REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

2

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 89-1689
6a. NAME OF PERFORMING ORGANIZATION University of California	6b. OFFICE SYMBOL (If applicable)
7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research	
6c. ADDRESS (City, State, and ZIP Code) Department of Mathematics Davis, CA 95616	7b. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR	8b. OFFICE SYMBOL (If applicable) NM
9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-85-0267	
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB, DC 20332-6448	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102F	2304	A1	

11. TITLE (Include Security Classification)
Observer Based Compensators for Nonlinear Systems, with additional reports at Appendix B, listed on reverse of this form

12. PERSONAL AUTHOR(S)
Arthur J. Krener

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
FINAL	FROM 30 Sep 85 TO 29 Mar 89		

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report is a compendium of all published documents which have been supported by the Air Force Office of Scientific Research under Grant no. 85-0267. The main body of the report is taken from the Ph.D. dissertation of Sinan Karahan (submitted to the Graduate Division of the University of California at Davis on June, 1989) and presents the most current status and results of the research. Included as an appendix is a compilation of the following publications, those which preceded and led up to the Ph.D. dissertation, as well as additional papers supported by the grant:

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT □ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT □ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt Col James Crowley	22b. TELEPHONE (Include Area Code) (202) 767-5028	22c. OFFICE SYMBOL NM

APPENDIX B:

ADDITIONAL PUBLICATIONS

The following publications are supported by the grant and they are included in this appendix:

PUBLISHED:

1. 1986 Krener, A. J., "Normal forms for linear and nonlinear systems. In Differential Geometry," *The Interface between Pure and Applied Mathematics*, M. Luksik, C. Martin and W. Shadwick, eds. Contempary Mathematics V.68, American Mathematical Society, Providence, 157-189.
2. 1987 Krener, A. J., S. Karahan, M. Hubbard and R. Frezza, "Higher order linear approximations to nonlinear control systems," *Proceedings, IEEE Conf. on Decision and Control*, Los Angeles, 519-523.
3. 1987 Karahan, S., "Determining Torque and Velocity Limits on Joint Actuators for Robot Arms with Coupled Joint Motion," *Proceedings of the Ninth IASTED International Symposium on Robotics and Automation*, Santa Barbara, CA, 96-99.
4. 1988 Frezza, R., S. Karahan, A. J. Krener and M. Hubbard, "Application of an efficient nonlinear filter," in *Analysis and Control of Nonlinear Systems*, C.I. Byrnes, C.F. Martin and R.E. Saeks, eds. North Holland, Amsterdam, 223-238.
5. 1988 Phelps, A.R. and A.J. Krener, "Computation of observer normal form using MACSYMA," in *Analysis and Control of Nonlinear Systems*, C.I. Byrnes, C.F. Martin and R.E. Saeks, eds. North Holland, Amsterdam, 475-482.
6. 1988 Krener, A.J., "Reciprocal processes, second order stochastic differential equations and PDE's of conservation and balance," in *Analysis and Control of Nonlinear Systems*, C.I. Byrnes, C.F. Martin and R.E. Saeks, eds. North Holland, Amsterdam, 579-590.
7. 1988 Krener, A.J. and H. Schaettler, "The structure of small-time reachable sets in low dimensions," *SIAM Journal on Control and Optimization*, 27:120-147.
8. 1988 Krener, A.J., "Reciprocal diffusions and stochastic differential equations of second order," *Stochastics* 24:393-422.

9. 1988 Krener, A.J., S. Karahan and M. Hubbard, "Approximate normal forms of nonlinear systems," *Proceedings, IEEE Conf. on Decision and Control*, San Antonio,1223-1229.

IN PRESS:

1. Krener, A. J., "Nonlinear controller design via approximate normal forms," *Proceedings of the IMA Summer Institute on Signal Processing*, University of Minnesota, 1988.
2. Clark, J.M.C., "A local characterization of reciprocal diffusions," *Stochastics*, to appear.

SUBMITTED:

1. Krener, A.J. and Y. Zhu, "The fractional representation of a class of nonlinear systems."
2. Kang, W. and A.J. Krener, "Observation of a rigid body from measurements of a principle axis."

AFOSR-TR- 89-1689

Observer Based Compensators for Nonlinear Systems

Final Technical Report

Grant No. AFOSR-85-0267

for the period

1 October 1986 - 31 March 1989

Principal Investigator: Arthur J. Krener

Co-Principal Investigator: Mont Hubbard

Research Associates:

Sinan Karahan

Andrew R. Phelps

Ruggero Frezza

Wei Kang

J. Martin Clark

Department of Mathematics

Department of Mechanical Engineering

Institute of Theoretical Dynamics

University of California

Davis, CA 95616

TABLE OF CONTENTS

PUBLISHED:

1. Krener, A. J., "Normal forms for linear and nonlinear systems."
2. Krener, A. J., S. Karahan, M. Hubbard and R. Frezza, "Higher order linear approximations to nonlinear control systems."
3. Karahan, S., "Determining Torque and Velocity Limits on Joint Actuators for Robot Arms with Coupled Joint Motion."
4. Frezza, R., S. Karahan, A. J. Krener and M. Hubbard, "Application of an efficient nonlinear filter."
5. Phelps, A.R. and A.J. Krener, "Computation of observer normal form using MACSYMA."
6. Krener, A.J., "Reciprocal processes, second order stochastic differential equations and PDE's of conservation and balance."
7. Krener, A.J. and H. Schaettler, "The structure of small-time reachable sets in low dimensions."
8. Krener, A.J., "Reciprocal diffusions and stochastic differential equations of second order."
9. Krener, A.J., S. Karahan and M. Hubbard, "Approximate normal forms of nonlinear systems."

IN PRESS:

1. Krener, A. J., "Nonlinear controller design via approximate normal forms."
2. Clark, J.M.C., "A Local characterization of reciprocal diffusions."

SUBMITTED:

1. Krener, A.J. and Y. Zhu, "The fractional representation of a class of nonlinear systems."
2. Kang, W. and A.J. Krener, "Observation of a rigid body from measurements of a principle axis."

Observer Based Compensators for Nonlinear Systems

Abstract

This report is a compendium of all published documents which have been supported by the Air Force Office of Scientific Research under Grant no. 85-0267. The main body of the report is taken from the Ph.D. dissertation of Sinan Karahan (submitted to the Graduate Division of the University of California at Davis on June, 1989) and presents the most current status and results of the research. Included as an appendix is a compilation of the following publications, those which preceded and led up to the Ph.D. dissertation, as well as additional papers supported by the grant:

PUBLISHED:

1. Krener, A. J., "Normal forms for linear and nonlinear systems."
2. Krener, A. J., S. Karahan, M. Hubbard, and R. Frezza, "Higher order linear approximations to nonlinear control systems."
3. Karahan, S., "Determining Torque and Velocity Limits on Joint Actuators for Robot Arms with Coupled Joint Motion."
4. Frezza, R., S. Karahan, A. J. Krener and M. Hubbard, "Application of an efficient nonlinear filter."
5. Phelps, A.R. and A.J. Krener, "Computation of observer normal form using MACSYMA."
6. Krener, A.J., "Reciprocal processes, second order stochastic differential equations and PDE's of conservation and balance."
7. Krener, A.J. and H. Schaettler, "The structure of small-time reachable sets in low dimensions."
8. Krener, A.J., "Reciprocal diffusions and stochastic differential equations of second order."
9. Krener, A.J., S. Karahan, and M. Hubbard, "Approximate normal forms of nonlinear systems."

IN PRESS:

1. Krener, A. J., "Nonlinear controller design via approximate normal forms."
2. Clark, J.M.C., "A Local characterization of reciprocal diffusions."

SUBMITTED:

1. Krener, A.J. and Y. Zhu, "The fractional representation of a class of nonlinear systems."

2. Kang, W. and A.J. Krener, "Observation of a rigid body from measurements of a principle axis."

Please see Appendix B for publications and proceedings in which the above papers appeared.

In the main body of the report, entitled "Higher Degree Linear Approximations of Nonlinear Systems," we develop a new method for obtaining higher degree linear approximations of a certain class of nonlinear control systems. The standard approach in the analysis and synthesis of nonlinear systems is a first order approximation by a linear model. This is usually performed by obtaining a series expansion of the system at some nominal operating point and retaining only the first degree terms in the series. Obviously, the accuracy of this approximation depends on how far the system moves away from the nominal point, and on the relative magnitudes of the higher degree terms in the series expansion. In the report, we seek an approximation for a nonlinear system by a linear model up to higher degrees than one. This is achieved by finding an appropriate nonlinear coordinate transformation-nonlinear feedback pair to perform the higher degree linearization. With the proposed method, one can improve the accuracy of the approximation up to arbitrarily higher degrees, provided certain solvability conditions are satisfied. The Hunt-Su linearizability theorem makes these conditions precise. Our approach to the solution of this linearization problem is similar to Poincaré's Normal Form Theorem in formulation, but different in its solution method. After some mathematical background we derive a set of equations (called the Homological Equations) based on the goal of obtaining a model accurate to a higher degree in the series expansion. A solution to this system of linear equations is equivalent to the solution to the problem of linearization up to higher degrees by coordinate change and feedback. However, it is generally not possible to solve the system of equations exactly. We outline a method for systematically finding approximate solutions to these equations using singular value decomposition, while minimizing an error with respect to some defined norm. The solution thus found minimizes the error between the approximate linearization of the given system and a "nearby" one that is exactly linearizable (in the sense of Hunt-Su) up to the specified degree of approximation. We present a computer program written in the MATLAB language that automates the solution of the considerably large system of equations. Finally, we demonstrate the applications of the method and the efficiency of the results by several examples and simulations.

In Appendix A we demonstrate the usage of the program with an example session recorded during running MATLAB.

Appendix B contains the additional publications supported by the grant.

1. INTRODUCTION

1.1 Background

The state space approach to linear control systems matured into a well–defined and powerful discipline by the 70's. Contrary to linear systems, nonlinear control systems have been studied with many different methods of approach, usually depending on the type of nonlinearites involved. In this framework, many tools of analysis involving qualitative, quantitative and computer–aided methods were developed as diverse as perturbation methods, limit cycle analysis, describing functions, and graphical phase–portrait methods [5,6]. Most of these approaches were suited only to the specific types of nonlinearities for which they were developed. Until the mid–seventies, a coherent theory of nonlinear systems did not seem possible for such diverse types of systems.

However, in recent years, a rich theory for nonlinear systems has been developed using differential geometric methods. We can now say that a theory for nonlinear control systems exists. In fact, the differential geometric setting allows the generalization of many known classical results in linear systems theory to nonlinear systems. With the introduction of differential geometric tools, many interesting results have been obtained for nonlinear controllability, observability, equivalence, decomposition, optimality, control synthesis, linearization and many others. In this broader sense, it can be stated that linear systems are a special case of the more general class of nonlinear control systems.

The first major developments in the general theory for nonlinear systems were through the introduction of differential geometry by Hermann [15], the analysis of linear multivariable systems using a geometric approach by Wonham [46], and one of the first important results in the transformation of a nonlinear system into a linear system by Krener [27,28]. We refer the reader to Sussmann [44] for an excellent survey and bibliography.

1.2 Motivation

In the analysis of scientific and engineering systems, one often encounters situations which do not lend themselves to exact solutions by conventional methods. The assumption of linearity in most control system models, for example, is an oversimplification at best. This assumption, of course, reflects the difficulties one would rather avoid in dealing with an otherwise nonlinear model. Indeed, one can seldom find a technique to solve a given nonlinear problem exactly. Since the control system designer is equipped with powerful methods and tools for attacking linear control systems, the motivation for "linearizing" a given nonlinear problem is clearly very strong.

Therefore, whenever possible, the control problem posed must be suitably transformed to bring it into an appropriate form that enables the implementation of linear control design techniques. However, the systematics underlying such modifications by transformation are by no means self-evident. The simplest of these modifications is a first degree linear approximation of the nonlinear model by calculating a series expansion at a nominal operating point. The validity of this approximation depends on the relative size of the higher degree terms. In systems where nonlinearities are strong, these higher degree terms cannot be neglected, and the approximation fails.

The question of whether a nonlinear system can be equivalent to a linear system under some group of transformations such as change of coordinates will be one of the main issues of this report. This question has been addressed by many researchers. The earliest example in this area was solved by Poincaré [42], who gave a sufficient condition for the linearizability of a vector field around a critical point by changing state coordinates. With the introduction of differential geometric techniques, the method of linearizing transformations under a nonlinear change of state coordinates and nonlinear state feedback was developed by various researchers.

Krener [27] discussed the question of when a nonlinear system can be transformed into a linear system by a change of state coordinates. Jakubczyk and Respondek [25], and Hunt and Su [17] independently considered the full state feedback and coordinate change problem. This problem, solved in a slightly more general setting by Hunt and Su, has been since coined as the "Hunt-Su Linearization Method", and it is one of the most important developments in the field. The Hunt-Su linearization method gives necessary and sufficient conditions for there to exist locally a coordinate transformation and feedback that carries a nonlinear system into a linear system. In [29], Krener considered the case in which one can find an approximate linearization by considering the second and higher degree terms in the truncated series expansion of the vector field, and proved a weakened version of the Hunt–Su linearization condition. In [35] and more recently in [34], further results in the solution of the resulting transformations were presented.

Many fine applications of the above developments have since appeared in literature. In [38,39] Meyer, Su, and Hunt have successfully applied these techniques to automatic flight control. Krener in [30] has suggested a new approach to compensator design in the same framework. Freund [13] applied these methods in robot control even before the theory was fully developed. Other applications in robotics appeared in [11,45]. In [24], Isidori and Ruberti have solved the input–output linearization problem for a system with output, where the goal is to find a state feedback law such that the input-dependent part of the output of the closed-loop system is linear in the new input.

Important theoretical results and applications were also developed on the dual problem of nonlinear observers. In order to construct an observer for a nonlinear system, a suitable coordinate change is first found which transforms the system into an observer canonical form. Then an observer with linearizable error dynamics is constructed in the new coordinates. This approach to the nonlinear observer problem was first proposed independently by Krener and Isidori [33] and Bestle and Zeitz [8]. Krener and Respondek in [36] extended the problem to multi-input cases. An application of this method appeared in [12].

In the next chapter of this report, the Normal Form Theorem of Poincaré, the Hunt–Su linearization condition, Krener's results on approximate linearization, and the results of [35] will be reviewed in detail. These topics are the most relevant prior developments on which this report is based.

The objectives of this report are to:

(1) Review the current developments and the mathematical background necessary to establish a good understanding of the approximate linearization of nonlinear control systems. Our approach here will be from an engineering point of view, and explicit reference to advanced mathematics will be kept to a minimum.

(2) Given a nonlinear control system in state-space form that consists of n first-order differential equations, find a general solution to the problem of approximate linearization by state feedback and coordinate change. Here we derive the set of linear equations from the Homological equations. Solution to this system of linear equations is equivalent to the solution of the linearization problem.

(3) Present an efficient method of solution to the Generalized Homological equations. In general, the solution will not be unique. A solution is found that is optimal in some statistical sense.

(4) Incorporate the method of solution in a computer program. The derivation and the solution of the homological equations is extremely tedious. The structure of the equations is dependent on the order of the

system being linearized. A computer program that automates the procedure has been written in the MATLAB program language .

(5) Illustrate the method using examples. Using the computer program that solves for the transformations as an aid, linearize an example control system up to second degree terms in its series approximation, thus demonstrating the program as a control system design tool. By simulations, compare the response of a control system that has been linearized up to higher degree against the response of the same system linearized only up to first degree.

1.3 Outline of the Report

The report is organized as follows:

In Chapter 2 some mathematical preliminaries that are directly relevant to this work are reviewed. Various mathematical tools such as Lie derivatives, Lie brackets, and distributions are introduced. Controllable, controller, observable and observer normal forms for linear and nonlinear systems are presented. The importance of the nonlinear controller form in our work is emphasized. The Hunt–Su linearization condition [17] and the extension of this result to the approximate linearization of control systems by state feedback and coordinate change [29] are explained. The results of these two papers are central to the report.

In Chapter 3 Poincaré's Normal Form Theory, and its relevance to higher degree approximations of control systems is discussed. The problem of higher degree approximations to nonlinear control systems is stated, and references [34,35], which consist of some preliminary work of this report, are reviewed. The Homological equations of Poincaré are the starting point for the derivation of the Generalized Homological Equations. A solution to the generalized homological equations yields the equivalent solution to the linearization problem with coordinate change and feedback. For the sake of analysis, an appropriate basis is introduced for the expression of higher degree monomials in the series expansion of a vector field. The generalized homological equations are evaluated with the aid of this basis. An equivalent system of linear equations to be solved is obtained. Properties of this linear mapping are discussed. An optimal solution which provides the best approximation in some statistical sense is presented for the case when there is no exact solution to the linearization problem.

Chapter 4 presents the analysis of the kernel and the co-kernel of the linear mapping that is equivalent to the generalized homological equations, and derivation of the linear system of equations that is equivalent to the homological equations. For the second-degree linearization problem of a single-input control system it is shown that the kernel of the mapping is always of dimension one. The analysis of the co-kernel of the mapping is far more complicated. The co-kernel equations are derived using a formula for repeated Lie derivatives. These equations represent the relationships that have to be satisfied among the coefficients of the second degree terms in the vector field in order the system to be exactly linearizable (up to second degree for the case we have analyzed). Next, the method of solution for the system of linear equations derived in Ch. 3 is presented. All of the terms that appear in the homological equations are expressed in a suitable basis. These expressions enable us to numerically calculate the coefficient matrix in the linear system of equations, which is implemented in the computer program.

