s) - V(s)y(s)\}\, ds \qquad 3.5.13$$

for $y(\cdot)$ in $\mathcal{E}([0,1],R^p)$. Then (i) U_{KL}^J maps $\mathcal{E}([0,1],R^p)$ into $\mathcal{E}([0,1],R^p)$ and Γ into Γ ; (ii) the operator $I-U_{KL}^J$ has a bounded linear inverse on Γ with

$$[I-U_{KL}^J]^{-1}\, y = [I-V_{MN}^{\tilde{J}}]y \qquad 3.5.14$$

for y in Γ and

$$V_{MN}^{\tilde{J}}\, y = H^{\tilde{J}}(t)\{-My(0) - Ny(1) + Ky(0) + Ly(1)\} + \int G^{\tilde{J}}(t,s)\{V(s)y(s) - W(s)y(s)\}ds \qquad 3.5.15$$

(iii) if $y(\cdot)$ is in $\mathcal{E}([0,1],R^p)$, then

$$T^{\tilde{J}}(y) = [I-U_{KL}^J]^{-1}\ [T^J(y) - U_{KL}^J(y)] \qquad 3.5.16$$

and (iv) under the differentiability assumptions

$$(T_y^{\tilde{J}})' = [I-U_{KL}^J]^{-1}\ [(T_y^{\ J})' - U_{KL}^J]. \qquad 3.5.17$$

Proof: (See [F1]).

We shall limit our future discussions to operators $U = U^J_{KL}$ of the form given by (3.5.13). It then follows from Lemma 3.5.12 that the modified contraction mapping method when applied to the equation $y = T^J(y)$ with modifying operator $U = U^J_{KL}$, is equivalent to the contraction mapping method applied to the equation $y = T^{\tilde{J}}(y)$.

The importantce of this point will become clearer as we develop techniques for estimating $(T_y{}^J)'$. We shall now indicate the approach that will be considered. Suppose that J is a boundary compatible set for which the corresponding Green's matrices are easy to evaluate and estimate. Then if $\|U^J_{KL}\| \leq q < 1$ so that $\|[U^J_{KL}]^{-1}\| \leq \frac{1}{1-q}$, we can obtain an estimate of (3.5.17) which involves only the Green's matrices corresponding to J. This advantage may well offset the loss of accuracy resulting from using (3.5.11). We now have the following.

Theorem 3.5.18

Let $y_0(\cdot)$ be an element of $\mathscr{C}([0,1],R^p)$ and let $\bar{S} = \bar{S}(y_0,r)$. Suppose that (i) $J = \{V(t),M,N\}$ is a boundary compatible set for which

$$\dot{y} = F(y,t) \quad g(y(0)) + h(y(1)) = c \tag{3.5.19}$$

is differentiable on $\bar{S}$; (ii) $\tilde{J} = \{U(t),K,L\}$ is a boundary compatible set; and (iii) there are real numbers η,q,β, and α with $\eta \geq 0$, $0 \leq q < 1$, $\beta \geq 0$, and $\alpha = \beta/(1-q) < 1$ such that

$$\|T^{\tilde{J}}(y_0) - y_0\| = \sup_i \sup_t \{|T^{\tilde{J}}(y_0)_i(t) - y_{0,i}(t)|\} \leq \eta \tag{3.5.20}$$

$$\|U^J_{KL}\| \leq q \tag{3.5.21}$$

$$\sup_{y \in \bar{S}} \{\|(T_y{}^J)' - U^J_{KL}\|\} \leq \beta \tag{3.5.22}$$

$$\frac{1}{1-\alpha}\ \eta \leq r. \tag{3.5.23}$$

Then the MCM sequence $\{y_n(\cdot)\}$ for T^J based on $y_0(\cdot)$ and U^J_{KL} converges uniformly to the unique solution $y^*(\cdot)$ of (3.5.19) in $\bar{S}$ and the rate of convergence is given by

$$\| y^*(\cdot) - y_n(\cdot) \| \leq \frac{\alpha^n}{1-\alpha} \| y_1(\cdot) - y_0(\cdot) \| . \qquad 3.5.24$$

Proof: Apply Theorem 3.5.4.

In order to illuminate the preceding discussion, let us consider an example utilizing the previous concepts.

Example 3.5.25.

Let us consider the iterative solution of the differentiable TPBVP given as

$$\dot{y}(t) = F(y,t) \quad Ky(0) + Ly(1) = c. \qquad 3.5.26$$

We shall discuss the choice of the boundary compatible set $\tilde{J} = \{W(t),M,N\}$ to be used in the integral representation of the TPBVP. Since the boundary conditions of (3.5.26) are linear, we shall choose $M = K$ and $N = L$. Let us suppose that $y_0(t)$ is a good initial estimate for the solution of (3.5.26). Then as indicated, let us choose $W(t)$ as

$$W(t) = (\partial F/\partial y)(y_0(t)), \qquad 3.5.27$$

assuming this choice of $\tilde{J} = \{W(t),K,L\}$ is boundary compatible. However, this general time varying choice for $W(t)$ makes it extremely difficult, if not impossible, to analytically calculate the fundamental matrix $\Phi^W(t,s)$ and the Green's functions.

Let us now decompose the $W(t)$ matrix as

$$W(t) = V + \delta V(t) \qquad 3.5.28$$

where V is a constant matrix of simple structure, e.g., diagonal, which is boundary compatible with K and L. Then for the boundary compatible set $J = \{V,K,L\}$ containing the simple V matrix, it is often possible to analytically calculate the Green's matrices. We now have

$$\tilde{T}^J(y) = [I-U^J_{KL}]^{-1}[T^J(y)-U^J_{KL}\, y] \qquad 3.5.29$$

where

$$T^J(y) = H^J(t)C + \int_0^1 G^J(t,s)\{F(y(s),s) - Vy(s)\}ds \qquad 3.5.30$$

$$U^J_{KL}\, y = \int_0^1 G^J(t,s)\{W(s)\, y(s) - Vy(s)\}ds \qquad 3.5.31$$

or

$$U^J_{KL}\, y = \int_0^1 G^J(t,s)\delta V(s)ds, \qquad 3.5.32$$

and finally we note

$$T^J(y) - U^J_{KL}\, y = H^J(t)c + \int_0^1 G^J(t,s)\{F(y(s),s) - W(s)y(s)\}ds \qquad 3.5.33$$

so that

$$[(T_y{}^J)' - U^J_{KL}]\, u = \int_0^1 G^J(t,s)\{(\partial F/\partial y)(y(s),s) - W(s)\}u(s)ds. \qquad 3.5.34$$

Hence we obtain the convergence benefits of choosing a general matrix W(t) while being able to calculate the Green's matrices using the V matrix of simple structure.

3.6. Applications of Contraction Mappings

In this section we shall investigate the application of the contraction mappings method to the iterative solution of the TPBVP's arising from the regulation of nonlinear systems. In particular, using Theorem 3.4.14 we shall present the general form of the translated convergence theorems for the iterative solution of these TPBVP's.

Let us first consider the application to the system presented in Example 2.3.1. Recall that this nonlinear system contained a nonlinear form containing only the state variable. We have for this case the following translated convergence theorem.

Theorem 3.6.1.

Let $y_0(\cdot)$ be an element of $\mathscr{E}([0,1],R^p)$ and let $\bar{S} = \bar{S}(y_0,r)$. Suppose that (i) $J = \{V(t),M,N\}$ is a boundary compatible set, and (iii) there are real numbers η and α with $\eta \geq 0$ and $0 \leq \alpha < 1$ such that

$$1)\quad \|T^J(y_0) - y\| = \left\| H^J(t)\begin{bmatrix} x_0 \\ 0 \end{bmatrix} + \int_0^1 G^J(t,s)\left\{\begin{bmatrix} A(s) & -B(s)R^{-1}(s)B'(s) \\ -Q(s) & -A'(s) \end{bmatrix}\begin{bmatrix} x_0(s) \\ p_0(s) \end{bmatrix}\right.\right. \tag{3.6.2}$$

$$\left.\left. - V(s)\begin{bmatrix} x_0(s) \\ p_0(s) \end{bmatrix} + \begin{bmatrix} \psi(x_0(s)) \\ -\frac{\partial\psi}{\partial x}'(x_0(s))p_0(s) \end{bmatrix}\right\} ds - \begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix}\right\| \leq \eta ,$$

$$2)\quad \sup_{y\in\bar{S}} \left\{\|(T_y^J)'\|\right\} = \sup_{y\in\bar{S}} \sup_{\|u\|\leq 1} \left\| \int_0^1 G^J(t,s)\left\{\begin{bmatrix} A(s) & -B(s)R^{-1}(s)B(s) \\ -Q(s) & -A'(s) \end{bmatrix} - V(s)\right.\right. \tag{3.6.3}$$

$$\left.\left. + \begin{bmatrix} \left(\frac{\partial\psi}{\partial x}\right)(x(s)) & 0 \\ D(x(s),p(s)) & -\left(\frac{\partial\psi}{\partial x}\right)'(x(s)) \end{bmatrix}\right\} u(s)ds \right\| \leq \alpha$$

where $D(s(s),p(s)) = [D_{ik}(x(s),p(s)] = \sum_{j=1}^{n} \left(\frac{\partial^2 \psi_j}{\partial x_k \partial x_i}\right)(x(s))\, p_j(s)$,

$$3)\quad \frac{1}{1-\alpha}\eta \leq r \tag{3.6.4}$$

Then the CM sequence $\{y_n(\cdot)\}$ for the TPBVP based on y_0 and J converges uniformly to the unique solution y^* in $\bar{S}$ and the rate of convergence is given by

$$\|y^*(\cdot) - y_n(\cdot)\| \leq \frac{\alpha^n}{1-\alpha} \|y_1(\cdot) - y_0(\cdot)\| \qquad 3.6.5$$

Proof: Apply Theorem 3.4.14 to the TPBVP of Example 2.3.1.

From this general theorem statement, the performance of the numerical algorithm is difficult to predict. However, in the sequel, we shall develop coarse estimates for the convergence criteria contained in Theorem 3.6.1.

We shall now apply the contraction mappings convergence theorem to the operator equation corresponding to the regulation of a system containing a general formulation for the nonlinearity, i.e., the TPBVP presented in Example 2.3.45. The nonlinearity contained in that TPBVP is given as

$$f(y(t)) = \begin{bmatrix} \phi(x(t),p(t)) + B(t)\ \xi(x(t),p(t)) \\ -Z'(x(t),p(t))p(t) \end{bmatrix} \qquad 3.6.6$$

where we defined

$$u = \xi(x(t),p(t))$$

$$\phi(x(t),p(t)) = \psi(x(t),\xi(x(t),p(t))$$

and

$$Z(x(t),p(t)) = (\partial\psi/\partial x)(x(t),\xi(x(t),p(t)).$$

Before applying the CM theorem, we shall first calculate an expression for $(\partial f/\partial y)(y(t))$ where y is the composite 2n vector [x,p]. We have

$$(\partial f/\partial y)(y) = \begin{bmatrix} (\partial\phi/\partial x)(x,p) + B(t)(\partial\xi/\partial x)(x,p) & (\partial\phi/\partial p)(x,p) + B(t)(\partial\xi/\partial p)(x,p) \\ - \partial/\partial x\ [Z'(x,p)p] & - \partial/\partial p\ [Z'(x,p)p] \end{bmatrix} \qquad 3.6.7$$

Now defining the matrix functions $D(x,p)$ and $W(x,p)$ to be composed of the elements

$$D_{ij}(x,p) = \sum_{k=1}^{n} (\partial z_{ki}/\partial x_j)(x,p)p_k \qquad 3.6.8$$

and

$$W_{ij}(x,p) = \sum_{k=1}^{n} (\partial z_{ki}/\partial p_j)(x,p)p_k \quad , \qquad 3.6.9$$

we have $(\partial f/\partial y)(x,p)$ given as

$$(\partial f/\partial y)(x,p) = \begin{bmatrix} Z(x,p) + B(t)(\partial\xi/\partial x)(x,p) & B(t)(\partial\xi/\partial p)(x,p) + (\partial\phi/\partial p(x,p) \\ -D(x,p) & -W(x,p) - Z'(x,p) \end{bmatrix} \qquad 3.6.10$$

Using equation (3.6.10) we have the following theorem.

Theorem 3.6.11.

Let y_0 be an element of $\mathscr{C}([0,1],R^p)$ and let $\bar{S} = \bar{S}(y_0,r)$. Suppose that (i) $J = \{V(t),M,N\}$ is a boundary compatible set for which (2.3.46) is differentiable, and (ii) there are real numbers η and α with $\eta \geq 0$ and $0 \leq \alpha < 1$ such that

$$1)\ \left\|T^J(y_0)-y_0\right\| = \left\| H^J(t)\begin{bmatrix}x_0\\0\end{bmatrix} + \int_0^1 G^J(t,s)\left\{\begin{bmatrix}A(s) & 0\\ -Q(s) & -A'(s)\end{bmatrix}\begin{bmatrix}x_0(s)\\p_0(s)\end{bmatrix}\right.\right.$$

$$\left. -V(s)\begin{bmatrix}x_0(s)\\p_0(s)\end{bmatrix} + \begin{bmatrix}\phi(x_0(s),p_0(s)) + B(t)\ \xi(x_0(s),p_0(s))\\ -Z'(x_0(s),p_0(s))\end{bmatrix}\right\} ds \qquad 3.6.12$$

$$\left. - \begin{bmatrix}x_0(t)\\p_0(t)\end{bmatrix}\right\| \leq \eta\ ,$$

$$2)\ \sup_{y\in\bar{S}}\left\{\left\|(T_y^J)'\right\|\right\} = \sup_{y\in\bar{S}}\ \sup_{\|u\|\leq 1}\left\| \int_0^1 G^J(t,s)\left\{\begin{bmatrix}A(s) & 0\\ -Q(s) & -A'(s)\end{bmatrix} -V(s)\right.\right. \qquad 3.6.13$$

$$+ \begin{bmatrix}Z(x(s),p(s))+B(s)(\partial\xi/\partial x)(x(s),p(s)) & B(s)(\partial\xi/\partial p)(x(s),p(s))+(\partial\phi/\partial p)(x(s),p(s))\\ -D(x(s),p(s)) & -W(x(s),p(s)) - Z'(x(s),p(s))\end{bmatrix}\Bigg\} u(s)ds\Bigg\| \leq \alpha$$

$$3)\ \frac{1}{1-\alpha}\ \eta \leq r$$

Then the CM sequence $\{y_n(\cdot)\}$ for the TPBVP based on y_0 and J converges uniformly to the unique solution y^* in $\bar{S}$ and the rate of convergence is given by

$$y^*(\cdot) - y_n(\cdot) \leq \frac{\alpha^n}{1-\alpha}\ y_1(\cdot) - y_0(\cdot)\ . \qquad 3.6.14$$

As we have indicated, cursory examination of Theorems 3.6.1, 3.6.10, and 3.6.21 yields limited information converning the convergence of the CM sequence.

The difficulty to a great extent lies in the intricacy of evaluating the integral containing the Green's function, $G^J(t,s)$, and the derivative term of the form $(\partial F/\partial y)(y(s)) - V(s)$. In the next chapter, we shall consider techniques for alleviating these difficulties so that meaningful convergence analysis can be made without extensive computation.

Page intentionally left blank

CHAPTER 4

CALCULATION OF CONVERGENCE CRITERIA

4.1. Introduction

For the boundary compatible set $J = \{V(t),M,N\}$, we consider the iterative solution of the operator equation

$$y = T^J(y) \tag{4.1.1}$$

where $T^J(y)$ is given by

$$\begin{aligned} y(t) = T^J(y)(t) &= H^J(t)\{c-g(y(0) - h(y(1)) + My(0) + Ny(1)\} \\ &+ \int_0^1 G^J(t,s)\{F(y(s),s) - V(s)y(s)\}ds \end{aligned} \tag{4.1.2}$$

and the Green's functions H^J, G^J are given by

$$H^J(t) = \Phi^V(t,0)[M + N\Phi^V(1,0)]^{-1}$$

and

$$G^J(t,s) = \begin{cases} G_I^J(t,s) = \Phi^V(t,0)[M + N\Phi^V(1,0)]^{-1}M\Phi^V(0,s) \ , \ 0 \le s \le t \\ G_{II}^J(t,s) = -\Phi^V(t,0)[M + N\Phi^V(1,0)]^{-1}N\Phi^V(1,s) \quad t < s \le 1. \end{cases} \tag{4.1.3}$$

Theorem 3.4.14 specified conditions necessary for convergence of the CM sequence $y_{n+1} = T^J(y_n)$. In this chapter, we discuss in detail the evaluation of the convergence criteria. In particular, we discuss two general schemes that may be

used to lessen the analytical difficulties involved in calculating the convergence parameters η and α.

The first scheme is simply that of selecting very simple V matrices for use in the representation. For example, one might select V as the zero matrix or a constant diagonal matrix. For these matrices the fundamental matrix is readily obtained and the Green's function matrices are often easily calculated.

The second scheme involves the use of a similarity transformation. In this approach, a more general constant V matrix is selected and transformed into a canonical form. Then using the canonical form, the fundamental matrix is obtained. However, for this approach, the calculation of the Green's function matrices is somewhat complicated by the transformation matrices. In conclusion, an approximate technique is developed which often yields accurate estimates.

4.2. Estimates of Convergence Criteria

Before considering specific boundary compatible sets, we first specify those estimates of the convergence parameters which are desired. As indicated in Theorem 3.4.14, the numbers to be calculated are estimates for $\|T^J(y_0)-y_0\|$ and $\|(T_y{}^J)'\|$.

First consider the estimation of $\|T^J(y_0)-Y_0\|$. At this point, it will be useful to discuss an effective technique for obtaining the initial estimate of the solution. Consider the iterative solution of the nonlinear TPBVP

$$\dot{y} = F(y,t) \qquad 4.2.1$$
$$Ky(0) + Ly(1) = c,$$

and the choice of the boundary compatible set $J = \{W(t),M,N\}$ to be used in the integral representation. Since the boundary conditions of (4.2.1) are linear,

we choose M=K, N=L in the representation. If we now choose W(t) based upon $(\partial F/\partial y)(y,t)$, i.e., a linearization of the system, then the solution to the linear TPBVP

$$\dot{y} = W(t)y \qquad 4.2.2$$
$$Ky(0) + Ly(1) = c$$

is often a good initial estimate for the solution of (4.2.1). Moreover, this choice considerably simplifies the calculation of $T^J(y_0)-y_0$ since $y_0(t)=H^J(t)c$ and

$$T^J(y_0)-y_0 = \int_0^1 G^J(t,s)\{F(y_0(s),s) - W(s)y_0(s)\}ds \qquad 4.2.3$$

for the boundary compatible set $J = \{W(t),M,N\}$.

The other norm which must be calculated is the derivative norm $\|(T_y{}^J)'\|$. As presented previously in (3.3.7), a coarse estimate for $\|(T_y{}^J)'\|$ is given as

$$\begin{aligned}\|(T_y{}^J)'\| &\le \sup_{\|u\| \le 1} \|\{ (T_y{}^J)'u \}\| \\ &\le \sup_{i\in P} \sup_t \left\{ \sum_{j=1}^p \left(\int_0^1 |G_{ij}^J(t,s)|ds \right) \left(\sup_s \left\{ \sum_{k=1}^p |(\partial F_j/\partial y_k)(y(s),s) \right.\right.\right. \\ &\left.\left.\left. - v_{jk}(s)| \right\}\right)\right\} \qquad 4.2.4\end{aligned}$$

Let us make the following definitions.

Definition 4.2.5.

Let $P(t) = [p_{ij}(t)]$ be a matrix with entries

$$p_{ij}(t) = \int_0^1 |g_{ij}^J(t,s)|ds \qquad 4.2.6$$

or

$$p_{ij}(t) = \int_0^t |g_{I_{ij}}^J(t,s)|ds + \int_t^1 |g_{II_{ij}}^J(t,s)|ds \qquad 4.2.7$$

where $g^J_{I_{ij}}$ and $g^J_{II_{ij}}$ are elements of $G_I{}^J(t,s)$ and $G_{II}{}^J(t,s)$ as given in (4.1.3).

Definition 4.2.8

Let $z_0 = [z_{0_i}]$ be a vector with elements

$$z_{0_i} = \sup_{t \in [0,1]} \left\{ \left| F_i(y_0(t),t) - \sum_{j=1}^{P} v_{ij}(t)\, y_{0_j}(t) \right| \right\} \qquad 4.2.9$$

Definition 4.2.10

Let $z = [z_i]$ be a vector with elements

$$z_i = \sup_{t \in [0,1]} \sup_{y \in S} \left\{ \sum_{j=1}^{P} \left| (\partial F_i/\partial y_j)(y(t),t) - v_{ij}(t) \right| \right\} . \qquad 4.2.11$$

From (4.2.3) and (4.2.4) it follows that conservative values for the convergence parameters η and α are given by

$$\| P(\cdot) z_0 \| = \sup_i \sup_t \left\{ \sum_{j=1}^{P} p_{ij}(t) z_{0_j} \right\} \leq \eta \qquad 4.2.12$$

and

$$\| P(\cdot) z \| = \sup_i \sup_t \left\{ \sum_{j=1} p_{ij}(t) z_j \right\} \leq \alpha \qquad 4.2.13$$

In the remainder of this chapter we shall be primarily concerned with techniques for determining the matrix $P(t)$ for boundary compatible sets containing simple V matrices.

4.3. Boundary Value Sets of Interest

In this section we shall briefly specify the form of those pairs of boundary condition matrices which are of interest. The necessary conditions for regulation of nonlinear systems reduced to TPBVP's of the form

$$\dot{y} = Sy + f(y) \qquad 4.3.1$$

$$My(0) + Ny(1) = c \qquad 4.3.2$$

where the matrices M and N depended on the quadratic cost functional being used. Specifically we had the following cases.

Definition 4.3.3.

For quadratic cost functionals including a terminal state penalty of the form $\langle x(T), Kx(T)\rangle$, the boundary condition matrices were

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \tag{4.3.4}$$

Since we have rank [M N] = 2n, Lemma 3.2.8 assures a matrix V exists so the set J = {V,M,N} is boundary compatible. We shall henceforth refer to set (4.3.4) as boundary value set {1}.

Example 4.3.5.

For quadratic cost functionals which do not include a terminal penalty, the boundary condition matrices are given as

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \tag{4.3.6}$$

Again rank [M N] = 2n, so a V matrix exists such that J = {V,M,N} is boundary compatible. The set (4.3.6) shall be referred to as boundary value set {2}.

4.4. Boundary Set for Regulation with Terminal Cost

In this section the use of simple V matrices with boundary value set {1} will be considered. The requirements for boundary compatibility of the various sets J = {V,M,N} will be noted in particular.

Boundary value set {1} is given specifically as

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \tag{4.4.1}$$

and a general 2n ×2n V matrix is represented as

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \tag{4.4.2}$$

The fundamental matrix for V is represented as

$$\Phi^V(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & \Omega_{12}(t,s) \\ \Omega_{21}(t,s) & \Omega_{22}(t,s) \end{bmatrix} \tag{4.4.3}$$

The matrix $[M + N\Phi^V(1,0)]$ is now formed explicitly as

$$M+N\Phi^V(1,0) = \begin{bmatrix} I & 0 \\ -K\Omega_{11}(1,0)+\Omega_{21}(1,0) & -K\Omega_{12}(1,0)+\Omega_{22}(1,0) \end{bmatrix} \tag{4.4.4}$$

and the inverse, if it exists, may be written as

$$[M+N\Phi^V(1,0)]^{-1} = \begin{bmatrix} I & 0 \\ -[-K\Omega_{12}(1,0)+\Omega_{22}(1,0)]^{-1}[-K\Omega_{11}(1,0)+\Omega_{21}(1)] & [-K\Omega_{12}(1,0)+\Omega_{22}(1,0)]^{-1} \end{bmatrix} \tag{4.4.5}$$

For this inverse to exist, the matrix $[-K\Omega_{12}(1,0)+\Omega_{22}(1,0)]$ must be nonsingular. It is noted that for V equal to the zero matrix or a diagonal matrix, the set $J = \{V,M,N\}$ is boundary compatible. The core of the Green's function is given by the matrices $[M+N\Phi^V(1,0)]^{-1}M$ and $[M+N\Phi^V(1,0)]^{-1}N$ which are explicitly given as

$$[M+N\Phi^V(1,0)]^{-1}M = \begin{bmatrix} I & 0 \\ -[-K\Omega_{12}(1,0)+\Omega_{22}(1,0)]^{-1}\ [-K\Omega_{11}(1,0)+\Omega_{21})1,0)] & 0 \end{bmatrix} \qquad 4.4.6$$

and

$$[M+N\Phi^V(1,0)]^{-1}N = \begin{bmatrix} 0 & 0 \\ -[-K\Omega_{12}(1,0)+\Omega_{22}(1,0)]^{-1}K & [-K\Omega_{12})1,0)+\Omega_{22}(1,0)]^{-1} \end{bmatrix} \qquad 4.4.7$$

We shall now consider specific choices for the V matrix.

Example 4.4.8.

Consider the choice of the simplest V matrix, i.e., assume V = 0. The fundamental matrix is then given as

$$\Phi^V(t,s) = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \qquad 4.4.9$$

Now using (4.4.6) and (4.4.7),

$$[M+N\Phi^V(1,0)]^{-1}M = \begin{bmatrix} I & 0 \\ K & 0 \end{bmatrix} \qquad 4.4.10$$

and

$$[M+N\Phi^V(1,0]^{-1}N = \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix}. \qquad 4.4.11$$

The Green's function matrices are calculated as

$$G_I^J(t,s) = \Phi^V(t,0)[M+N\Phi^V(1,0)]^{-1}\ M\Phi^V(0,s) = \begin{bmatrix} I & 0 \\ K & 0 \end{bmatrix} \qquad 4.4.12$$

and

$$G_{II}^J(t,s) = -\Phi^V(t,0)[M+N\Phi^V(1,0)]^{-1}\ N\Phi^V(1,s) = \begin{bmatrix} 0 & 0 \\ K & -I \end{bmatrix}, \qquad 4.4.13$$

and the $2n \times 2n$ $P(t)$ matrix defined as

$$P(t) = \int_0^t |G_I^J(t,s)|ds + \int_t^1 |G_{II}^J(t,s)|ds \qquad 4.4.14$$

is calculated to be

$$P(t) = \begin{bmatrix} tI & 0 \\ |K| & (1-t)I \end{bmatrix} \qquad 4.4.15$$

where elements of $|K|$ are given as $|k_{ij}|$.