In Chapter 5, the computer program used for solving the equations is explained in detail. This program takes the first and second degree parts of a nonlinear control system as input, and solves the equivalent set of linear equations that are obtained from the generalized homological equations. It is written in the MATLAB application program and incorporates into the solution of the problem all the results of Chapters 3 and 4 in an efficient way. Next, various nonlinear control systems are considered as examples to demonstrate the efficiency of the approximate linearization method. The examples chosen are control systems that are either exactly linearizable up to second degree, or systems that yield only an approximate linearization. Comparisons are made in the time-domain responses against first-degree approximations of the same systems. The response plots, the performance and effectiveness of the method, and their significance are discussed.

Finally, in Conclusion, the results are summarized and the significance and implications of this study in control system science are discussed. Possible future research topics in this area are suggested.

The appendix presents a sample session of the MATLAB program for Approximate Linearization of Control Systems that was recorded during running the program.

2. MATHEMATICAL BACKGROUND

This section will aim to clarify the connections between the classical treatment of linear control systems and the more recent results in nonlinear control systems. To this end, we will introduce some mathematical definitions and results that are not traditionally used in control system analysis. Our central focus in presenting the mathematical background will be toward a precise statement of the linearization problem which this report addresses. We will closely follow prior work by Isidori [21,22], Banks [7], and Krener [27,29,30,32].

2.1 Preliminaries

We introduce some notations and definitions:

$\mathbb{R}^n$: n–dimensional Euclidian space.

M: a paracompact, connected C^∞ manifold of dimension n.

$V(M)$: the real linear space of C^∞ vector fields on M.

$C^\infty(M)$: set of real–valued functions on M.

$T_x(M)$: tangent space to M at $x \in M$ (a copy of $\mathbb{R}^n$).

TM: the tangent bundle over M; the union $\bigcup_{x \in M} T_x M$ of tangent spaces.

$V^*(M)$: the real linear space of C^∞ one-forms on M, dual to V.

Δ: a distribution on M. Δ_x $\underline{\text{fi}}T_x (M)$, $\forall\ x \in M$.

The differential operator d: $C^\infty(M) \rightarrow V^*(M)$ is defined by

$$\mathrm{d}h = (\partial h/\partial x)\mathrm{d}x_1 + \cdots + (\partial h/\partial x)\mathrm{d}x_n \qquad \text{for } h \in C^\infty(M).$$

For any one form

$$w = w_1 \mathrm{d}x_1 + \cdots + w_n \mathrm{d}x_n,$$

and vector field

$$f = f_1(\partial/\partial x_1) + \cdots + f_n(\partial/\partial x_n)$$

the dual product $<w,f>$ is defined as the scalar function

$$w_1 f_1 + \cdots + w_n f_n.$$

Usually, a vector field $f \in V(M)$ is denoted as a column vector

$$f = (f_1 , \cdots, f_n)^T$$

and a one form $w \in V^*(M)$ is denoted as a row vector

$$w = (w_1, \cdots, w_n).$$

A function $h \in C^\infty(M)$ defines a one form

$$\mathrm{d}h = (\partial h/\partial x_1, \cdots, \partial h/\partial x_n).$$

Definition 2.1.1 Let $f \in V(M)$. There are three kinds of Lie derivatives related to f, which are expressed as follows:

(i) The Lie (or directional) differentiation of a scalar function h with respect to f:

$$h \in C^{\infty}(M); \quad L_f : C^{\infty}(M) \to C^{\infty}(M),$$

$$L_f(h) = <\mathrm{d}h, f> = \sum_{i=1}^{n} f_i(x) \frac{\partial h}{\partial x_i}.$$

This is more precisely the derivative of the function h in the direction of the vector f.

(ii) The Lie derivative of a vector field g with respect to f:

$$g \in V(M); \quad L_f : V(M) \to V(M),$$

$$L_f(g) = [f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g.$$

This is called the Lie bracket, and it is also denoted by $ad_f g$.

(iii) The Lie derivative of a one-form w with respect to f:

$$w \in V^*(M); \quad L_f : V^*(M) \to V^*(M),$$

$$L_f(w) = \mathrm{d}<w, f> = \left(\frac{\partial w^T}{\partial x} f \right)^T + w \frac{\partial f}{\partial x}$$

where the superscript T denotes the transpose operator.

Higher derivatives can also be defined by induction as follows:

(a) $L_f^0 h = h, \qquad L_f^1 h = <\mathrm{d}h, f>, \ \dots, \ L_f^i(h) = L_f^1(L_f^{i-1} h)$

(b) $ad_f^0 g = g, \qquad ad_f^1 g = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g, \ldots , ad_f^i\, g = [\, f, ad_f^{i-1} g\ ]$

(c) $L_f^0 w = w, \qquad L_f^1 w = \mathrm{d}\!<\! w, f \!>, \ldots , L_f^i w = L_f^1 (L_f^{i-1} w\ )$

These three types of Lie derivatives are related by the Leibnitz formula:

$$<\! L_f w,\ g \!> \ = \ <\! w,\ ad_f g \!> + L_f <\! w,\ g \!> \tag{2.1}$$

<u>*Definition* 2.1.2</u> A set of C^∞ vector fields $\{X^1, \ldots, X^d\}$ on M is *involutive* if there exist C^∞ functions $c_k^{ij}(x)$ such that

$$[\ X^i, X^j\ ] = \sum_{k=1}^{d} c_k^{ij}\,(x) X^k\,(x), \qquad 1 \le i, j \le d\ ;\ i \ne j. \tag{2.2}$$

<u>*Definition* 2.1.3</u> A *distribution* Δ on a manifold M is a mapping assigning to each point p of M a *subspace* $\Delta(p)$ of the tangent space $T_p(M)$ to M at p.

<u>*Definition* 2.1.4</u> A distribution is nonsingular on U, an open subset of M, if there exists an integer d such that dim $\Delta(p) = d$ for all $p \in U$.

2.2 Normal Forms for Linear and Nonlinear Systems

A state space description of a controllable linear system can be transformed to controllable or controller form by a linear change of state space. Similarly, a state space description of an observable linear system can be transformed to observable or observer form by a linear change of state variables. In the context of this section, controllable, observable, controller, and observer canonical forms are called "Normal Forms." The definition of normal forms slightly differ in literature. We follow that of [26].

In this section we will present normal forms for linear and nonlinear systems. This will be done in a systematic way that is very suitable for the extension of these concepts to nonlinear systems. The treatment closely follows [32].

We introduce some notations and definitions. The *indices* $l_1, \ldots, l_q$ are positive integers summing to n.

<u>*Definition* 2.2.1</u> A *prime triple* (A, B, C) with indices $l_1, \ldots, l_q$ is a triple of block diagonal matrices of dimension $n \times n$, $n \times q$, and $q \times n$ of the form

$$A = \begin{bmatrix} A_1 & & & & \\ & \cdot & & & \\ & & \cdot & 0 & \\ & 0 & & \cdot & \\ & & & & \cdot & \\ & & & & & A_q \end{bmatrix} ; \text{ where } A_i = \begin{bmatrix} 0 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ \cdot & & & & \cdot & \\ \cdot & & & & & 1 \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix} l_i \times l_i \tag{2.3a}$$

$$B = \begin{bmatrix} B_1 & & & & \\ & \cdot & & & \\ & & \cdot & 0 & \\ & 0 & & \cdot & \\ & & & & \cdot & \\ & & & & & B_q \end{bmatrix} ; \text{ where } B_i = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} l_i \times 1 \tag{2.3b}$$

and

$$C = \begin{bmatrix} C_1 & & & 0 \\ & \cdot & & \\ & & \cdot & \\ 0 & & & C_q \end{bmatrix} ; \text{ where } C_i = [\, 1 \ 0 \ \cdot \ \cdot \ \cdot \ 0 \,]^{1 \times l_i} \qquad (2.3c)$$

Consider a linear system described by

$$\dot{\xi} = F\xi + Gu \qquad (2.4a)$$

$$y = H\xi \qquad (2.4b)$$

where $F^{n \times n}$, $G^{n \times m}$, and $H^{p \times n}$, $\xi \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$. The system is said to be *controllable* if

$$\text{rank } \{F^{r-1}G^j : j = 1, \ldots, m;\ r = 1, \ldots, n\} = n. \qquad (2.5)$$

The system (2.4) is *observable* if

$$\text{rank } \{H_i F^{r-1} : i = 1, \ldots, p;\ r = 1, \ldots, n\} = n. \qquad (2.6)$$

Note that in Eqns. (2.5) and (2.6) while F^r denotes the r'th power of F, G^j and H_i denote the j'th column of G and i'th row of H, respectively.

The *controllable form* of a linear system is

$$\dot{x} = Ax - \alpha Cx + Bu \qquad (2.7a)$$

$$y = \gamma x \qquad (2.7b)$$

where (A, B, C) is a prime triple with indices $l_1, \ldots, l_m$, and α and γ are arbitrary matrices of dimensions $n \times m$ and $p \times n$. For example, when $m = 1$, i.e. for a single–input system, (2.7a) takes the form

$$\dot{x} = \begin{bmatrix} -\alpha_1 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ \cdot & & & & \cdot & \\ \cdot & & & & & 1 \\ -\alpha_n & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} u.$$

The *observable form* of a linear system is

$$\dot{x} = Ax - B\alpha x + \beta u \tag{2.8a}$$

$$y = Cx \tag{2.8b}$$

where (A, B, C) is a prime triple with indices $l_1, \ldots, l_p$, and α and β are arbitrary matrices of dimensions $p \times n$ and $n \times m$.

A system (2.4) can be transformed into controllable form (2.7) by a linear change of state coordinates if and only if it is controllable. Similarly, a system (2.4) can be transformed into observable form (2.8) by a linear change of state coordinates iff it is observable.

The *controller form* of a linear system is

$$\dot{x} = Ax - B\alpha x + B\beta u \tag{2.9a}$$

$$y = \gamma x \tag{2.9b}$$

where (A, B, C) is a prime triple with indices $l_1, \ldots, l_p$, and α, β and γ are arbitrary matrices of dimensions $m \times n$, $m \times m$, $p \times n$ except β must be nonsingular. Define a pseudo–output ψ for the system (2.4)

$$\psi = K\xi \tag{2.10}$$

where K is an $m \times n$ matrix such that

$$K_i F^{r-1} G^j = \begin{cases} 0 & 1 \le r < l_j \\ \delta_i^j & r = l_j \end{cases} \tag{2.11}$$

The observable form realization of (2.4a) and (2.10) is a controller form realization of (2.4). In other words, the controller form (2.9) is actually the observable form of the original system (2.4) with respect to the pseudo–output (2.10).

The *observer form* of a linear system is

$$\dot{x} = Ax - \alpha Cx + \beta u \tag{2.12a}$$

$$y = \gamma Cx \tag{2.12b}$$

where (A, B, C) is a prime triple with indices $l_1, \ldots, l_p$, and α, β, γ are arbitrary matrices of dimensions $n \times p$, $n \times m$ and $p \times p$ except γ must be nonsingular. One can define a pseudo–input μ

$$\dot{\xi} = F\xi + Q\mu \tag{2.13}$$

where Q is an n × p matrix defined as

$$H_i F^{r-1} Q^j = \begin{cases} 0 & 1 \le r < l_i \\ \delta_i^j & r = l_i \end{cases} \tag{2.14}$$

The controllable form realization (2.12) of (2.4) is an observer form realization of (2.4) with respect to the pseudo–input μ.

Next we consider a nonlinear system described by

$$\dot{\xi} = f(\xi) + g(\xi)u \tag{2.15a}$$

$$y = h(\xi) \tag{2.15b}$$

where $\xi \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, and f, g, h are smooth C^∞ functions. There are four normal forms for the nonlinear system (2.15), defined as follows:

Observable Form :

$$\dot{x} = Ax - B\alpha(x) + \beta(x)u \tag{2.16a}$$

$$y = Cx \tag{2.16b}$$

where (A, B, C) is a prime ripie with indices $l_1, \ldots, l_p$, and α and β are smooth matrix valued functions of x with dimensions $m \times 1$ and $n \times m$. To illustrate how one obtains the observable form, we write (2.16a) in explicit form:

$$y_i = x_{i:1}$$

$$\dot{x}_{i:1} = x_{i:2} + \beta_{i:1}(x)u$$

$$\vdots$$

$$\dot{x}_{i:l_i} = -\alpha_i(x) + \beta_{i:l_i}(x)u$$

for $i = 1, \ldots, p$. We make note of the notation in which we are using the colon ":" to separate the first index, which represents the component of the output, from the succeeding subscript, which represents the number of derivatives from the output. Thus $x_{i:j}$ is the $(j-1)$-st time derivative of the

ith output y_i. Since the output y_i is equal to $h_i(\xi)$ in ξ coordinates (Eqn (2.15b)), the following coordinate transformation relates (2.15) and (2.16):

$$x_{i:1} = h_i(\xi)$$

$$x_{i:2} = L_f(h_i)(\xi)$$

$$\cdot$$

$$\cdot$$

$$x_{i:j} = L_f^{j-1}(h_i)(\xi) \quad \text{for } j = 1, \ldots, l_i \tag{2.17a}$$

and

$$\beta^j_{ir} = L_{g^j}(L_f^{r-1}(h_i))(\xi) \quad \text{for } r = 1, \ldots, l_i \; ; \; j = 1, \ldots, l_i \,.$$

$$-\alpha_i = L_f^{l_i}(h_i)(\xi) \tag{2.17b}$$

Next we present the *Controller Form*:

$$\dot{x} = Ax - B\alpha(x) + B\beta(x)u \tag{2.18a}$$

$$y = \gamma(x) \tag{2.18b}$$

where (A, B, C) is a prime triple with indices $l_1, \ldots, l_m$ and α, β and γ are smooth matrix valued functions of x with dimensions $m \times 1$, $m \times m$ and $p \times 1$. To obtain the nonlinear controller form, one chooses a pseudo-output $\psi = k(\xi)$, where $\psi \in \mathbb{R}^m$ and construct the observable form relative to ψ such that in x coordinates $\psi = Cx$. The observability indices of ψ are $l_1, \ldots, l_m$ and the coordinates are chosen as derivatives of this "output":

$$x_{i:j} = L_f^{j-1}(\psi_i)(\xi) \quad \text{for } j = 1, \ldots, l_i; \; i = 1, \ldots, m. \tag{2.19a}$$

and

$$\beta_{i:j} = L_{g_j}(L_f^{l_i-1}(\psi_i\ ))(\xi)$$

$$-\alpha_i = L_f^{l_i}(\psi_i\ )(\xi) \tag{2.19b}$$

We note the similarity of controller form to observable form.

Controllable Form :

$$\dot{x} = Ax - \alpha(Cx) + Bu \tag{2.20a}$$

$$y = \gamma(x) \tag{2.20b}$$

where (A, B, C) is a prime triple with indices $l_1, \ldots, l_m$ and α, γ are smooth matrix valued functions of x with dimensions $n \times 1$ and $p \times 1$. We emphasize an important property of the nonlinear controllable form: while α is a function of a pseudo-output $\psi = Cx$, the output γ is a function of x. If $\alpha(\psi)$ is a linear function of ψ then the dynamics (2.20) of the nonlinear controllable form agrees with the dynamics (2.7) of the linear controllable form. Therefore the question of the existence of a nonlinear controllable form is closely related to the question of linearizing the dynamics (2.20a) by a coordinate transformation.

Observer Form :

$$\dot{x} = Ax - \alpha(Cx) + \beta(Cx)u \tag{2.21a}$$

$$y = \gamma(Cx) \tag{2.21b}$$

where (A, B, C) is a prime triple with indices $l_1, \ldots, l_p$ and α, β and γ are smooth matrix valued functions of x with dimensions $n \times 1$, $n \times m$ and $p \times 1$. To obtain the observer form, one defines a pseudo–input $q(\xi)\mu$, and finds the controllable form with respect to this input. This explains the similarity of controllable form to observer form.

2.3. The Hunt-Su Linearization Theorem

In this section we introduce the Hunt-Su linearization condition [17,25], and also present the approximate version of this theorem by Krener [29]. This question is equivalent to the existence of the nonlinear controller form. We present the following theorem for $m = 1$.

Theorem. There exists a change of coordinates of the nonlinear system (2.15) to the controller form (2.18) around the nominal point ξ° iff

(i) Controllability condition: $\{g(\xi^\circ), \ldots, ad_{-f}^{n-1}\, g(\xi^\circ)\}$ span $T_{\xi^\circ}\mathbb{R}^n$,

(ii) Integrability condition: $\{g(\xi^\circ), \ldots, ad_{-f}^{n-2}\, g(\xi^\circ)\}$ is involutive,

(iii) (i), (ii) $\Rightarrow$ $\{g(\xi^\circ), \ldots, ad_{-f}^{n-1} g(\xi^\circ)\}$ is involutive.

For a complete proof, we refer the reader to [17,18,25]. In the following, we present a systematic method to find the change of coordinates to transform a nonlinear system into controller form. The procedure has been directly adopted from the proof of the theorem.