Example 4.4.16

The use of a $p \times p$ $(2n \times 2n)$ diagonal V matrix is now considered. Let V be represented as

$$V = \left[\begin{array}{ccc|ccc} \lambda_1 & & & & & \\ & \ddots & & & 0 & \\ & & \lambda_n & & & \\ \hline & & & \mu_1 & & \\ & 0 & & & \ddots & \\ & & & & & \mu_n \end{array}\right] \qquad 4.4.16$$

so the fundamental matrix is then simply

$$\Phi^V(t,s) = \left[\begin{array}{ccc|ccc} e^{\lambda_1(t-s)} & & & & & \\ & \ddots & & & 0 & \\ & & e^{\lambda_n(t-s)} & & & \\ \hline & & & e^{\mu_1(t-s)} & & \\ & 0 & & & \ddots & \\ & & & & & e^{\mu_n(t-s)} \end{array}\right] \qquad 4.4.17$$

which shall be denoted as

$$\Phi^V(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & 0 \\ 0 & \Omega_{22}(t,s) \end{bmatrix} . \qquad 4.4.18$$

This yields using (4.4.6) and (4.4.7),

$$[M+N\Phi^V(1,0)]^{-1} M = \begin{bmatrix} I & 0 \\ \Omega_{22}^{-1}(1,0)K\,\Omega_{11}(1,0) & 0 \end{bmatrix} \qquad 4.4.19$$

and

$$[M+N\Phi^V(1,0)]^{-1} N = \begin{bmatrix} 0 & 0 \\ -\Omega_{22}^{-1}(1,0)K & \Omega_{22}^{-1}(1,0) \end{bmatrix} . \qquad 4.4.20$$

The Green's function matrices are determined to be

$$G_I^J(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & 0 \\ \Omega_{22}(t,0)\Omega_{22}^{-1}(1,0)K\Omega_{11}(1,s) & 0 \end{bmatrix} \qquad 4.4.21$$

and

$$G_{II}^J(t,s) = \begin{bmatrix} 0 & 0 \\ \Omega_{22}(t,0)\Omega_{22}^{-1}(1,0)K\Omega_{11}(1,s) & -\Omega_{22}(t,s) \end{bmatrix} . \qquad 4.4.22$$

In many instances the K matrix associated with the terminal cost is a diagonal matrix. Let us now assume K diagonal with elements k_i. Then using (4.4.21) and (4.4.22), the P(t) matrix is found to be

$$P(t) = \left[\begin{array}{ccc|ccc} \frac{1}{\lambda_1}(e^{\lambda_1 t}-1) & & & & & \\ & \ddots & & & 0 & \\ & & \frac{1}{\lambda_n}(e^{\lambda_n t}-1) & & & \\ \hline \frac{|K_1|}{\lambda_1} e^{-\mu_1(1-t)}(1-e^{\lambda_1}) & & & \frac{1}{\mu_1}(1-e^{-\mu_1(1-t)}) & & \\ & \ddots & & & \ddots & \\ & & \frac{|k_n|}{\lambda_n} e^{-\mu_n(1-t)}(1-e^{\lambda_n}) & & & \frac{1}{\mu_n}(1-e^{-\mu_n(1-t)}) \end{array}\right] \qquad 4.4.23$$

Example 4.4.24.

Many nonlinear systems of interest have an underlying oscillator structure. For this reason we shall consider a choice of V matrix containing linear oscillator elements. This choice is represented as

$$V = \left[\begin{array}{ccc|ccc} S_1 & & & & & \\ & \ddots & & & 0 & \\ & & S_j & & & \\ \hline & & & S_k & & \\ & 0 & & & \ddots & \\ & & & & & S_n \end{array}\right] \qquad 4.4.24$$

where the S_i are 2×2 matrices of the form

$$S_i = \begin{bmatrix} \sigma_i & \omega_i \\ -\omega_i & \sigma_i \end{bmatrix} . \qquad 4.4.25$$

The fundamental matrix for this choice of V is given as

$$\Phi^V(t,s) = \left[\begin{array}{ccc|ccc} \phi_1(t,s) & & & & & \\ & \ddots & & & 0 & \\ & & \phi_j(t,s) & & & \\ \hline & & & \phi_k(t,s) & & \\ & 0 & & & \ddots & \\ & & & & & \phi_n(t,s) \end{array}\right] \qquad 4.4.26$$

or

$$\Phi^V(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & 0 \\ 0 & \Omega_{22}(t,s) \end{bmatrix} , \qquad 4.4.27$$

where the $\phi_i(t,s)$ are 2×2 matrices of the form

$$\phi_i(t,s) = \begin{bmatrix} e^{\sigma_i(t-s)} \cos \omega_i(t-s) & e^{\sigma_i(t-s)} \sin \omega_i(t-s) \\ -e^{\sigma_i(t-s)} \sin \omega_i(t-s) & e^{\sigma_i(t-s)} \cos \omega_i(t-s) \end{bmatrix} . \qquad 4.4.28$$

From (4.4.5), the matrix $[M+N\Phi^V(1,0)]$ is nonsingular if $\Omega_{22}(1,0)$ is nonsingular. We now have

$$\Omega_{22}^{-1}(1,0) = \begin{bmatrix} \phi_k^{-1}(1,0) & & 0 \\ & \ddots & \\ 0 & & \phi_n^{-1}(1,0) \end{bmatrix} \qquad 4.4.29$$

where

$$\phi_i^{-1}(1,0) = \begin{bmatrix} e^{-\sigma_i} \cos \omega_i & -e^{-\sigma_i} \sin \omega_i \\ e^{-\sigma_i} \sin \omega_i & e^{-\sigma_i} \cos \omega_i \end{bmatrix} , \qquad 4.4.30$$

so this choice leads to a boundary compatible set.

The Green's functions are found to be given as

$$G_I^J(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & 0 \\ \Omega_{22}(t,0)\ \Delta\ \Omega_{11}(o,s) & 0 \end{bmatrix} \qquad 4.4.31$$

and

$$G_{II}^J(t,s) = \begin{bmatrix} 0 & 0 \\ -\Omega_{22}(t,0)\ \Delta\ \Omega_{11}(0,s) & -\Omega_{22}(t,s) \end{bmatrix}, \qquad 4.4.32$$

where the matrix Δ is given as

$$\Delta = -\Omega_{22}^{-1}(1,0)\ K\ \Omega_{11}(1,0). \qquad 4.4.33$$

The matrix P(t) is then given as

$$P(t) = \begin{bmatrix} \int_0^t |\Omega_{11}(t,s)|ds & 0 \\ \int_0^1 |\Omega_{22}(t,0)\ \Delta\ \Omega_{11}(0,s)|ds & \int_0^1 |\Omega_{22}(t,s)|ds \end{bmatrix} \qquad 4.4.34$$

Due to the oscillatory nature of the elements of $G^J(t,s)$, the integration of the absolute values somewhat complicates an analytic solution for P(t). However, in a future section we shall consider approximate techniques for obtaining this P(t) matrix.

4.5. Boundary Set for Regulation with No Terminal Cost

In this section the use of simple V matrices with boundary value set {2} is considered. The requirements for boundary compatibility of the various sets J = {V,M,N,} shall be noted in particular. Boundary value set {2} is given specifically as

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \qquad 4.5.1$$

A general $2n \times 2n$ matrix is represented as

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \qquad 4.5.2$$

and the corresponding fundamental matrix is given as

$$\Phi^V(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & \Omega_{12}(t,s) \\ \Omega_{21}(t,s) & \Omega_{22}(t,s) \end{bmatrix} . \qquad 4.5.3$$

The matrix $[M+N\Phi^V(1,0)]$ is formed as

$$[M+N\Phi^V(1,0)] = \begin{bmatrix} I & 0 \\ \Omega_{21}(1,0) & \Omega_{22}(1,0) \end{bmatrix} , \qquad 4.5.4$$

and the inverse, if it exists, is given by

$$[M+N\Phi^V(1,0)]^{-1} = \begin{bmatrix} I & 0 \\ -\Omega_{22}^{-1}(1,0)\Omega_{21}(1,0) & \Omega_{22}^{-1}(1) \end{bmatrix} . \qquad 4.5.5$$

For this inverse to exist, $\Omega_{22}(1,0)$ must be nonsingular. It is noted that for V equal to the zero or diagonal matrix, the set $J = \{V,M,N\}$ is boundary compatible. At this point we shall begin to take advantage of the fact that the remaining results desired in this section may be obtained from the results of the previous section with K equal to zero. These results are now presented for V matrices of simple structure.

Example 4.5.6.

The first selection for the V matrix is the zero matrix, i.e., $V = 0$. Using the results of Example 4.4.8 with $K = 0$, the $P(t)$ matrix is given as

$$P(t) = \begin{bmatrix} tI & 0 \\ 0 & (1-t)I \end{bmatrix} . \qquad 4.5.6$$

Example 4.5.7

For the second choice of the V matrix, a p × p (2n ×2n) diagonal matrix is selected, i.e.,

$$V = \left[\begin{array}{ccc|ccc} \lambda_1 & & & & & \\ & \ddots & & & 0 & \\ & & \lambda_n & & & \\ \hline & & & \mu_1 & & \\ & 0 & & & \ddots & \\ & & & & & \mu_n \end{array}\right] \qquad 4.5.8$$

Now specializing the results of Example 4.4.16 with K = 0, the P(t) matrix is obtained as

$$P(t) = \left[\begin{array}{ccc|ccc} \frac{1}{\lambda_1}(e^{\lambda_1 t}-1) & & & & & \\ & \ddots & & & 0 & \\ & & \frac{1}{\lambda_n}(e^{\lambda_n t}-1) & & & \\ \hline & & & \frac{1}{\mu_1}(1-e^{-\mu_1(1-t)}) & & \\ & 0 & & & \ddots & \\ & & & & & \frac{1}{\mu_n}(1-e^{-\mu_n(1-t)}) \end{array}\right] \qquad 4.5.9$$

4.6. Application of Similarity Transformations

As an introduction to the use of similarity transformations, consider the linear TPBVP

$$\dot{y} = Vy \quad My(0) + Ny(1) = c . \qquad 4.6.1$$

If the set $J = \{V,M,N\}$ is boundary compatible, the solution to (4.6.1) is given by Theorem 3.23 as

$$y(t) = \Phi^V(t,0)[M+N\Phi^V(1,0)]^{-1}\, c. \qquad 4.6.2$$

In an attempt to ease the calculation of the fundamental matrix $\Phi^V(t,0)$, consider the use of the nonsingular linear transformation

$$\Lambda z = y\ . \qquad 4.6.3$$

From (4.6.1) , the transformed TPBVP is given as

$$\dot{z} = \Lambda^{-1}V\Lambda\, z,\quad M\Lambda z(0) + N\Lambda z(1) = c\ . \qquad 4.6.4$$

If the set $\tilde{J} = \{\Lambda^{-1}V\Lambda,\ M\Lambda,\ N\Lambda\}$ is boundary compatible, the solution for (4.6.4) may be written as

$$z(t) = \Phi^{\Lambda^{-1}V\Lambda}(t)[M\Lambda + N\Lambda\Phi^{\Lambda^{-1}V\Lambda}(1,0)]^{-1}\, c. \qquad 4.6.5$$

In passing, it may be quickly shown that if the set $J = \{V,M,N\}$ is boundary compatible, the transformed set $\tilde{J} = \{\Lambda^{-1}V\Lambda,\ M\Lambda, N\Lambda\}$ is also boundary compatible. With the matrix $[M+N\Phi^V(1,0)]$ nonsingular, post multiplication by Λ yields the nonsingular matrix $[M\Lambda+N\Phi^V(1,0)\Lambda]$. The fundamental matrices are related by $\Phi^V(1,0) = \Lambda\Phi^{\Lambda^{-1}V\Lambda}(1,0)\Lambda^{-1}$ so the nonsingular matrix $[M\Lambda+N\Phi^V(1,0)\Lambda]$ may be written as $[M\Lambda+N\Phi^{\Lambda^{-1}V\Lambda}(1,0)]$ indicating the transformed set $J = \{\Lambda^{-1}V\Lambda, M\Lambda, N\Lambda\}$ is boundary compatible. If the transformation $\Lambda^{-1}V\Lambda$ reduces V to a canonical form, the fundamental matrix $\Phi^{\Lambda^{-1}V\Lambda}(t,s)$ is of simple structure.

Now consider the nonlinear TPBVP

$$\dot{y} = Sy + f(y) \qquad My(0) + Ny(1) = c\ . \qquad 4.6.6$$

Again consider the nonsingular linear transformation

$$\Lambda z = y \qquad 4.6.7$$

and let

$$D = \Lambda^{-1} V \Lambda \ . \qquad 4.6.8$$

Then (4.6.6) becomes the transformed TPBVP

$$\dot{z} = \Lambda^{-1} S \Lambda z + \Lambda^{-1} f(\Lambda z) \qquad 4.6.9$$

$$M\Lambda z(0) + N\Lambda z(1) = c \ .$$

If the set $\tilde{J} = \{\Lambda^{-1}V\Lambda,\ M\Lambda,\ N\Lambda\}$ is boundary compatible, the integral representation for (4.6.9) is

$$T^{\tilde{J}}(y) = H^{\tilde{J}}(t)c + \int_0^1 G^{\tilde{J}}(t,s)\{\Lambda^{-1}S\Lambda z + \Lambda^{-1}f(\Lambda z(s)) - Dz\}ds \qquad 4.6.10$$

where the Green's functions are given as

$$H^{\tilde{J}}(t) = \Phi^D(t,0)[M\Lambda + N\Lambda\Phi^D(1,0)]^{-1} \qquad 4.6.11$$

and

$$G^{\tilde{J}}(t,s) = \begin{cases} \Phi^D(t,0)[M\Lambda + N\Lambda\Phi^D(1,0)]^{-1}M\Phi^D(0,s) & , \ 0 \le s \le t \\ -\Phi^D(t,0)[M\Lambda + N\Lambda\Phi^D(1,0)]^{-1}N\Phi^D(1,s) & , \ t < s \le 1 \ . \end{cases} \qquad 4.6.12$$

If it is desired to investigate the iterative solution of the operator equation

$$z = T^{\tilde{J}}(z) \ , \qquad 4.6.13$$

the operator derivative $(T_z^{\tilde{J}})'$ is given as

$$(T_z^{\tilde{J}})'u = \int_0^1 G^{\tilde{J}}(t,s)\{\Lambda^{-1}S\Lambda z(s) + \Lambda^{-1}(\partial f/\partial y)(\Lambda z(s))\Lambda - D\}u(s)ds \qquad 4.6.14$$

if the TPBVP is differentiable. However, rather than using (4.6.14), another approach may be taken. It may easily be shown that a direct relationship exists between the Green's functions for the boundary compatible set $J = \{V,M,N\}$ and the transformed boundary compatible set $\tilde{J} = \{D,M\Lambda,N\Lambda\}$. In particular,

$$H^J(t,s) = \Lambda H^{\tilde{J}}(t,s) \qquad 4.6.15$$

and

$$G_I^J(t,s) = \Lambda G_I^{\tilde{J}}(t,s)\Lambda^{-1} \qquad 4.6.16$$

$$G_{II}^J(t,s) = \Lambda G_{II}^{\tilde{J}}(t,s)\Lambda^{-1} \qquad 4.6.17$$

Hence the integral representation for (4.6.6) may be written as

$$y(t) = T^J(y)(t) = \Lambda H^{\tilde{J}}(t,s)c + \int_0^1 \Lambda G^{\tilde{J}}(t,s)\Lambda^{-1}\{Sy(s)+f(y(s),s)-Vy(s)\}ds \qquad 4.6.18$$

Then if the matrix $\Lambda^{-1}V\Lambda$ is a canonical form, $\Phi^D(t,s)$ and $G^{\tilde{J}}(t,s)$ are often much easier to calculate, and it may very well be easier to calculate estimates for $(T_y{}^J)'$.

The theory of canonical forms has received great attention in the past years. General books of interest include Gantmacher [G1], Bodweig [B2],Turnbull [T1], and Ferrar [F2]. Of interest to control analysts are the books of Bellman [B1] and Ogata [O1]. In particular, we now present a well known theorem concerning the diagonalization of matrices.

Theorem 4.6.19

If the characteristic roots λ_i of the matrix V are distinct, there exists a matrix Λ such that

$$\Lambda^{-1}V\Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \qquad 4.6.20$$

Proof. (See Bellman or Ogata).

However, if a $p \times p$ matrix V does not possess p linearly independent eigenvectors, then V is not similar to a diagonal matrix. In this case, it can be proved rigorously that a $p \times p$ matrix, V, possessing less than p linearly independent characteristic vectors is similar to the Jordan canonical form, where the elements in the main diagonal are the characteristic roots and the elements immediately above the main diagonal are either one or zero and all other elements are zero. (The proof of this statement may be found in Turnbull.) However, rather than using the more involved Jordan canonical form, we shall make use of the following result from Bellman.

Theorem 4.6.21

Given any matrix W, we can find a matrix V with distinct characteristic roots such that $\|W-V\| \leq \mathcal{E}$, where $\mathcal{E}$ is any preassigned quantity.

Proof. (See Bellman.)

The importance of Theorem 4.6.21 is as follows. Assume analysis of the convergence conditions indicates the matrix W is a good choice for use in the integral representation. If W contains multiple characteristic roots, it is not similar to a diagonal form and the advantages of this simple form are not available. However, since we are free to choose the matrix, we may use Theorem 4.6.21 and "perturb" the W matrix to a V matrix "close to W" (i.e., $\|W-V\| \leq \mathcal{E}$) which does have distinct characteristic roots. We may then determine a matrix

Λ such that $\Lambda^{-1}V\Lambda$ is a diagonal form. We shall consider one further special case and that is the system whose distinct characteristic roots involve complex conjugate paris. We shall follow Ogata.

Assume for convenience that the system involves only one pair of complex conjugate characteristic roots. Extension to the case where there are more than one pair of complex conjugate characteristic roots is obvious. Assume that the eigenvalues λ_1 and $\bar{\lambda}_1$ are complex conjugates and are given by

$$\lambda_1 = \sigma + j\omega \quad \bar{\lambda}_1 = \sigma - j\omega \ . \qquad 4.6.22$$

Assume also that the eigenvalues $\lambda_3,\ldots,\lambda_p$ are real and distinct. The diagonal matrix is then of the form

$$D = \Lambda^{-1}V\Lambda = \begin{bmatrix} \sigma+j\omega & & & & 0 \\ & \sigma-j\omega & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ 0 & & & & \lambda_p \end{bmatrix} \qquad 4.6.23$$

which may be transformed into the modified diagonal form

$$\hat{D} = \begin{bmatrix} \sigma & \omega & & & 0 \\ -\omega & \sigma & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ 0 & & & & \lambda_p \end{bmatrix} \qquad 4.6.24$$

by means of the transformation matrix defined by

$$K = \begin{bmatrix} 1/2 & -j/2 & & & 0 \\ 1/2 & j/2 & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{bmatrix} \qquad 4.6.25$$

Namely the modified diagonal form D is given as

$$\hat{D} = K^{-1}\ DK = K^{-1}\ \Lambda^{-1}V\Lambda K \ . \qquad 4.6.26$$

Not only does $\hat{D}$ have only real elements but, more significantly, $K^{-1}\Lambda^{-1}$ and ΛK have only real elements. Now setting

$$\hat{\Lambda} = \Lambda K \qquad 4.6.27$$
$$\hat{\Lambda}^{-1} = K^{-1}\Lambda^{-1} ,$$

the transformation is given the standard form of

$$D = \hat{\Lambda}^{-1} V \hat{\Lambda} . \qquad 4.6.28$$

As a result of this discussion, in the sequel the term canonical form shall specifically refer either to diagonal or to the modified diagonal form as in (4.6.24).

We shall now choose several forms for the V matrix to illustrate the use of the similarity transformation with the boundary value sets of interest.

Example 4.6.29.

Let us consider the boundary value set

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} \qquad 4.6.30$$

and the $2n \times 2n$ V matrix with distinct characteristic roots

$$V = \begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix} . \qquad 4.6.31$$

The similarity transformation has the form

$$\Lambda = \begin{bmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{bmatrix} \quad \Lambda^{-1} = \begin{bmatrix} \Lambda_{11}^{-1} & 0 \\ 0 & \Lambda_{22}^{-1} \end{bmatrix} \qquad 4.6.32$$

and the canonical matrix D is given as

$$D = \Lambda^{-1} V \Lambda \ . \qquad 4.6.33$$

The fundamental matrix of the canonical matrix has the general form

$$\Phi^D(t,s) = \begin{bmatrix} \phi_{11}(t,s) & 0 \\ 0 & \phi_{22}(t,s) \end{bmatrix} \ . \qquad 4.6.34$$

The matrix $[M\Lambda + N\Lambda\Phi^D(1,0)]$ is obtained as

$$[M\Lambda + N\Lambda\Phi^D(1,0)] = \begin{bmatrix} \Lambda_{11} & 0 \\ -K\Lambda_{11}\phi_{11}(1,0) & \Lambda_{22}\phi_{22}(1,0) \end{bmatrix} , \qquad 4.6.35$$

and if the inverse exists,

$$[M\Lambda + N\Lambda\Phi^D(1,0)]^{-1} = \begin{bmatrix} \Lambda_{11}^{-1} & 0 \\ \phi_{22}^{-1}(1,0)\Lambda_{22}^{-1} K\Lambda_{11}\phi_{11}(1,0) & \phi_{22}^{-1}(1,0)\Lambda_{22}^{-1} \end{bmatrix} \qquad 4.6.36$$

The Green's functions are then found as

$$\tilde{G}_I^J(t,s) = \begin{bmatrix} \phi_{11}(t,s) & 0 \\ \phi_{22}(t,0)\Delta\phi_{11}(0,s) & 0 \end{bmatrix} \qquad 4.6.37$$

and

$$\tilde{G}_{II}^J(t,s) = \begin{bmatrix} 0 & 0 \\ \phi_{22}(t,0)\Delta\phi_{11}(0,s) & -\phi_{22}(t,s) \end{bmatrix} \qquad 4.6.38$$

where the $n \times n$ matrix Δ is given by

$$\Delta = -\phi_{22}^{-1}(1,0)\Lambda_{22}^{-1}\, K\Lambda_{11}\phi_{11}(1,0) \ . \qquad 4.6.39$$

Then for $\tilde{P}(t)$ defined as

$$\tilde{P}(t) = \int_0^t |G_I^{\tilde{J}}(t,s)|ds + \int_t^1 |G_{II}^{\tilde{J}}(t,s)|ds \ ,$$

we have

$$\tilde{P}(t) = \begin{bmatrix} \int_0^t |\phi_{11}(t,s)|ds & 0 \\ \int_0^1 |\phi_{22}(t,0)\Delta\phi_{11}(0,s)|ds & \int_t^1 |\phi_{22}(t,s)|ds \end{bmatrix} . \qquad 4.6.40$$

It should be noticed that $P(t)$ may be obtained as

$$P(t) = \Lambda\tilde{P}(t)\Lambda^{-1} \qquad 4.6.41$$

Example 4.6.42.

Let us now consider the use of a $2n \times 2n$ V matrix of the form

$$V = \begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix} \qquad 4.6.42$$

with the boundary value set corresponding to no terminal penalty, i.e.,

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} . \qquad 4.6.43$$

Defining the canonical form $D = \Lambda^{-1}V\Lambda$, we can calculate $\tilde{P}(t)$ by specializing the results of Example 4.6.29 with $K = 0$. $\tilde{P}(t)$ is obtained as

$$\tilde{P}(t) = \begin{bmatrix} \int_0^t |\phi_{11}(t,s)|\,ds & 0 \\ 0 & \int_t^1 |\phi_{22}(t,s)|\,ds \end{bmatrix} . \qquad 4.6.44$$

This is an especially nice result if the V_{11} and V_{22} matrices may be diagonalized. Specifically, if D has the form

$$D = \left[\begin{array}{ccc|ccc} \lambda_n & & & & & \\ & \ddots & & & 0 & \\ & & \lambda_n & & & \\ \hline & & & \mu_1 & & \\ & 0 & & & \ddots & \\ & & & & & \mu_n \end{array}\right] , \qquad 4.6.45$$

Then $\tilde{P}(t)$ has the particularly simple form

$$\tilde{P}(t) = \left[\begin{array}{ccc|ccc} \frac{1}{\lambda_1}(e^{\lambda_1 t}-1) & & & & & \\ & \ddots & & & 0 & \\ & & \frac{1}{\lambda_n}(e^{\lambda_n t}-1) & & & \\ \hline & & & \frac{1}{\mu_1}\left(1-e^{-\mu_1(1-t)}\right) & & \\ & 0 & & & \ddots & \\ & & & & & \frac{1}{\mu_n}\left(1-e^{-\mu_n(1-t)}\right) \end{array}\right] \qquad 4.6.46$$

In this section, the use of similarity transformations was introduced in an attempt to simplify the calculation of P(t). For some cases the technique worked very well yielding simple expressions for P(t). However, in some instances the matrix P(t) is very awkward to calculate. Consideration of the similarity transformation led to the development of an approximate technique which is presented in the next section.

4.7. Approximate Technique

We shall now introduce a technique which, though not mathematically rigorous, allows one to obtain estimates for P(t) in a much simpler fashion. Using the canonical transformation Λ, define the matrix D as $D = \Lambda^{-1}V\Lambda$ and write $G_I^J(t,s)$ and $G_{II}^J(t,s)$ as

$$G_I^J(t,s) = \{\Lambda\Phi^D(t,0)\Lambda^{-1}\}\{[M+N\Phi^V(1,0)]^{-1}M\}\{\Lambda\Phi^D(0,s)\Lambda^{-1}\} \qquad 4.7.1$$

$$G_{II}^J(t,s) = -\{\Lambda\Phi^D(t,0)\Lambda^{-1}\}\{[M+N\Phi^V(1,0)]^{-1}N\Phi^V(1,0)\}\{\Lambda\Phi^D(0,s)\Lambda^{-1}\}. \qquad 4.7.2$$

The terms have been separated by brackets to indicate the factors contributing to the magnitude of P(t), namely the inversion and the integration of the fundamental matrices. Consider the general 2n × 2n V matrix and the resultant fundamental matrix to be given as

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \qquad \Phi^V(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & \Omega_{12}(t,s) \\ \Omega_{21}(t,s) & \Omega_{22}(t,s) \end{bmatrix} . \qquad 4.7.3$$

For the boundary value matrices

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ -K & I \end{bmatrix} , \qquad 4.7.4$$

we have

$$[M+N\Phi^V(1,0)] = \begin{bmatrix} I & 0 \\ -K\Omega_{11}(1,0)+\Omega_{21}(1,0) & -K\Omega_{12}(1,0)+\Omega_{22}(1,0) \end{bmatrix} \qquad 4.7.5$$

and if the inverse exists,

$$[M+N\Phi^V(1,0)]^{-1} = \begin{bmatrix} I & 0 \\ -[-K\Omega_{12}(1,0)+\Omega_{22}(1,0)]^{-1}[-K\Omega_{11}(1,0)+\Omega_{21}(1,0)] & [-K\Omega_{12}(1,0)+\Omega_{22}(1,0)]^{-1} \end{bmatrix} . \qquad 4.7.6$$

The center bracketed terms are then found as

$$[M+N\Phi^V(1,0)]^{-1} M = \begin{bmatrix} I & 0 \\ \Delta & 0 \end{bmatrix} \qquad 4.7.7$$

and

$$[M+N\Phi^V(1,0)]^{-1} N\Phi^V(1,0) = \begin{bmatrix} 0 & 0 \\ -\Delta & I \end{bmatrix} \qquad 4.7.8$$

where the $n \times n$ matrix Δ is given as

$$\Delta = -[K\Omega_{12}(1,0)+\Omega_{22}(1,0)]^{-1}[-K\Omega_{11}(1,0)+\Omega_{21}(1,0)]. \qquad 4.7.9$$

Now assuming that $\Phi^D(t,s)$ represents the primary magnitude characteristics of $\Lambda\Phi^D(t,s)\Lambda^{-1}$, we shall form

$$\{\Phi^D(t,0)\}\{[M+N\Phi^V(1,0)]^{-1} M\}\{\Phi^D(0,s)\} \qquad 4.7.10$$

and

$$-\{\Phi^D(t,0)\}\{[M+N\Phi^V(1,0)]^{-1} N\Phi^V(1,0)\}\{\Phi^D(0,s)\} \qquad 4.7.11$$

as approximations to $G_I^J(t,s)$ and $G_{II}^J(t,s)$. Following the discussion in Section 4.6 concerning canonical forms, the fundamental matrix $\Phi^D(t,s)$ may be represented in the form

$$\Phi^D(t,s) = \begin{bmatrix} \phi_{11}^D(t,s) & 0 \\ 0 & \phi_{22}^D(t,s) \end{bmatrix} . \qquad 4.7.12$$

Using (4.7.10) and (4.7.11), we obtain the approximations

$$G_I^J(t,s) \approx \begin{bmatrix} \phi_{11}^D(t,s) & 0 \\ \phi_{22}^D(t,0)\Delta\phi_{11}^D(0,s) & 0 \end{bmatrix} \qquad 4.7.13$$

and

$$G_{II}^J(t,s) \approx \begin{bmatrix} 0 & 0 \\ \phi_{22}^D(t,0)\Delta\phi_{11}^D(0,s) & \phi_{22}^D(t,s) \end{bmatrix} . \qquad 4.7.14$$

Finally, this yields

$$P(t) \begin{bmatrix} \int_0^t |\phi_{11}^D(t,s)|ds & 0 \\ \int_0^1 |\phi_{22}^D(t,0)\Delta\phi_{11}^D(0,s)|ds & \int_t^1 |\phi_{22}^D(t,s)|ds \end{bmatrix} . \qquad 4.7.15$$

For the boundary value set

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad N = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} , \qquad 4.7.16$$

we may specialize the results of (4.7.15) with K = 0 to obtain

$$P(t) \quad \begin{bmatrix} \int_0^t |\phi_{11}^D(t,s)|ds & 0 \\ \int_0^1 |\phi_{22}^D(t,0)\hat{\Delta}\phi_{11}^D(0,s)|ds & \int_t^1 |\phi_{22}^D(t,s)|ds \end{bmatrix} \qquad 4.7.17$$

where

$$\hat{\Delta} = -\Omega_{22}^{-1}(1,0)\Omega_{21}(1,0) \ . \qquad 4.17.18$$

These approximations greatly simplify the calculation of P(t), and moreover, they capture the primary quantitative behavior of the P(t) matrix. The concepts and techniques introduced in this chapter will be illustrated in several numerical examples in Chapter 6.

Page intentionally left blank

CHAPTER 5

CONTROLLABILITY FOR NONLINEAR SYSTEMS

5.1. Introduction

The concept of null controllability is a natural aspect of the study of optimal regulation for nonlinear systems. Whereas the optimal regulator attempts to drive the system from its initial state into a region near the origin, null controllability is concerned with driving the system precisely to the origin. Historically, the issues of regulation and controllability are closely intertwined. The study of linear regulator problems in a general framework served to uncover some of the underlying relationships that exist between the structure of the optimal system and the fundamental concept of controllability [K1], [K3]. Much of the effort to date concerning null controllability of nonlinear systems has involved determination of feedback controllers such that the driven systems satisfy certain Lyapunov-type stability arguments [B4], [G3]. In this chapter the integral representation of TPBVP's and the contraction mapping theorem will be used to investigate the controllability of nonlinear systems via existence of solutions arguments.

5.2. Controllability for Linear Systems

As an introduction to the controllability issue and the approach to be taken in the study, we shall first consider the controllability of linear systems.

Definition 5.2.1.

The autonomous linear control process

$$\dot{x}(t) = Ax(t) + Bu(t) \ . \qquad 5.2.2$$

with $u \in \Omega = R^m$, is (completely) controllable in case: for each pair of points x_0 and x_1 in R^n, there exists a bounded measurable controller $u(t)$ on some finite interval $0 \leq t \leq T$, which steers x_0 to x_1 .

Theorem 5.2.3.

The autonomous linear system

$$\dot{x}(t) = Ax(t) + Bu(t) \qquad 5.2.4$$

with $u \in \Omega = R^m$, is (completely) controllable if and only if a solution exists to the linear TPBVP

$$\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix} \qquad 5.2.5$$

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x\ (0) \\ p\ (0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \begin{bmatrix} x(1) \\ p(1) \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \ .$$

Proof: Assume a solution $x^*(t)$, $p^*(t)$ exists to the TPBVP (5.2.5). Now consider the optimization problem of determining a control $u(t)$ to drive the system (5.2.4) from the initial state x_0 to the terminal state x_1 such that the cost functional

$$J = \frac{1}{2} \int_0^1 \langle u(t), u(t) \rangle \, dt \qquad 5.2.6$$

is minimized. Application of the Pontryagin principle yields precisely the TPBVP (5.2.5). Then the control

$$u(t) = -B'p^*(t) \qquad 5.2.7$$

drives the system from x_0 to x_1 .

Conversely, assume that the system (5.2.4) is completely controllable. We shall show that a solution exists to the TPBVP. The linear TPBVP

$$\dot{y}(t) = Vy(t) + f(t) \qquad 5.2.8$$

$$My(0) + Ny(1) = c$$

with

$$V = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix} \quad M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad N = \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \qquad 5.2.9$$

has a solution for every f(t) and c if $\det[M+N\Phi^V(1,0)] \neq 0$. The fundamental matrix for V is given in the form

$$\Phi^V(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & \Omega_{12}(t,s) \\ 0 & \Omega_{22}(t,s) \end{bmatrix} , \qquad 5.2.10$$

and the matrix $[M+N\Phi^V(1,0)]$ is obtained in the form

$$[M+N\Phi^V(1,0)] = \begin{bmatrix} I & 0 \\ \Omega_{11}(1,0) & \Omega_{12}(1,0) \end{bmatrix} . \qquad 5.2.11$$

The inverse, if it exists, is given as

$$[M+N\Phi^V(1,0)]^{-1} = \begin{bmatrix} I & 0 \\ -\Omega_{12}^{-1}(1,0)\Omega_{11}(1,0) & \Omega_{12}^{-1}(1,0) \end{bmatrix} \qquad 5.2.12$$

and $\det[M+N\Phi^V(1,0)] = \det[\Omega_{12}(1,0)]$. Now investigating the differential equation describing $\Phi^V(t,0)$, we have

$$\begin{aligned} \dot{\Omega}_{12}(t,0) &= A\Omega_{12}(t,0) - BB'\Omega_{22}(t,0) \quad , \quad \Omega_{12}(0,0) = 0 \\ \dot{\Omega}_{22}(t,0) &= -A'\Omega_{22}(t,0) \quad , \quad \Omega_{22}(0,0) = I \end{aligned} \qquad 5.2.13$$

These equations yield

$$\Omega_{22}(t,0) = \Phi^{-A'}(t,0) = \Phi^A(0,t)' \qquad 5.2.14$$

and

$$\Omega_{12}(t,0) = \Phi^{-A}(t,0) \int_0^t \Phi^A(0,\sigma)BB'\Phi^A(0,\sigma)'d\sigma \ . \qquad 5.2.15$$

Hence for the existence of a solution to the TPBVP (5.2.5), we must have

$$\det[\Phi^A(1,0) \int_0^1 \Phi^A(0,\sigma)BB'\Phi^A(0,\sigma)'d\sigma] \neq 0 \ . \qquad 5.2.16$$

However, the assumption of complete controllability specifies

$$\det[\int_0^1 \Phi^A(0,\sigma)BB'\Phi^A(0,\sigma)'d\sigma] \neq 0 \ , \qquad 5.2.17$$

therefore a solution to the TPBVP must exist.

Hence the issue of invertibility of $\Omega_{12}(1,0)$ leads to the well known controllability Grammian and the approach is seen to yield conditions compatible with previously derived results. The following corollary will often prove useful when selecting a V matrix for a controllability investigation.

Corollary 5.2.18

Let the constant $2n \times 2n$ matrices V,M, and N be of the form

$$V = \begin{bmatrix} V_{11} & V_{12} \\ 0 & -V_{11}' \end{bmatrix} \qquad M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \tag{5.2.19}$$

with $V_{12} = -BB'$ where B is an $n \times r$ matrix. Then the set $J = \{V,M,N\}$ is boundary compatible if and only if

$$\text{rank } [B, V_{11}B, \ldots, V_{11}^{n-1} B] = n \tag{5.2.20}$$

Proof. See proof of theorem 5.2.3 and Lee and Markus [L1] for the relationship between (5.2.20) and the controllability Grammian.

The obvious advantage provided by Corollary 5.2.18 is that it removes the calculation of $\Phi^V(1,0)$ when determining the boundary compatibility of a set J in the form of (5.2.19). With this background, we shall now consider nonlinear controllability.

5.3. Nonlinear Controllability

In this section we shall extend the approach of Section 5.2 to include nonlinear systems. However, rather than considering global controllability as for linear systems, we shall consider local null controllability, i.e., the problem of regulating an initial state, near the origin, to the origin.

Definition 5.3.1. [L1]

Consider the control process in R^n

$$\dot{x} = f(x,u) \quad \text{in} \quad C^2 \quad \text{in} \quad R^n \times \Omega \tag{5.3.2}$$

where Ω is a restraint set in R^m. The domain $\mathcal{N}$ of null controllability is

defined as the set of all points $x_0 \in R^n$, each of which can be steered to $x_1 = 0$ by some bounded measurable controller $u(t) \subset \Omega$ in finite time. If $\mathcal{N}$ contains an open neighborhood of $x_1 = 0$, then (5.3.2) is said to be locally controllable (near the origin). We shall now consider the null controllability of nonlinear systems by means of integral representations.

Theorem 5.3.3.

Consider the control process in R^n

$$\dot{x} = f(x,u) \quad \text{in} \quad C^2 \quad \text{in} \quad R^n \times \Omega \qquad 5.3.4$$

with $u = 0$ interior to the restraint set $\Omega \subset R^m$.

Assume

(a) $f(0,0) = 0$ 5.3.5

(b) $\text{rank } [B, AB, \ldots, A^{n-1}B] = n$ 5.3.6

where $A = (\partial f/\partial x)(0,0)$ and $B = (\partial f/\partial u)(0,0)$ 5.3.7

Then the domain $\mathcal{N}$ of null controllability is open in R^n.

Proof. Let us define the function $\psi(x,u)$ as

$$\psi(x,u) = f(x,u) - Ax - Bu \qquad 5.3.8$$

so that

$$\psi(0,0) = 0 \qquad 5.3.9$$

$$\left(\frac{\partial \psi}{\partial x}\right)(0,0) = 0 \qquad 5.3.10$$

and

$$\left(\frac{\partial \psi}{\partial u}\right)(0,0) = 0 \ . \qquad 5.3.11$$

Now consider the optimization problem composed of the system

$$\dot{x} = Ax + Bu + \psi(x,u) \ , \qquad 5.3.12$$

the boundary conditions,

$$x(0) = x_0 \ , \ x(1) = 0 \ , \qquad 5.3.13$$

and the cost functional

$$J = \frac{1}{2} \int_0^1 \langle u(t),u(t)\rangle \ dt. \qquad 5.3.14$$

The Hamiltonian for this problem is given by

$$H = \frac{1}{2} \langle u(t),u(t)\rangle + \langle Ax(t),p(t)\rangle + \langle Bu(t),p(t)\rangle + \langle \psi(x,u),p(t)\rangle \qquad 5.3.15$$

and the costate variable is described by the differential equation

$$\dot{p} = A'p - (\frac{\partial\psi}{\partial x})' \ (x,u)p \ . \qquad 5.3.16$$

Now assume a control of the form

$$u = -B'p \qquad 5.3.17$$

accomplishes the desired transfer. The canonical system of equations is now given as

$$\dot{x} = Ax - BB'p + \psi(x,u(p)) \qquad 5.3.18$$

$$\dot{p} = -A'p - (\frac{\partial\psi}{\partial x})' \ (x,u(p))p \qquad 5.3.19$$

subject to the boundary conditions

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(0) \\ p(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \begin{bmatrix} x(1) \\ p(1) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 5.3.20$$

This may be expressed as

$$\begin{aligned} \dot{y} &= Sy + F(y) \\ My(0) &+ Ny(1) = c \end{aligned} \qquad 5.3.21$$

where

$$S = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix} \quad M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad N = \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \quad c = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 5.3.22$$

and

$$F(y) = \begin{bmatrix} \psi(x,u(p)) \\ -\left(\frac{\partial\psi}{\partial x}\right)'(x,u(p))p \end{bmatrix} . \qquad 5.3.23$$

For the boundary compatible set $J = \{V,M,N\}$ where M and N are given in (5.3.22), the solution to (5.3.20) may be written as

$$y(t) = H^J(t)c + \int_0^1 G^J(t,s)\{Sy(s) + F(y(s)) - V(s)y(s)\}ds \qquad 5.3.24$$

or as an operator equation

$$y = T^J(y). \qquad 5.3.25$$

Clearly with $c = 0$, $y(\cdot) = 0$ is a fixed point of T^J. We are now interested in the existence of a solution, $y(\cdot)$, if c is varied in a neighborhood of the origin.

Define the variable z as

$$z = P^J(c,y) . \qquad 5.3.27$$

It is seen that a zero of P^J is a fixed point of T^J. We shall now show that $(P_y^J)'\ (0,0)$ has a bounded inverse and the desired conclusions concerning the existence of a solution y may be deduced from the implicit function theorem. (For presentations of the implicit function theorem, see Kantorovich [K4] and Holtzman [H1]). The operator derivative is given as

$$(P_y^J)'\ (0,0)v(t) = v(t) - \int_0^1 G^J(t,s)\{S + (\frac{\partial F}{\partial y})(0) - V(s)\}v(s)ds \qquad 5.3.28$$

or

$$(P_y^J)'\ (0,0)v = [I-D^J]v\ . \qquad 5.3.29$$

If the set $\tilde{J} = \{S + (\frac{\partial F}{\partial y})(0),\ M,N,\}$ is boundary compatible, then $[I-D^J]^{-1}$ is bounded and is given as (see Falb [F1])

$$[I-D^J]^{-1}\ v = [I-R^{\tilde{J}}]\ v \qquad 5.3.30$$

where

$$R^{\tilde{J}}v = \int_0^1 G^{\tilde{J}}(t,s)\{V(s) - S - (\frac{\partial F}{\partial y})(0)\}\ v(s)ds. \qquad 5.3.31$$

All that now remains is to show that $\tilde{J} = \{S + (\frac{\partial F}{\partial y})(0),M,N\}$ is boundary compatible. We have

$$F(y) = \begin{bmatrix} \psi(x,u(p)) \\ -(\frac{\partial \psi}{\partial x})'(x,u(p))p \end{bmatrix} \qquad 5.3.32$$

and

$$(\frac{\partial F}{\partial y})(y) = \begin{bmatrix} (\frac{\partial \psi}{\partial x})(x,u(p)) & (\frac{\partial \psi}{\partial p})(x,u(p)) \\ -\frac{\partial}{\partial x}[(\frac{\partial \psi}{\partial x})'(x,u(p))p] & -\frac{\partial}{\partial p}[(\frac{\partial \psi}{\partial x})'(x,u(p))p] \end{bmatrix} \qquad 5.3.33$$

We have from (5.3.10) that $(\partial\psi/\partial x)(0,0) = 0$, and $(\partial\psi/\partial p)(x,u(p))$ may be obtained as

$$(\frac{\partial\psi}{\partial p})(x,u(p)) = (\frac{\partial\psi}{\partial u})(\frac{\partial u}{\partial p})(x,u(p)) \doteq -[(\frac{\partial f}{\partial u})(x,u(p)) - B]B' \qquad 5.3.34$$

which evaluated for $y(\cdot) = 0$ yields

$$(\frac{\partial\psi}{\partial p})(0,0) = 0 \ . \qquad 5.3.35$$

Defining the $n \times n$ matrix $D(x,p)$ as

$$D(x,p) = (\frac{\partial\psi}{\partial x})'(x,u(p)) \ , \qquad 5.3.36$$

it only remains to calculate

$$\frac{\partial}{\partial x}[D(x,p)p] \quad \text{and} \quad \frac{\partial}{\partial p}[D(x,p)p] \ . \qquad 5.3.37$$

If the $n \times n$ matrix $Q(x,p)$ is defined as

$$Q(x,p) = \frac{\partial}{\partial x}[D(x,p)p] \ , \qquad 5.3.38$$

Then the elements of the matrix are given as

$$q_{k_i}(x,p) = \sum_{j=1}^{n} (\frac{\partial D_{kj}}{\partial x_i})(x,p)p_j \qquad 5.3.39$$

and then,

$$Q(0,0) = 0 \ .$$

Similarly if the $n \times n$ matrix $\Gamma(x,p) = \frac{\partial}{\partial p}[D(x,p)p]$ 5.3.41

then the elements of the matrix are given as

$$\gamma_{ki}(x,p) = \sum_{j=1}^{n} (\frac{\partial D_{kj}}{\partial p_i})(x,p)p_j + D(x,p)_{ki} \ . \qquad 5.3.42$$

Since $D(0,0) = 0$, then from (5.3.42)

$$\Gamma(0,0) = 0 \ . \qquad 5.3.43$$

As a result,

$$(\frac{\partial F}{\partial y})(0) = 0 \ , \qquad 5.3.44$$

and $\tilde{J}$ is given simply as $\tilde{J} = \{S,M,N\}$ where

$$S = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix} . \qquad 5.3.45$$

Then from the assumption that the set $\{A,B\}$ is controllable and the result of Corollary 5.2.18, the set $\tilde{J} = \{S,M,N\}$ is boundary compatible, and consequently the inverse is bounded. Hence for c in a neighborhood of the origin, a solution y exists to the TPBVP and the system is null controllable in a neighborhood of the origin as was to be proved. In addition, we note that the terminal state is not required to be the origin, but may be any point x_1 in a neighborhood of the origin.

The previous theorem does not specify the size of $\mathcal{N}$, the controllable region, only that the system is null controllable in a region near the origin. In addition, the condition that the linearized system be controllable about the origin is not necessary for nonlinear null controllability. The use of the contraction mapping theorem allows us to consider the domain of x_0, x_1 such that a solution exists to the TPBVP, and moreover, the theorem is stated without specifying linearized controllability. As an example of the use of the contraction mappings theorem for controllability investigation, the broad class of systems described as

$$\dot{x} = f(x) + Bu \qquad 5.3.46$$

will be considered.

Theorem 5.3.47.

Consider the control process in R^n

$$\dot{x} = f(x) + Bu \quad \text{in} \quad C^2 \quad \text{in} \quad R^n \times \Omega \ . \qquad 5.3.48$$

Let $y_0(\cdot)$ be an element of $\mathcal{C}([0,1],R^p)$ and let $\bar{S} = \bar{S}(y_0,r)$. Suppose that (i) $J = \{V,M,N\}$ is a boundary compatible set, and (ii) there are real numbers η and α with $\eta \geq 0$ and $0 \leq \alpha < 1$ such that

$$1) \quad \| T^J(y_0) - y_0 \| = \left\| H^J(t) \begin{bmatrix} x_0 \\ 0 \end{bmatrix} + \int_0^1 G^J(t,s) \left\{ \begin{bmatrix} -BB'p_0(s) + f(x_0(s)) \\ -(\frac{\partial f}{\partial x})'(x_0(s))p_0(s) \end{bmatrix} - V(s) \begin{bmatrix} x_0(s) \\ p_0(s) \end{bmatrix} \right\} ds - \begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix} \right\| \leq \eta \qquad 5.3.49$$

$$2) \quad \sup_{y \in \bar{S}} \left\{ \| (T_y^J)' \| \right\} = \sup_{y \in \bar{S}} \sup_{\|u\| \leq 1} \left\| \int_0^1 G^J(t,s) \left\{ \begin{bmatrix} (\frac{\partial f}{\partial x})(x(s)) & 0 \\ -\frac{\partial}{\partial x}[(\frac{\partial f}{\partial x})'(x(s))p(s)] & -(\frac{\partial f}{\partial x})'(x(s)) \end{bmatrix} - V(s) \right\} u(s) ds \right\| \leq \alpha \qquad 5.3.50$$

$$3) \quad \frac{1}{1-\alpha} \eta \leq r \ . \qquad 5.3.51$$

Then the CM sequence $\{y_n(\cdot)\}$ for the TPBVP based on y_0 and J converges uniformly to the unique solution y^* in $\bar{S}$ and a control exists, $u = -B'p^*$, to steer the

system (5.3.48) from x_0 to the origin.

Proof. Consider the optimization problem consisting of the system (5.3.48), the cost functional

$$J = \frac{1}{2} \int_0^1 \langle u(t), u(t) \rangle \, dt, \qquad 5.3.52$$

and the boundary conditions

$$x(0) = x_0 \ , \ x(1) = 0 \ . \qquad 5.3.53$$

Application of the Pontryagin principle to the posed optimization problem reduces the necessary conditions for optimality to the TPBVP

$$\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} -BB'p + f(x) \\ -\left(\frac{\partial f}{\partial x}\right)'(x)p \end{bmatrix} \qquad 5.3.54$$

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(0) \\ p(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \begin{bmatrix} x(1) \\ p(1) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \qquad 5.3.55$$

For the boundary compatible set $J = \{V,M,N\}$, the solution to the TPBVP may be written under certain smoothness conditions as

$$\begin{bmatrix} x(t) \\ p(t) \end{bmatrix} = T^J(y)(t) = H^J(t) \begin{bmatrix} x_0 \\ 0 \end{bmatrix} + \int_0^1 G^J(t,s) \left\{ \begin{bmatrix} -BB'p(s) + f(x(s)) \\ -\left(\frac{\partial f}{\partial x}\right)'(x(s))p(s) \end{bmatrix} - V(s) \begin{bmatrix} x(s) \\ p(s) \end{bmatrix} \right\} ds \qquad 5.3.56$$

Applying contraction mappings Theorem 3.4.14 to the operator (5.3.56) yields the conditions to be proved in Theorem 5.3.47.

For the general system formulation, $\dot{x} = f(x,u)$, the canonical equations of 2.3.54 are used subject to the boundary conditions (5.3.55). Techniques for calculating the criteria of Theorem 5.3.47 will now be considered.

5.4. Evaluation of Controllability Convergence Parameters

In Section 4.6, the variables z_0, z, and $P(t)$ were defined such that coarse estimates for η and α were obtained as

$$\eta = \|P(\cdot)z_0\| \qquad 5.4.1$$

and

$$\alpha = \|P(\cdot)z\| \ . \qquad 5.4.2$$

In this section we shall consider the determination of z_0, z, and $P(t)$ for the controllability Theorem 5.3.47. In particular, the conditions for boundary compatibility and the use of simple V matrices and similarity transformations will be considered.