Assume a change of coordinates exist. Define a pseudo-output $\psi(\xi) = Cx(\xi)$ and note that

$$<d\psi(\xi^\circ), ad_{-f}^{r-1}\, g(\xi^\circ)> = L_{ad_{-f}^{r-1}g}(\psi) = \begin{cases} 0 & r < n \\ \beta \neq 0 & r = n \end{cases} \tag{2.22}$$

In particular we have, for $d\psi \neq 0$:

$$d\psi \perp \{g(\xi^\circ), \dots, ad_{-f}^{n-2}\, g(\xi^\circ)\}$$

Then, since for any two vectors X^i, X^j

$$L_{X^i}(\psi) = 0 \Rightarrow L_{X^i}L_{X^j}(\psi) - L_{X^j}L_{X^i}(\psi) = L_{[X^i, X^j]}(\psi) = 0$$

we get (noting that $L_f(d\psi) = dL_f(\psi)$):

$$d\psi,\ dL_f(\psi) \perp \{g(\xi^\circ), \dots, ad_{-f}^{n-3}\, g(\xi^\circ)\}$$

and, continuing in this fashion, we obtain

$$d\psi,\ dL_f(\psi),\ dL_f^2(\psi) \perp \{g(\xi^\circ), \dots, ad_{-f}^{n-4}\, g(\xi^\circ)\}$$

$$\vdots$$

$$d\psi,\ dL_f(\psi),\ dL_f^2(\psi), \dots, dL_f^{n-1}(\psi)$$

Note that the n one-forms $d\psi$, $dL_f(\psi)$, $dL_f^2(\psi)$, ..., $dL_f^{n-1}(\psi)$ in the above are independent, and as in (2.19a)

$$x_i = L_f^{i-1}(\psi)$$

defines the coordinate change. Once the controller form

$$\dot{x} = Ax - B\alpha(x) + B\beta(x)u \tag{2.18a}$$

is obtained, the choice of a feedback

$$u = -\alpha(x) + \frac{1}{\beta(x)}v \tag{2.23}$$

will linearize the dynamics as:

$$\dot{x} = Ax + Bv \tag{2.24}$$

where v is a new open loop feedback. Obviously, $\beta(x)$ is assumed to be nonzero.

Krener in [29] has extended this proof to the approximate linearization of control systems using a series expansions of the nonlinear terms. Following [29], we first introduce some definitions as follows:

A distribution Δ has an *order* ρ *local basis* around ξ^0 if there exist vector fields $X^1, \ldots, X^d$ which are linearly independent at ξ^0 and such that for every $Y \in \Delta$ there exist functions c_k such that

$$Y = \sum_{k=1}^{d} c_k X^k + O(\xi - \xi^0)^{\rho+1} \tag{2.25}$$

The integer d is the *order* ρ *dimension* of Δ at ξ^0. Such a distribution is said to be *order* ρ *involutive* at ξ^0 if there exist functions c_k^{ij} such that:

$$[X^i, X^j] = \sum_{k=1}^{d} c_k^{ij} X^k + O(\xi - \xi^0)^{\rho}. \tag{2.26}$$

Such a distribution is said to be *order* ρ *integrable* at ξ^0 if there exist $n - d$ independent functions $h_{d+1}, \ldots, h_n$ such that

$$<dh_i, X^j> = O(\xi)^{\rho}. \tag{2.27}$$

Theorem. (Frobenius with remainder) (Krener) *Let* Δ *be a distribution with order* ρ *basis* $\{ X^1, \ldots, X^d \}$ *at* ξ^0. Δ *is order* ρ *integrable at* ξ^0 *iff* Δ *is order* ρ *involutive at* ξ^0.

A proof of the above theorem can be found in [Krener1984a]. Given the nonlinear system (2.15) we define distributions

$$\Delta^k = C^\infty \text{ span } \{ad_f^l\, g^j : 0 \leq l < k; j = 1, \ldots, m\}.$$

Now we state the central result of [29]:

Theorem. *The nonlinear system* (2.15) *can be transformed with a coordinate change*

$$x = x(\xi) \tag{2.28a}$$

and feedback

$$u = u(\xi, v) = \alpha(\xi) + \beta(\xi)v \tag{2.28b}$$

into the order ρ *linear system*

$$\dot{x} = Ax + Bv + O(x,v)^{\rho+1} \tag{2.29}$$

where (A, B) *is a controllable pair with controllability indices* $l_1 \geq \ldots \geq l_n$ *iff*

(i) Δ^k *has an order* ρ *local basis at* ξ° *consisting of*

$\{ad_f^l\, g^j : 0 \leq l < \min(k_j, k); j = 1, \ldots, m\}$.

(ii) $\Delta^{k_j - 1}$ *is order* ρ *involutive at* ξ° *for* $j = 1, \ldots, m$.

The proof for the general case as stated in the theorem can be found in [29]. Here we will present a simplified proof for $m = 1$. First, we restate the theorem for $m = 1$:

Theorem. *The nonlinear system* (2.15) *with a single input u can be transformed into the order* ρ *linear system*

$$\dot{x} = Ax + Bv + O(x,v)^{\rho + 1} \tag{2.30}$$

where (A, B) *is a controllable pair iff*

(i) $\Delta^n = C^\infty$ span $\{g(\xi^\circ), \ldots, ad_{-f}^{n-1}\, g(\xi^\circ)\}$ *has an order* ρ *local basis at* ξ° *consisting of*

$\{g(\xi^\circ), \ldots, ad_{-f}^{n-1}\, g(\xi^\circ)\}$,

(ii) $\Delta^{n-1} = C^\infty$ span $\{g(\xi^\circ), \ldots, ad_{-f}^{n-2}\, g(\xi^\circ)\}$ *is order* ρ *involutive.*

Note the similarity of the above statement to the Hunt-Su linearization conditions.

Proof. Assume the change of coordinates and feedback (2.28) exist. Let $\bar{f}(\xi)$ and $\bar{g}(\xi)$ denote the transforms of Ax and B into ξ coordinates. It is straightforward to verify that the distribution

$\bar{\Delta}^n = C^\infty$ span $\{\, \bar{g}(\xi), \ldots, ad_{-\bar{f}}^{n-1}\ \bar{g}(\xi)\,\}$

satisfies (i) and (ii) with phrase "order ρ" deleted. Moreover from the form of (2.28) one can verify that $\bar{\Delta}^n$ and Δ^n agree to order ρ at ξ°, i.e. any vector field of one agrees with a vector field of the other to order ρ. Hence (i) and (ii) follow.

On the other hand suppose (i) and (ii) hold. By the controllability assumption Δ^n is of codimension zero. Δ^{n-1} is of codimension one and by (ii) is order ρ involutive. Therefore we can find a scalar function $h(\xi)$

which will annihilate it to order ρ as in (2.27). This function and its Lie derivatives define the desired linearizing coordinates (2.28a)

$$x_i = L_f^{i-1}(h) \ ; \ i = 1, \ldots, n.$$

In these coordinates the nonlinear system (2.15) becomes

$$\dot{x}_i = \begin{cases} x_{i+1} + O(\xi, v)^{\rho+1} & \text{if } i < n, \\ L_f^i(h) + L_g L_f^{i-1}(h) v + O(\xi, v)^{\rho+1} & \text{if } i = n. \end{cases} \qquad (2.31)$$

The linearizing feedback (2.28b) is given by

$$u = L_f^i(h) + L_g L_f^{i-1}(h) v. \qquad (2.32)$$

3. HIGHER DEGREE LINEAR APPROXIMATIONS OF NONLINEAR CONTROL SYSTEMS

3.1. Introduction

In Chapter 2, we introduced normal forms, the Hunt–Su linearization results, and Krener's extension of the Hunt–Su method by approximate linearization (via the series expansion of the vector field around a nominal point). In this chapter, the normal form theorem of Poincaré will be introduced first. This is an approximate linearization of an autonomous vector field by a nonlinear (local) change of coordinates. Next, we will present the approximate linearization problem that will be investigated in this report, which is formulated in a very similar spirit.

The fundamental difference between the Poincaré linearization and Krener's approximate linearization method needs some emphasis. As will be discussed in the next section, a sufficient condition for a linearizing transformation to exist for an autonomous nonlinear system is the so called "resonance condition" (see Sec. 3.2 of this chapter) for the eigenvalues of the linear part, or the Jacobian, of the vector field at the nominal point. In contrast, one of the necessary conditions for finding a coordinate transform and feedback pair that linearizes a nonlinear *control* system exactly (the Hunt–Su linearization) or approximately (Krener's linearization) is that the system be *locally controllable.* Obviously, this requirement implies that one is able to freely assign the eigenvalues of the linear part of the vector field. Thus, unlike the Poincaré problem, a resonance condition for the eigenvalues does not exist for this case. Basically, one has the capability to *change* the flow of the vector field through the input (though only locally within the framework we presented; global controllability of a nonlinear system has much more stringent requirements). The fundamental difference, is that, while in Poincaré's problem one only wants to be able to understand and predict the behavior of the flow of a vector field by means of a closer approximation through coordinate transformations, in the problem of linearization by feedback and coordinate change the goal is a much more ambitious one: We first seek to *linearize* the nonlinear system and ultimately to *control* its behavior. Reminiscent of the connection between the two problems, however, we continue to use the term "Homological Equations" (after Arnold [4]) for the set of equations that we will develop in the solution to the linearization problem.

There are cases in which the two approaches might in fact be used together. When a nonlinear system has both controllable and uncontrollable modes, one can decouple the state space locally into controllable and uncontrollable submanifolds by some suitable transformation. In the controllable submanifold, the approximate linearization problem may be solved based on the results presented here. On the other hand, one can decompose the uncontrollable submanifold into stable, critical, and unstable parts. An uncontrollable-unstable mode is, of course, beyond help. One does not need to worry about the uncontrollable–stable modes, and can only hope they will decay sufficiently fast. The uncontrollable–critical modes (in which the eigenvalues are on the imaginary axis), however, can be analyzed using Poincaré's method. In fact, in the framework of the analysis of nonlinear oscillations and bifurcations [14] the purpose of the "Center Manifold Theorem" is precisely that. In control systems literature there are some fine examples of work towards *controlling* the *stability* of uncontrollable-critical modes of a control system by means of the Center Manifold Theorem and bifurcation analysis (also called "Bifurcation Control") [1,2]. Although this problem may be formulated as an extension of our work, this topic is beyond the scope of this report.

3.2. Higher Degree Approximations of Autonomous Systems

Let us consider an autonomous system:

$$\dot{x} = f(x) \tag{3.1a}$$

$$x(0) = x^{\circ}. \tag{3.1b}$$

where $x \in \mathbb{R}^n$ and the system is assumed to be at rest at the origin, i.e. $f(0) = 0$. Without loss of generality we will assume $x^{\circ} = 0$. The calculations can be easily repeated for $x^{\circ} \neq 0$. First, consider the linearization of (3.1) at x°:

$$\dot{x} = F x \tag{3.2a}$$

$$F = \frac{\partial f}{\partial x}(0) \tag{3.2b}$$

We will seek a coordinate change for (3.1) of the form identity plus higher degree terms, such that the resulting system will agree with (3.1) up to an error of degree $O(x)^{\rho+1}$ where ρ is the degree of approximation.

Obviously, we obtain Eqn. (3.2) when $\rho = 1$. We will now derive the case for $\rho = 2$ and the results will be generalized to any arbitrary degree ρ by induction.

We assume a transformation of the form:

$$z = x - \phi^{(2)}(x) \tag{3.3}$$

where z denotes a new set of coordinates. $\phi^{(2)}(x)$ is a polynomial of degree 2, the monomial coefficients of which are to be found. The function $f(x)$ is expanded in a series:

$$\begin{aligned} f(x) &= f^{(1)}(x) + f^{(2)}(x) + O(x)^3 \\ &= Fx + f^{(2)}(x) + O(x)^3 \end{aligned} \tag{3.4}$$

The goal of the transformation (3.3) is to choose $\phi^{(2)}(x)$ such that in z coordinates the dynamics of the system is represented by

$$\dot{z} = Fz \quad + O(x)^3 \tag{3.5}$$

in other words, the second degree terms in the series expansion (3.4) vanish under the coordinate change. We take the time derivative of (3.3):

$$\dot{z} = \dot{x} - \frac{\partial \phi^{(2)}(x)}{\partial x}\dot{x}$$

Using (3.1a), (3.3), (3.4) and (3.5) evaluate each side in the above:

$$F(x - \phi^{(2)}(x)) = Fx + f^{(2)}(x) - \frac{\partial \phi^{(2)}(x)}{\partial x}Fx + \ O(x)^3 \tag{3.6}$$

After some cancellation, ignoring $O(x)^3$ and higher terms, and using the Lie bracket notation we obtain

$$f^{(2)}(x) = [Fx,\phi^{(2)}(x)] \tag{3.7}$$

Equation (3.7) is called the *Homological Equation* [4]. A similar derivation can also be found in [14]. In Eqn. (3.7), $f^{(2)}(x)$ and $\phi^{(2)}(x)$ are n-dimensional functions of homogeneous polynomials of degree 2. The Lie bracket operation defines a mapping

$$[Fx, \cdot\,] : \phi^{(2)}(x) \rightarrow [Fx,\phi^{(2)}(x)] \tag{3.8}$$

Obviously, (3.8) represents a linear mapping from $n^2(n + 1)/2$ dimensional parameter space of the coefficients of $\phi^{(2)}(x)$ to an $n^2(n + 1)/2$ dimensional parameter space that is the result of the Lie bracket operation. The question is whether $f^{(2)}(x)$ in the range of this mapping. In other words, is it always possible to find $\phi^{(2)}(x)$ that will satisfy (3.7)? This problem was first solved by Poincaré. In the following, we present a slightly modified proof [4,14]:

Suppose F has a full set of linearly independent eigenvectors. Then we can take these eigenvectors as a basis, which are defined by

$$Fv^k = \lambda_k v^k \tag{3.9}$$

where $v^k \in \mathbb{C}^{n \times 1}$ and $\lambda_k \in \mathbb{C}$. Similarly there exists a basis of eigenvectors of F^T defined by

$$w_i F = \lambda_i w_i \tag{3.10}$$

where $w_i \in \mathbb{C}^{1 \times n}$. We define a basis for n-dimensional functions of homogeneous polynomials of degree 2 as follows:

$$\varphi^k_{ij}(x) = v^k(w_i x)(w_j x) \qquad 1 \le i \le j \le n \,;\, 1 \le k \le n. \tag{3.11}$$

Then, using this basis to express the degree 2 polynomials in Eqn. (3.7), we evaluate the Lie bracket

$$[Fx,\varphi_{ij}^{k}(x)] = \frac{\partial\varphi_{ij}^{k}(x)}{\partial x}Fx \; - F\varphi_{ij}^{k}(x)$$

$$= v^{k}((w_{i}x)w_{j} + (w_{j}x)w_{i})Fx - Fv^{k}(w_{i}x)(w_{j}x)$$

and introducing (3.9) and (3.10) we obtain

$$[Fz,\varphi_{ij}^{k}(x)] = v^{k}(\lambda_{j}(w_{i}x)(w_{j}x) + \lambda_{i}(w_{i}x)(w_{j}x) - \lambda_{k}(w_{i}x)(w_{j}x))$$

$$= (\lambda_{i} + \lambda_{j} - \lambda_{k})\phi_{ij}^{k}(x) \qquad (3.12)$$

So the mapping (3.8) is onto if $(\lambda_i + \lambda_j - \lambda_k) \neq 0$ for all j, $k = 1, \ldots, n$; $i = 1, \ldots, j$. In the literature, this is called the *resonance condition*. We note that this is only a sufficient condition. A generalization of the proof for the case when F does not have a full set of independent eigenvectors may be found in [4].

The above proof can easily be extended to an arbitrary degree of linearization as follows. Given an autonomous system that has degree ρ and higher nonlinear terms

$$\dot{x} = Fx \, + f^{(\rho)}(x) \qquad (3.13)$$

one seeks a coordinate transformation of the form

$$z = x - \phi^{(\rho)}(x) \qquad (3.14)$$

with the goal to obtain

$$\dot{z} = Fz \; + \mathrm{O}(z)^{\rho+1}. \qquad (3.15)$$

This leads to a new homological equation

$$[Fx, \phi^{(\rho)}(x)] = f^{(\rho)}(x). \tag{3.16}$$

With the assumption that the matrix F has n linearly independent eigenvectors, and using Eqns. (3.9) and (3.10) we choose a basis for n–dimensional functions of homogeneous polynomials of degree ρ as

$$\varphi^k_{i_1,\ldots,i_\rho}(z) = v^k(w_{i_1}z)\ldots(w_{i_\rho}z) \qquad \text{for } 1 \le i_1 \le \ldots \le i_\rho \le n\ ;\ 1 \le k \le n. \tag{3.17}$$

Then a similar calculation yields

$$[Fx, \varphi^k_{i_1,\ldots,i_\rho}(x)] = (\lambda_{i_1} + \cdots + \lambda_{i_\rho} - \lambda_k)\varphi^k_{i_1,\ldots,i_\rho}(x). \tag{3.18}$$

From the above we conclude that the condition of no resonance requires that $(\lambda_{i_1} + \cdots + \lambda_{i_\rho} - \lambda_k) \neq 0$.