The boundary value set for controllability problems is given as

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \ . \qquad 5.4.3$$

Assuming a general form for the $2n \times 2n$ V matrix and the fundamental matrix $\Phi^V(t,s)$, i.e.,

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \qquad \Phi^V(t,s) = \begin{bmatrix} \Omega_{11}(t,s) & \Omega_{12}(t,s) \\ \Omega_{21}(t,s) & \Omega_{22}(t,s) \end{bmatrix} . \qquad 5.4.4$$

The matrix $[M+N\Phi^V(1,0)]$ is obtained as

$$[M+N\Phi^V(1,0)] = \begin{bmatrix} I & 0 \\ \Omega_{11}(1,0) & \Omega_{12}(1,0) \end{bmatrix} . \qquad 5.4.5$$

This yields $\det[M+N\Phi^V(1,0)] = \det[\Omega_{12}(1,0)]$. Hence the condition for boundary compatibility reduces to the nonsingularity of $\Omega_{12}(1,0)$. In passing, it is seen that neither the zero matrix nor a diagonal matrix (nor a modified diagonal) may be used in the integral representation. If $\det[\Omega_{12}(1,0)] \neq 0$, the inverse of $[M+N\Phi^V(1,0)]$ is given as

$$[M+N\Phi^V(1,0)]^{-1} = \begin{bmatrix} I & 0 \\ -\Omega_{12}^{-1}(1,0)\ \Omega_{11}(1,0) & \Omega_{12}^{-1}(1,0) \end{bmatrix} \qquad 5.4.6$$

and then the core matrices of the Green's function are given as

$$[M+N\Phi^V(1,0)]^{-1}M = \begin{bmatrix} I & 0 \\ -\Omega_{12}^{-1}\ (1,0)\ \Omega_{11}(1,0) & 0 \end{bmatrix} . \qquad 5.4.7$$

and

$$[M+N\Phi^V(1,0)]^{-1}N\Phi^V(1,0) = \begin{bmatrix} 0 & 0 \\ \Omega_{12}^{-1}(1,0)\ \Omega_{11}(1,0) & I \end{bmatrix} . \qquad 5.4.8$$

Since the boundary value set (5.4.3) disallows the use of particularly simple V matrices, we shall consider an approximate technique for calculating $G^J(t,s)$ and $P(t)$ for V matrices of general structure.

Example 5.4.9.

Using the canonical transformation

$$D = \Lambda^{-1}V\Lambda , \qquad 5.4.10$$

the Green's function matrices are given as

$$G_I^J(t,s) = \{\Lambda\Phi^D(t,0)\Lambda^{-1}\}\{[M+N\Phi^V(1,0)]^{-1}M\}\{\Lambda\Phi^D(0,s)\Lambda^{-1}\} \qquad 5.4.11$$

$$G_{II}^J(t,s) = -\{\Lambda\Phi^D(t,0)\Lambda^{-1}\}\{[M+N\Phi^V(1,0)]^{-1}N\Phi^V(1,0)\}\{\Lambda\Phi^D(0,s)\Lambda^{-1}\} \qquad 5.4.12$$

for the boundary compatible set $J = \{V,M,N\}$. From (5.4.7) and (5.4.8), the center bracketed terms in (5.4.11), (5.4.12) are given as

$$[M+N\Phi^V(1,0)]^{-1}M = \begin{bmatrix} I & 0 \\ \Delta & 0 \end{bmatrix} \qquad 5.4.13$$

and

$$[M+N\Phi^V(1,0)]^{-1}N\Phi^V(1,0) = \begin{bmatrix} 0 & 0 \\ -\Delta & I \end{bmatrix} \qquad 5.4.14$$

where

$$\Delta = -\Omega_{12}^{-1}(1,0)\ \Omega_{11}(1,0) \qquad 5.4.15$$

Assuming that $\Phi^D(t,s)$ represents the primary magnitude characteristics of $\Lambda\Phi^D(t,s)\Lambda^{-1}$, approximations for $G_I^J(t,s)$ and $G_{II}^J(t,s)$ are formed as

$$G_I^J\ (t,s) \approx \Phi^D(t,0)[M+N\Phi^V(1,0)]^{-1}M\Phi^D(0,s) \qquad 5.4.16$$

and

$$G_{II}^J(t,s) \approx -\Phi^D(t,0)[M+N\Phi^V(1,0)]^{-1}N\Phi^V(1,0)\Phi^D(0,s)\ . \qquad 5.4.17$$

Following the discussion in Section 4.6, V is chosen such that $\Phi^D(t,s)$ is diagonal or modified diagonal and may be represented as

$$\Phi^D(t,s) = \begin{bmatrix} \phi_{11}^D(t,s) & 0 \\ 0 & \phi_{22}^D(t,s) \end{bmatrix}\ . \qquad 5.4.18$$

Using this form of $\Phi^D(t,s)$ in (5.4.16) and (5.4.17), the approximations for $G_I^J(t,s)$ and $G_{II}^J(t,s)$ are given as

$$G_I^J(t,s) \approx \begin{bmatrix} \phi_{11}^D(t,s) & 0 \\ \phi_{22}^D(t,0)\ \Delta\phi_{11}^D(0,s) & 0 \end{bmatrix} \qquad 5.4.19$$

and

$$G_{II}^J(t,s) \approx \begin{bmatrix} 0 & 0 \\ \phi_{22}^D(t,0)\ \Delta\phi_{11}^D(0,s) & \phi_{22}^D(t,s) \end{bmatrix} \qquad 5.4.20$$

The approximation for P(t) is then given as

$$P(t) \approx \begin{bmatrix} \int_0^t |\phi_{11}^D(t,s)|ds & 0 \\ \int_0^1 |\phi_{22}^D(t,0)\ \Delta\phi_{11}^D(0,s)|ds & \int_t^1 |\phi_{22}^D(t,s)|ds \end{bmatrix} \qquad 5.4.21$$

where

$$\Delta = -\Omega_{12}^{-1}(1,0)\ \Omega_{11}(1,0)\ . \qquad 5.4.22$$

For D a diagonal matrix, P(t) given by (5.4.21) becomes a particularly simple form. Several variables are now defined which will be used with P(t) to calculate estimates for η and α in Theorem 5.3.47.

Definition 5.4.22.

Using the boundary compatible initial estimate, $y_0(t) = H^J(t)c$, define the 2n vector z_0 as

$$z_o = \sup_s \begin{Bmatrix} |-BB'p_0(s) - V_{11}x_0(s) - V_{12}p_0(s) + f(x_0(s))| \\ |-V_{21}x_0(s) - V_{22}p_0(s) - (\partial f/\partial x)'(x_0(s))p_0(s)| \end{Bmatrix} . \qquad 5.4.23$$

A conservative estimate for η is given as

$$\eta = \|P(\cdot)z_0\| . \qquad 5.4.24$$

Definition 5.4.25.

Define the real numbers ν_{ij}, ξ_{ij}, and σ_{ij} as

$$\nu_{ij} = \sup_{x \in \bar{S}} \left\{ \left| \left(\frac{\partial f_i}{\partial x_j}\right)(x) - V_{11_{ij}} \right| \right\} ; \quad i,j=1,n \qquad 5.4.26$$

$$\xi_{ij} = \left| (-BB')_{ij} - V_{12_{ij}} \right| \quad ; \quad i,j=1,n \qquad 5.4.27$$

and

$$\sigma_{ij} = \sup_{x,p \in \bar{S}} \left\{ \left| \sum_{k=1}^{n} \left(\frac{\partial^2 f_k}{\partial x_i \partial x_j}\right)(x)\ p_k \right| \right\} , \qquad 5.4.28$$

and define the n vectors z_I and z_{II} to be composed of the elements

$$z_{I_i} = \sum_{j=1}^{n} (\nu_{ij} + \xi_{ij}) \qquad 5.4.29$$

and

$$z_{II_i} = \sum_{j=1}^{n} (\sigma_{ij} + \nu_{ij}) \qquad 5.4.30$$

Then the 2n vector z defined as

$$z = \begin{bmatrix} z_I \\ z_{II} \end{bmatrix}, \qquad 5.4.31$$

may be used with P(t) to obtain a coarse estimate for α as

$$\alpha = \|P(\cdot)z\| \quad . \qquad 5.4.32$$

Numerical evaluation of the convergence criteria is presented for various examples in Chapter 6.

Page Intentionally Left Blank

CHAPTER 6

NUMERICAL EXAMPLES

6.1. Introduction

We examine the regulation and control of several nonlinear systems to demonstrate the usefulness of contraction mappings and to illustrate the practical applications of the major theorems. There are many well known and very powerful iterative methods for the solution of optimal control problems. However, practical convergence criteria are few and far between. In this chapter we demonstrate that general results may be obtained via the application of the contraction mappings convergence theorem. In addition, the practical application of the contraction mapping algorithm demonstrates that in many cases it is an efficient, straightforward technique for the solution of optimal control problems. The examples demonstrate that practical application has a much broader range than the theoretical results might imply. This is primarily due to the coarse estimates which are used to evaluate the convergence parameters.

The first example to be considered is the regulation of the well known Van der Pol equation. As an illustrative exercise, both contraction mappings and modified contraction mappings are applied to this problem. The results obtained are compared with previously published data. The second example begins a two part sequence investigating the null controllability of nonlinear systems. The first member of the sequence is a particularly simple system which serves to introduce bounded control problems. The final example of the

chapter considers the null controllability of the pitching motion of a satellite with bounded control thrust.

6.2. Van der Pol's Equation

In this rather long example, we consider in detail many of the concepts essential to the contraction mappings theory. In particular, the choice of the boundary compatible set J and the calculation of the convergence parameters will be investigated closely. The system to be considered is the driven, second order nonlinear oscillator studied by Van der Pol.

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + \varepsilon(1-x_1^2)x_2 + u \ . \end{aligned} \qquad 6.2.1$$

The cost functional to be minimized is taken from Bullock [B6] as

$$J = \frac{1}{2} \int_0^5 [x_1^2(t) + x_2^2(t) + u^2(t)]dt \ , \qquad 6.2.2$$

and the boundary conditions for the optimal regulator problem are given as

$$\begin{aligned} x_1(0) &= 1.0 \qquad x_1(5) \text{ unspecified} \\ x_2(0) &= 0.0 \qquad x_2(5) \text{ unspecified} \ . \end{aligned} \qquad 6.2.3$$

From Example 2.3.10, the necessary conditions for optimality reduce to the TPBVP

$$\begin{aligned} \dot{y} &= Sy + f(y) \\ Ky(0) &+ Ly(1) = c \end{aligned} \qquad 6.2.4$$

where

$$S = \begin{bmatrix} 0 & 5 & 0 & 0 \\ -5 & 0 & 0 & -5 \\ -5 & 0 & 0 & 5 \\ 0 & -5 & -5 & 0 \end{bmatrix} \qquad f(y) = \varepsilon \begin{bmatrix} 0 \\ 5(1-x_1^2)x_2 \\ 10x_1x_2p_2 \\ -5(1-x_2^2)p_2 \end{bmatrix} \qquad 6.2.5$$

and

$$K = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad L = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad c = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad 6.2.6$$

Using a boundary compatible set $J = \{V(t), M, N\}$, (6.2.4) may be expressed in integral form as

$$y(t) = H^J(t)\{c-Ky(0) - Ly(1) + My(0) + Ny(1)\} + \int_0^1 G^J(t,s)\{Sy(s) + f(y(s)) - V(s)y(s)\}ds \qquad 6.2.7$$

The iterative solution of (6.2.7) by contraction mappings is now considered. We begin with the selection of the boundary compatible set $J = \{V(t),M,N\}$.

Since the boundary conditions of (6.2.4) are linear, the natural choice for the matrices M and N are M=K, N=L. If the initial estimate of the solution is then chosen as $H^J(t)c$, every member of the contraction mapping sequence satisfies the boundary conditions. As indicated previously, it is often advantageous to choose the matrix V in such a way that $\{S + (\partial f/\partial y)(y)-V(s)\}$ is small. Following this guideline generally requires inclusion of time varying functions in the V matrix, thus complicating the convergence analysis. However, if ε is small in (6.2.5), an acceptable choice for V is simply the linear part of (6.2.4), i.e., V = S. For larger values of ε, it may become necessary to include an effect of the nonlinearity in the choice of the V matrix, but for now, consider V to be chosen as

$$V = \begin{bmatrix} 0 & 5 & 0 & 0 \\ -5 & 0 & 0 & -5 \\ -5 & 0 & 0 & 5 \\ 0 & -5 & -5 & 0 \end{bmatrix} . \qquad 6.2.8$$

The variables $P(t)$, z_0, and z, defined in Chapter 4 for use in convergence analysis, will now be obtained for this example. From (4.2.9), (4.2.11), (6.2.5) and (6.2.8), the vectors z_0 and z are

$$z_0 = \sup_t \mathcal{E}\left\{\begin{bmatrix} 0 \\ |5\ (1-x_1^2(t)_0)x_2(t)_0| \\ |10\ x_1(t)_0 x_2(t)_0 p_2(t)_0| \\ |5\ (1-x_1^2(t)_0)p_2(t)_0| \end{bmatrix}\right\} \qquad 6.2.9$$

and

$$z = \sup_t \sup_{y \in \bar{S}} \mathcal{E}\left\{\begin{bmatrix} 0 \\ |2\ x_1(t)x_2(t)| + |(1-x_1^2(t))| \\ |2\ x_2(t)p_2(t)| + |2\ x_1(t)p_2(t)| + |2\ x_1(t)x_2(t)| \\ |2x_1(t)p_2(t)| + |(1-x_1^2(t))| \end{bmatrix}\right\} \qquad 6.2.10$$

for V given by (6.2.8), the calculation and structure of the fundamental matrix is somewhat involved, thus complicating the calculation of $P(t)$. Since the characteristic roots of V are two pairs of complex conjugates, the techniques of Example 4.4.24 are useful for evaluating the $P(t)$ matrix. The canonical transformation

$$D = \Lambda^{-1} V \Lambda \qquad 6.2.11$$

transforms the V matrix into the block diagonal form

$$D = \begin{bmatrix} \sigma_1 & \omega_1 & 0 & 0 \\ -\omega_1 & \sigma_1 & 0 & 0 \\ 0 & 0 & \sigma_2 & \omega_2 \\ 0 & 0 & -\omega_2 & \sigma_2 \end{bmatrix}, \qquad 6.2.12$$

where $\sigma_1 = -3.35$, $\omega_1 = 4.91$, and $\sigma_2 = 3.35$, $\omega_2 = 4.91$. It is then straightforward to calculate the matrix P(t) from (4.7.1) and (4.7.2).

The parameter ε is included in the example so that the general case may be considered. In particular we are interested in determining the range of ε for which the contraction mappings theorem is valid. Before proceeding with the convergence analysis, the sphere $\bar{S}(y_0,r)$ must be defined. The initial estimate of the solution, $y_0(\cdot)$ is taken to be the boundary compatible initial estimate, i.e., the solution to the linear TPBVP

$$\dot{y} = Vy \qquad My(0) + Ny(1) = c$$

or 6.2.13

$$y(t) = H^J(t)\ c.$$

(It should be noticed that this choice for y_0 does not require additional computation since the terms are necessary for the CM algorithm.) With this choice of y_0, the radius of $\bar{S}$ is taken as $r = 0.1$. This sphere $\bar{S}(y_0,r)$ is illustrated in Figure 6.1.

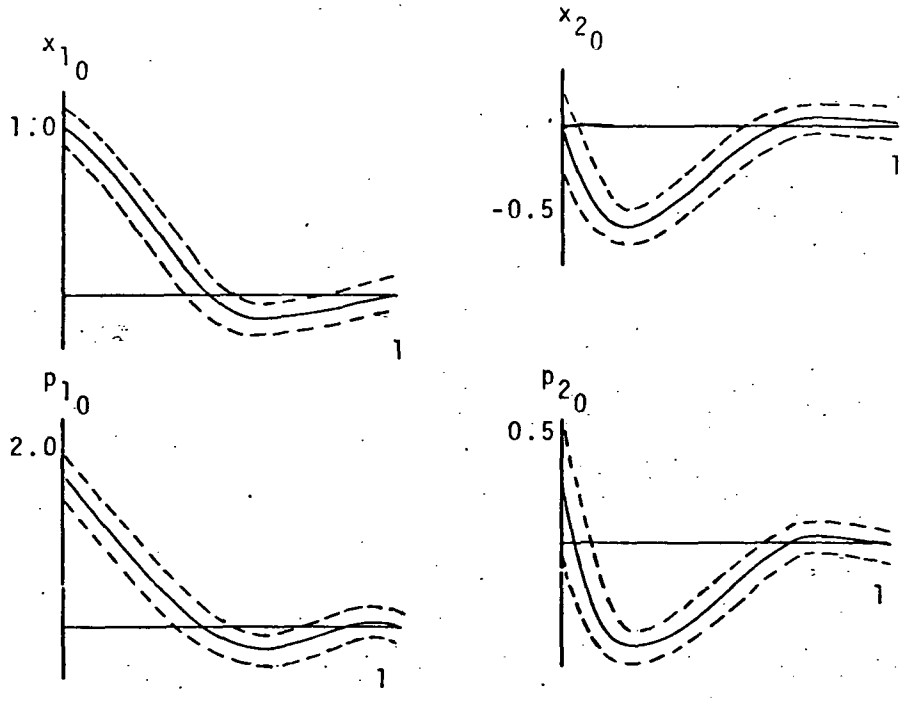


Figure 6.1 The Sphere $\bar{S}(y_0,r)$

From (6.2.9) and (6.2.10), the vectors z_0 and z are calculated as

$$z_0 = \begin{bmatrix} 0.0 \\ 2.1 \\ 2.5 \\ 3.8 \end{bmatrix} \mathcal{E} \qquad z = \begin{bmatrix} 0.0 \\ 5.8 \\ 12.6 \\ 7.9 \end{bmatrix} \mathcal{E} \qquad 6.2.14$$

Conservative estimates for the convergence parameters η and α are obtained as

$$\eta = \sup_t \{P(t)z_0\} = 2.1\mathcal{E} \qquad 6.2.15$$

and

$$\alpha = \sup_t \{P(t)z\} = 6.7\mathcal{E} \qquad 6.2.16$$

Using (6.2.15) and (6.2.16), the requirements of Theorem 3.4.14 are specified as

$$6.7\mathcal{E} < 1 \qquad 6.2.17$$

and

$$\frac{2.1\,\mathcal{E}}{1-6.7\mathcal{E}} \leq 0.1 \qquad 6.2.18$$

Analysis of (6.2.17) and (6.2.18) shows that for $\mathcal{E} \leq 0.034$ the convergence conditions of the theorem are satisfied.

The case for $\mathcal{E} = 1.0$ is treated in the paper "A Second-Order Feedback Method for Optimal Control Computations", by Bullock and Franklin [B6]. In the paper, the optimization problem presented by (6.1,2,3) is solved by the techniques of steepest descent and second variation. We now consider the application of contraction mappings to (6.2.4) with $\mathcal{E} = 1.0$. Again the V matrix is chosen V = S. Now rather than taking y_0 as the solution to the linear TPBVP, $y_0(t)$ shall be given by the fifth iteration of the CM algorithm begun with the initial guess $H^J(t)c$. This choice for y_0 is made so that the region $\bar{S}(y_0,r)$ is more likely to include the solution $y(t)$ to the nonlinear TPBVP. We again take r = 0.1. This sphere is illustrated in Figure 6.2.

We shall first determine the convergence rate factor α. Taking the supremum over $\bar{S}$, z is given as

$$z = \begin{bmatrix} 0.0 \\ 6.1 \\ 12.9 \\ 8.1 \end{bmatrix},$$

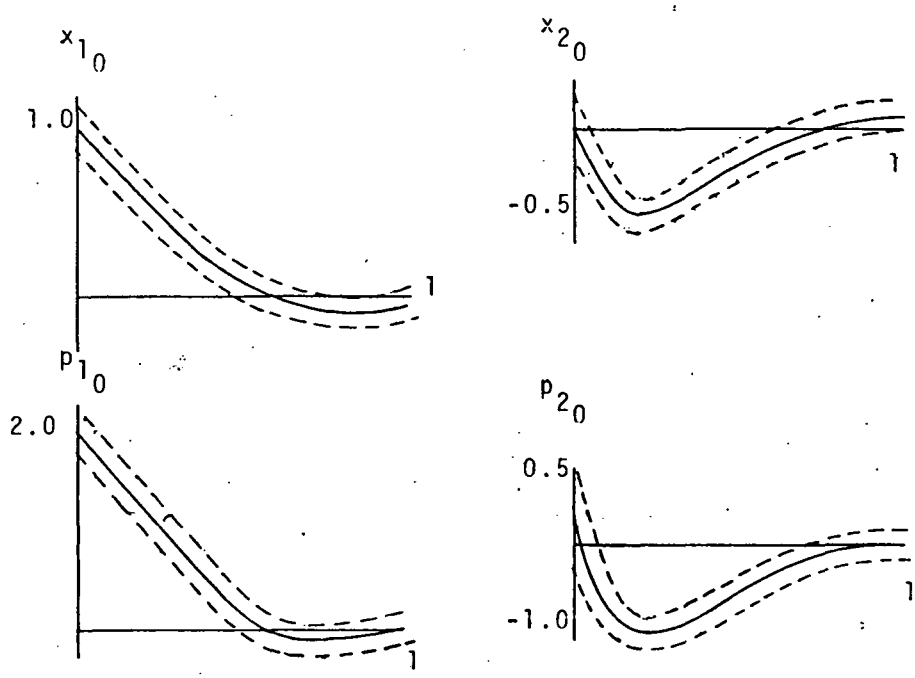


Figure 6.2 The Sphere $\bar{S}(y_0,r)$

and a coarse estimate for α is

$$\alpha = \sup_t\{P(t)z\} = 6.8 \qquad 6.2.19$$

With $\alpha > 1$, the conditions of the theorem are not satisfied and convergence is not guaranteed by the theorem. However, these theoretical results are only guidelines for the practical application of contraction mappings. In fact,

the CM algorithm reduced the convergence norm (given as $\sup_{i \in p} \sup_{t} |y_{i,n+1}(t) - y_{i,n}(t)|$) to 10^{-4} in fourteen iterations. In order to compare these results with those presented in [B6], we note that the norm of the cost function (given as $|J_{n+1} - J_n|$) was reduced to 10^{-5} in fourteen iterations by the CM algorithm. In [B6], the computed cost agreed with the optimal in only two significant figures after eighteen iterations for the steepest descent procedure. The more complicated second order technique obtained five place accuracy in the cost after five iterations.

Using the results of Section 3.5, we now consider a technique which is often effective in reducing α and the number of iterations required by the CM algorithm. In this approach, a more complicated boundary compatible set $J = \{W(t), M, N\}$ is used in the integral representation. The matrix $W(t)$ is designed to include time varying terms attempting to model the effects of the nonlinearity. For example, model $[1 - x_1^2(t)]$ as $[1 - (1-t)^2]$ and select the $W(t)$ matrix as

$$W(t) = \begin{bmatrix} 0 & 5 & 0 & 0 \\ -5 & 5\mathcal{E}[1-(1-t)^2] & 0 & -5 \\ -5 & 0 & 0 & 5 \\ 0 & -5 & -5 & -5\mathcal{E}[1-(1-t)^2] \end{bmatrix} . \qquad 6.2.20$$

However, using the equivalence relation (3.5.29), we have

$$T^{\tilde{J}}(y) = [I - U^J_{MN}]^{-1}[T^J(y) - U^J_{MN}\, y] \qquad 6.2.21$$

for the boundary compatible sets $J = \{V,M,N\}$ and $\tilde{J} = \{W(t),M,N\}$. Hence the Green's function may be calculated using the simpler set $J = \{V,M,N\}$ where V is given by (6.2.8) and $P(t)$ is calculated using the D matrix (6.2.12). We previously found that with V given by (6.2.8), the conditions of the contraction mappings theorem

are satisfied for $\mathcal{E} \leq 0.034$. A similar analysis is now done for $\tilde{J} = \{W(t),M,N\}$.

The vectors z_0 and z are given as

$$z_0 = \sup_t \left\{ \begin{bmatrix} 0 \\ |5x_2(t)_0[1-x_1^2(t)_0) - (1-(1-t)^2)]| \\ |10x_1(t)_0x_2(t)_0p_2(t)_0| \\ |5p_2(t)_0[1-x_1^2(t)_0) - (1-(1-t)^2)]| \end{bmatrix} \right\} \mathcal{E} \qquad 6.2.22$$

and

$$z = \sup_{y \in \bar{S}} \left\{ \begin{bmatrix} 0 \\ |2x_1(t)x_2(t)| + |(1-x_1^2(t)) - (1-(1-t)^2)| \\ |2x_2(t)p_2(t)| + |2x_1(t)p_2(t)| + |2x_1(t)x_2(t)| \\ |2x_1(t)p_2(t)| + |(1-x_1^2(t)) - (1-(1-t)^2)| \end{bmatrix} \right\} \mathcal{E} \qquad 6.2.23$$

Now using $\tilde{J} = \{W(t), M, N\}$ with $y_0(t) = H^{\tilde{J}}(t)c$, $r = 0.1$, we find following Example 3.5.25 and (6.2.22), (6.2.23) that conservative values for the convergence parameters are

$$\eta = \sup_t \{P(t)z_0\} = 0.64\mathcal{E}$$

and

$$\alpha = \sup_t \{P(t)z\} = 5.0\mathcal{E}$$

The requirements of Theorem 3.4.14 are then

$$5.0\mathcal{E} < 1 \qquad 6.2.24$$

and

$$\frac{0.64\mathcal{E}}{1-5.0\mathcal{E}} \leq 0.1. \qquad 6.2.25$$

Analysis of (6.2.24), (6.2.25) shows that the convergence conditions of the theorem are satisfied for

$$\mathcal{E} < 0.092 \qquad 6.2.26$$

a three fold increase over the previous value. These results are guidelines, but indicate the improvement due to use of the better designed, though more complicated, W(t) matrix.

Using the boundary compatible set $\tilde{J}$ = {W(t),M,N} for $\mathcal{E}$ = 1.0, a conservative value for α is α = 5.1, an improvement over (6.2.14), but again violating the theoretical specifications. However, the practical application of the CM algorithm reduced the convergence norm to 10^{-5} in eight iterations, a significant improvement over the algorithm using J = {V,M,N}. The iterative sequence for the control function is shown in Figure 6.3.

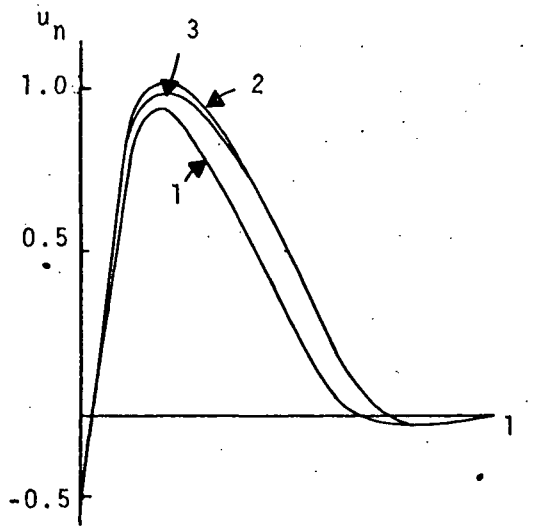


Figure 6.3 Control Iterations
(Numbers indicate iteration sequence.)