3.3. Higher Degree Approximations of Control Systems

Generally speaking, there are two different goals for linearizing a nonlinear system. One may seek a linearization for the purpose of designing a control, or alternatively the linearization may be tailored for the purposes of estimation. In this section we will attempt to find a solution for the problem of linearization for control. This, of course, assumes full state observability. Let us consider a nonlinear system in which the control u enters the dynamics in a linear fashion:

$$\dot{x} = f(x) + g(x)u \tag{3.19a}$$

$$x(0) = x^\circ. \tag{3.19b}$$

where $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. The system is assumed to be at rest at the nominal operating point (x°; $u^\circ = 0$). Again, we will assume $x^\circ = 0$. First, consider the linearization of (3.19) at x°:

$$\dot{x} = Fx + Gu \tag{ɜ.20a}$$

$$F = \frac{\partial f}{\partial x}(0), \quad G = g(0). \tag{3.20b}$$

We will seek a coordinate change for (3.19) of the form identity plus higher degree terms, such that the resulting linear plant will agree with (3.19) up to an error of degree $O(x, u)^{\rho+1}$ (i.e. terms of $O(x)^{\rho+1}$and $O(x, u)^{\rho}$) where ρ is the degree of approximation. When $\rho = 1$, the first degree approximation (3.20) is obtained. Similar to the previous section, the case for $\rho = 2$ will be derived first, and the results will be generalized to any arbitrary degree ρ by induction. Before proceeding further, the nonlinear functions f and g are expanded in a series:

$$\begin{aligned} f(x) &= f^{(1)}(x) + f^{(2)}(x) + O(x)^3 \\ &= Fx + f^{(2)}(x) + O(x)^{(3)} \end{aligned} \tag{3.21}$$

$$\begin{aligned} g(x) &= g^{(0)}(x) + g^{(1)}(x) + O(x)^2 \\ &= G + g^{(1)}(x) + O(x)^2 \end{aligned} \tag{3.22}$$

and the nonlinear system (3.19a) is rewritten as

$$\dot{x} = Fx + f^{(2)}(x) + \left(G + g^{(1)}(x)\right)u + O(x,u)^3 \tag{3.23}$$

We assume the same transformation proposed in Sec. 3.1:

$$z = x - \phi^{(2)}(x) \tag{3.3}$$

where z denotes the new set of coordinates. $\phi^{(2)}(x)$ is a polynomial of degree 2. In addition, a new input, denoted as v, is chosen as

$$v = \alpha(x) + \beta(x)u \tag{3.24a}$$

where $\alpha(x) = \alpha^{(2)}(x)$, an $n \times 1$ vector of second degree polynomials, and $\beta(x) = \mathrm{I} + \beta^{(1)}(x)$, an identity matrix plus first degree terms, both of dimension $m \times m$. So (3.24a) becomes:

$$v = \alpha^{(2)}(x) + \left(\mathrm{I} + \beta^{(1)}(x)\right)u. \tag{3.24b}$$

We note the slightly different form of Eqn. (3.24); in the literature (as well as in the treatment presented in Chapter 2 of this report), the feedback that accompanies a coordinate change for linearization problems is usually given as $u = \alpha(x) + \beta(x)v$. One can always obtain one expression from the other, since by definition $\beta(x)$ is nonsingular. The above choice simplifies the algebra, as we will see in the following derivations. Now we want the system to become, in z coordinates,

$$\dot{z} = Fz + Gv + \mathrm{O}(z,v)^3 \tag{3.25}$$

The time derivative of (3.3) yields:

$$\dot{z} = \dot{x} - \frac{\partial\phi^{(2)}(x)}{\partial x}\, x^{\cdot} \tag{3.26}$$

We introduce the transformation (3.3), and Eqns. (3.23), (3.24b) and (3.27) into the above:

$$F(x - \phi^{(2)}(x)) + G\,(\alpha^{(2)}(x) + (\mathrm{I} + \beta^{(1)}(x))u)$$

$$= (\mathrm{I} - \frac{\partial\phi^{(2)}(x)}{\partial x})(Fx + f^{(2)}(x) + Gu + g^{(1)}(x)u\,)$$

Then by expanding and cancelling terms on each side we obtain

$$-F\phi^{(2)}(x) + G\,\alpha^{(2)}(x) + G\,\beta^{(1)}(x)u$$

$$= f^{(2)}(x) + g^{(1)}(x)u - \frac{\partial\phi^{(2)}(x)}{\partial x}Fx - \frac{\partial\phi^{(2)}(x)}{\partial x}Gu + \mathrm{O}(x,u)^3$$

Rearranging and ignoring the $\mathrm{O}(x,u)^3$ terms we get

$$f^{(2)}(x) + g^{(1)}(x)u$$

$$= \frac{\partial\phi^{(2)}(x)}{\partial x}Fx - F\phi^{(2)}(x) + \frac{\partial\phi^{(2)}(x)}{\partial x}Gu + G\alpha^{(2)}(x) + G\beta^{(1)}(x)u \qquad (3.27)$$

Define Lie brackets as follows:

$$\frac{\partial\phi^{(2)}(x)}{\partial x}Fx - F\phi^{(2)}(x) = [Fx,\phi^{(2)}(x)] \qquad (3.28a)$$

$$\frac{\partial\phi^{(2)}(x)}{\partial x}Gu = [Gu,\phi^{(2)}(x)] \qquad (3.28b)$$

Using (3.28), Eqn. (3.27) can be written as

$$f^{(2)}(x) + g^{(1)}(x)u = [Fx,\phi^{(2)}(x)] + G\alpha^{(2)}(x) + [Gu,\phi^{(2)}(x)] + G\beta^{(1)}(x)u \qquad (3.29)$$

or, since terms of $\mathrm{O}(x)^2$ are independent of terms of $\mathrm{O}(x,u)^2$

$$f^{(2)}(x) = [Fx,\phi^{(2)}(x)] + G\alpha^{(2)}(x) \qquad (3.30a)$$

$$g^{(1)}(x)u = [Gu,\phi^{(2)}(x)] + G\beta^{(1)}(x)u \qquad \forall \text{ constant } u. \qquad (3.30b)$$

Because of its similarity to the homological equations derived in the previous section, we call (3.30) the *Generalized Homological Equations*.

In order to find an approximation of the next higher degree, we rewrite (3.25) by reverting to the variables x and u for convenience:

$$\dot{x} = Fx + Gu + O(x,u)^3 \tag{3.31}$$

At this point, we assume that the second degree terms in a given nonlinear system, if any, have been already removed as outlined above. Then we seek a new transformation of the form:

$$z = x - \phi^{(3)}(x) \tag{3.32}$$

Note that except for the linear part, transformation (3.32) will not introduce any terms of degree less than 3. Then the same procedure outlined above is repeated, with the feedback:

$$v = \alpha^{(3)}(x) + (I + \beta^{(2)}(x))u \tag{3.33}$$

which, after a series of similar calculations, results in a new set of generalized homological equations:

$$f^{(3)}(x) = [Fx, \phi^{(3)}(x)] + G\alpha^{(3)}(x) \tag{3.34a}$$

$$g^{(2)}(x)u = [Gu, \phi^{(3)}(x)] + G\beta^{(2)}(x)u \qquad \forall \text{ constant } u. \tag{3.34b}$$

These results can be generalized as follows. Given a system which is accurate to only degree $\rho - 1$, i. e.

$$\dot{x} = Fx + Gu + O(x,u)^\rho \tag{3.35}$$

a coordinate change is sought as:

$$z = x - \phi^{(\rho)}(x) \tag{3.36}$$

along with feedback:

$$v = \alpha^{(\rho)}(x) + (\mathrm{I} + \beta^{(\rho-1)}(x))u \tag{3.37}$$

which yields the generalized homological equations to be solved:

$$f^{(\rho)}(x) = [Fx,\phi^{(\rho)}(x)] + G\alpha^{(\rho)}(x) \tag{3.38a}$$

$$g^{(\rho-1)}(x)u = [Gu,\phi^{(\rho)}(x)] + G\beta^{(\rho-1)}(x)u \qquad \forall \text{ constant } u. \tag{3.38b}$$

In (3.38), $\phi^{(\rho)}, f^{(\rho)}, \alpha^{(\rho)}, g^{(\rho-1)}$ and $\beta^{(\rho-1)}$ are, respectively, homogeneous vector fields of degrees corresponding to their superscripts. The resulting system is accurate up to degree ρ:

$$\dot{z} = Fz + Gv + O(z,v)^{\rho+1} \tag{3.39}$$

Once a higher degree linear approximation is obtained for a nonlinear system one of the important issues is the stability of the closed loop system. Thus one may choose, for instance, a linear state feedback for the approximate model

$$\dot{z} = Fz + Gv \tag{3.40}$$

by setting $v = Kz$. The gain matrix K is chosen such that in the closed loop the system gives the desired performance. If we assume that the model has been linearized up to second degree, the feedback v is, from Eqn. (3.24b)

$$v = \alpha^{(2)}(x) + \left(\mathrm{I} + \beta^{(1)}(x)\right)u. \tag{3.41}$$

Using Eqn. (3.41), the feedback $v = Kz$, and transformation (3.3) we calculate the feedback law u:

$$Kx - K\phi^{(2)}(x) = \alpha^{(2)}(x) + \left(\mathrm{I} + \beta^{(1)}(x)\right)u \tag{3.42}$$

and

$$\begin{aligned} u &= \left(\mathrm{I} + \beta^{(1)}(x)\right)^{-1}\left(Kx - K\phi^{(2)}(x) - \alpha^{(2)}(x)\right) \\ &= \left(\mathrm{I} - \beta^{(1)}(x) + \cdots\right)\left(Kx - K\phi^{(2)}(x) - \alpha^{(2)}(x)\right) \\ &= Kx - \left\{\beta^{(1)}(x)Kx + K\phi^{(2)}(x) + \alpha^{(2)}(x)\right\} + \mathrm{O}(x,u)^3. \end{aligned} \tag{3.43}$$

From (3.43), the purpose of the feedback u becomes immediately clear. In addition to a linear feedback, there are *second degree correction terms* (placed inside curly brackets in (3.43) for emphasis). While one chooses a closed loop feedback $u = Kx$ to achieve stability, pole placement, etc. for the first degree approximation (3.20a) (i.e. accurate up to first degree in comparison with a linear model) to get

$$\dot{x} = (F + GK)x \tag{3.44}$$

the feedback (3.43) introduces certain second degree terms and achieves a second degree approximation (i.e. accurate up to second degree in comparison with a linear model in the z coordinates) toward the same feedback design goals:

$$\begin{aligned} \dot{x} = (F + GK)x + f^{(2)}(x) + g^{(1)}(x)Kx - G\{\beta^{(1)}(x)Kx + K\phi^{(2)}(x) \\ + \alpha^{(2)}(x)\} + \mathrm{O}(x,u)^3. \end{aligned} \tag{3.45}$$

One important feature of the feedback (3.43) and the resulting closed loop system (3.45) is that one need not transform the state variables into the new coordinates z that were introduced for the sake of calculations. Feedback design can be performed in the natural coordinates in which the system is originally presented. Obviously, the above scheme assumes a priori that all states are available for feedback. If some of the states are not observable, one can estimate the unavailable state variables by means of an observer, and apply the same procedure. This case will be treated in Section 3.5 of this chapter.

3.4. Approximate Linearization for Systems with Small Parameters

In this section, we consider a control system of the form:

$$\dot{x} = f(x,\varepsilon) + g(x,\varepsilon)u \tag{3.46a}$$

$$x(0) = x^{\circ}. \tag{3.46b}$$

where ε is a small parameter which characterizes the way parasitic effects or disturbances enter into the system. We will develop a method of linearizing transformation for this type of system, similar to that of Section 3.2. First, (3.46) is expanded as follows:

$$\dot{x} = Fx + Gu + \varepsilon(f^{(1)}(x) + g^{(1)}(x)u) + O(\varepsilon)^2 \tag{3.47}$$

In (3.47), the nonlinear function is expanded and grouped in powers of ε. Thus, the superscripts of f and g now represent the powers of ε multiplying these functions. Note that this notation is different than that of Section 3.2. A coordinate change is assumed of the following form:

$$z = x - \varepsilon\phi^{(1)}(x) \tag{3.48}$$

where neither the coefficients, nor the polynomial degree of the function $\phi^{(1)}(x)$ is yet determined (i.e. the superscript in this context denotes degrees in ε). After the transformation and feedback, we want the system to become

$$\dot{z} = Fz + G\,v + O(\varepsilon)^2 \tag{3.49}$$

where

$$v = \; u + \varepsilon\{\alpha^{(1)}(x) + \beta^{(1)}(x)u\} \tag{3.50}$$

Note that the superscripts of α and β in (3.49), as well as the superscript of ϕ in (3.48) represent the power of ε these terms are multiplied with. Thus, the notation used is different than that of Sections 3.2 and 3.3. Repeating the calculations similar to Section 3.3 yields the homological equations:

$$f^{(1)}(x) = [Fx,\phi^{(1)}(x)] + G\alpha^{(1)}(x) \tag{3.51a}$$

$$g^{(1)}(x)u \;= [Gu,\phi^{(1)}(x)] + G\beta^{(1)}(x)u \qquad \forall \text{ constant } u. \tag{3.51b}$$

This result can be generalized for an arbitrary power of ε in the same fashion: A solution to

$$f^{(\rho)}(x) = [Fx,\phi^{(\rho)}(x)] + G\alpha^{(\rho)}(x) \tag{3.52a}$$

$$g^{(\rho)}(x)u \;= [Gu,\phi^{(\rho)}(x)] + G\beta^{(\rho)}(x)u \qquad \forall \text{ constant } u. \tag{3.52b}$$

will yield a coordinate transform-feedback pair that will transform the system into:

$$\dot{z} = Fz + Gv + O(\varepsilon)^{\rho+1} \tag{3.53}$$

Even though Eqns. (3.52) and (3.38) look very similar, there are some fundamental differences. All the variables in Eqn. (3.52) have different definitions than those of Eqn. (3.38), as mentioned earlier in this section.

Moreover, the solvability conditions of (3.52) are not the same as the conditions of Eqn. (3.38). Actually, (3.52) may represent an infinite family of equations as opposed to the finite dimensional set of expressions that arise from (3.38).

Any nonlinear system expressed in the form of in Eqn. (3.19) can always be transformed into the form of (3.46) (and vice versa) as follows: First, consider the expanded form of (3.19), i.e. Eqn. (3.23):

$$\dot{x} = Fx + f^{(2)}(x) + \big(G + g^{(1)}(x)\big)u + \mathrm{O}(x,u)^3 \tag{3.23}$$

Scale the coordinates and the input with:

$$\zeta = \varepsilon^{-1}x$$

$$\mu = \varepsilon^{-1}u$$

Introducing the above into (3.23) yields

$$\dot{\zeta} = F\zeta + G\mu + \varepsilon(\bar{f}^{(2)}(\zeta) + \bar{g}^{(1)}(\zeta)\mu) + \mathrm{O}(\varepsilon)^2 \tag{3.54}$$

This equation is of the form of Eqn. (3.47), except for the difference in the fashion the expansions of *f* and *g* are defined. We use the overbar notation to emphasize this point. The input

$$v = \alpha^{(2)}(x) + (\mathrm{I} + \beta^{(1)}(x))u \tag{3.24}$$

is also transformed with an additional scaling $\eta = \varepsilon^{-1}v$:

$$\eta = \mu + \varepsilon(\bar{\alpha}^{(2)}(\zeta) + \bar{\beta}^{(1)}(\zeta)\mu) \tag{3.55}$$

With the above scaling of coordinates, a linearization problem given as in Section 3.2 can be alternatively solved with the procedure outlined in this section.

3.5. Analysis of the Linear Mapping for Linearization for Control

For the case of linearization for control, we derived the generalized homological equations (3.30) in Section 3.3. In these equations, the second degree terms $f^{(2)}(x)$ and $g^{(1)}(x)u$ can be cancelled by proper choice of $\phi^{(2)}(x)$, $\alpha^{(2)}(x)$, and $\beta^{(1)}(x)$ under certain solvability conditions. When the *coefficients* of the like terms in (3.30) are set equal, a linear mapping is obtained as

$$\begin{Bmatrix} \phi^{(2)}(x) \\ \alpha^{(2)}(x) \\ \beta^{(1)}(x) \end{Bmatrix} \longrightarrow \begin{Bmatrix} f^{(2)}(x) \\ g^{(1)}(x) \end{Bmatrix} \qquad (3.56)$$

A simple dimension count yields the dimensions of the domain and the range:

$$\frac{n^2(n+1)}{2} + \frac{mn(n+1)}{2} + m^2 n \quad \longrightarrow \quad \frac{n^2(n+1)}{2} + n^2 m \qquad (3.57)$$

To analyze the mapping, we first make a table for the dimensions of the domain and the range, where n is the dimension of the state space and m is the dimension of the input space:

For $m = 1$:

Dimension	Dimensions	
of	of	
State Space	Domain	Range
$n = 2$	11	10
$n = 3$	27	27
$n = 4$	54	56
·	·	·
·	·	·

For $m = 2$:

Dimension	Dimensions	
of	of	
State Space	Domain	Range
$n = 2$	20	14
$n = 3$	42	36
$n = 4$	76	72
$n = 5$	125	125
·	·	·

Dimensions of the domain and the range become equal whenever $n = 2m + 1$. However, this does not imply that the mapping is of full rank. For example, when $m = 1$, $n = 3$ the rank is 26, not 27. In general, when $m = 1$, the rank of the mapping is always one less than the dimension of the domain for $n \geq 3$.