A comparison of the convergence behavior for J and $\tilde{J}$ is shown in Figure 6.4.

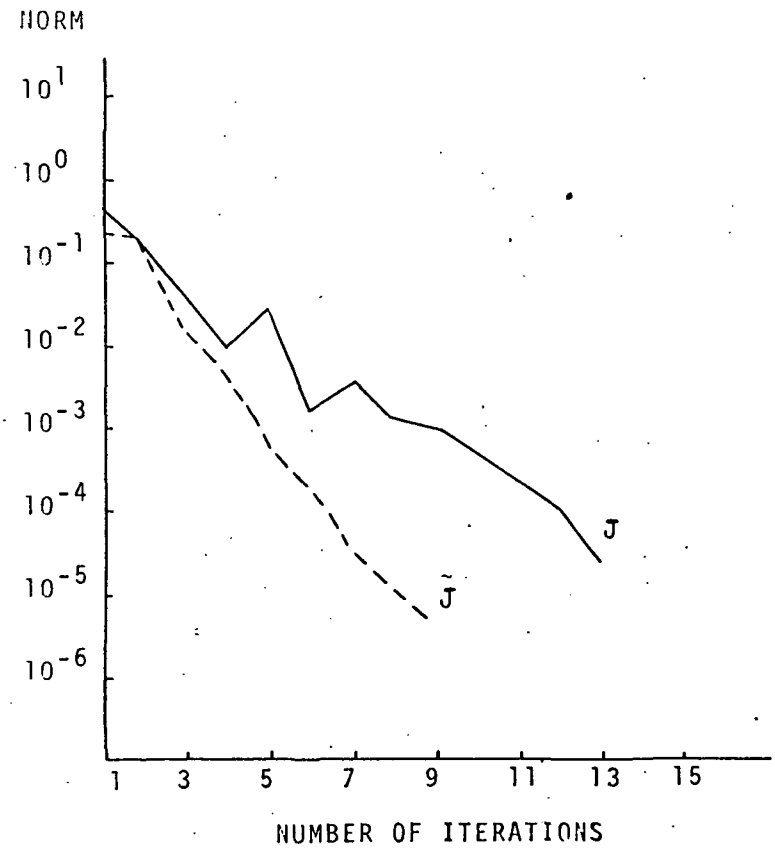


Figure 6.4 Comparison of Performance for Contraction Mappings and Modified Contraction Mapptions.

6.3. Null Controllability with Bounded Control

The first example of system null controllability involves a simple linear system with bounded input control. The example is included primarily as an introduction to the techniques of dealing with a bounded control. Consider the system

$$\dot{x} = Ax + Bu \qquad 6.3.1$$

where

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{6.3.2}$$

and the control magnitude is constrained to satisfy

$$|u(t)| \leq 1 \quad , \quad 0 \leq t \leq T \ . \tag{6.3.3}$$

The initial conditions are

$$x_1(0) = 1 \ , \ x_2(0) = 1 \ ,$$

and the final state of the system is required to be the origin, i.e.,

$$x_1(t) = 0 \ , \ x_2(T) = 0 \ , \tag{6.3.5}$$

where T is a prescribed fixed terminal time.

The linear system (6.3.1) is clearly controllable since rank [B,AB] = 2. However there do exist combinations of T and x_0 such that the system cannot be driven to the origin by the bounded control in time T. We shall investigate the null controllability of this system by considering the optimization problem composed of the system (6.4.1), the cost functional

$$J = \frac{1}{2} \int_0^T u^2(t)dt, \tag{6.3.6}$$

and the boundary conditions (6.3.4),(6.3.5).

Analytical investigation of this optimization problem yields the information that the minimum time required for the system to be driven from (1,1) to the origin is $1 + \sqrt{6}$, and at this value, the H-minimal control is bang-bang. As T is increased from $1 + \sqrt{6}$, the optimal control becomes a

saturating function, and when T is sufficiently great, the H-minimal control never saturates, i.e., it never takes on its maximum allowable magnitude. These points concerning null controllability are now illustrated by applying contraction mappings to the TPBVP associated with the posed optimization problem.

Application of the minimum principle and a change of time variable transforms the optimization problem into the TPBVP

$$\begin{aligned} \dot{x}_1 &= ax_2 \\ \dot{x}_2 &= -a\ \mathrm{SAT}\{p_2\} \\ \dot{p}_1 &= 0 \\ \dot{p}_2 &= ap_1 \end{aligned} \qquad 6.3.7$$

with boundary conditions

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \\ p_1(0) \\ p_2(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(1) \\ x_2(1) \\ p_1(1) \\ p_2(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad 6.3.8$$

or

$$\begin{aligned} \dot{y} &= f(y) \\ My(0) + Ny(1) &= c \end{aligned} \qquad 6.3.9$$

where SAT(·) is defined in (2.2.17), and where the time variable has been changed so that t = as where s $\in$ [0,1] and a = T. [(·) now indicates differentiation with respect to s.] We shall consider the case with a = 5.0.

Using the boundary compatible set J = {V,M,N}, the solution to (6.3.9), if it exists, may be written as

$$y(t) = H^J(t)c + \int_0^1 G^J(t,s)\{f(y(s)) - V(s)\ y(s)\}ds\ . \qquad 6.3.10$$

From Corollary 5.2.18, the V matrix given by

$$V = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix} \qquad 6.3.11$$

is boundary compatible with M and N given in (6.3.8). Using V specified in (6.3.11), (6.3.10) may be written explicitly as

$$y(t)=T^J(y)(t)=H^J(t)c + \int_0^1 G^J(t,s) \begin{bmatrix} 0 \\ ap_2(s)-aSAT\{p_2(s)\} \\ 0 \\ 0 \end{bmatrix} ds \ . \qquad 6.3.12$$

We shall now investigate the convergence conditions for the contraction mappings algorithm when applied to this non-differentiable TPBVP. Instead of deriving conditions satisfied by the Frechet derivative, we shall be concerned rather with conditions on the Lipschitz norm of the operator T^J. The initial estimate of the solution and the center of the region $\bar{S}(y_0,r)$ is taken to be $H^J(t)c$. To complete the definition of $\bar{S}$, the radius is set as r=0.2. This region is illustrated in Figure 6.5.

Values for $\|T^J(y_0)-y_0\|$ and the Lipschitz norm $\|T^J\|_{\bar{S}}$ must now be calculated. From (6.3.12), it is seen that the nonlinearity is contained in only the second component of the forcing function. Hence we shall investigate the Lipschitz norm of the operators

$$T_i^J(u) = \int_0^1 G_{i2}^J (t,s)[au(s) - aSAT\{u(s)\}]ds \qquad 6.3.13$$

where G_{i2}^J is the i,2 element of the Green's matrix $G^J(t,s)$. The Lipschitz norm is formally defined as

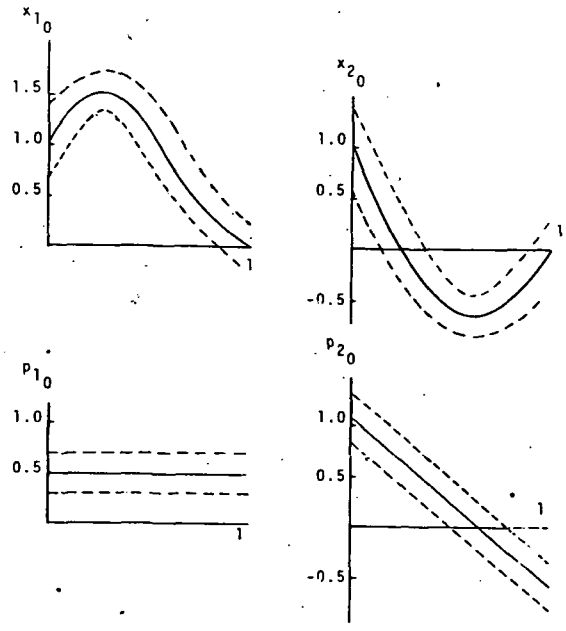


Figure 6.5 The Sphere $\bar{S}(y_0,r)$

$$\| T_i^J \|_{\bar{S}} = \sup_{\substack{u,v \in \bar{S} \\ u \neq v}} \left\{ \frac{\|T_i(u) - T_i(v)\|}{\|u - v\|} \right\} \qquad 6.3.14$$

and for the operator in (6.4.13) ,

$$\frac{\|T_i(u)-T_i(v)\|}{\|u-v\|} = \frac{\left\| \int_0^1 G_{i2}^J (t,s)[a(u(s)-v(s))+aSAT\{v(s)\}-aSAT\{u(s)\}]ds \right\|}{\|u(\cdot) - v(\cdot)\|} \qquad 6.3.15$$

or

$$\frac{\|T_i(u)-T_i(v)\|}{\|u - v\|} \leq \frac{a \sup_t \int_0^1 |G^J_{i2}(t,s)\{[u(s)-v(s)]+[SAT\{v(s)\}-SAT\{u(s)\}]\}|ds}{\sup_\rho |u(\rho)-v(\rho)|} \qquad 6.3.16$$

Now noting

$$\sup_\rho |u(\rho)-v(\rho)| \geq |u(s)-v(s)| \qquad 6.3.17$$

and

$$\sup_\rho |u(\rho)-v(\rho)| \geq |SAT\{u(s)\}-SAT\{v(s)\}| , \qquad 6.3.18$$

we have

$$\frac{\|T_i(u)-T_i(v)\|}{\|u - v\|} \leq a \sup_t \int_0^1 |G^J_{i2}(t,s) \left\{ \frac{u(s)-v(s)}{|u(s)-v(s)|} + \frac{SAT\{v(s)\}-SAT\{u(s)\}}{|u(s)-v(s)|} \right\}|ds \qquad 6.3.19$$

It may be shown that with $r = 0.2$

$$\left\{ \frac{u(s)-v(s)}{|u(s)-v(s)|} + \frac{SAT\{v(s)\}-SAT\{u(s)\}}{|u(s)-v(s)|} \right\} \leq 2 . \qquad 6.3.20$$

The required Lipschitz norm may now be evaluated in several ways, each of varying degrees of accuracy. We now consider one of the more accurate techniques.

Note that as a result of the choice of $\bar{S}(y_0,r)$, saturation can only occur for $0 \leq s \leq 0.25$. We may now write

$$\frac{\|T_i(u)-T_i(v)\|}{\|u - v\|} \leq 2\, a \sup_t \left\{ \int_0^{.25} |G^J_{i2}(t,s)|ds \right\} \qquad 6.3.21$$

which may be approximated as

$$\|T^J\|_{\bar{S}} \leq \sup_i \sup_t \{2a(.25)P_{i2}(t)\} = 5/16 < 1 \qquad 6.3.22$$

where

$$P_{i2}(t) = \int_0^t |G^J_{I_{i,2}}(t,s)|ds + \int_t^1 |G^J_{II_{i,2}}(t,s)|ds \ . \qquad 6.3.23$$

Now determining $\|T^J(y_0) - y_0\|$, we have

$$T^J_i(y_0) - y_{0i} = \int_0^1 G^J_{i2}(t,s)[ap_2(s)_0 - aSAT\{p_2(s)_0\}]ds \ . \qquad 6.3.24$$

From $y_o(\cdot)$

$$\sup_t\{|ap_2(t)_0 - aSAT\{p_2(t)_0\}|\} \leq 0.04a \ , \qquad 6.3.25$$

and then

$$\|T^J(y_0)-y_0\| \leq \sup_i \sup_t\{0.04a \int_0^{.25} |G_{i2}(t,s)|ds\} \qquad 6.3.26$$

$$\leq (0.04a)(0.25)\sup_i \sup_t \{P_{i2}(t)\} = \frac{1}{60} \ . \qquad 6.3.27$$

Taking conservative values $\alpha = 5/16$, $\eta = 1/60$,

$$\frac{\eta}{1-\alpha} = 0.1 < r = 0.2$$

so that the theoretical application of contraction mappings is successful. Hence a solution exists to the TPBVP and a control exists to accomplish the desired transfer. These concepts will now be applied to a nonlinear system.

6.4. Controllability of Satellite Pitch Motion

The pitch motion of a satellite in circular orbit can be described by the normalized differential equation

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -\sin x_1 + u \end{bmatrix} \qquad 6.4.1$$

whenever a principal axis of the satellite remains normal to the orbit plane [R1]. The controlling torque u(t) is bounded ($|u(t)| \leq 1$) and $x_1(t)$ is twice the pitch coordinate. In investigating the null controllability of the system, our goal is to find an acceptable control u(t) which zeros the pitch and pitch rate in a prescribed fixed time T .

In a neighborhood of the origin, system (6.4.1) behaves as

$$\dot{x} = Ax + Bu \qquad 6.4.2$$

where

$$A = (\partial f/\partial x)(0,0) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad B = (\partial f/\partial u)(0,0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad 6.4.3$$

For the linear system (6.4.2), rank [B,AB] = 2, and from Theorem 5.3.3, the nonlinear system (6.4.1) is controllable in a region of the origin. As in the previous example, null controllability is investigated by considering the optimization problem consisting of the system (6.4.1), the specified boundary conditions, and the cost functional

$$J = \frac{1}{2} \int_0^T u^2(t)dt \qquad 6.4.4$$

Application of the minimum principle and a change of time variable transforms the optimization problem into a TPBVP of the form

$$\dot{y} = Sy + f(y) \qquad 6.4.5$$

$$My(0) + Ny(1) = c$$

where

$$S = \begin{bmatrix} 0 & a & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -a & 0 \end{bmatrix} \qquad f(y) = \begin{bmatrix} 0 \\ -a \sin x_1 - aSAT\{p_2\} \\ ap_2 \cos x_1 \\ 0 \end{bmatrix} \qquad 6.4.6$$

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad c = \begin{bmatrix} x_{1_0} \\ x_{2_0} \\ 0 \\ 0 \end{bmatrix} \qquad 6.4.7$$

where the differentiation is now with respect to s where $t = as$, $s \in [0,1]$.

If a solution to (6.4.5) exists, it may be represented as

$$y(t) = H^J(t)c + \int_0^1 G^J(t,s)\{Sy(s) + f(y(s)) - V(s)y(s)\}ds \qquad 6.4.8$$

where $J = \{V(t), M, N\}$ is a boundary compatible set. Since the linear system (6.4.2) is controllable, Corollary 5.2.18 states that the $2n \times 2n$ V matrix given as

$$V = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix} \qquad 6.4.9$$

is boundary compatible with M and N given by (6.4.7). Choosing V from (6.4.9) yields

$$V = \begin{bmatrix} 0 & a & 0 & 0 \\ -a & 0 & 0 & -a \\ 0 & 0 & 0 & a \\ 0 & 0 & -a & 0 \end{bmatrix} \qquad 6.4.10$$

We shall now investigate the null controllability of the system for various initial conditions and time intervals.

Example 6.4.11.

Consider the initial condition for the system to be 60° for the pitch angle and zero for the pitch rate. It is desired to regulate the system to zero in one half period, i.e., $T = \pi$. The initial estimate of the solution and the center of the sphere $\bar{S}$ is taken to be $H^J(t)c$. The region $\bar{S}(y_0,r)$ is defined by setting $r = 0.1$ and is illustrated in Figure 6.6.

It is seen that for this $\bar{S}(y_0, r)$ that $|p_2|$ is always less than one so that saturation never occurs. Hence the forcing function for (6.4.8) may be considered as

$$F(y) = Sy + f(y) - Vy = \begin{bmatrix} 0 \\ -a(\sin x_1 - x_1) \\ a(\cos x_1 - 1)p_2 \\ 0 \end{bmatrix}$$

Estimates for the convergence parameters are calculated using the variables defined as

$$P(t) = \int_0^1 |G^J(t,s)|ds \qquad 6.4.13$$

$$z_{0_i} = \sup_s \{|F_i(y_0(s))|\} \qquad 6.4.14$$

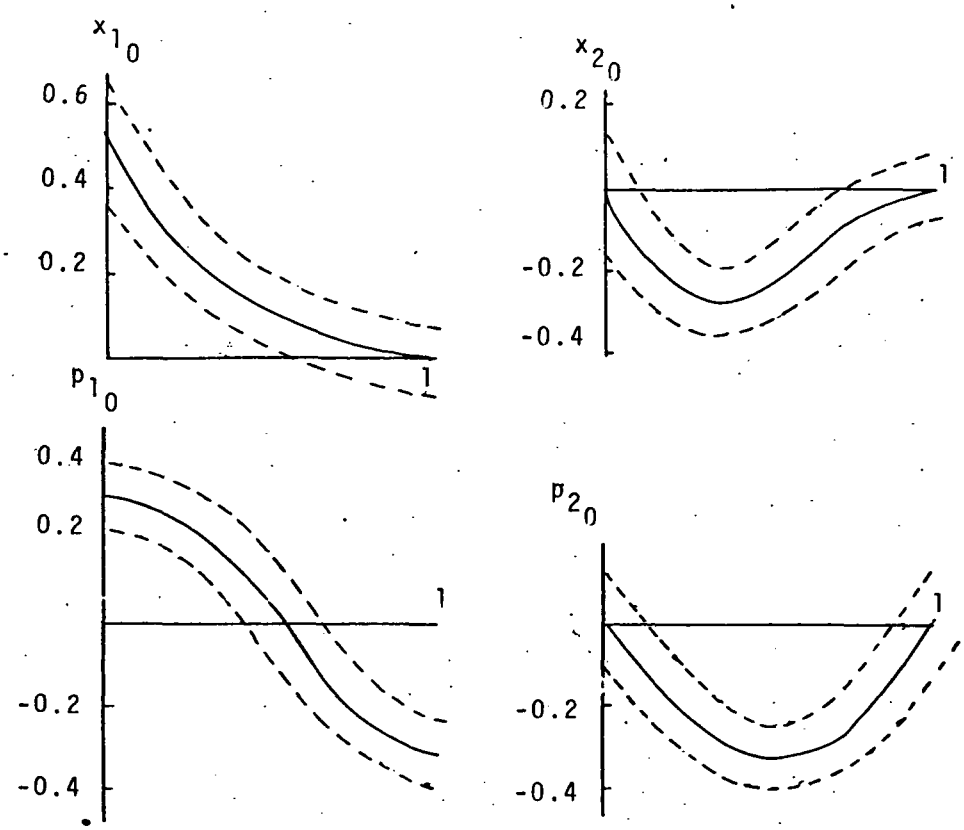


Figure 6.6 The Sphere $\bar{S}(y_0,r)$ for $T = \pi$.

and

$$z_i = \sup_{y \in \bar{S}} \left\{ \sum_{j=1}^{p} \left| \left(\frac{\partial F_i}{\partial y_j} \right) (y(s)) \right| \right\} \qquad 6.4.15$$

where $F(y)$ is given by (6.4.12). The matrix V given by (6.4.10) may be transformed into the canonical matrix D given as

$$D = \begin{bmatrix} \sigma_1 & \omega_1 & 0 & 0 \\ -\omega_1 & \sigma_1 & 0 & 0 \\ 0 & 0 & \sigma_2 & \omega_2 \\ 0 & 0 & -\omega_2 & \sigma_2 \end{bmatrix} \qquad 6.4.16$$

where $\sigma_1 = 0$, $\omega_1 = a$, $\sigma_2 = 0.05$, $\omega_2 = a$. From (6.4.14), the matrix P(t) may then be obtained by integration of the expression

$$P(t) = \int_0^t |\{\Lambda\Phi^D(t,0)\Lambda^{-1}\}\{[M+N\Phi^V(1,0)]^{-1}M\}\{\Lambda\Phi^D(0,s)\Lambda^{-1}\}|ds$$
$$+ \int_t^1 |\{\Lambda\Phi^D(t,0)\Lambda^{-1}\}\{[M+N\Phi^V(1,0)]^{-1}N\Phi^V(1,0)\}\{\Lambda\Phi^D(0,s)\Lambda^{-1}\}|ds$$

6.4.17

Using (6.4.12), (6.4.15), and taking the supremum over $\bar{S}$ yields

$$z = \sup_{y \in \bar{S}} \begin{bmatrix} 0 \\ |a(\cos x_1 - 1)| \\ |ap_2 \sin x_1| + |a(\cos x_1 - 1)| \\ 0 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.418 \\ 0.653 \\ 0.0 \end{bmatrix} \qquad 6.4.18$$

The vector z_0 is found from (6.4.12) and (6.4.14) as

$$z_0 = \begin{bmatrix} 0.0 \\ 0.095 \\ 0.028 \\ 0.0 \end{bmatrix} \qquad 6.4.19$$

Using (6.4.17), (6.4.18), and (6.4.19), conservative estimates for the convergence parameters η and α are found as

$$\eta = \sup_t \{P(t)z_0\} = 0.047 \qquad 6.4.20$$

$$\alpha = \sup \{P(t)z\} = 0.38 \qquad 6.4.21$$

Now testing $\eta/(1-\alpha) \leq r$, we have

$$\frac{\eta}{1-\alpha} = 0.075 < r = 0.1 \ . \qquad 6.4.22$$

Hence the convergence conditions of the contraction mappings theorem are satisfied and the theoretical application of contraction mappings is successful. Moreover, a solution exists to the TPBVP and a control exists to accomplish the desired transfer.

Example 6.4.23

Consider the initial condition for the system to be 60° for the pitch angle and zero for the pitch rate. It is desired to regulate the system to zero in one quarter period, i.e., $T = \pi/2$. The initial estimate of the solution and the center of the sphere $\bar{S}$ is taken to be $H^J(t)c$. The region $\bar{S}(y_0,r)$ is defined by setting $r = 0.1$. This region is illustrated in Figure 6.7.

It is seen that $\bar{S}(y_0,r)$ contains a saturating region for p_2. Hence the forcing function for (6.4.8) must be considered as

$$F(y) = Sy + f(y) - Vy = \begin{bmatrix} 0 \\ -a(\sin x_1 - x_1) - a(SAT\{p_2\} - p_2) \\ ap_2(\cos x_1 - 1) \\ 0 \end{bmatrix} \qquad 6.4.24$$

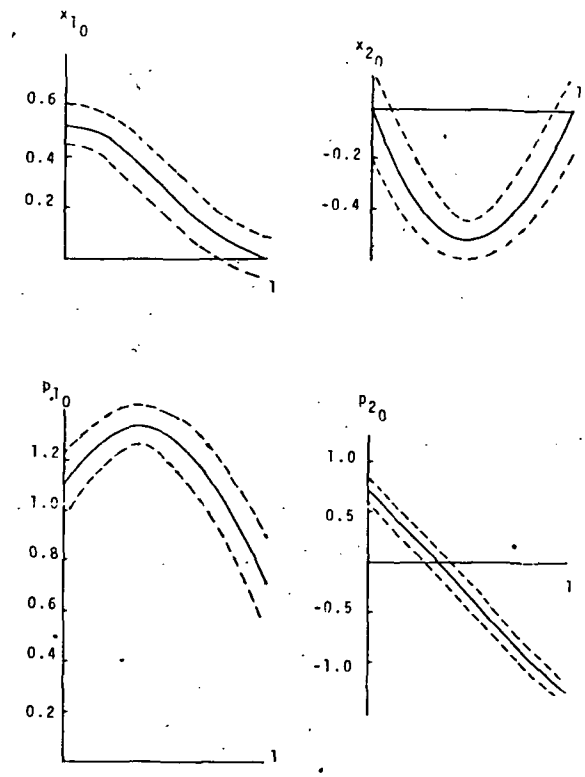


Figure 6.7 The Sphere $\bar{S}(y_0,r)$ for $T = \pi/2$.

As in Section 6.3, the Lipschitz norm of the operator $T^J(y)$, not the Frechet derivative, must be investigated. For $\bar{S}(y_0,r)$, the Lipschitz condition for $F(y)$ is given as

$$|F(y)-F(y')| \leq \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.39 & 0 & 0 & 3.14 \\ 0.75 & 0 & 0 & 0.34 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} |x_1-x'_1| \\ |x_2-x'_2| \\ |p_1-p'_1| \\ |p_2-p'_2| \end{bmatrix} \qquad 6.4.25$$

or

$$\|F(y) - F(y')\| \leq C \|y - y'\| . \qquad 6.4.26$$

Using P(t) from (6.4.17) and defining the 2n vector z as composed of the elements

$$z_i = \sum_{j=1}^{2n} [C_{ij}] , \qquad 6.4.27$$

a conservative estimate for α is $\|P(\cdot)z\|$. However, because the saturation occurs only over a short interval of $\bar{S}(y_0, r)$, this estimate would tend to be quite inaccurate. Hence we deal with the saturation effect separately. Now let

$$z_1 = \begin{bmatrix} 0 \\ 0.34 \\ 1.14 \\ 0 \end{bmatrix} \quad \text{and} \quad z_2 = \begin{bmatrix} 0 \\ 3.14 \\ 0 \\ 0 \end{bmatrix} \qquad 6.4.28$$

where z_1 arises from the differentiable part and z_2 from the saturating effect. Then as in Section 6.3 ,

$$\alpha = \sup_t \{P(t)z_1\} + \sup_t \left\{ \int_{0.85}^{1.0} |G^J(t,s)z_2| ds \right\} \qquad 6.4.29$$

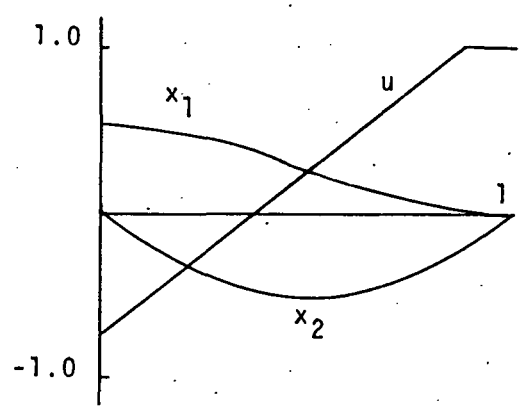
or

$$\alpha = \sup_t \{P(t)z_1\} + \sup_t \{(3.14)(0.15)P_{i2}(t)\} . \qquad 6.4.30$$

Using the values of P(t) from (6.4.17), α is evaluated as

$$\alpha = 1.56 . \qquad 6.4.31$$

Hence the requirement that $\alpha < 1.0$ is violated, and convergence is not guaranteed. However, as indicated previously, these coarse estimates are used as guidelines for the practical application of contraction mappings. Indeed, the CM algorithm reduced the convergence norm to 10^{-5} in ten iterations. Figure 6.8 illustrates the state and control history.



6.8 State and Control History

Page Intentionally Left Blank

CHAPTER 7

PRELIMINARY STUDY ON THE DYNAMICS OF DRUG USAGE WITHIN A COMMUNITY

7.1. Introduction

The modeling of complex socio-economic systems has recently received considerable attention. Arising initially as an aid to management decision making, [F7], [R2], system modeling is now applied to many systems of public concern [F5], [R3]. The primary objective of the modeling effort is the formulation of improved administrative control policies. Typically, once a model is developed, the process of designing improved policies is largely a trial and error process. That is, the behavior of the system is first simulated with the model using one control policy and then another. The simulation results are then compared to determine which policy yielded the "best" behavior, clearly an inexact and inefficient technique of analysis.

In this chapter we consider the feasibility of applying the systematic techniques of optimal control theory to the determination of policies for social systems. Specifically, a dynamic model attempting to represent the causal, feedback structure of community drug usage is developed. Then using optimization theory, we attempt to gain insight into how a community might best respond to a rapidly growing heroin addiction problem. The initial phase of the study is the creation of a dynamic model which reflects the modes of behavior of the system being investigated.

7.2. Development of a Dynamic Model

The development of a model for a complex system such as drug usage is in itself a major effort. The subtle interrelationships and multi-feedback loops are often difficult to conceptualize. Likewise, the determination of various parameters within the model is a difficult task involving much data analysis. The main thrust of this chapter is not in the modeling direction. Rather, we shall develop a simple model which hopefully reflects inpart the basic behavior of a very complex system. Similarly, parameter values are chosen after discussions and readings and are believed to be reasonable. In this spirit, the development of the model is begun. (For the development of a more comprehensive model, see Roberts [R3]).

The model concerns itself with three groups of people within the community. These groups represent the three levels of drug usage which will be considered in the model. These three pools of people are:

i) potential drug users

ii) drug users

iii) heroin addicts

Of course, much finer lines may be drawn, but these three are sufficient for this study. The dynamical nature of this problem is reflected in the constantly changing population of each level and the inherent relationships between these changes. The multiple interrelationships are often difficult to conceptualize, but are critical to the feedback, multiloop structure of the system. Figure 7.1 represents how one might initially conceive this system as simply involving transitions of people from various stages of drug usage.

In Figure 7.1, the double lines represent the flow of people between levels and the values controlling these flows are determined by the variables alongside.

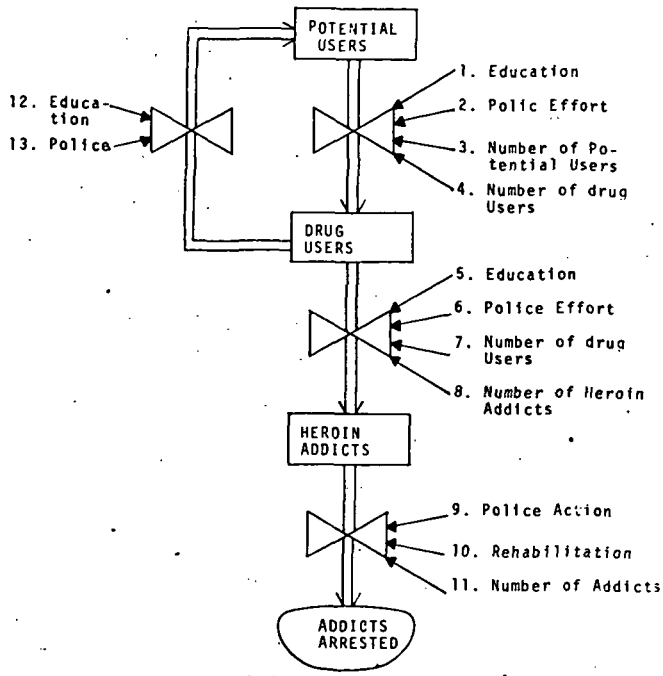


Figure 7.1 Levels of Drug Usage

However, upon recognizing the feedback structure of the system, Figure 7.2 is a more accurate representation.

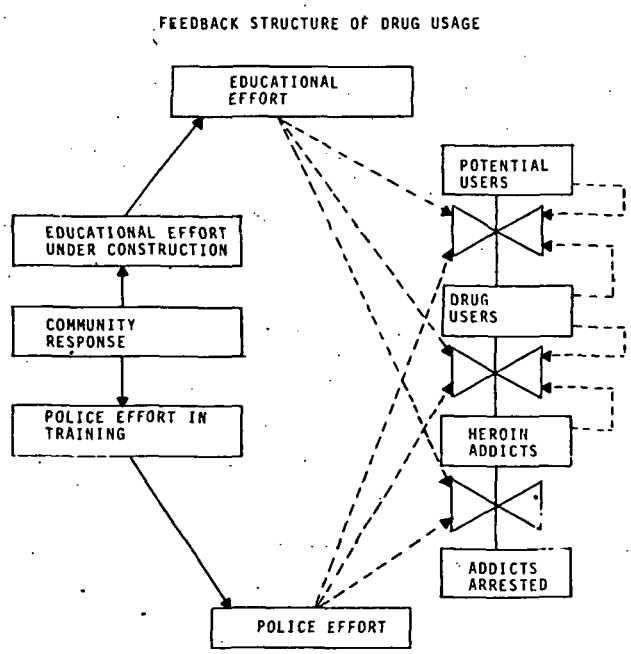


Figure 7.2 Feedback Structure of Drug Usage

Note that in reality there are feedback paths from the drug users level and the heroin addicts level back to the community response. However, we shall be attempting to determine how a community might best respond rather than modeling the present reaction.

Let us first consider the flow between potential drug users and drug users. In this study, "Potential Users" will represent the community population between the ages of ten and thirty who are not using drugs illegally. The next level of drug usage, "Drug Users," represents that group of people who occasionally participate in the illegal use of drugs, but who are not addicted to heroin. The flow between these levels is determined by the drug education program, the police effort, the number of potential users, and the number of drug users. Of these four variables, the number of drug users might be considered the dominant. This is simply due to the fact that the users tend to share their supply, turn-on their friends, and in general, tend to increase their numbers. The level of the drug education program and the fear of arrest may tend to deter some potential users, but these are not the dominant effects. The flow rate from potential users to users depends on the number of potential users in the sense of availability, i.e., if there are few potential users remaining, the inflow into drug users will wither, and, conversely, if there are many potential users, the self induced growth rate of drug usage is unimpeded. Some drug users revert back to potential users through the efforts of police and education, but this is considered a minor effect.

The flow from drug users to addicts is of the same form as the flow from potential users to users. Again, education and police effort tend to deter the flow and a self growth rate is again present via the number of addicts. The flow

from Drug Users to Heroin Addicts simply reflects the fact that most addicts previously used "soft" drugs; it is not a causal indication. Addicts are removed from the street primarily by police action which is a result of the community response to the number of addicts manifested by the rising crime rate. The drug education program and police effort are created by community spending for these programs. In this model, the community spending for police and education are considered as the two control variables to be determined.

For simulation and optimization studies, the general description of the model must be transformed into a system of equations characterizing the dynamics of the system. A convenient procedure for developing equations describing the dynamics of a general system is the DYNAMO format [P3]. Developed by the Industrial Dynamics Group at the Sloan School, M.I.T., DYNAMO is both a simulation language and a discrete equation representation for the system dynamics. We now develop the DYNAMO equations which describe the dynamics of the drug usage model.

As indicated in the general description of the system, the number of drug users determines the nominal growth rate of drug usage, i.e., the "recruitment" rate. This is represented as

$$NGRU.K = \left(\frac{1}{AODC}\right) DU.K \qquad 7.2.1$$

where

NGRU Nominal Growth Rate of Drug Usage $\left(\frac{men}{month}\right)$

.K a postscript indicating that NGRU.K refers to nominal growth rate at the present time K

DU Drug Users

AODC a constant determining the growth rate.

The availability of potential users is included as a multiplier of the nominal growth rate and is a function of the difference between the initial number of potential users and the present number of drug users. The nonlinear relationship has the general form illustrated in Figure 7.3 where APUM is the availability of potential users multiplier and IPU is the initial number of potential users.

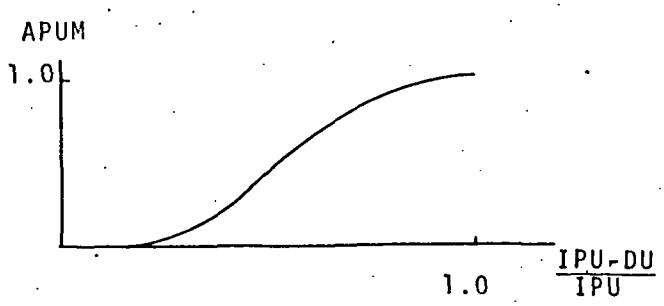


Figure 7.3 Availability of Potential Users Multiplier

The total flow from potential users to drug users is then given as

$$GRU.KL = (APUM.K)(NGRU.K) \qquad 7.2.2$$

where

GRU Growth Rate of Usage $(\frac{men}{month})$

.KL postscript indicating that GRU.KL refers to the rate of growth of drug usage during the time increment from K to L.

The nominal growth rate of addiction is determined by the number of addicts as

$$NGRA.K = (\frac{1}{AOD})\ AD.K \qquad 7.2.3$$

where

NGRA Nominal Growth Rate of Addiction $(\frac{men}{month})$

AD Addicts (men)

AOD Constant determining the growth rate.

The number of drug users influences the growth rate of addiction as an availability multiplier of the form illustrated in Figure 7.4 where ADUM is the availability of drug users multiplier.

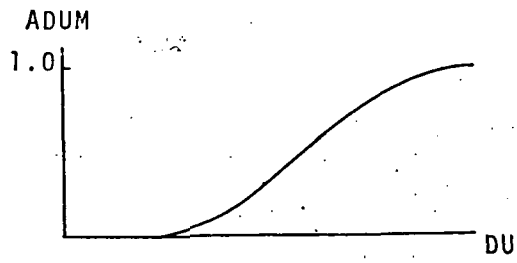


Figure 7.4 Availability of Drug Users Multiplier

The drug education level acts to deter the flow rate and is included as a multiplier which decreases with increasing education effort. The form of the function is illustrated in Figure 7.5 where AEDM is the addiction education multiplier.

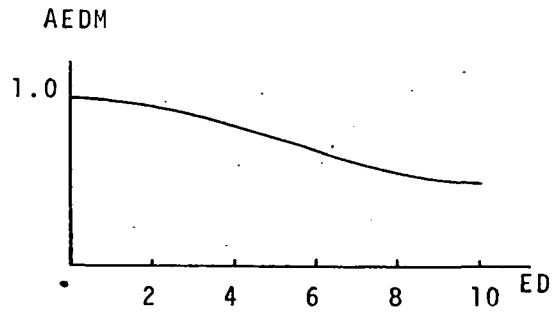


Figure 7.5 Effect of Education Upon Addiction Growth Rate

The total flow from Drug Users to Addicts is the growth rate of the addiction level and is given as

$$GRA.KL = (AEDM.K)(ADUM.K)(NGRA.K). \qquad 7.2.4$$

The population of the drug usage level is then given by

$$DU.K = DU.J + (DT)(GRU.JK - GRA.JK) \qquad 7.2.5$$

where GRU is the growth rate of drug usage and GRA is the growth rate of addiction, i.e., the flow rate from drug usage. DT is delta time, the discrete time increment.

The removal rate of addicts depends on the number of police, the effectiveness of police action, and the number of addicts. If it is assumed that each policeman

arrests a certain number of addicts per month, the nominal removal rate is given as

$$NRRPE.K = (GAIN.K)(PE.K) \qquad 7.2.6$$

where

NRRPE Nominal Removal Rate due to Police Effort $(\frac{men}{month})$

GAIN The effectiveness of police

PE Police Effort (men).

The variable "GAIN" in (7.2.6) is not a constant because addicts are increasingly careful as police effort increases and, as a result, police effectiveness in making arrests decreases. The nonlinear form of the GAIN multiplier is shown in Figure 7.6.

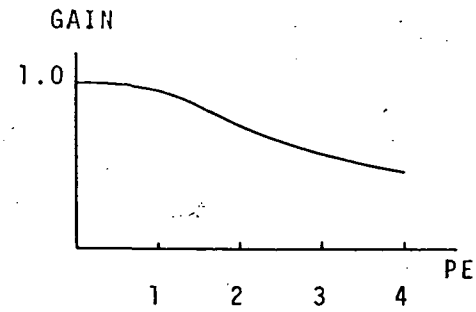


Figure 7.6 Police Effectiveness

The removal rate of addicts is also influenced by the availability of addicts to arrest. This effect is included as a multiplier which decreases with decreasing numbers of addicts, reflecting the difficulty in finding the addicts. The form

of this relationship is illustrated in Figure 7.7 where AAM is the availability of addicts multiplier.

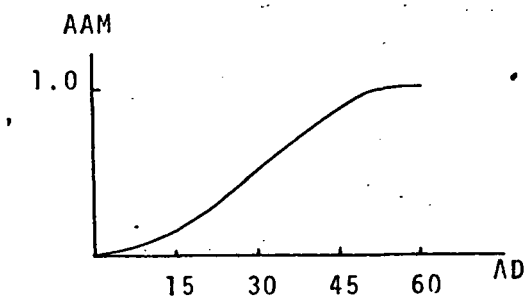


Figure 7.7 Availability of Addicts Multiplier

The total removal rate is then given as

$$RRPE.KL = (AAM.K)(GAIN.K)(NRRPE.K) \qquad 7.2.7$$

where RRPE is the removal rate due to police effort. The number of addicts is then the integration of the inflow and outflow rates, i.e.,

$$AD.K = AD.J + (DT)(GRA.JK - RRPE.JK). \qquad 7.2.8$$

Police effort and the drug education program are considered to be first order responses to community spending. In DYNAMO this is represented as

$$PE.K = PE.J + (DT)(\frac{1}{DAP})(CSPE.JK - PE.J) \qquad 7.2.9$$

$$ED.K = ED.J + (DT)(\frac{1}{DAE})(CSED.JK - ED.J) \qquad 7.2.10$$

where PE represents the Police Effort (men), DAP the Delay in Adjusting the

Police (months), CSPE the Community Spending on Police Effort (men), CSPE the Community Spending on Police Effort (men), ED the Education program (men), DAP the Delay in Adjusting the Education program (months), and CSED the Community Spending on Education (men).

This completes the development of the system equations, however a simplification is now considered. The three states "Potential Users", "Drug Users", and "Heroin Addicts" are included in the model equations. These three states modeled the changing population for three divisions of the youth population. However, in many communities, especially those in which heroin addiction is becoming a problem, the time dynamics of the first two variables have been completed. That is, the percentage of the youth population which falls into the extremely broad category "Drug Users" is relatively fixed or slowly time varying, the major growth phase being essentially complete. For these reasons, only the variable "Heroin Addicts" is included as a dynamic variable. This assumption yields the following equations describing the system:

i) **Addicts:** $AD.K = AD.J + (DT)(GRA.JK - RRPE.JK)$ 7.2.11

ii) **Police:** $PE.K + PE.J + (DT)(\frac{1}{DAP})(CSPE.JK - PE.J)$ 7.2.12

iii) **Education:** $ED.K = ED.J + (DT)(\frac{1}{DAE})(CSED.JK - ED.J).$ 7.2.13

The growth rate of addiction, GRA, is given as

$$GRA.KL = (AEDM.K)(\frac{1}{AOD})AD.K \qquad 7.2.14$$

where AEDM is the effect of drug education and AOD is the nominal growth rate factor of addiction. The removal rate of addicts due to police effort is given

$$RRPE.KL = (AAM.K)(GAIN.K)PE.K \qquad 7.2.15$$

where AAM is the availability of addicts multiplier and GAIN is the effectiveness of police effort.

These discrete representations may easily be transformed into the form of continuous differential equations as

$$\dot{x}_1 = f_1(x_1,x_3) - f_2(x_1,x_2) = f(x_1,x_2,x_3)$$

$$\dot{x}_2 = -(\frac{1}{DAP})x_2 + (\frac{1}{DAP})u_1 \qquad 7.2.16$$

$$\dot{x}_3 = -(\frac{1}{DAE})x_3 + (\frac{1}{DAE})u_2$$

where x_1 represents addicts, x_2 police effort, x_3 drug education program, u_1 community spending on police effort, u_2 community spending on drug education, (DAP) delay in adjusting police effort, and (DAE) the delay in adjusting the education program. f_1 and f_2 represent respectively the growth rate of addiction and the removal rate of addicts. This system belongs to the broad class of nonlinear systems described as

$$\dot{x} = Ax + Bu + \psi(x). \qquad 7.2.17$$

The results obtained in Chapters 2 and 3 regarding the optimal regulation of (7.2.17) will now be applied to the drug usage model.

7.3. Optimal Regulation of the Nonlinear System

The cost functional for the optimization problem is designed to regulate the number of addicts yet maintain public expenditures at a reasonable level. Consider the cost functional to be of the form

$$J = \frac{1}{2} \int_0^T [qx_1^2(t) + (CP)u_1^2(t) + (CE)u_2^2(t)]dt \qquad 7.3.1$$

where x_1 represents addicts, u_1 community spending for police, and u_2 community spending for drug education. Appropriate choices for the cost parameters q, CP, and CE must be made to obtain "acceptable" levels of x(t) and u(t). A choice that is often quite reasonable [B7] is given as

$$\begin{aligned} \frac{1}{q} &= \text{maximum allowable } (x_1)^2 \\ \left(\frac{1}{CP}\right) &= \text{maximum allowable } (u_1)^2 \\ \left(\frac{1}{CE}\right) &= \text{maximum allowable } (u_2)^2 \end{aligned} \qquad 7.3.2$$

Using (2.3.8) and (2.3.9), the necessary conditions of optimality for the optimization problem consisting of the system (7.2.1,2,3), the cost functional (7.3.1), and the initial condition $x(0) = x_0$ reduce to the TPBVP

$$\dot{y} = Sy + \psi(y) \qquad 7.3.3$$

$$My(0) + Ny(1) = c \qquad 7.3.4$$

where y is the 2n composite vector

$$y = \begin{bmatrix} x \\ \hline p \end{bmatrix}$$

$$S = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{a}{(DAP)} & 0 & 0 & -\frac{a}{(CP)(DAP)^2} & 0 \\ 0 & 0 & -\frac{a}{(DAP)} & 0 & 0 & -\frac{a}{(CE)(DAE)^2} \\ -qa & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{a}{(DAP)} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{a}{(DAE)} \end{bmatrix} \qquad 7.3.5$$

$$\psi(y) = \begin{bmatrix} af(x_1, x_2, x_3) \\ 0 \\ 0 \\ -a(\partial f/\partial x_1)(x_1,x_2,x_3)p_1 \\ -a(\partial f/\partial x_2)(x_1,x_2,x_3)p_1 \\ -a(\partial f/\partial x_3)(x_1,x_2,x_3)p_1 \end{bmatrix} \qquad 7.3.6$$

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad N = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \quad C = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$$

and where a is the change of time scale variable.

Values must now be assigned to the various system parameters. In a sense these parameters depend on the community and environment being discussed. We assume that the community of interest is neither an extremely wealthy suburb nor the extremely poor section of an inner city. We assume the community has a population of 50,000. The youth population of such a community roughly comprises 30% of the population [S3]. Since we are primarily interested in regulating the early phases of heroin usage, we assume that initially the community has a low level of heroin addiction, say one per thousand of the youth population. Communities generally have a police force composed of approximately one policeman per thousand of population, [S3]. We assume that initially the police force has no effort directed specifically at heroin. The community is also assumed to initially have no drug education program. A reasonable value for police effectiveness is one conviction per month per policeman but decreasing in a nonlinear manner as police effort increases due to increasing caution among addicts. The

The delay in adjusting the police effort is essentially a training delay and is assumed to be six months. The delay in adjusting the drug education program is assumed to be one year. The boundary compatible set J = {V,M,N} must now be chosen for the integral representation. The boundary matrices are chosen directly from (7.3.7). The V matrix is chosen in the form

$$V = \begin{bmatrix} ac & ad & ae & 0 & 0 & 0 \\ 0 & -\frac{a}{(DAP)} & 0 & 0 & -\frac{a}{(CP)(DAP)^2} & 0 \\ 0 & 0 & -\frac{a}{(DAE)} & 0 & 0 & -\frac{a}{(CE)(DAE)^2} \\ -aq & 0 & 0 & -ac & 0 & 0 \\ 0 & 0 & 0 & -ad & \frac{a}{(DAP)} & 0 \\ 0 & 0 & 0 & -ae & 0 & \frac{a}{(DAE)} \end{bmatrix} \quad 7.3.8$$

where c,d, and e may be chosen to model the nonlinearity f. The characteristic roots of this matrix are real, distinct, and readily evaluated, thus easing the determination of the P(t) matrix for convergence analysis. Numerical cases are now considered as examples.

Example 7.3.9.

In this example we consider the rather short time interval of one year. Specific values are selected for the cost parameters q, CP, CE, and the contraction mapping method is applied to the TPBVP arising from the optimal regulator problem. If it is desired to prevent addiction from growing greatly from its initial value, q may be selected as 0.04. This represents the maximum desired number of addicts as 5 in (7.3.2). If the police can allocate a maximum

nonlinear effectiveness curve is illustrated in Figure 7.8.

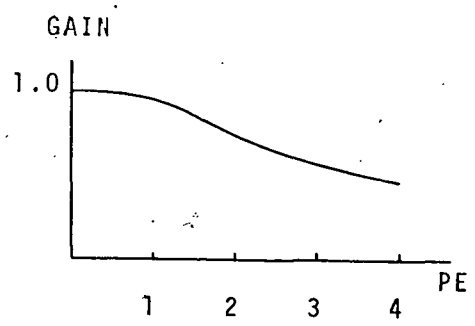


Figure 7.8 Police Effectiveness

The effectiveness of the drug education program is assumed to reduce the addiction growth rate by a maximum of 50% for a highly effective education program. The effectiveness is modeled as a function of the number of people involved in the drug education program. This is illustrated in Figure 7.9.

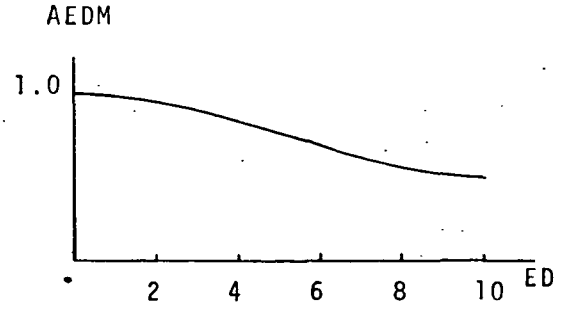


Figure 7.9 Effect of Education Upon Addiction Growth Rate

of two men to the control of addiction, CP may be chosen as 0.25. Similarly, if the school committee believes that ten teachers are sufficient for the drug education program, CE may be chosen as CE = 0.01.

The contraction mapping algorithm is begun with $H^J(t)c$; y_0, the center of $\bar{S}(y_0,r)$, is chosen as the third member of the CM sequence, and r is set as r = 0.2. The center of $\bar{S}(y_0,r)$ is shown in Figure 7.10.

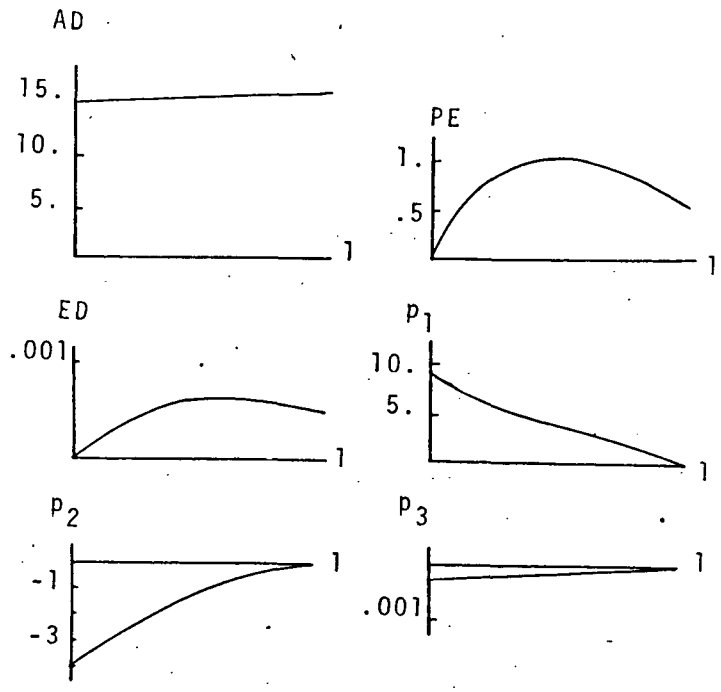


Figure 7.10 The Function $y_0(t)$ for T = 12 Months.