As described in Chapter 2, a necessary condition for finding a coordinate change-feedback pair for a nonlinear control system is the local controllability condition at the nominal point. Thus, for the system (3.23) with a scalar input u , i.e. $m = 1$, local controllability implies

$$\text{rank } \{G \;\; FG \; \ldots F^{n-1}G\} = n. \tag{3.58}$$

On the other hand, we define a $1 \times n$ matrix K such that

$$KF^{i-1}G^{j} = \begin{cases} 0 & 1 \leq i < n \\ 1 & i = n \end{cases} \tag{3.59}$$

Then,

$$\text{rank } \{K \;\; KF \ldots KF^{n-1}\} = n. \tag{3.60}$$

(3.58) and (3.60) together imply that we can define a basis for n-dimensional second and first degree monomials as follows. First define as a basis

$$v^k = F^{k-1}G \tag{3.61a}$$

and a co-basis

$$w_i = KF^{i-1} \tag{3.61b}$$

Then we define a basis for second degree monomials as

$$\varphi^k_{ij}(x) = v^k(w_ix)(w_jx) \qquad \text{for } j, k = 1, \ldots, n \; ; \; i = 1, \ldots, j. \tag{3.62}$$

and a basis for first degree monomials as

$$\varphi^k_i(x) = v^k(w_ix) \qquad \text{for } k = 1, \ldots, n \; ; \; i = 1, \ldots, n. \tag{3.63}$$

Using the basis definitions (3.62) and (3.63) is a great convenience for calculating the Lie bracket expressions that appear in the generalized homological equations (3.30). Calculation of (3.30a) gives

$$[Fx, \varphi^k_{ij}(x)\,] = \begin{cases} \varphi^k_{i+1,j} + \varphi^k_{i,j+1} - \varphi^{k+1}_{i,j} & 1 \le i \le j < n \; ; \; 1 \le k < n \\ \varphi^k_{i+1,j} - \varphi^{k+1}_{i,j} & 1 \le i < j = n \; ; \; 1 \le k < n \\ -\varphi^{k+1}_{i,j} & i = j = n \; ; \; 1 \le k < n \end{cases} \tag{3.64}$$

In the evaluation of (3.64), when $k = n$, the expressions become slightly more complicated. For multi-input problems, the above calculations become even more involved. In the following chapters, we will present a further simplified way to calculate the bracket in the most general case that is much more suitable for numerical implementation. Next, we calculate (3.30b)

$$[G,\varphi_{ij}^{k}(x)\ ] = \begin{cases} 0 & i,\, j < n \\ \varphi_i^n & i < j = n \\ 2\varphi_n^n & i = j = n \end{cases} \tag{3.65}$$

We can use these two formulas to compute the kernel and co-kernel of the mapping

$$\begin{Bmatrix} \phi^{(2)}(x) \\ \alpha^{(2)}(x) \\ \beta^{(1)}(x) \end{Bmatrix} \longrightarrow \begin{Bmatrix} f^{(2)}(x) \\ g^{(1)}(x) \end{Bmatrix} \tag{3.56}$$

such that, we now obtain a set of linear equations expressed in matrix form:

$$L \begin{bmatrix} \phi^{(2)} \\ \alpha^{(2)} \\ \beta^{(1)} \end{bmatrix} = \begin{bmatrix} f^{(2)} \\ g^{(1)} \end{bmatrix} \tag{3.66}$$

In (3.66), L is a constant coefficient matrix of $n^2(n+1)/2 + n^2$ rows by $n^2(n+1)/2 + n(n+1)/2 + n$ columns that is found from the above evaluation of the Lie brackets of the mapping. $\begin{bmatrix} \phi^{(2)} \\ \alpha^{(2)} \\ \beta^{(1)} \end{bmatrix}$ and $\begin{bmatrix} f^{(2)} \\ g^{(1)} \end{bmatrix}$ are the constant coefficients of their corresponding terms, stacked in a consistent lexicographic ordering. For the single input linearization problem, the column rank of L is $(n^2(n+1)/2 + n(n+1)/2 + n - 1)$.

A solution to the linearization problem is developed as follows. First, we note that since the mapping (3.66) is deficient in rank for $n > 2$, a given control system with nonlinear terms $f^{(2)}(x)$ and $g^{(1)}(x)u$ will not, in general, have an exact solution to yield a second degree linearization. In fact, the Hunt-Su linearization result [17] (or Krener's extension of the same to the approximate linearization case in [29]) is a test for precisely this condition. Consequently, Eqn. (3.66) will not usually have an exact solution for $n > 2$. Then, it is reasonable to seek an approximate solution which will minimize the error in the linearization with respect to some norm. In order to give a precise meaning to this problem, first assume that we have adequate knowledge about the operating regime of the control system and the desired accuracy as determined by

$\rho(x,u)$: A probability density function; typically uniform over some compact set, or Gaussian.

Q : A sensitivity matrix, positive definite.

Then define the "error"

$$\left\|\begin{pmatrix} f^{(2)} \\ g^{(1)} \end{pmatrix} - \begin{pmatrix} \tilde{f}^{(2)} \\ \tilde{g}^{(1)} \end{pmatrix}\right\|^2 \overset{\text{def}}{=} \iint \left| f^{(2)} - \tilde{f}^{(2)} + (g^{(1)} - \tilde{g}^{(1)})u \right|_Q^2 \rho(x,u)\mathrm{d}x\,\mathrm{d}u \tag{3.67}$$

In Eqn. (3.67) the norm $|x|_Q^2$ is defined as x^TQx. We want to choose the terms $\tilde{f}^{(2)}$ and $\tilde{g}^{(1)}$ such that the norm of the above error is minimized. Note that $\begin{pmatrix} \tilde{f}^{(2)} \\ \tilde{g}^{(1)} \end{pmatrix}$ is in the range of the mapping, i.e. it satisfies the homological equations

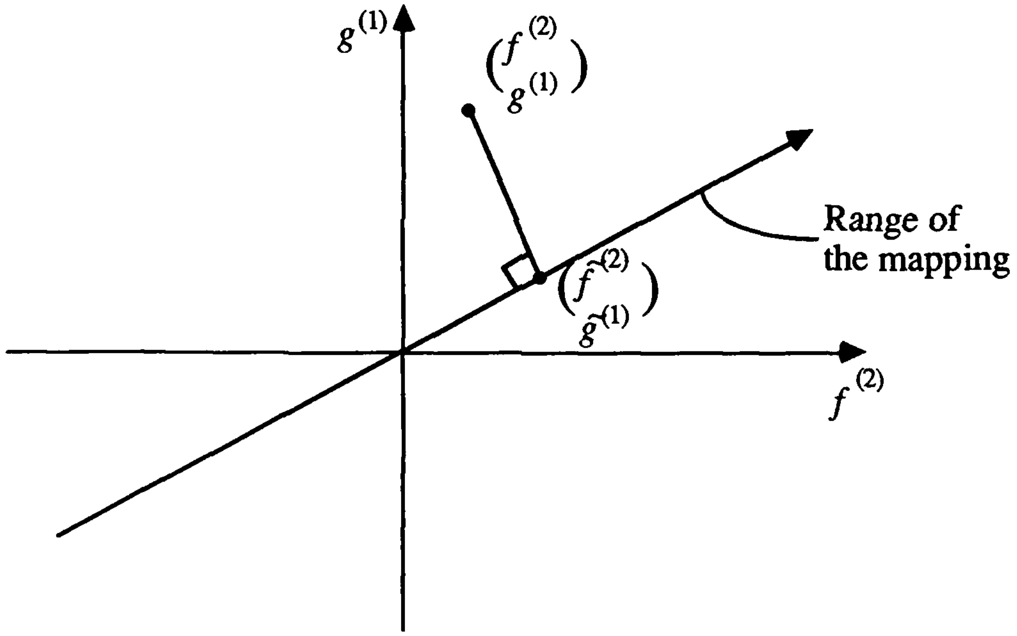
$$\tilde{f}^{(2)}(x) = [Fx, \phi^{(2)}(x)] + G\alpha^{(2)}(x) \tag{3.68a}$$

$$\tilde{g}^{(1)}(x)u = [Gu, \phi^{(2)}(x)] + G\beta^{(1)}(x)u. \qquad \forall \text{ constant } u. \tag{3.68b}$$

Furthermore, we wish to choose the smallest $\begin{pmatrix}\phi^{(2)}\\ \alpha^{(2)}\\ \beta^{(1)}\end{pmatrix}$ that will achieve the above. Again, we choose positive definite matrices S, R and minimize

$$\iint |\phi^{(2)}|_S^2 + |\alpha^{(2)}(x) + \beta^{(1)}(x)u|_R^2 \rho(x,u)\mathrm{d}x\mathrm{d}u \qquad (3.69)$$

or one can take a weighted combination of the above. In Figs. (2.1) and (2.2) we illustrate the above:



$g^{(1)}$
$\begin{pmatrix}f^{(2)}\\ g^{(1)}\end{pmatrix}$
Range of the mapping
$\begin{pmatrix}\tilde{f}^{(2)}\\ \tilde{g}^{(1)}\end{pmatrix}$
$f^{(2)}$

Fig. 2.1 The range space of the mapping.

Fig. 2.1 represents the $n^2(n + 1)/2 + n^2$ dimensional parameter space for the range of the mapping. The second degree terms in the given control system define a point in this space, denoted by $\begin{pmatrix}f^{(2)}\\ g^{(1)}\end{pmatrix}$. The range of L is represented by a straight line going through the origin. Those points in the range space of L that exactly satisfy (3.66) will lie on this line. Among these infinitely many points we want to find the one (shown as $\begin{pmatrix} \tilde{f}^{(2)} \\ \tilde{g}^{(1)} \end{pmatrix}$ on the figure) which will minimize,with respect to a norm defined earlier, the error between the actual system that is being approximately linearized and a model which is exactly linearizable (up to degree 2) by the coordinate change and feedback.

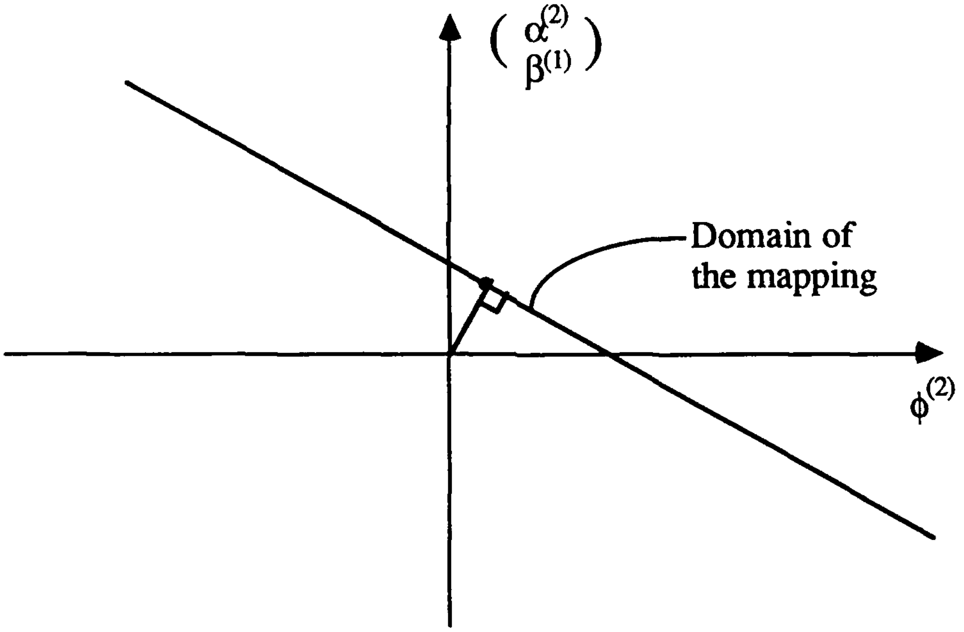


Fig. 2.2 Domain of the mapping

Fig. 2.2 illustrates the $n^2(n+1)/2 + n(n+1)/2 + n$ dimensional domain space of the mapping , and the minimization done in the domain space.

The central issue in this problem is how to define the appropriate metric to minimize the error. To this end, we assume that the states and the input have been scaled by their characteristic values. Then, the probability distribution function $\rho(x,u)$ in the integrals may be assumed to have zero mean and unit covariance. With this assumption, the matrices Q, R, and S in the integrals of (3.67) and (3.69) can be approximated by identity matrices. We assume a basis for vector valued monomials of degree 2 and express the vector monomials as

$$\phi^{(2)}(x) = \phi_k^{ij}\varphi_{ij}^k(x) \tag{3.70}$$

where $\varphi_{ij}^k = v^k x_i x_j$ and v^k is the unit column vector along the kth coordinate direction (see eqn. 4.13). Now the error terms in the integrals (3.67) and (3.69) may be easily computed. For instance, considering a term in (3.69)

$$||\phi^{(2)}||_S^2 = (\phi_k^{ij}\varphi_{ij}^k)^T S(\phi_k^{ij}\varphi_{ij}^k)$$

the integral (3.69) becomes:

$$\int S^{k\bar{k}}\phi_k^{ij}\phi_{\bar{k}}^{\bar{i}\,\bar{j}}x_i x_j x_{\bar{i}} x_{\bar{j}}\rho(x,u)\mathrm{d}x\mathrm{d}u$$

If we assume $\bar{x} = 0$ and $S = I$, the identity matrix, the above integral is:

$$\int \phi_k^{ij}\phi_{\bar{k}}^{\bar{i}\,\bar{j}}x_i x_j x_{\bar{i}} x_{\bar{j}}\rho(x,u)\mathrm{d}x\mathrm{d}u = P_{ij}P_{\bar{i}\bar{j}} + P_{i\bar{i}}P_{j\bar{j}} + P_{i\bar{j}}P_{j\bar{i}} \tag{3.71}$$

The evaluation of the integral (3.71) reduces to the simple calculation of the fourth central moments of a probability density function around zero mean since $\rho(x,u)$ is assumed to have zero mean and unit covariance in the scaled coordinates. Even though the matrix obtained from the above fourth moments is not diagonal, it is assumed to be approximately equal to the identity. This assumption greatly simplifies the numerical calculations, and it does not introduce a large deviation from the actual value of the integral being minimized.

To solve the linear set of equations

$$L \begin{bmatrix} \phi^{(2)} \\ \alpha^{(2)} \\ \beta^{(1)} \end{bmatrix} = \begin{bmatrix} f^{(2)} \\ g^{(1)} \end{bmatrix} \tag{3.66}$$

we use a singular value decomposition procedure as follows. First the matrix L in (3.66) is decomposed as

$$L = U \Sigma V^T \tag{3.72}$$

U and V in the above are orthogonal matrices and

$$\Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$$

where the matrix Σ_r is

$$\Sigma_r = \begin{bmatrix} \sigma_1 & & & & \\ & \cdot & & & \\ & & \cdot & 0 & \\ & 0 & & \cdot & \\ & & & & \cdot \\ & & & & & \sigma_r \end{bmatrix}$$

i.e. it contains the r singular values of L along its main diagonal and is zero elsewhere. Obviously, the above development assumes that the rank of L is equal to r, which is in general less than the number of columns in the matrix L. If we denote the equation (3.66) to be solved as as $Lx = b$ for brevity, we have:

$$\begin{aligned} ||Lx - b\,||^2 &= ||\;U\Sigma V^T x - b\,||^2 \\ &= ||\;U(\Sigma V^T x - d)\,||^2 \end{aligned} \tag{3.73}$$

where d is defined by $Ud = b$. Also define a new unknown

$$y = V^T x. \tag{3.74}$$

We note that multiplting a vector with an orthogonal matrix (U and V in this problem) leaves the norm of the vector invariant. Then the above becomes:

$$||Lx - b\,||^2 = ||\;U(\Sigma y - d\,)||^2 = ||\Sigma y - d\,||^2 \tag{3.75}$$

The minimizing solution is found as:

$y_i = d_i\,/\sigma_i$ for $i \leq r$

$y_i = 0$ for $i > r$.

Once the values of y are calculated, the original unknowns x are obtained by an additional matrix multiplication by V (see (3.74)). This yields the least square solution for the nonlinear coordinate change and feedback.

3.6. Linearization for Tracking and Estimation

In this section we will consider a slightly different control problem. The development included in this section is beyond the scope of this report, and is included here for the sake of completeness. This work is due to Arthur J. Krener and Andrew Phelps. Suppose that the control problem is the tracking of a reference signal, i.e. we have a p-dimensional output signal y (or its series expansion at the nominal point) and a reference $r(t)$:

$$y = h(x) = Hx + h^{(2)}(x) + O(x)^{(3)} \tag{3.76}$$

and the goal is to achieve

$$y(t) - r(t) \rightarrow 0. \tag{3.77}$$

In the following, the linearization problem for degree 2 terms is treated. We assume that in the problem the estimation of the states is also required. Along with Eqn. (3.23) we have (3.76):

$$\dot{x} = Fx + f^{(2)}(x) + \left(G + g^{(1)}(x)\right)u + O(x,u)^3 \tag{3.78a}$$

$$y = Hx + h^{(2)}(x) + O(x)^{(3)} \tag{3.78b}$$

Then we consider a coordinate change on the states as well as on the output:

$$z = x - \phi^{(2)}(x) \tag{3.79a}$$

$$w = y - \psi^{(2)}(y) \tag{3.79b}$$

and a new input v is defined as

$$v = \alpha^{(2)}(y) + \left(G + \beta^{(1)}(y)\right)u \tag{3.80}$$

Note that in (3.80), only the available output is used for feedback. With the above feedback-coordinate transform pair we want to obtain

$$\dot{z} = Fz + v + O(z,v)^3 \tag{3.81a}$$

$$w = Hz. \tag{3.81b}$$

in the new state coordinates z and the new outputs w. The following development is similar to the derivations done in Sec. 3.2. We take the time derivative of the coordinate transformation (3.79a) and introduce (3.79a,b), (3.80), (3.81) on each side:

$$F(x - \phi^{(2)}(x)) + \alpha^{(2)}(y) + (G + \beta^{(1)}(y))u$$
$$= (\mathrm{I} - \frac{\partial\phi^{(2)}(x)}{\partial x})(Fx + f^{(2)}(x) + (G + g^{(1)}(x))u\,)$$

A calculation for the output gives, using (3.79a,b) and (3.81)

$$H(x - \phi^{(2)}(x)) = Hx\ + h^{(2)}(x) - \psi^{(2)}(Hx) + O(x)^3$$

By expanding and rearranging the terms in the above equations, and using the Lie bracket notation we obtain the Generalized Homological equations:

$$f^{(2)}(x) = [Fx,\phi^{(2)}(x)] + \alpha^{(2)}(Hx) \tag{3.82a}$$

$$g^{(1)}(x)u\ = [Gu,\phi^{(2)}(x)] + \beta^{(1)}(Hx)u \qquad \forall \text{ constant } u. \tag{3.82b}$$

$$h^{(2)}(x) = \psi^{(2)}(Hx) - H\,\phi^{(2)}(x) \tag{3.82c}$$

These equations define a linear mapping as follows:

$$\begin{Bmatrix} \phi^{(2)}(x) \\ \psi^{(2)}(Hx) \\ \alpha^{(2)}(Hx) \\ \beta^{(1)}(Hx) \end{Bmatrix} \longrightarrow \begin{Bmatrix} f^{(2)}(x) \\ g^{(1)}(x) \\ h^{(2)}(x) \end{Bmatrix} \tag{3.83}$$

A dimension count gives the dimensions of the domain and the range:

$$\frac{n^2(n+1)}{2}+\frac{p^2(p+1)}{2}+\frac{np(p+1)}{2}+nmp \longrightarrow \frac{n^2(n+1)}{2}+n^2m+\frac{pn(n+1)}{2} \qquad (3.84)$$

The following table shows the dimensions of the domain and the range for various cases:

For $m = 1, p = 1$:

State Space	Domain	Range
$n = 2$	11	13
$n = 3$	25	33
$n = 4$	49	66
.	.	.

For $m = 1, p = 2$:

State Space	Domain	Range
$n = 2$	22	16
$n = 3$	39	39
$n = 4$	66	76
.	.	.

For $m = 1, p = 3$:

State Space	Domain	Range
$n = 2$	–	–
$n = 3$	63	45
$n = 4$	94	86
$n = 5$	138	145
.	.	.
.	.	.

For $m = 2, p = 2$:

State Space	Domain	Range
$n = 2$	26	20
$n = 3$	45	48
$n = 4$	74	92
.	.	.
.	.	.

As seen in the tables, for larger state space dimensions and fewer input-output pairs the mapping is rank deficient. Again, one sets up a least square problem for finding a solution, similar to Sect. 3.4.

The linearization method can be extended to the problem of estimation of the unavailable states as follows. Suppose we have

$$\dot{z} = Fz + \alpha^{(2)}(w) + (G + \beta^{(1)}(w))u \tag{3.85a}$$

$$w = Hz. \tag{3.85b}$$

Then we set up an observer

$$\dot{\hat{z}} = F\hat{z} - L(w - H\hat{z}) + \alpha^{(2)}(w) + \big(G + \beta^{(1)}(w)\big)u \tag{3.86}$$

so that the error is

$$\dot{\tilde{z}} = (F + LH)\,\tilde{z} \tag{3.87}$$

and the gain L is chosen to achieve a stable dynamics for the observer. Considering the estimated states, we define a coordinate transformation

$$\hat{x} = \hat{z} + \phi^{(2)}(\hat{z}) \tag{3.90}$$

Then, with the above, the error dynamics in x coordinates becomes

$$\dot{\hat{x}} = \big(\mathrm{I} + \frac{\partial\phi^{(2)}(\hat{z})}{\partial\hat{z}}\big)\big(F\hat{z} - L(w - H\hat{z}) + \alpha^{(2)}(w) + \big(G + \beta^{(1)}(w)\big)\big)u\big)$$

or

$$\dot{\hat{x}} = F(\hat{x} + \phi^{(2)}(\hat{x})) - L(y - \psi^{(2)}(y) - H(\hat{x} + \phi^{(2)}(\hat{x}))) + \alpha^{(2)}(y) + Gu$$
$$+ \beta^{(1)}(y)u + \frac{\partial\phi^{(2)}(\hat{x})}{\partial\hat{x}}(F\hat{x} - L(y - \psi^{(2)}(y) - H\hat{x}) + Gu) \tag{3.91}$$

Rewriting the above with the Lie bracket notation we finally get

$$\dot{\hat{x}} = F\hat{x} - L(y - H\hat{x}) + Gu + [F\hat{x},\phi^{(2)}(\hat{x})] - [L(y - H\hat{x}),\phi^{(2)}(\hat{x})]$$

$$+ [Gu,\phi^{(2)}(\hat{x})] + \alpha^{(2)}(y) + \beta^{(1)}(y)u + L\psi^{(2)}(y) \tag{3.92}$$

Eqn. (3.92) is the observer for

$$\dot{x} = Fx + Gu + f^{(2)}(x) + g^{(1)}(x)u \tag{3.93a}$$

$$y = Hx + h^{(2)}(x) \tag{3.93b}$$

Notice that the linear part of the observer agrees with standard practice and the second degree part is the correction. In the closed loop we have the feedback

$$u = K\hat{x} - \{\beta^{(1)}(\hat{x})K\hat{x} + K\phi^{(2)}(\hat{x}) + \alpha^{(2)}(\hat{x})\} \tag{3.94}$$

as expected.

4. ANALYSIS OF THE TRANSFORMATIONS

4.1. Introduction

In Chapter 3, we derived the generalized homological equations for the second degree linearization problem and introduced an equivalent set of linear equations. A solution to this set of linear equations, if it exists, will yield the coefficients of the polynomials in the coordinate transformation-feedback pair. In this section, we first show that the mapping is rank deficient. We compute both the kernel and the co-kernel of the mapping. We then describe a solution method in detail aided by the insight gained through this analysis.

A computer program written in the MATLAB package that solves for the transformations based on the analysis of this chapter is presented in Ch. 5.

4.2. Kernel of the mapping

In Section 3.2, the generalized homological equations were derived as:

$$f^{(2)}(x) = [Fx,\phi^{(2)}(x)] + G\alpha^{(2)}(x) \tag{4.1a}$$

$$g^{(1)}(x)u = [Gu,\phi^{(2)}(x)] + G\beta^{(1)}(x)u \qquad \forall \text{ constant } u. \tag{4.1b}$$

It was shown that these equations describe a mapping for the monomial coefficients of the terms in (4.1):

$$\left\{\begin{matrix}\phi^{(2)}(x)\\ \alpha^{(2)}(x)\\ \beta^{(1)}(x)\end{matrix}\right\}\longrightarrow\left\{\begin{matrix}f^{(2)}(x)\\ g^{(1)}(x)\end{matrix}\right\} \tag{4.2}$$

Furthermore, we obtained a set of linear equations equivalent to this mapping, expressed in matrix form:

$$L\begin{bmatrix}\phi^{(2)}\\ \alpha^{(2)}\\ \beta^{(1)}\end{bmatrix}=\begin{bmatrix}f^{(2)}\\ g^{(1)}\end{bmatrix} \tag{4.3}$$

The derivation of the exact form of the constant coefficient matrix L in the above equation will be discussed later. At this point we assume that we have a nonlinear system in which the linear part of the vector field is in Brunovsky canonical form:

$$\dot{x}=Ax+f^{(2)}(x)+\left(B+g^{(1)}(x)\right)u+\mathrm{O}(x,u)^3 \tag{4.4}$$

where A, B are matrices of the prime triple (A, B, C) (see Eqns. (2.3a,b)). As a matter of fact, if the linear part of a given system is controllable, one can always obtain (4.4) with some appropriate linear coordinate change and feedback. Furthermore, we also assume that the system (4.4) is linearizable up to second degree in accordance with Krener's Theorem for approximate linearization [29] as presented in Chapter 2. Then one can linearize this control system with a coordinate transformation and feedback pair

$$z = x - \phi_1^{(2)}(x) \tag{4.5}$$

$$v = \alpha_1^{(2)}(x) + \left(\mathrm{I} + \beta_1^{(1)}(x)\right)u \tag{4.6}$$

to obtain

$$\dot{z} = Az + B\,v + \mathrm{O}(z,v)^3. \tag{4.7}$$

We pose the following question at this point: Can one find some other coordinate transformation-feedback pair:

$$\zeta = z - \phi_2^{(2)}(z) \tag{4.8}$$

$$\mu = \alpha_2^{(2)}(z) + \left(\mathrm{I} + \beta_2^{(1)}(z)\right)v \tag{4.9}$$

similar in form to those of Eqns. (4.5) and (4.6) such that, after (4.7) is transformed by the above one obtains, in ζ coordinates

$$\zeta = A\zeta + B\,\mu + \mathrm{O}(\zeta,\mu)^3 \tag{4.10}$$

i.e. another linear system (up to degree 2)? If such a transformation and feedback pair (4.8)-(4.9) exist, then the original linearizing pair (4.5)-(4.6) is obviously not unique, since a combination of both will yield (ignoring cubic terms)

$$\zeta = x - \phi_3^{(2)}(x) \tag{4.11}$$

$$\mu = \alpha_3^{(2)}(x) + \left(\mathrm{I} + \beta_3^{(1)}(x)\right)u \tag{4.12}$$

To show such a transformation is indeed possible, we first choose the "natural" basis for expressing the first and second degree monomials. Choose as a basis the unit column vectors v^k, $k = 1, \ldots, n$:

$$z = x - \phi_1^{(2)}(x) \tag{4.5}$$

$$v = \alpha_1^{(2)}(x) + \left(\mathrm{I} + \beta_1^{(1)}(x)\right)u \tag{4.6}$$

to obtain

$$\dot{z} = Az + B\ v + \mathrm{O}(z,v)^3. \tag{4.7}$$

We pose the following question at this point: Can one find some other coordinate transformation-feedback pair:

$$\zeta = z - \phi_2^{(2)}(z) \tag{4.8}$$

$$\mu = \alpha_2^{(2)}(z) + \left(\mathrm{I} + \beta_2^{(1)}(z)\right)v \tag{4.9}$$

similar in form to those of Eqns. (4.5) and (4.6) such that, after (4.7) is transformed by the above one obtains, in ζ coordinates

$$\zeta = A\zeta + B\ \mu + \mathrm{O}(\zeta,\mu)^3 \tag{4.10}$$

i.e. another linear system (up to degree 2)? If such a transformation and feedback pair (4.8)-(4.9) exist, then the original linearizing pair (4.5)-(4.6) is obviously not unique, since a combination of both will yield (ignoring cubic terms)

$$\zeta = x - \phi_3^{(2)}(x) \tag{4.11}$$

$$\mu = \alpha_3^{(2)}(x) + \left(\mathrm{I} + \beta_3^{(1)}(x)\right)u \tag{4.12}$$

To show such a transformation is indeed possible, we first choose the "natural" basis for expressing the first and second degree monomials. Choose as a basis the unit column vectors v^k, $k = 1, \ldots, n$:

$$v^k = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \leftarrow k \text{ th entry} \tag{4.13}$$

Similarly choose as a co-basis the unit row vectors w_i, $i = 1, \ldots, n$:

$$w_i = [0 \cdot \cdot \ 1 \cdot \cdot \ 0] \tag{4.14}$$

with unity in the i th entry and zero everywhere else. As in Chapter 3, we define a basis for n-dimensional vector valued functions of homogeneous polynomials of degree 2 and degree 1 as follows:

$$\varphi^k_{ij}(z) = v^k(w_i z)(w_j z) \qquad \text{for } j, k = 1, \ldots, n \ ; \ i = 1, \ldots, j. \tag{4.15a}$$

$$\varphi^j_i(z) = v^j(w_i z) \qquad \text{for } j = 1, \ldots, n \ ; \ i = 1, \ldots, n. \tag{4.15b}$$

Similarly, a basis for scalar-valued second degree polynomials is chosen as:

$$h_{ij}(z) = (w_i z)(w_j z) \qquad \text{for } j = 1, \ldots, n \ ; \ i = 1, \ldots, j. \tag{4.16}$$

also for the first degree terms:

$$h_i(z) = w_i z \qquad \text{for } i = 1, \ldots, n. \tag{4.17}$$

Now we recall the proof of Krener's theorem for approximate linearization. We choose a scalar function $h(z)$, and take this function and its Lie derivatives as the desired new coordinates in (4.8):

$$h(z) = \sum_{i=1}^{n} a^i h_i(z) + \sum_{1 \le i \le j}^{n} a^{ij} h_{ij}(z) \tag{4.18}$$

$$\begin{aligned} \zeta_1 &= h(z) \\ \zeta_2 &= L_f h(z) \\ &\cdot \\ &\cdot \\ \zeta_n &= L_f^{n-1} h(z) \end{aligned} \tag{4.19}$$

Comparison of the above with (4.5) implies that, since the two coordinates z and ζ have to agree in their first degree terms, in (4.18) $a^1 = 1$ and $a^i = 0$ for $i = 2, \ldots, n$. The coefficients of the second degree terms are yet to be determined. Obviously, if all of the terms $a^{i,j}$ in (4.18) are zero, then the only transformation that takes (4.7) into (4.10) is the identity, with $\phi_2^{(2)}(z) = 0$ in (4.8). We will show that one can find a scalar function $h(z)$ with nonzero second degree terms (and therefore nonzero $\phi_2^{(2)}(z)$) to achieve the above transformation. According to the Hunt-Su linearization theorem a transformation will exist if (see Chapter 2)

$$L_g(L_f^{r-1}h(z)) = \begin{cases} 0 & 1 \leq r < n \\ \neq 0 & r = n \end{cases} \tag{4.20}$$

With the aid of the chosen basis, we proceed to calculate the first $n - 1$ equations in (4.20). First note that, since there are no second degree terms in the given control system (4.7), f and g are equal to Az and B, respectively. It is also clear that (4.20) will be satisfied if h and its first $n - 1$ Lie derivatives along f are not a function of z_n, i.e.

$$L_g(\psi(z)) = \frac{\partial\psi(z)}{\partial z} g = \left(\frac{\partial\psi(z)}{\partial z_1} \ \frac{\partial\psi(z)}{\partial z_2} \ \cdots \ \frac{\partial\psi(z)}{\partial z_n}\right) \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} = 0 \tag{4.21}$$

if $\psi(z) = \psi(z_1, z_2, \ldots, z_{n-1})$ because the n th entry in the above one-form will then vanish. If we assume that

$$h(z) = z_1 + a^{11}z_1{}^2 \tag{4.22}$$

one can easily show by calculating the repeated Lie derivatives that the term z_n will not appear until the n – 1st derivative. In fact, any quadratic term in h that is a function of z_i will cause $L_f^{r-1}h(z)$ to be a function of z_n at the $n - i$ th derivative. Therefore we establish that there is a one-parameter family of solutions dependent on the choice of $a^{1,1}$ in (4.22) that will yield additional solutions. Note that this solution does not redefine the input, and both $\alpha_2^{(2)}(z)$ and $\beta_2^{(1)}(z)$ in Eqn. (4.9) are zero. To make the explanation more precise, we have found a nonzero vector $\begin{bmatrix} \phi^{(2)} \\ \alpha^{(2)} \\ \beta^{(1)} \end{bmatrix}$ which, when added to the original solution in Eqn. (4.3), will not change the right-hand side of (4.3). In other words the solution found belongs to the kernel of the mapping. For a single input system, (4.22) is the only possible function which results in a transformation that belongs to the kernel of the mapping, and the dimension of the kernel is equal to one.

4.3. Co-kernel of the mapping

In Chapter 2, it was shown (via Krener's approximate linearization theorem) that a controllable system

$$\dot{x} = f(x) + g(x)u = Ax + f^{(2)}(x) + \left(B + g^{(1)}(x)\right)u + O(x,u)^3 \tag{4.23}$$

is linearizable up to degree 2 around $x = x^{\circ}$ iff the distribution

$$\Delta^{n-1} = C^{\infty} \text{ span } \{g(x^{\circ}), \ldots, ad_{f}^{n-2}\, g(x^{\circ})\} \tag{4.24}$$

is involutive up to degree 2. In [Krener1984a] it is also proven that this is equivalent to the integrability condition. For a single input system there exists a nonzero function $h(x)$ (precisely of the form (4.18)) such that

$$<dh, ad_{f}^{r}\, g(x)> = O(x)^{2}. \tag{4.25}$$

i.e. constant and first degree terms in (4.25) must vanish for a nonzero h. When (4.25) is not satisfied, a system is not exactly linearizable up to degree 2, and consequently an exact solution to

$$L \begin{bmatrix} \phi^{(2)} \\ \alpha^{(2)} \\ \beta^{(1)} \end{bmatrix} = \begin{bmatrix} f^{(2)} \\ g^{(1)} \end{bmatrix} \tag{4.3}$$

does not exist. In this section, we will attempt to find which terms (or linear combinations of terms) in the second degree part of the vector field cause the system to be not linearizable (or equivalently, non-involutive). In linear algebraic terms, we are looking for a specific vector or vectors $\begin{bmatrix} f^{(2)} \\ g^{(1)} \end{bmatrix}$ (the entries of this vector are the coefficients of the second degree terms of the control system in a particular basis) which will satisfy the adjoint equation

$$L^{\mathrm{T}} \begin{bmatrix} f^{(2)} \\ g^{(1)} \end{bmatrix} = 0. \tag{4.26}$$

The precise definition of the co-kernel of the mapping is given by (4.26). However, once the generalized homological equations are expressed in their equivalent linear form (4.3), the insight given by the integrability (or involutiveness) condition becomes lost. Therefore we will try to derive the co-kernel expressions using Lie derivatives and by finding the conditions that satisfy (4.25).