Now determing the convergence parameters η and α , the vectors z_0 and z composed of the elements

$$z_{0_i} = \sup_t \left\{ \left| \sum_{j=1}^{2n} s_{ij}(t) y_{0_j}(t) + \psi_i(y_0(t)) - \sum_{j=1}^{2n} v_{ij}(t) y_{0_j}(t) \right| \right\}$$

$$z_i = \sup_{\bar{S}} \left\{ \sum_{j=1}^{2n} \left| s_{ij}(t) + (\partial\psi_i/\partial y_j)(y(t)) - v_{ij}(t) \right| \right\} \qquad 7.3.10$$

are evaluated as

$$z_0 = \begin{bmatrix} 0.12 \\ 0 \\ 0 \\ 0.01 \\ 0.11 \\ 0.02 \end{bmatrix} \qquad z = \begin{bmatrix} 0.13 \\ 0 \\ 0 \\ 0.02 \\ 0.14 \\ 0.03 \end{bmatrix} \qquad 7.3.11$$

Using the distinct characteristic roots, (4.7.17) is evaluated for P(t) yielding conservative values for η and α as

$$\eta = \sup_t \{P(t)z_0\} = 0.14$$

$$\alpha = \sup_t \{P(t)z\} = 0.16 \ . \qquad 7.3.12$$

We then have

$$\frac{\eta}{1-\alpha} = 0.17 < r = 0.2 .$$

Hence the conditions of the theorem are satisfied and the theoretical application of contraction mappings is successful and convergence of the CM sequence is indicated. The practical application of the CM algorithm reduced the convergence norm to 10^{-3} in ten iterations. The time histories for the state variables addicts, police effort, and drug education are illustrated in Figure 7.11.

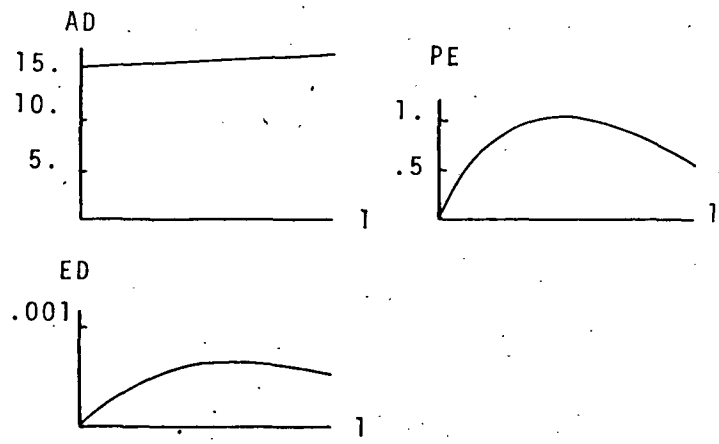


Figure 7.11 Addicts, Police, and Education for T = 12 Months

We shall delay discussing the implications of these results until the next example is presented.

Example 7.3.13.

In this example we consider longer term behavior and let the time interval of interest be four years. If it is desired to prevent addiction from growing over 20 in the four year period, q may be selected as 0.0025. If the police can

allocate only one man to the control of addiction, CP may be chosen as 1.0. Similarly, if the school committee believes that ten teachers are sufficient for the drug education program, CE may be selected as 0.01.

The contraction mapping algorithm is begun with $H^J(t)c$; y_0, the center of $\bar{S}(y_0,r)$, is chosen as the fifth member of the CM sequence; and r is set as r = 0.2. The center of $\bar{S}(y_0,r)$ is illustrated in Figure 7.12.

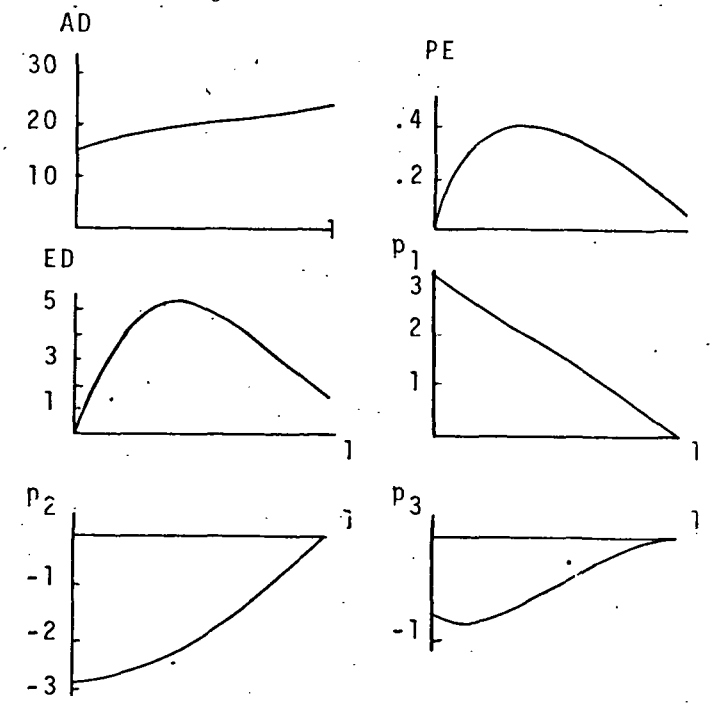


Figure 7.12 Addicts, Police, and Education for T = 12 Months

Using (7.3.10) and $\bar{S}(y_0,r)$, the vectors z_0 and z are evaluated as

$$z_0 = \begin{bmatrix} 0.55 \\ 0 \\ 0 \\ 0.06 \\ 0.52 \\ 0.09 \end{bmatrix} \qquad z = \begin{bmatrix} 0.59 \\ 0 \\ 0 \\ 0.08 \\ 0.54 \\ 0.11 \end{bmatrix} \qquad 7.3.14$$

Using a = 48, (4.7.17) is evaluated for P(t) yielding conservative values for η and α as

$$\eta = \sup_t \{P(t)z_0\} = 0.62$$
$$\alpha = \sup_t \{P(t)z\} = 0.73 \qquad 7.3.15$$

We have $\alpha < 1.0$, however

$$\frac{\eta}{1-\alpha} = 2.5 > r = 0.2 \qquad 7.3.16$$

so the theoretical application of contraction mappings does not guarantee convergence. However, these results are only guidelines for the practical application of the CM algorithm. In fact, the CM algorithm reduced the convergence norm to 10^{-3} in twelve iterations. The time histories for the state variables addicts, police effort, and drug education are illustrated in Figure 7.13.

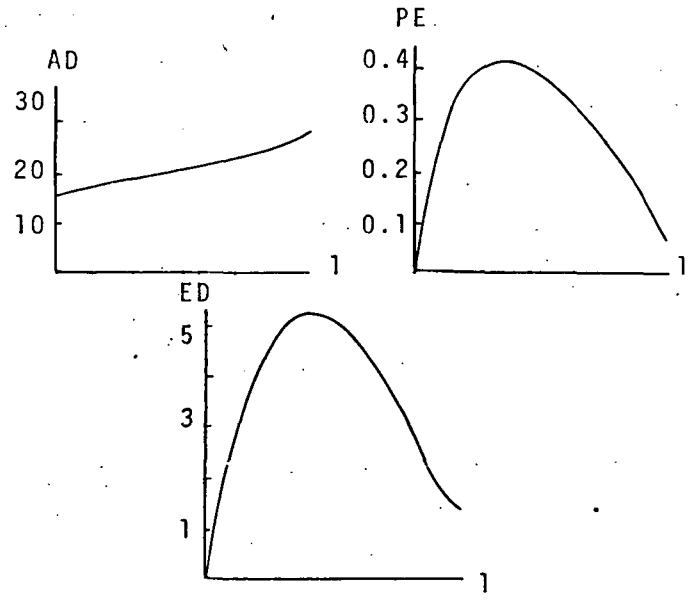


Figure 7.13 Addicts, Police, and Education for T = 48 Months

Discussion of Results

Although the stated purpose of this chapter is illustrative in nature, perhaps one or two broad qualitative implicatons may be drawn from the results of the two sample cases. First, the need for prompt action is clearly indicated. With addiction growing at an exponential rate, any delay in dealing with the problem is critical. In both examples, the police effort with a short reaction time is used to begin removing the addiction core as quickly as possible. In the first example, it is seen that the controller responds to the short term situation with basically only a police effort. This is primarily due to the fact that the controller does not have the time to establish a viable drug education program. The second example is a longer term situation and the control response is seen to be reasonably balanced, i.e., the optimal regulator responds with both police effort and an education program. Again the police effort is the first to be utilized, but the education program is brought into play as quickly as possible and tends to deter long term growth. The drawing of quantitative conclusions from these examples would be of dubious value. However, the chapter illustrates that system modeling and optimal control theory may be jointly utilized to obtain information and insight into policy formulation for complex systems. Moreover, the chapter demonstrates that contraction mappings is a useful concept and tool for both the theoretical and practical investigation of nonlinear system control.

CHAPTER 8

SUMMARY, CONTRIBUTIONS AND RECOMMENDATIONS

8.1. Summary

In the broadest sense, the objective of this dissertation was to study the theoretical and applied aspects of contraction mappings for the solution of nonlinear control problems. This objective was achieved by considering the theoretical and practical application of contraction mappings to the particular issues of optimal regulation and controllability of nonlinear dynamical systems.

It was shown in the study that application of the Pontryagin principle to the optimal regulator problem yielded necessary conditions for optimality in the form of a two point boundary value problem. Optimal system regulation was considered for both unconstrained and bounded controls and results were derived for the optimal regulation of linear dynamical systems and several classes of nonlinear systems. By an appropriate selection of boundary conditions, it was shown that the issue of controllability for dynamical systems may also be reduced to the study of two point boundary value problems.

The representation of two point boundary value problems by an integral equation was then introduced and made it possible to consider the solution of two point boundary value problems as the solution of corresponding operator equations. The joint application of the integral representation and the implicit function theorem provided new insight into the controllability of nonlinear systems. The methods of contraction mappings and modified contraction mappings were then presented for the solution of operator equations. Convergence theorems

were presented for both methods, and translated convergence theorems were derived for those operators arising from the optimal regulation of nonlinear systems. A detailed investigation of the calculation of the theoretical convergence criteria was conducted. Upper bounds were presented for the Lipschitz norm and derivative norm, and various techniques for evaluating these bounds were introduced. In particular, the use of simply structured matrices and similarity transformations were considered. The use of partitioned matrices in these developments provide considerable insight into the generic structure of the Green's matrices contained within the integral representation.

Several numerical examples were presented to illustrate the theoretical and practical application of contraction mappings to the regulation and control of nonlinear systems. In particular, an example involving the regulation of Van der Pol's equation was used to illustrate the calculation of the convergence parameters and to demonstrate the manner in which the modified contraction mappings method may be used to extend the range of applicability of contraction mappings. An example considering the null controllability of the pitch motion of a satellite with bounded control thrust was then presented. This example illustrated the application of contraction mappings to an operator which did not satisfy differentiability conditions. The Lipschitz norm rather than the derivative norm was then used for the theoretical convergence analysis and to prove null controllability from the initial point. The final example involved the development of a dynamic model attempting to represent the causal, feedback structure of community drug usage. Optimal regulator theory and contraction mappings were then used to gain insight into how a community might best respond to a rapidly growing heroin addiction problem. The various examples demonstrate that contraction mappings is a useful tool for both the theoretical and practical investigation of nonlinear system control.

8.2. Contributions

The author considers the following items to constitute the original contributions of this dissertation.

1. The determination of Green's functions in explicit form using simply structured matrices and similarity transformations.

2. The development of insight into the generic structure of broad classes of Green's functions by the use of partitioned matrices.

3. The development of a controllability theory for nonlinear dynamical systems based on an integral representation of TPBVP's, the implicit function theorem, and contraction mappings.

4. The theoretical and practical application of contraction mappings to a nonlinear control problem with bounded input control and the subsequent use of the Lipschitz norm to prove convergence for the nondifferentiable operator equation.

5. The theoretical and practical application of contraction mappings to the optimal regulation of a dynamic model of a socio-economic system.

In addition, convergence theorems are presented for operators arising from the optimal regulation of several classes of nonlinear systems. However, these results are translations of the general theorems presented in Falb [F1] and in that sense are not completely original.

8.3. Recommendations

In this section some areas of possible future research will be briefly outlined. As indicated in the summary, the main thrust of this dissertation has been directed toward the application of contraction mappings. However, Falb and de Jong [F1] have succinctly revealed the close relationship which exists between contraction mappings, modified contraction mappings, and Newton's method. The

first area for possible additional research lies in exploiting this relationship and applying Newton's method to those operators arising from the optimal regulation of nonlinear systems. Investigation of the convergence criteria for Newton's method should yield additional insight into the theory of state regulation for nonlinear systems. The second area of research lies in the extension of the controllability results of Chapter 5. These results for the controllability of nonlinear systems are essentially local in nature, i.e., they consider controllability near the origin. However with additional analysis using the integral representation, it should be possible to identify classes of problems for which global results may be proved. The third and final area of recommended research involves an in-depth analysis into the relationship between the drug system model and the results of optimization. In particular, the data base for the model, parameter identification, and a sensitivity analysis deserve significant attention. In this manner, critical issues of the problem may be identified for additional social investigation and data collection.

APPENDIX A

The contraction mappings program consists of a main program and several subroutines. A brief description of the function of each part is now presented.

MAIN essentially directs the program and performs no actual computation. MAIN first calls the subroutine STTRM which calculates the fundamental matrices $\Phi^V(t,0)$ and $\Phi^V(0,s)$. To accomplish this task, STTRM calls AFCT and VELEMS, and the integration to calculate $\Phi^V(\cdot,\cdot)$ is performed by DIFEQ. The resultant fundamental matrices are stored by OUTP and OUTT. MAIN next calls CALC, the major subroutine of the algorithm. CALC computes the Green's functions and directs the solution of the successive members of the CM sequence. VCAL and VELEMS are used to calculate $V(t)$, and SBFN calculates $\{c - g(y(0)) - h(y(0)) + My(0) + Ny(1)$ and $F(y)\}$. FINT then calls DQSF to integrate the expression

$$\int_0^t G_I^J(t,s)\{F(y_n(s),s)-V(s)y_n(s)\}ds + \int_t^1 G_{II}^J(t,s)\{F(y_n(s),s)-V(s)y_n(s)\}ds.$$

CONV is then called to test for convergence. If the test for convergence is successful, the program returns to MAIN and ends. If the test for convergence fails, the algorithm remains in CALC and calculates the successive solutions until either convergence is attained or a stop condition is reached. All computations are done in double precision arithmetic. To use the contraction mappings program, the user must modify only two subroutines, VELEMS and SBFN. In VELEMS, the user specifies the choice of the $V(t)$ matrix. In SBFN,

the user specifies the differential equation $\dot{y} = F(y,t)$ and the boundary condition $g(y(0)) + h(y(1)) = c$. The program contains many comment statements to ease application.

```
C  CONTRACTION MAPPING ALGORITHM
C
C MAIN
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION FNINT
      DOUBLE PRECISION UNITY
      DOUBLE PRECISION UN
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION UNITY(15,15)
C NDIM IS THE DIMENSION OF THE PROBLEM VECTOR
      READ(5,2) NDIM
    2 FORMAT(I5)
      NDM=NDIM
      NSQ=NDIM*NDIM
C THE INCREMENT OF SOLUTION IS NOW READ IN.
      READ(5,3) DELT
    3 FORMAT(D10.2)
C THE NUMBER OF INCREMENTS IS NOW CALCULATED.
      FNINT=1.0D0/DELT+1.1D0
      NINT=IDINT(FNINT)
C THE SUBROUTINE STTRM WILL NOW BE CALLED TO CALCULATE THE STATE
C TRANSITION MATRIX OF THE SPECIFIED LINEAR SYSTEM AND ITS ADJOINT
      CALL STTRM(NDIM)
C THE MATRIX UNITY IS FORMED TO CHECK THE ACCURACY IN CALCULATING
C PHI AND PHIS.
      DO 663 J=1,NDM
      DO 663 I=1,NDM
      UNITY(I,J)=0.0D0
      DO 663 K=1,NDM
  663 UNITY(I,J)=UNITY(I,J)+PHI(I,K,21)*PHIS(K,J,21)
      DO 665 I=1,NDM
      WRITE(6,664) (UNITY(I,J), J=1,NDM)
```

```
  664 FORMAT(' ',5X,D15.8)
  665 CONTINUE
      UN=0.0D0
      DO 333 I=1,NDIM
      UN=UNITY(I,I)+UN
  333 CONTINUE
      IF(UN .GT. 1.5D0*NDIM) GO TO 606
C NOW THE MAJOR SUBROUTINE CALC IS CALLED TO CALCULATE AND STORE THE
C NEW SOLUTION.
      CALL CALC(NDIM)
C A STOP CONDITION IS CHECKED.
      IF(ITER .EQ. 15) GO TO 606
      ITER1=ITER+1
      DO 19 K=1,ITER1
      WRITE(6,16) K
   16 FORMAT('0',5X,4HK = ,I3)
      DO 18 J=1,NDIM
      WRITE(6,20) J
   20 FORMAT(' ',10X,4HJ = ,I3)
      WRITE(6,17) (YS(J,NDS,K), NDS=1,NINT)
   17 FORMAT(' ',15X,D15.8)
   18 CONTINUE
   19 CONTINUE
  606 STOP
      END
```

```
      SUBROUTINE SBFN(NDIM)
C
C THIS SUBROUTINE CALCULATES THE VALUES OF FN=F(Y)-V*Y FOR VALUES OF NDT
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION YV,F,Z,YI,YT,G,H,TT,TTT
      DOUBLE PRECISION X1,X2,X3,X4,X5,X6,DFDXT,GAIN,GMAX,DGDP
      DOUBLE PRECISION GTAU,ADD,AEDM,PI,AAM,AVB,DRDA,PCS,FF
      DOUBLE PRECISION DAP,DAE,CP,CE,Q1,Q2,Q3
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION YV(15),F(15),Z(15),YI(15),YT(15),G(15),H(15)
      DIMENSION TT(15),TTT(15)
      DIMENSION DFDXT(4,4),PCS(4),FF(9)
C THE FOLLOWING VARIABLES ARE USED TO CALCULATE THE NONLINEAR
C FORCING FUNCTION F(I).
      PI=3.141593D0
      AVB=60.0D0
      DT=12.0D0
      GMAX=1.0D0
      GTAU=.255D0
      DAP=6.0D0
      DAE=12.0D0
      CE=.01D0
      CP=.25D0
      Q1=.04D0
      Q2=0.0D0
      Q3=0.0D0
      DO 600 NDT=1,NINT
C THE NONLINEAR EQUATION IS A FUNCTION OF THE STATE AT THE CURRENT
C TIME.  A VECTOR OF THE STATE AT THE NDT IS CREATED AND IS USED TO
C CALCULATE F AT NDT.
      DO 599 I=1,NDIM
  599 YV(I)=YS(I,NDT,ITER)
```

```
      X1=YV(1)
      X2=YV(2)
      X3=YV(3)
      X4=YV(4)
      X5=YV(5)
      X6=YV(6)
      DO 102 I=1,3
      DO 102 J=2,3
  102 DFDXT(I,J)=0.0D0
      GAIN=DCOS(0.36D0*X2)
      DGDP=-0.36D0*DSIN(0.36D0*X2)
      AOD=50.0D0
      IF(X3 .GT. 10.0D0) GO TO 110
      AEDM=.75D0+.25D0*DCOS(PI*X3/10.0D0)
      DFDXT(3,1)=-(.25D0*PI/10.0D0)*DSIN(PI*X3/10.0D0)*X1/AOD
      GO TO 111
  110 AEDM=.50D0
      DFDXT(3,1)=0.0D0
  111 CONTINUE
      IF(X1 .GT. 60.0D0) GO TO 112
      AAM=.5D0+.5D0*DSIN(PI*(X1-AVB/2.0D0)/AVB)
      DRDA=GAIN*X2*.5D0*(PI/AVB)*DCOS(PI*(X1-AVB/2.0D0)/AVB)
      GO TO 113
  112 AAM=1.0D0
      DRDA=0.0D0
  113 CONTINUE
      DFDXT(1,1)=AEDM/AOD-DRDA
      DFDXT(2,1)=-AAM*GAIN-AAM*X2*DGDP
      DO 109 K=1,3
  109 PCS(K)=YV(K+3)
      DO 108 I=1,3
      FF(I)=0.0D0
      DO 108 K=1,3
  108 FF(I)=FF(I)+DFDXT(I,K)*PCS(K)
      GRA=AEDM*X1/AOD
      RRPE=AAM*GAIN*X2
```

```
      F(1)=DT*(GRA-RRPF)
      F(2)=DT*(-1.0D0*X2/DAP-1.0D0*X5/(DAP*DAP*CP))
      F(3)=DT*(-X3/DAE-X6/(DAE*DAE*CE))
      F(4)=DT*(-Q1*X1-FF(1))
      F(5)=DT*(-Q2*X2+X5/DAP-FF(2))
      F(6)=DT*(-Q3*X3+X6/DAE-FF(3))
C THE PRODUCT  V(NDT)*Y(NDT) IS NOW OBTAINED AND STORED AS A
C FUNCTION OF TIME.
      DO 598 I=1,NDIM
      Z(I)=0.0D0
      DO 598 K=1,NDIM
  598 Z(I)=Z(I)+V(I,K,NDT)*YV(K)
C FN IS NOW OBTAINED AND STORED AS A FUNCTION OF TIME
      DO 597 IL=1,NDIM
  597 FN(IL,NDT)=F(IL)-Z(IL)
C THIS PROCEDURE IS REPEATED FOR INCREASING NDT
  600 CONTINUE
C THIS SUBROUTINE ALSO CALCULATES THE EXPRESSION,
C   D=C-G(Y)-H(Y)+XM*Y(0)+XN*Y(1).  THE INITIAL AND TERMINAL STATE
C VECTORS ARE GENERATED BELOW.
      DO 601 I=1,NDIM
      YI(I)=YS(I,1,ITER)
  601 YT(I)=YS(I,NINT,ITER)
      G(1)=YI(1)
      G(2)=YI(2)
      G(3)=YI(3)
      G(4)=0.0D0
      G(5)=0.0D0
      G(6)=0.0D0
      H(1)=0.0D0
      H(2)=0.0D0
      H(3)=0.0D0
      H(4)=YT(4)
      H(5)=YT(5)
      H(6)=YT(6)
C THE PRODUCTS M*Y(0) AND N*Y(1) ARE NOW OBTAINED AND THE RESULT D
```

```
C IS FORMED.
      DO 602 I=1,NDIM
      TT(I)=0.0D0
      DO 602 K=1,NDIM
  602 TT(I)=TT(I)+XM(I,K)*YT(K)
      DO 603 I=1,NDIM
      TTT(I)=0.0D0
      DO 603 K=1,NDIM
  603 TTT(I)=TTT(I)+XN(I,K)*YT(K)
      DO 604 M=1,NDIM
  604 D(M)=C(M)-G(M)-H(M)+TT(M)+TTT(M)
      RETURN
      END
```

```
      SUBROUTINE VELEMS(X,A)
C THIS SUBROUTINE CALCULATES THE LINEAR SYSTEM MATRIX IN VECTOR FORM
      DOUBLE PRECISION X,A
      DOUBLE PRECISION TF
      DOUBLE PRECISION A1,A2,A3,DT,DAP,DAE,Q1,Q2,Q3,CP,CE
      DIMENSION A(225)
      A1=.02D0
      A2=-.15D0
      A3=-.005D0
      DT=12.0D0
      DAP=6.0D0
      DAE=12.0D0
      CE=.01D0
      CP=.25D0
      Q1=.04D0
      Q2=0.0D0
      Q3=0.0D0
      A(1)=A1*DT
      A(2)=0.0D0
      A(3)=0.0D0
      A(4)=-Q1*DT
      A(5)=0.0D0
      A(6)=0.0D0
      A(7)=A2*DT
      A(8)=-DT/DAP
      A(9)=0.0D0
      A(10)=0.0D0
      A(11)=-Q2*DT
      A(12)=0.0D0
      A(13)=A3*DT
      A(14)=0.0D0
      A(15)=-DT/DAE
      A(16)=0.0D0
      A(17)=0.0D0
      A(18)=-Q3*DT
      A(19)=0.0D0
```

```
A(20)=0.0D0
A(21)=0.0D0
A(22)=-A1*DT
A(23)=-A2*DT
A(24)=-A3*DT
A(25)=0.0D0
A(26)=-DT/(DAP*DAP*CP)
A(27)=0.0D0
A(28)=0.0D0
A(29)=DT/DAP
A(30)=0.0D0
A(31)=0.0D0
A(32)=0.0D0
A(33)=-DT/(DAE*DAE*CE)
A(34)=0.0D0
A(35)=0.0D0
A(36)=DT/DAE
RETURN
END
```

```
      SUBROUTINE VCAL
C THIS SUBROUTINE CALCULATES AND STORES THE V MATRIX IN TIME.
      DOUBLE PRECISION X,A
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION A(225)
      DO 100 J=1,21
      X=(J-1)*DELT
      CALL VELEMS(X,A)
      DO 555 IQ=1,NDM
      DO 554 JQ=1,NDM
      KQ=(IQ-1)*NDM+JQ
  554 V(JQ,IQ,J)=A(KQ)
  555 CONTINUE
  100 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE AFCT(X,SM,DERV)
C THIS SUBROUTINE IS USED TO CALCULATE V(.) IN THE INTEGRATION
C FOR PHI.
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION X,SM,DERV,A,TMP
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION DERV(15),A(225),SM(15)
      CALL VELEMS(X,A)
      DO 555 IQ=1,NDM
      TMP=0.0D0
      DO 554 JQ=1,NDM
      KQ=(JQ-1)*NDM+IQ
  554 TMP=TMP+A(KQ)*SM(JQ)
  555 DERV(IQ)=TMP
      RETURN
      END
```

```
      SUBROUTINE ATEC(X,SM,DERV)
C THIS SUBROUTINE IS USED TO CALCULATE -V'(.) IN THE INTEGRATION
C FOR PHIS.
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION X,SM,DERV,A,TMP
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,ITI
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION DERV(15),A(225),SM(15)
      CALL VELEMS(X,A)
      DO 555 IQ=1,NDM
      TMP=0.0D0
      DO 554 JQ=1,NDM
      KQ=(IQ-1)*NDM+JQ
  554 TMP=TMP+A(KQ)*SM(JQ)
  555 DERV(IQ)=-TMP
      RETURN
      END
```

```
      SUBROUTINE DIFEQ(N,PMODE,T,DT,CTR,VAR,RHS)
C THIS SUBROUTINE IS USED TO INTEGRATE FOR PHI AND PHIS.
C THE TECHNIQUE IS A FOURTH ORDER RUNGE-KUTTA AS MODIFIED BY GILL.
      DOUBLE PRECISION VAR(6),RHS(2),QLAM(50),CCC1,CCC2,CCC3
      DOUBLE PRECISION UGHLY,ROOT2,MNUS,PLUS
      DOUBLE PRECISION T,DT
   30 FORMAT(43HIMPROPER COUNTER SETTING IN THE DIFEQ SUBRO)
      INTEGER CTR,PMODE
      IF(PMODE) 99,1,2
    1 DO 4 J=1,N
    4 QLAM(J)=0.
      ROOT2=1.41421356237309D0
      MNUS=1.D0-1.D0/ROOT2
      PLUS=1.D0+1.D0/ROOT2
      PMODE=1
      CTR=0
    2 IF(CTR) 99,3,5
    3 CCC1=.5D0
      CCC2=1.D0
      CCC3=DT*.5D0
      T=T+CCC3
      GO TO 20
    5 IF(CTR-2) 6,7,8
    6 CCC1=MNUS
   14 CCC2=CCC1
      CCC3=CCC1*DT
      GO TO 20
    7 CCC1=PLUS
      T=T+DT*.5D0
      GO TO 14
    8 CCC1=.16666666666666667D0
      CCC2=.33333333333333333D0
      CCC3=DT*.5D0
      CTR=-1
   20 CTR=CTR+1
      CCC1=CCC1*DT
```

```
      DO 22 J=1,N
      UGHLY=CCC1*RHS(J)-CCC2*QLAM(J)
      QLAM(J)=QLAM(J)+UGHLY+UGHLY+UGHLY-CCC3*RHS(J)
   22 VAR(J)=VAR(J)+UGHLY
      RETURN
   99 WRITE(6,30)
      RETURN
      END
```

```
      SUBROUTINE STTRM(NDIM)
C THIS SUBROUTINE COMPUTES THE STATE TRANSITION MATRIX OF THE LINEAR
C SYSTEM AND ITS ADJOINT AND STORES THEM AS FUNCTIONS OF TIME.
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION Y,DERY,TF,T,DT
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION Y(15),DERY(15)
      INTEGER CTR,PMODE
      READ(5,1) DT
    1 FORMAT(D10.2)
      WRITE(6,2) DT
    2 FORMAT('0',5X,19HINTEGRATION STEP = ,D15.8)
      DO 7 II=1,NDIM
      DO 3 J=1,NDIM
      IF((II-J) .EQ. 0) Y(J)=1.0D0
      IF((II-J) .NE. 0) Y(J)=0.0D0
    3 CONTINUE
      T=0.0D0
      TF=1.0D0
      PMODE=0
      CTR=0
      KK=0
      CALL OUTP(T,Y,NDIM)
    4 CONTINUE
      CALL AFCT(T,Y,DERY)
      CALL DIFEQ(NDIM,PMODE,T,DT,CTR,Y,DERY)
      IF(CTR .EQ. 0) GO TO 5
      GO TO 4
    5 CALL OUTP(T,Y,NDIM)
      IF(T .GE. TF) GO TO 6
      GO TO 4
    6 CONTINUE
```

```
 7 CONTINUE
   DO 9 NN=1,NDIM
   WRITE(6,8) (PHI(NN,NM,21), NM=1,NDIM)
 8 FORMAT('0',5X,D15.8)
 9 CONTINUE
   DO 14 III=1,NDIM
   DO 10 J=1,NDIM
   IF((III-J) .EQ. 0) Y(J)=1.0D0
   IF((III-J) .NE. 0) Y(J)=0.0D0
10 CONTINUE
   T=0.0D0
   TF=1.0D0
   PMODE=0
   CTR=0
   LL=0
   CALL OUTT(T,Y,NDIM)
11 CONTINUE
   CALL ATFC(T,Y,DERY)
   CALL DIFEQ(NDIM,PMODE,T,DT,CTR,Y,DERY)
   IF(CTR .EQ. 0) GO TO 12
   GO TO 11
12 CALL OUTT(T,Y,NDIM)
   IF(T .GE. TF) GO TO 13
   GO TO 11
13 CONTINUE
14 CONTINUE
   DO 16 NN=1,NDIM
   WRITE(6,15) (PHIS(NN,NM,21), NM=1,NDIM)
15 FORMAT('0',5X,D15.8)
16 CONTINUE
   RETURN
   END
```

```
      SUBROUTINE OUTP(T,Y,NDIM)
C THIS SUBROUTINE STORES THE MATRIX PHI(.,0) AT THE APPROPRIATE
C INCREMENTS OF TIME.
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION T,Y
      DOUBLE PRECISION DELT,QQ,TEST,DABS
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION Y(9)
      DELT=.05D0
      QQ=FLOAT(KK)
      TEST=DABS(QQ*DELT-T)
      IF(TEST .GT. .0001D0) GO TO 100
      WRITE(6,101) T
  101 FORMAT(' ',4HT = ,D15.8)
      KK=KK+1
      DO 99 J=1,NDIM
      PHI(J,II,KK)=Y(J)
   99 CONTINUE
  100 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE OUTT(T,Y,NDIM)
C THIS SUBROUTINE STORES THE MATRIX PHIS(.,0) AT APPROPRIATE
C INCREMENTS OF TIME.
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION T,Y
      DOUBLE PRECISION DELT,RR,TEST,DABS
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION Y(9)
      DELT=.05D0
      RR=FLOAT(LL)
      TEST=DABS(RR*DELT-T)
      IF(TEST .GT. .000100) GO TO 100
      WRITE(6,101) T
  101 FORMAT(' ',4HT = ,D15.8)
      LL=LL+1
      DO 99 J=1,NDIM
      PHIS(III,J,LL)=Y(J)
   99 CONTINUE
  100 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE CALC(NDIM)
C THIS IS THE MAJOR SUBROUTINE IN THE PROGRAM.  HERE THE INTEGRAL
C EQUATIONS ARE SOLVED FOR THE ITERATED SOLUTIONS AND THE TEST FOR
C CONVERGENCE IS MADE.
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION SI,XC,T,TM,TN,TEMP1,TEMP2,TEMP3
      DOUBLE PRECISION EPS
      DOUBLE PRECISION DET
      DOUBLE PRECISION VSI,SIIN
      DOUBLE PRECISION TSI,TSIIN
      DOUBLE PRECISION RF
      DOUBLE PRECISION COST,QIV,Z
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION TEMP1(15),TEMP2(15),TEMP3(15),L(15),M(15)
      DIMENSION T(15,15,21),TM(15,15,21),TN(15,15,21)
      DIMENSION XC(15,15),VSI(225),SIIN(15,15),SI(15,15)
      DIMENSION TSI(225),TSIIN(225)
      DIMENSION QIV(21),Z(21)
      EQUIVALENCE (SI(1,1),TSI(1))
      EQUIVALENCE (SIIN(1,1),TSIIN(1))
C THE RELAXATION FACTOR IS READ IN. NORMALLY IT IS ONE.
      READ(5,555) RF
  555 FORMAT(D10.2)
C THE BOUNDARY CONDITION MATRICES IN THE BOUNDARY COMPATIBLE
C SET J=(V,M,N) ARE NOW READ IN.
      DO 2 I=1,NDIM
      READ(5,1) (XM(I,J), J=1,NDIM)
    1 FORMAT(D10.2)
    2 CONTINUE
      DO 4 I=1,NDIM
      READ(5,3) (XN(I,J), J=1,NDIM)
    3 FORMAT(D10.2)
```

```
    4 CONTINUE
C EPS, THE CONVERGENCE MEASURE IS NOW READ IN.
      READ(5,9) EPS
    9 FORMAT(D10.2)
C THE BOUNDARY CONDITION VECTOR C IS NOW READ IN.
      READ(5,10) (C(I), I=1,NDIM)
   10 FORMAT(D10.2)
C THE CONSTANT ISM IS NOW READ IN. IF ISM IS ONE, THE PROGRAM
C COMPUTES THE INITIAL BOUNDARY COMPATIBLE GUESS.
C IF ISM IS NOT ONE, THE INITIAL SOLUTION IS NOW READ IN.
      READ(5,666) ISM
  666 FORMAT(I10)
C NOW FORMING THE PRODUCT OF N*PHI(1,0)
      DO 7 J=1,NDIM
      DO 7 I=1,NDIM
      XC(I,J)=0.0D0
      DO 7 K=1,NDIM
    7 XC(I,J)=XC(I,J)+XN(I,K)*PHI(K,J,NINT)
C THE MATRIX SUM (M+N*PHI(1,0))) IS NOW FORMED.
      DO 8 J=1,NDIM
      DO 8 I=1,NDIM
    8 SI(I,J)=XM(I,J)+XC(I,J)
      WRITE(6,12)
   12 FORMAT('0',2X,2HSI)
      DO 15 I=1,NDIM
      DO 14 J=1,NDIM
      WRITE(6,13) SI(I,J)
   13 FORMAT(' ',10X,D15.8)
   14 CONTINUE
   15 CONTINUE
      MODE=2
      CALL ARRAY(MODE,NDIM,NDIM,15,15,VSI,TSI)
      CALL MINV(VSI,NDIM,DET,L,M)
      MODE=1
      CALL ARRAY(MODE,NDIM,NDIM,15,15,VSI,TSIIN)
      WRITE(6,112)
```

```
112 FORMAT('0',2X,4HSIIN)
    DO 115 I=1,NDIM
    DO 114 J=1,NDIM
    WRITE(6,113) SIIN(I,J)
113 FORMAT('0',10X,D15.8)
114 CONTINUE
115 CONTINUE
    DO 400 NDT=1,NINT
    DO 397 J=1,NDIM
    DO 397 I=1,NDIM
    T(I,J,NDT)=0.0D0
    DO 397 K=1,NDIM
397 T(I,J,NDT)=T(I,J,NDT)+PHI(I,K,NDT)*SIIN(K,J)
    DO 398 J=1,NDIM
    DO 398 I=1,NDIM
    TM(I,J,NDT)=0.0D0
    DO 398 K=1,NDIM
398 TM(I,J,NDT)=TM(I,J,NDT)+T(I,K,NDT)*XM(K,J)
    DO 399 J=1,NDIM
    DO 399 I=1,NDIM
    TN(I,J,NDT)=0.0D0
    DO 399 K=1,NDIM
399 TN(I,J,NDT)=TN(I,J,NDT)+T(I,K,NDT)*XC(K,J)
400 CONTINUE
    ITER=0
401 ITER=ITER+1
    IF(ITER .EQ. 15) GO TO 909
    IF(ITER .GT. 1) GO TO 782
    IF(ISM .EQ. 1) GO TO 669
    DO 668 I=1,NDIM
    READ(5,667) (YS(I,J,1), J=1,NINT)
667 FORMAT(D10.2)
668 CONTINUE
    GO TO 670
669 DO 814 NDS=1,NINT
    DO 815 I=1,NDIM
```

```
      YS(I,NDS,1)=0.0D0
      DO 816 K=1,NDIM
  816 YS(I,NDS,1)=YS(I,NDS,1)+T(I,K,NDS)*C(K)
  815 CONTINUE
  814 CONTINUE
  670 CONTINUE
      DO 818 I=1,NDIM
      WRITE(6,817) (YS(I,N,1),N=1,NINT)
  817 FORMAT(' ',20X,D15.8)
  818 CONTINUE
C THE SUBROUTINE VCAL IS NOW CALLED TO CALCULATE AND STORE THE LINEAR
C SYSTEM MATRIX AS A FUNCTION OF TIME
      CALL VCAL
C SUBROUTINE SBFN WILL NOW BE CALLED TO CALCULATE FN=F(Y)-V*Y
  782 CALL SBFN(NDIM)
C SUBROUTINE FINT INTEGRATES PHI(0,S)*FN(S) FROM ZERO TO T AND
C STORES THE INTEGRAL AS A FUNCTION OF T, WHERE T VARIES FROM ZERO
C TO ONE. THESE VALUES ARE USED TO CALCULATE THE INTEGRAL FROM T TO ONE.
      CALL FINT(NDIM)
C THE NEXT SEQUENCE OF INSTRUCTIONS SOLVES FOR THE NEXT ITERATED
C SOLUTION. FIRST THE PRODUCT T(T)*D WILL BE CALCULATED.
      DO 304 NDS=1,NINT
      DO 300 I=1,NDIM
      TEMP1(I)=0.0D0
      DO 300 K=1,NDIM
  300 TEMP1(I)=TEMP1(I)+T(I,K,NDS)*D(K)
C NEXT, THE PRODUCT OF TM(NDS) AND THE INTEGRAL OF FN FROM ZERO TO
C NDS IS FORMED.
      DO 301 I=1,NDIM
      TEMP2(I)=0.0D0
      DO 301 K=1,NDIM
  301 TEMP2(I)=TEMP2(I)+TM(I,K,NDS)*QINT(K,NDS)
C NEXT, THE PRODUCT OF TN(NDS) AND THE INTEGRAL OF FN FROM NDS TO ONE
C IS FORMED.
      DO 302 I=1,NDIM
      TEMP3(I)=0.0D0
```

```
      DO 302 K=1,NDIM
  302 TEMP3(I)=TEMP3(I)+TN(I,K,NDS)*QQINT(K,NDS)
C THE THREE TEMPS ARE SUMMED TO GIVE THE VALUE OF THE NEW SOLUTION
C AT TIME NDS.
      DO 303 JJ=1,NDIM
  303 YS(JJ,NDS,ITER+1)=(1.0D0-RF)*YS(JJ,NDS,ITER)
     C +RF*(TEMP1(JJ)+TEMP2(JJ)-TEMP3(JJ))
C THIS PROCEDURE IS REPEATED FOR INCREASING NDS
  304 CONTINUE
      ITER1=ITER+1
      WRITE(6,404)
  404 FORMAT('0',5X,2HYS)
      DO 407 I=1,NDIM
      DO 406 N=1,NINT
      WRITE(6,405) YS(I,N,ITER1)
  405 FORMAT(' ',10X,D15.8)
  406 CONTINUE
  407 CONTINUE
C CONVERGENCE OF THE ITERATIONIS NOW TESTED.
      CALL CONV(NDIM,MM,EPS)
      IF(MM .EQ. 1) GO TO 401
  909 RETURN
      END
```

```
      SUBROUTINE FINT(NDIM)
C SUBROUTINE FINT INTEGRATES PHI(0,S)*FN(S) FROM ZERO TO T AND
C STORES THE INTEGRAL AS A FUNCTION OF T, WHERE T VARIES FROM ZERO
C TO ONE. THESE VALUES ARE USED TO CALCULATE THE INTEGRAL FROM T TO ONE.
      DOUBLE PRECISION PHI,PHIS,PHIDS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION QQ,Z,QIV
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION QQ(15,21),Z(21),QIV(21)
C CALCULATE AND STORE THE VECTOR PHI(0,NDS)*FN(NDS) AS A FUNCTION OF NDS
      DO 97 NDS=1,NINT
      DO 95 I=1,NDIM
      QQ(I,NDS)=0.0D0
      DO 94 K=1,NDIM
   94 QQ(I,NDS)=QQ(I,NDS)+PHIS(I,K,NDS)*FN(K,NDS)
   95 CONTINUE
   97 CONTINUE
C THE TIME HISTORY OF EACH COMPONENT IS PUT IS VECTOR FORM AND
C INTEGRATED BY DQSF.
      DO 100 KJ=1,NDIM
      DO 98 LJ=1,NINT
      QIV(LJ)=QQ(KJ,LJ)
   98 CONTINUE
      CALL DQSF(DELT,QIV,Z,NINT)
C THE INTEGRALS ARE STORED IN QINT AND QQINT.
      DO 99 NN=1,NINT
      QINT(KJ,NN)=Z(NN)
   99 CONTINUE
  100 CONTINUE
      DO 202 M=1,NDIM
      DO 201 MM=1,NINT
  201 QQINT(M,MM)=QINT(M,NINT)-QINT(M,MM)
  202 CONTINUE
```

Page intentionally left blank

```
      SUBROUTINE CONV(NDIM,MM,EPS)
C THIS SUBROUTINE TESTS FOR CONVERGENCE OF THE ITERATION.
      DOUBLE PRECISION PHI,PHIS,PHIOS,DELT,FN,D,YS,QINT,QQINT,V,C
      DOUBLE PRECISION XN,XM
      DOUBLE PRECISION DY,BIGC,BIG,EPS,CON
      DOUBLE PRECISION DABS
      DOUBLE PRECISION COST,QIV,Z
      COMMON PHI(9,9,21),PHIS(9,9,21),DELT,FN(9,21),D(9)
      COMMON YS(9,21,15),QINT(9,21),QQINT(9,21),V(9,9,21)
      COMMON C(9),XN(9,9),XM(9,9),II,III
      COMMON KK,LL,NDM,NINT,ITER
      DIMENSION DY(21),BIGC(15)
      DIMENSION QIV(21),Z(21)
      DO 700 I=1,NDIM
      DO 698 NDS=1,NINT
  698 DY(NDS)=DABS(YS(I,NDS,ITER+1)-YS(I,NDS,ITER))
C THE LARGEST ABSOLUTE DIFFERENCE IN THIS COMPONENT WILL NOW V
C THE LARGEST ABSOLUTE DIFFERENCE IN THIS COMPONENT WILL NOW BE FOUND
      BIG=DY(1)
      DO 699 M=2,NINT
      IF(DY(M) .LT. BIG) GO TO 699
      BIG=DY(M)
  699 CONTINUE
  700 BIGC(I)=BIG
      CON=BIGC(1)
      DO 701 L=2,NDIM
      IF(BIGC(L) .LT. CON) GO TO 701
      CON=BIGC(L)
  701 CONTINUE
      WRITE(6,755) CON
  755 FORMAT('0',15X,30HNORM OF FUNCTION DIFFERENCE = ,D15.8)
      IF(CON .LT. EPS) GO TO 999
      MM=1
      GO TO 998
  999 MM=0
  998 RETURN
```

Page intentionally left blank

BIBLIOGRAPHY

A1. Athans, M., and Falb, P. F., Optimal Control: An Introduction to the Theory and Its Application, McGraw-Hill Book Company, New York, 1966.

B1. Bellman, R., Introduction to Matrix Analysis, McGraw-Hill, New York, 1960.

B2. Bodewig, E., Matrix Calculus, Interscience Publishers, New York, 1956.

B3. Booten, R. C., "An Optimization Theory for Time-Varying Linear Systems with Non-Stationary Statistical Inputs", Proc. IRE, Vol. 40, 977-981, 1952.

B4. Brunovsky, P., "On Optimal Stabilization of Nonlinear Systems", in Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, Eds., Academic Press, New York, 1967.

B5. Burghart, J. H., "A Technique for Suboptimal Feedback Control of Nonlinear Systems", IEEE Trans. Automatic Control, October, 1969.

B6. Bullock, T. E., and Franklin, G. F., "A Second-Order Feedback Method for Optimal Control Computations", IEEE Trans. Automatic Control, December 1967.

C1. Coddington, E. A., and Levinson, N., Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955.

C2. Collatz, L., Functional Analysis and Numerical Mathematics, Academic Press, New York, 1966.

D1. Durbeck, R. C., "An Approximate Technique for Suboptimal Control", IEEE Trans. Automatic Control, Vol. AC-10, pp.144-149, April 1965.

D2. Dyer, Peter, and S. R. McReynolds, The Computation and Theory of Optimal Control, New York, Academic Press, 1970.

F1. Falb, P. L., and deJong, J. L., Some Successive Approximation Methods in Control and Oscillation Theory, Academic Press, New York, 1969.

F2. Ferrar, W. L., Finite Matrices, Oxford University Press, London, 1951.

F3. Forrester, J. W., Industrial Dynamics, M.I.T. Press, M.I.T., Cambridge, Mass., 1961.

F4. Forrester, J. W., Principles of Systems, Wright-Allen Press, Cambridge, Mass., 1968.

F5. Forrester, J. W., Urban Dynamics, M.I.T. Press, M.I.T., Cambridge, Mass., 1969.

F6. Forrester, J. W., World Dynamics, Wright-Allen Press, Cambridge, Mass., 1971.

F7. Forrester, J. W., "Industrial Dynamics: A Major Breakthrough for Decision Makers", Harvard Business Review, July-August 1958.

F8. Friedland, B., "A Technique of Quasi-Optimal Control", J. Basic Engrg. Vol. 88, June 1966.

G1. Gantmacher, F. R., The Theory of Matrices, New York, Chelsea Publishing Company, 1959.

G2. Garrard, W. L., et.al., "An Approach to Suboptimal Feedback Control of Nonlinear Systems", Internatl. J. Control, Vol. 5, pp.425-435, November 1967.

G3. Gershwin, S. B., and Jacobson, D. H., "A Controllability Theory for Nonlinear Systems", IEEE Trans. Automatic Control, Vol. AC-16, No. 1, February 1971.

H1. Holtzman, J. M., Nonlinear System Theory, Prentice Hall, Englewood Cliffs, New Jersey, 1970.

J1. Jazwinski, A. H., "Quadratic and Higher-Order Feedback Gains for Control of Nonlinear Systems", AIAA J., Vol 3, pp.925-935, May 1965.

K1. Kalman, R. E., "Contributions to the Theory of Optimal Control", Bol. Soc. Mat. Mexico, pp.102-119, 1960.

K2. Kalman, R. E., "When Is a Linear System Optimal?", J. Basic Engineering, (ASME Trans.), Vol. 86, pp. 1-10, 1964.

K3. Kalman, R. E., Ho, and Narenda, "Controllability of Linear Dynamical Systems", Contributions to Differential Equations, Vol. 1, 1962.

K4. Kantorovich, L. V., and Akilov, G. P., Functional Analysis in Normed Spaces, Macmillan, New York, 1964.

K5. Kleinman, D. L., On the Linear Regulator Problem and the Matrix Riccati Equation, Electronic Systems Laboratory Report ESL-R-271, M.I.T., Cambridge, 1966.

L1. Lee, B.L., and Markus, L., Foundations of Optimal Control Theory, John Wiley and Sons, Inc., New York, 1967.

L2. Long, R.S., "Newton-Raphson Operator; Problems with Undetermined Endpoints", AIAA J., 3, pp. 1351-1352, 1965.

L3. Longmuir, A. G., and Bohn, E. V., "The Synthesis of Suboptimal Feedback Control Laws", IEEE Trans. Automatic Control, Vol. AC-12, pp. 775-758, December 1967.

L4. Lukes, D. L., "Optimal Regulation of Nonlinear Dynamical Systems", SIAM J. Control, Vol. 7, No. 1, February 1969.

N1. Newton, Gould, and Kaiser, Analytical Design of Linear Feedback Controls, John Wiley and Sons, Inc., New York, 1957.

O1. Ogata, K., State Space Analysis of Control Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

P1. Pearson, J. D., "Approximate Methods in Optimal Control", J. Electron Control, Vol. 13, pp.453-469, 1962.

P2. Picard, E., Traite d'Analyse, 3rd ed., Vol. III, Gauthiers-Villars, Paris, 1928.

P3. Pugh, A. L., DYNAMO User's Manual, M.I.T. Press, Cambridge, Mass., 1963.

R1. Rees, F. J., and Flugge-Lotz, I., Minimum Fuel Control of a Pitch Motion of a Satellite in Circular Orbit, SUDAAR No. 352, Dept. of Aero. and Astro., Stanford Univ., Stanford, Cal., 1968.

R2. Roberts, E. B., The Dynamics of Research and Development, Harper and Row, New York, 1964.

R3. Roberts, E. B., et. al., "Narcotics and the Community: A System Simulation", To be published in the Journal of Health, May-June, 1972.

S1. Simmons, G. F., Topology and Modern Analysis, McGraw-Hill, New York, 1963.

S2. Schoenberger, M., "Optimization and Implementation of System Control Laws", Proc. 4th Ann. Allerton Conf. Circuit and System Theory, pp.557-566, 1966.

S3. Statistical Abstract of the United States: 1970, U. S. Bureau of the Census, U. S. Department of Commerce.

T1. Turnbull, H. W., and Aitkin, A. C., An Introduction to the Theory of Canonical Matrices, Blakies and Son, Ltd., London, 1932.

W1. Weiner, N., The Extrapolation, Interpolation and Smoothing of Stationary Time Series, Technology Press, M.I.T., Cambridge, Mass., 1949.

W2. Willis, B. H., "The Frequency Domain Solution of Regulator Problems", Presented at the 1965 JACC, Troy, N. Y., June 22-25, 1965.

Z1. Zadeh, L. A., and Ragazzini, J. R., "An Extension of Weiner's Theory of Predictions", J. Appl. Phys., Vol. 21, pp.945-955, 1950.

Page Intentionally Left Blank

BIOGRAPHICAL NOTE

William Robert Killingsworth, Jr. was born in [redacted] on [redacted]. He attended public schools in Birmingham and graduated from Banks High School in June, 1963. Mr. Killingsworth was awarded a Gorgas Science Foundation Scholarship and entered Auburn University in June, 1963. He graduated from Auburn in June, 1966, receiving the degree of Bachelor of Science in Aerospace Engineering with highest honors. The President's Award for the School of Engineering was awarded to Mr. Killingsworth, and he was elected to Phi Kappa Phi and Tau Beta Pi honorary societies.

Mr. Killingsworth entered the Department of Aeronautics and Astronautics at M.I.T. in September, 1966 and was elected to the Sigma Xi honorary society the following year. He received the degree of Master of Science in June, 1968. His S.M. thesis was entitled "Computation Frames for Strapdown Inertial Systems". Through June, 1966 his graduate study was supported by a National Science Foundation Graduate Fellowship. In September, 1968 Mr. Killingsworth became a research assistant at the M.I.T. Measurement Systems Laboratory, where he pursued his doctoral research.

During summers, Mr. Killingsworth has been employed by the Hayes International Corporation (1966), working on a range safety analysis; the Electronic Systems Laboratory, M.I.T. (1968), working on a computer display algorithm; and The Analytic Sciences Corporation (1969), working on a navigation and control project.

Mr. Killingsworth is married to the former Joyce Caryn[redacted] of Westwood, Massachusetts.