Eqn. (4.25) implies (see the Hunt-Su theorem in Chapter 2)

$$dh \perp \{g(x),\, ad_f\, g(x),\, \ldots,\, ad_f^{n-2}\, g(x)\} \quad \text{up to } O(x)^2. \tag{4.27}$$

Our derivation of the co-kernel equations will be based on calculation of the successive Lie derivatives of $g(x)$ along $f(x)$, which appear inside the bracket in (4.27). To simplify the expressions, we will first make a coordinate transformation that will eliminate only the terms $g^{(1)}(x)u$ in (4.23). We choose a coordinate change

$$z = x - \phi^{(2)}(x) \tag{4.28}$$

and no change in the feedback (which means $\alpha^{(2)}(x)$ and $\beta^{(1)}(x)$ will be zero) in order to satisfy only the second generalized homological equation (4.1b)

$$g^{(1)}(x) = [B, \phi^{(2)}(x)] \tag{4.29}$$

where the input u has been cancelled on both sides. We express the left-hand side of (4.29) in the basis defined in Section 4.1 (4.15b):

$$g^{(1)}(x) = \sum g_k^i\, \varphi_i^k \tag{4.30}$$

where g_k^i are the constant coefficients of $g^{(1)}(x)$ in this basis. Similarly, express $\phi^{(2)}(x)$ in the same basis defined in (4.15a):

$$\phi^{(2)}(x) = \phi_k^{ij}\,\varphi_{ij}^k \tag{4.31}$$

where ϕ_k^{ij} are the unknown coefficients of the second degree terms. The bracket on the right hand side of (4.29) is calculated using (4.31), noting that $B = v^n$ in the same basis is given by

$$[B,\phi^{(2)}(x)] = [v^n,\phi_k^{ij}\,\varphi_{ij}^k] = \frac{\partial\phi_k^{ij}\,\varphi_{ii}^k}{\partial x}v^n = \phi_k^{ij}\,v^k\Big((w_ix)w_j + (w_jx)w_i\Big)v^n$$

$$= \sum(1 + \delta_n^i\,)\,\phi_k^{in}\,\varphi_i^k \tag{4.32}$$

where δ_j^i is the Kronecker delta function, and a summation over the indices i, k is implied. Setting equal the coefficients of the basis elements φ_i^k in (4.32) and (4.30), we obtain:

$$\phi_k^{in} = g_k^i\,/(1 + \delta_n^i\,) \qquad \text{for } i, k = 1, \ldots, n. \tag{4.33a}$$

and

$$\phi_k^{ij} = 0 \qquad \text{for } j < n. \tag{4.33b}$$

Thus we have determined the form of (4.31), and the transformation (4.28), which will transform the system into:

$$\dot{z} = Az + Bu \ + f^{(2)}(z) + \ \mathrm{O}(z,u)^3 \tag{4.34}$$

Note that the coefficients of $f^{(2)}(z)$ in (4.34) are not the same as the coefficients of $f^{(2)}(x)$ in (4.23), because the transformation (4.28) contributes new quadratic terms to the vector field. These new terms may be calculated via the generalized homological equation (4.1a). Since the derivation of the co-kernel equations is identical, we shall not calculate the resulting change.

Using (4.34), Eqn. (4.27) may be rewritten as:

$$dh \perp \{B, ad_f B, \ldots, ad_f^{n-2} B\} \quad \text{up to } O(z)^2 \tag{4.35}$$

where $f = Az + f^{(2)}(z)$. Eqn. (4.35) is the same as

$$L_{ad_{-f}^{r-1}B}(h) = O(z)^2 \quad \text{for } r = 1, \ldots, n-1. \tag{4.36}$$

We use the well-known iterative formula:

$$L_{ad_{-f}^{r}g}(h) = L_{ad_{-f}^{r-1}g}\, L_f(h) - L_f L_{ad_{-f}^{r-1}g}(h) \tag{4.37}$$

which can be written as:

$$\frac{\partial}{\partial z}\left\{\frac{\partial h}{\partial z}\, ad_{-f}^{r-1}B\right\}(Az + f^{(2)}) - \frac{\partial}{\partial z}\left\{\frac{\partial h}{\partial z}\,(Az + f^{(2)})\right\} ad_{-f}^{r-1}B = 0. \tag{4.38}$$

First, the terms in (4.38) are expressed in the monomial basis:

$$f^{(2)} = f_k^{i,j}\, \varphi_{i,j}^k = f_k^{i,j}\, v^k(w_i z)(w_j z) \tag{4.39a}$$

$$B = v^n \tag{4.39b}$$

$$Av^i = \begin{cases} v^{i-1} & \text{for } 2 \le i \le n \\ 0 & \text{for } i = 1 \end{cases} \tag{4.39c}$$

$$w_i A = \begin{cases} w_{i+1} & \text{for } 1 \le i \le n-1 \\ 0 & \text{for } i = n \end{cases} \tag{4.39d}$$

$$h = h_1 + h^{i,j} = w_1 z + h^{i,j}(w_i z)(w_j z) \tag{4.39e}$$

Then we calculate the Lie brackets using (4.36) and (4.37):

$$ad_f B = [B, (Az \ + f^{(2)})] = [v^n, (Az \ + f_k^{i,j}\, v^k(w_i z)(w_j z))]$$

$$= \Big(A \ + f_k^{i,j}\, v^k((w_i z)w_j + (w_j z)w_i)\Big)v^n$$

$$= v^{n-1} + f_k^{i,n}\, v^k(1 + \delta_n^i\ )(w_i z) \tag{4.40}$$

$$ad_f^2 B = ad_f(ad_f B\ )$$

$$= \Big(A + f_k^{i,j}\, v^k((w_i z)w_j + (w_j z)w_i)\Big)\Big(v^{n-1} + f_k^{i,n}\, v^k(1 + \delta_n^i\ )(w_i z)\Big)$$

$$-f_k^{i,n}\, v^k(1 + \delta_n^i\ )w_i A z$$

$$= v^{n-2} + f_k^{i,n}\, v^{k-1}(1+\delta_n^i\ )(w_i z) + f_k^{i,n-1}\, v^k(1+\delta_{n-1}^{\ i}\ )(w_i z) - f_k^{i,n}\, v^k(1+\delta_n^i)(w_{i+1} z)$$

$$= v^{n-2} + v^k\Big(f_{k+1}^{i,n}(1+\delta_n^i\ ) + f_k^{i,n-1}(1+\delta_{n-1}^{\ i}\ ) - f\,_k^{i-1,n}(1+\delta_n^i\ )\Big)(w_i z) \tag{4.41}$$

When the above calculations are repeated for more steps, a general formula can be written for $ad_f^r B$ as follows:

$$ad_f^r B = v^{n-r} + \sum_{j=1}^{n}\sum_{k=1}^{n}\Big\{\ \sum_{i=1}^{r} v^{k+i-r} f_k^{j,n-i-1}(1 + \delta_{n-i+1}^{\ j}\ )$$

$$+ \ \sum_{q=1}^{r-1}\sum_{s=0}^{q-1}(-1)^{q+s}\binom{q}{s}\, v^{k-s}\, f\,_k^{j-q+s,n-r+q+1}\,(1 + \delta_{n-r+q+1}^{\ j-q+s}\ )\Big\}(w_j z)(4.42)$$

Note that in the calculation of (4.40) through (4.42) $O(z)^2$ and higher terms have been ignored. Since we are trying to calculate

$$<dh, ad_f^{r-1}\, B> = O(z)^2 \tag{4.43}$$

we evaluate dh using the expression (4.39e):

$$\frac{\partial h}{\partial z} = w_1 + h^{i,j}\Big((w_i z)w_j + (w_j z)w_i\Big) \tag{4.44}$$

Now, (4.43) must be obtained by multiplying (i.e. taking the inner product of) (4.44) and (4.42) and keeping only the constant and $O(z)^1$ terms. This yields:

$$<dh,ad_f^rB> = \sum_{j=1}^{n} \left\{ h^{j,n-r}(1+\delta_r^j) + \sum_{i=1}^{r} f_{r-i+1}^{j,n-i+1}(1+\delta_{n-i+1}^j) \right.$$

$$\left. + \sum_{q=1}^{r-1}\sum_{s=0}^{q-1}(-1)^{q+s}\binom{q}{s} f_{s+1}^{j-q+s,n-r+q+1}(1+\delta_{n-r+q+1}^{j-q+s}) \right\}(w_j z) \qquad (4.45)$$

Then, using the above equation, one sets the terms inside the curly brackets { } equal to zero for $j = 1, \ldots, n$. Solutions to these equations yield the co-kernel equations. Because of the complicated nature of Eqn. (4.45), the calculation of the co-kernel equations become more difficult as the order of the system increases. However, this fact does not make the numerical calculation of the *actual linearizing solution* (or an approximate linearization) any more difficult, since the exact form of the co-kernels is not necessary for a numerical solution.

4.4 Derivation of the equivalent linear system of equations for the solution with a computer program

In the computer program, we will solve the following linear system:

$$L\begin{bmatrix}\phi^{(2)}\\ \alpha^{(2)}\\ \beta^{(1)}\end{bmatrix} = \begin{bmatrix}f^{(2)}\\ \\ g^{(1)}\end{bmatrix} \qquad (4.46)$$

In (4.46), the coefficients of the second degree terms in the vector field $f^{(2)}(x)$ and $g^{(1)}(x)u$ are obtained using the natural basis introduced in

Section 4.2, and stacked in the right hand side vector in a consistent lexicographic ordering. Similarly, unknown coefficients of the terms $\phi^{(2)}(x)$, $\alpha^{(2)}(x)$, and $\beta^{(1)}(x)$ are stacked in the left hand side vector of unknowns in the same manner. The coefficient matrix L in (4.46) is constant, and it is obtained from the set of generalized homological equations:

$$f^{(2)}(x) = [Fx,\phi^{(2)}(x)] + G\alpha^{(2)}(x) \tag{4.47a}$$

$$g^{(1)}(x)u \; = [Gu,\phi^{(2)}(x)] + G\beta^{(1)}(x)u \qquad \forall \text{ constant } u. \tag{4.47b}$$

When the terms in (4.47) are expressed in the natural basis and calculated, we get the following:

$$[Fx,\phi^{(2)}(x)] = [Fx,\phi^{k}_{ij}] = v^k\Big((w_i x)w_j + (w_j x)w_i\Big)Fx - Fv^k(w_i z)(w_j z) \tag{4.48}$$

We note that:

$$w_j F = \sum_{l=1}^{n} F_j^l w_l \tag{4.49a}$$

$$Fv^k = \sum_{l=1}^{n} F_l^k v^l \tag{4.49b}$$

where F_i^j is the i, j th entry in the matrix F. Then one can write, for each element of the monomial basis φ^k_{ij},

$$\begin{aligned}[Fx,\varphi^k_{ij}] &= \sum_{l=1}^{n} v^k\Big(F_j^l(w_i x)(w_l\, x) + F_i^l(w_j x)(w_l\, x)\Big) - F_l^k v^l(w_i z)(w_j z) \\ &= \sum_{l=1}^{n} F_j^l\, \varphi^k_{il} + F_i^l\, \varphi^k_{jl} - F_l^k\, \varphi^l_{ij}\end{aligned} \tag{4.50}$$

The calculation of the constant coefficient matrix L in Eqn. (4.46) is partly based on the result obtained in (4.50) as follows: The dimension of the monomial basis φ^k_{ij} is equal to $\frac{n^2(n+1)}{2}$. We restrict the ordering of the indices as $1 \le i \le j \le n$; $1 \le k \le n$ and note that in (4.50) under the summation of index l we have to interchange the subscript indices i, l or j, l for the first two terms in the summation in order to keep the ordering of the basis elements consistent. This will yield the new expression:

$$[Fx,\varphi^k_{ij}] = \sum_{l=1}^{i} F_j^{\,l}\,\varphi^k_{li} + \sum_{l=i+1}^{n} F_j^{\,l}\,\varphi^k_{il} + \sum_{l=1}^{j} F_i^{\,l}\,\varphi^k_{lj} + \sum_{l=j+1}^{n} F_i^{\,l}\,\varphi^k_{jl} - \sum_{l=1}^{n} F_l^{\,k}\,\varphi^l_{ij} \tag{4.51}$$

where a summation over indices $1 \le i \le j \le n$; $1 \le k \le n$ is implied. Then, we collect the terms in the right hand side of Eqn. (4.51) under dummy indices with overbars $\bar{i}, \bar{j}, \bar{k}$ or, in other words, regroup the terms under a monomial $\varphi^{\bar{k}}_{\bar{i}\,\bar{j}}$ and sum over the indices $\bar{i}, \bar{j}, \bar{k}$ to get:

$$\begin{aligned}[Fx,\varphi^k_{ij}] = \{ (&\sum_{\bar{i}=1}^{i} F_j^{\bar{i}}\,\delta_i^{\bar{j}} + \sum_{\bar{j}=i+1}^{n} F_j^{\bar{j}}\,\delta_i^{\bar{i}} + \sum_{\bar{i}=1}^{j} F_i^{\bar{i}}\,\delta_j^{\bar{j}} + \sum_{\bar{j}=j+1}^{n} F_i^{\bar{j}}\,\delta_j^{\bar{i}})\delta^k_{\bar{k}} \\ &- F^k_{\bar{k}}\,\delta_i^{\bar{i}}\,\delta_j^{\bar{j}} \}\varphi^{\bar{k}}_{\bar{i}\,\bar{j}} \end{aligned} \tag{4.52}$$

where δ^j_i is the Kronecker delta function. Precisely speaking, given the reverse lexicographic ordering $1 \le i \le j \le n$, $1 \le k \le n$ for the entries in the linear operator representation of the homological equation, Eqn. (4.52) will yield the value of the $(\bar{i},\bar{j}.\bar{k})$th row-$(i,j,k)$th column entry in the matrix L of Eqn. (4.46). To be able to write the right-hand side of the linear set of equations, we express the second degree terms $f^{(2)}(x)$ as:

$$f^{(2)}(x) = \sum_{\substack{1 \le \bar{i} \le \bar{j} \\ 1 \le \bar{k}}}^{n} f_{\bar{k}}^{\bar{i}\,\bar{j}} \varphi_{\bar{i}\,\bar{j}}^{\bar{k}} \tag{4.53}$$

Note that Eqn. (4.52) will yield only some of the entries in L; for finding the entire matrix, we need to consider all the terms present in the generalized homological equations. To this end, we now define a unit vector for $m \times 1$ dimensional vector valued second degree monomials to aid in expressing $\alpha^{(2)}(x)$:

$$\theta_{ij}^{\lambda}(x) = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix} (w_i x)(w_j x) \qquad 1 \le i \le j \le n \;;\; 1 \le \lambda \le m. \tag{4.54}$$

with the $m \times 1$ column vector equal to unity in the λth entry, and zero elsewhere. The expression of $\alpha^{(2)}(x)$ in this basis is:

$$\alpha^{(2)}(x) = \sum_{\substack{1 \le i \le j \le n \\ 1 \le \lambda \le m}} \alpha_{\lambda}^{ij}\, \theta_{ij}^{\lambda} \tag{4.55}$$

Thus, a unit term corresponding to $G\alpha^{(2)}(x)$ in the homological equation (4.47a) becomes, with the aid of a summation,

$$G\theta_{ij}^{\lambda} = \sum_{\substack{1 \le \bar{i} \le \bar{j} \\ 1 \le \bar{k}}}^{n} (G_{\bar{k}}^{\lambda} \delta_{i}^{\bar{i}}\, \delta_{j}^{\bar{j}}\,) \varphi_{\bar{i}\,\bar{j}}^{\bar{k}} \tag{4.56}$$

Next we define the following unit vectors for a similar calculation of the second generalized homological equation (4.47b). Define the unit $n \times 1$ dimensional vector valued first degree monomials as before:

$$\varphi_i^k(x) = v^k(w_i x) \tag{4.57}$$

and the $m \times 1$ dimensional vector valued monomials in (x,u):

$$\pi_{i\mu}^{\lambda}(x,u) = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}^{m \times 1} (w_i x) u_\mu \qquad 1 \le i \le n \;;\; 1 \le \lambda, \mu \le m. \tag{4.58}$$

with the $m \times 1$ column vector equal to unity in the λth entry, and zero elsewhere. Equation (4.47b)

$$g^{(1)}(x)u = [Gu, \phi^{(2)}(x)] + G\beta^{(1)}(x)u \qquad \forall \text{ constant } u. \tag{4.47b}$$

can now be calculated by using the unit vectors as defined previously:

$$[Gu, \varphi_{ij}^k] = \sum_{\mu=1}^{m} G_i^{\mu} u_\mu \varphi_j^k + G_j^{\mu} u_\mu \varphi_i^k$$

$$= \sum_{\substack{1 \le \bar{i}, \bar{k} \le n \\ 1 \le \bar{\mu} \le m}} (G_i^{\bar{\mu}} \delta_j^{\bar{i}} \, \delta_{\bar{k}}^{k} + G_j^{\bar{\mu}} \delta_i^{\bar{i}} \, \delta_{\bar{k}}^{k}) \, \varphi_{\bar{i}}^{\bar{k}} \, u_{\bar{\mu}} \tag{4.59}$$

In addition, the left-hand side of Eqn.(4.47b) is expressed as:

$$g^{(1)}(x)u = \sum_{\substack{1 \le \bar{i}, \bar{k} \le n \\ 1 \le \bar{\mu} \le m}} g_{\bar{k}}^{\bar{i}\,\bar{\mu}} \varphi_{\bar{i}}^{\bar{k}} \, u_{\bar{\mu}} \tag{4.60}$$

We also rewrite the term $\beta^{(1)}(x)u$ as follows:

$$\beta^{(1)}(x)u = \sum_{\substack{1 \le i \le n \\ 1 \le \mu, \lambda \le m}} \beta_{\lambda}^{i\mu} \pi_{i\mu}^{\lambda} \tag{4.61}$$

and $G\beta^{(1)}(x)u$ is represented as:

$$G\pi^{\lambda}_{i\,\mu} = \sum_{\substack{1\le\bar{i},\bar{k}\le n \\ 1\le\bar{\mu}\le m}} G^{\lambda}_{\bar{k}}\delta^{\bar{i}}_{i}\delta^{\bar{\mu}}_{\mu}\varphi^{\bar{k}}_{\bar{i}}\,u_{\bar{\mu}} \tag{4.62}$$

Combining all of the above, we can calculate the coefficient matrix L. We summarize these results in tabular form:

	$\boxed{\phi^{ij}_{k}}$	$\boxed{\alpha^{ij}_{\lambda}}$	$\boxed{\beta^{i\mu}_{\lambda}}$
$\boxed{f^{\bar{i}\ \bar{j}}_{\bar{k}}}$	Eqn. (4.43)	$G^{\lambda}_{\bar{k}}\delta^{\bar{i}}_{i}\delta^{\bar{j}}_{j}$	0
$\boxed{g^{\bar{i}\ \bar{\mu}}_{\bar{k}}}$	$G^{\ \bar{\mu}}_{i}\delta^{\bar{i}}_{j}\delta^{k}_{\bar{k}} + G_{j}^{\ \bar{\mu}}\delta^{\bar{i}}_{i}\delta^{k}_{\bar{k}}$	0	$G^{\ \lambda}_{\bar{k}}\delta^{\bar{i}}_{i}\delta^{\bar{\mu}}_{\mu}$

A dimension count will yield the size of the above matrix as $\dfrac{n^2(n+1)}{2} + n^2 m$ rows by $\dfrac{n^2(n+1)}{2} + \dfrac{mn(n+1)}{2} + m^2 n$ columns, as found before. The above table yields the entries of the coefficient matrix in the linear system of equations.

5. COMPUTER PROGRAM AND SIMULATIONS

5.1. Introduction

A computer program consisting of a collection of routines, or M-files, written in the MATLAB application language was prepared to implement the results of this report. The package is intended for use along with the MATLAB program, and with the Control System Toolbox for MATLAB. It is used for calculating nonlinear coordinate transformation and feedback pairs to linearize control systems up to a specified degree in the series expansion of the nonlinear terms. The current version of the program calculates transformations up to degree two in the series expansion. After obtaining the transformation, one applies any of the standard control design procedures for the resulting model. In the program we have provided tools for feedback pole placement for the linear part of the closed loop system. In addition, one also has the option to choose various forms of inputs and different initial conditions for simulating the resulting system. Comparisons of the performance of the nonlinear feedback with a linear feedback design can be made.

In order to run the program, the folder (or directory) that contains the Approximate Linearization Toolbox for Nonlinear Control Systems should be present in the directory or folder of the MATLAB program and toolboxes. Either the MATLABPATH has to be set appropriately (see MATLAB User's Manual), or after starting MATLAB the subdirectory that contains the M-files of the program has to be the opened or set to be the current working directory. The program is menu-driven, and executing the M-file **mainmenu** by typing its name will present a menu from which the system parameters can be entered, the linearization problem can be solved, and the results can be simulated.

5.2. Using the Program

Prior to running the program, the nonlinear terms in the control system have to be expanded in a series t) obtain

$$\dot{x} = Fx + f^{(2)}(x) + \left(G + g^{(1)}(x)\right)u + O(x,u)^3 \tag{5.1}$$

where $Fx + Gu$ is the linear part of the plant, and $f^{(2)}(x) + g^{(1)}(x)u$ is the second degree part. A characteristic scale for each state should be obtained and entered into the program. For a complete listing of the program code see Appendix 1. An example session recorded during use is also provided in Appendix 2. After a coordinate change and feedback have been found, the user can perform feedback design and simulate the resulting models. The feedback gains are calculated in the program either by specifying the closed loop eigenvalues, or according to a quadratic optimal regulator design procedure to minimize a performance index C:

$$C = min \int_0^\infty (x^T Q x + u^T R u)\mathrm{d}t \tag{5.2}$$

Either procedure yields a set of feedback gains K. In the program, a total of threε different systems are simulated together:

1: A linear model with linear feedback (LMLF)

$$\dot{x} = (F + GK)x + Gr \tag{5.3}$$

where F and G are identical to those of the original nonlinear system, and the same gains K as found above are used to perform pole placement on the model. The term r is a reference input signal. This system is intended to serve as a benchmark to evaluate the performance of the nonlinear feedback synthesis procedure. It is expected that the nonlinear control will drive the system to behave approximately like (5.9).

2: The nonlinear system with a linear feedback law (NSLF):

$$\dot{x} = (F + GK)x + f^{(2)}(x) + g^{(1)}(x)Kx + (G + g^{(1)}(x))r \tag{5.4}$$

This is the original nonlinear system. After the standard first degree approximation, the same feedback gains K as in the LMLF are used. In the simulations, the response of this system is compared to that of the the second degree approximation.

3: The nonlinear system with the linear feedback and the linearizing quadratic feedback law (NSQF):

$$\dot{x} = Fx + f^{(2)}(x) + (G + g^{(1)}(x))(\mathrm{I} + \beta^{(1)}(x))^{-1}\{Kx - K\phi^{(2)}(x) - \alpha^{(2)}(x) + r\} \tag{5.5}$$

This is the nonlinear system on which a quadratic feedback law is applied in addition to the linear feedback with gain K. Note that in (5.5), the reference input r appears in the bracket. In other words, the above is equivalent to (in z coordinates)

$$\dot{z} = (F + GK)z + r \tag{5.6}$$

provided there is no residual (i.e. non-linearizable terms) after the coordinate change and feedback. The nonlinear feedback applied to (5.5) is:

$$u = (\mathrm{I} + \beta^{(1)}(x))^{-1}\{K(x - \phi^{(2)}(x)) - \alpha^{(2)}(x) + r\} \qquad (5.7)$$

In the program, one or more of the response curves of the three models may be plotted and the performances can be compared. The program allows the s'mulation of the above systems with:

a) Zero initial conditions,

b) Impulse input with zero or nonzero initial conditions,

c) Step input with zero or nonzero initial conditions,

d) A sinusoidal forcing funcion as the reference signal with zero or nonzero initial conditions.

In the next section we will present simulations of four different control systems, each chosen to illustrate a particular feature or aspect of the approximate linearization method.

5.3. Example Simulations:

In this chapter, four example systems will be simulated. In the simulations of Examples 1, 2, and 3, we will present cookbook nonlinear systems. The three examples are set up to possess specific properties. The first example is exactly linearizable; the second example is not exactly linearizable but in controller canonical form; and the third example is not exactly linearizable and it has open loop poles very close to the desired closed loop poles. None of these three examples represent physical systems, and each has some isolated property as mentioned above. They are intended to test and prove the effectiveness of the nonlinear control scheme within the framework of each of these special properties.

In contrast, Example 4 for the approximate linearization method is a physical system. The example considered is a satellite in planar earth orbit specified by its position and velocity in polar coordinates. The nonlinear equations of motion are expanded in a Taylor series at a nominal earth orbit, and truncated at the second degree term in the series.

Three response curves will be compared for every simulation: 1) An ideal linear model, 2) The nonlinear system with a linear feedback design, and 3) The nonlinear system with the quadratic feedback based on the method that has been developed. The reason for including the linear model in the comparison is to check the improvement achieved by nonlinear feedback toward making the system respond more linearly. In other words, our evaluation of the effectiveness of the approximate linearization will be based on how close the time response curves of a nonlinear system with quadratic feedback follows the responses of a purely linear system, and how superior this improvement is in comparison with the response of a system with a feedback design based on a first degree approximation.

The following table presents a list of the example simulations, the various initial conditions and disturbance inputs applied to each system, and the figures associated with each simulation.

Example System	Initial Conditions or disturbances	Figures
Example 1	*x* = [0.05 0.05 0.05]'	5.1 through 5.7
	Impulse with unit magnitude	5.8 through 5.14
	Step with magnitude = 0.2	5.15 through 5.14
	Sinusoidal with *r*(*t*) = sin(6*t*)	5.19 through 5.25
Example 2	*x* = [0.35 0.35 0.35]'	5.36 through 5.29
	Impulse with unit magnitude	5.30 through 5.33
	Step with magnitude = 0.5	5.34 through 5.37
	Sinusoidal: *r*(*t*) = sin(6*t*)	5.38 through 5.44
Example 3	*x* = [0.1 0 0]'	5.45 through 5.48
	x = [–0.1 0 0]'	5.49 through 5.52
	x = [0 0 0.1]'	5.53 through 5.56
	x = [0 0 –0.1]'	5.57 through 5.60
	Sinusoidal: *r*(*t*) = sin(6*t*)	5.61 through 5.64
Nearly Circular Satellite (Example 4)		
	x = [2 0 0 0]'	5.65 through 5.69
	x = [–2 0 0 0]'	5.70 through 5.74
	x = [0 0 1 0]'	5.75 through 5.79
	x = [0 0 –1 0]'	5.80 through 5.84
	Sinusoidal: *r*(*t*) = 30,000sin(6*t*)	5.85 through 5.89

5.3.1. Example 1:

For the first example simulation, the following nonlinear system has been used:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0&1&0\\0&0&1\\-3&2&-1\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}+\begin{bmatrix}0\\0\\1\end{bmatrix}u+\begin{bmatrix}x_1{}^2-x_1x_3\\-x_1x_2+x_2{}^2+x_3{}^2\\x_2x_3\end{bmatrix}+\begin{bmatrix}x_1\\0\\x_2\end{bmatrix}u \tag{5.8}$$

This system is exactly linearizable, i.e. it satisfies the conditions of the Hunt-Su theorem. The first and second degree parts of the system (5.14) were entered into the computer program, and the linearizing coordinate change was found to be:

$$z_1 = x_1 - 0.19697x_1{}^2 + x_1x_3 - 0.5x_2{}^2 \tag{5.9a}$$

$$z_2 = x_2 + 4x_1{}^2 - 2.3939x_1x_2 \tag{5.9b}$$

$$z_3 = x_3 + 7x_1x_2 - 2.3939x_1x_3 - 1.3939x_2{}^2 + x_3{}^2 \tag{5.9c}$$

Second and first degree terms in the nonlinear feedback were:

$$\alpha^{(2)}(x) = -1.4091x_1{}^2 + 7x_1x_2 - 2x_1x_3 + 7.1061x_2{}^2 - 0.18182x_2x_3 - x_3{}^2 \tag{5.10}$$

$$\beta^{(1)}(x) = -2.3939x_1 + x_2 + 2x_3 \tag{5.11}$$

The above coordinate change and feedback transform the system (5.14) into:

$$\begin{bmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\end{bmatrix}=\begin{bmatrix}0&1&0\\0&0&1\\-3&2&-1\end{bmatrix}\begin{bmatrix}z_1\\z_2\\z_3\end{bmatrix}+\begin{bmatrix}0\\0\\1\end{bmatrix}v \tag{5.12}$$

with the input v defined as

$$v = \alpha^{(2)}(x) + \left(I + \beta^{(1)}(x)\right)u. \tag{5.13}$$

A closed loop feedback was designed by the quadratic optimal regulator procedure with $Q = I$ and $R = I$. The feedback gains were found as:

$K = [\ 0.1623 \quad 6.9158 \quad 2.9789\ ]$

which resulted in the closed-loop eigenvalues:

$\lambda_{1,2} = -0.7245 \pm 0.8515i$

$\lambda_3 = -2.5299.$

For the following graphs, the curves are:

LMLF (————): Reference Linear Model with Linear Feedback.

NSLF (– · – · – ·): Nonlinear System with the Linear Feedback design.

NSQF (- - - - - - -): Nonlinear System with Quadratic Feedback design.

Figures 5.1, 5.2, and 5.3 show the response curves of the above three systems for states x_1, x_2, and x_3 respectively. Initial conditions are $x_1 = 0.05$; $x_2 = 0.05$; $x_3 = 0.05$ and there is no forcing input. The solid curve of the LMLF behaves as expected from a linear model. The curve of the NSLF goes unstable, while the NSQF tracks LMLF quite successfully. The NSQF rapidly approaches the LMLF at steady state. This is not surprising, since in z coordinates the system is exactly linear. We note a somewhat larger overshoot in the transient of the NSQF, especially apparent in the curve of x_3 in Fig. 5.3. This is a tendency of the NSQF to become unstable in this neighborhood. When the magnitude of the initial conditions were further increased, the response of the NSQF went unstable (not plotted).

In Fig. 5.4, the magnitude of the control effort used for driving each system is shown. The quantity is equal to the linear feedback for LMLF and NSLF, and for the NSQF it includes the quadratic terms as well. This graph offers some good insight in both the transient and the steady state response of the NSQF. An unstable behavior, if any, is more readily apparent in the graph of the control effort. In this first simulation, we notice that the magnitude of this input goes to zero at steady state. In other words, when the disagreement between the linear model and the nonlinear system is small, the magnitude of the input necessary to drive the system is likewise small.

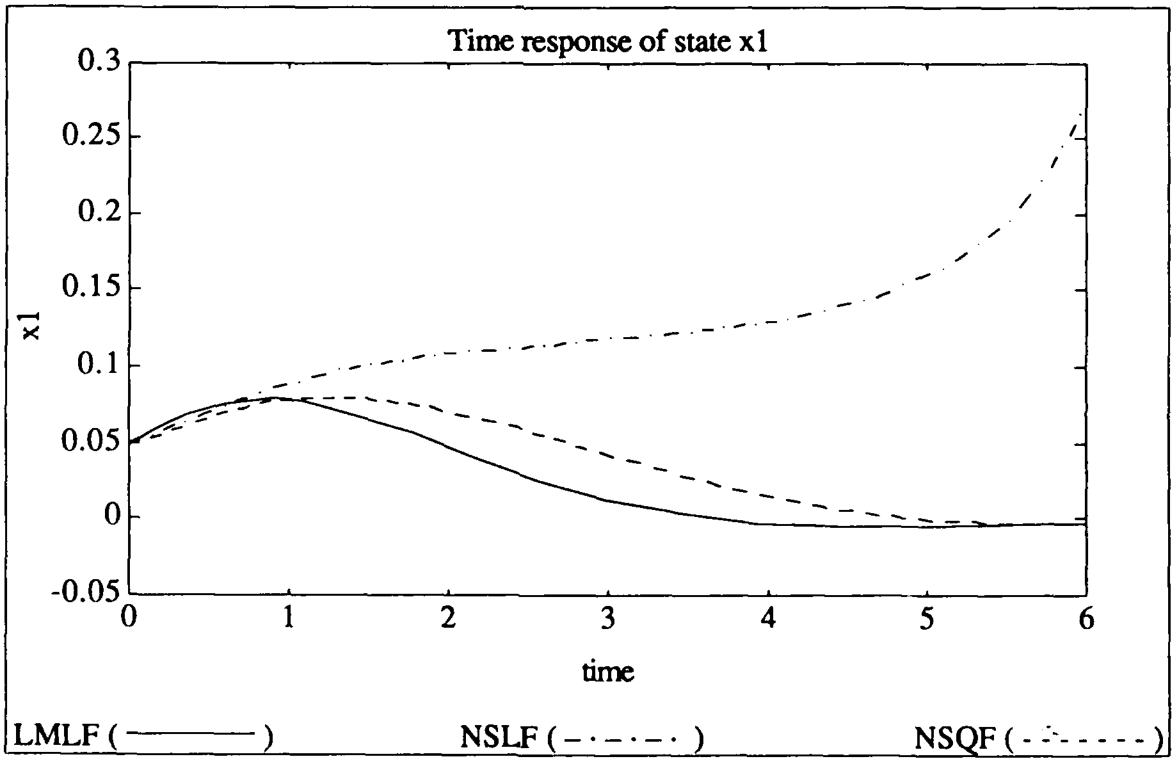


Fig. 5.1. Free response of state x_1 of Example 1 with nonzero initial conditions.

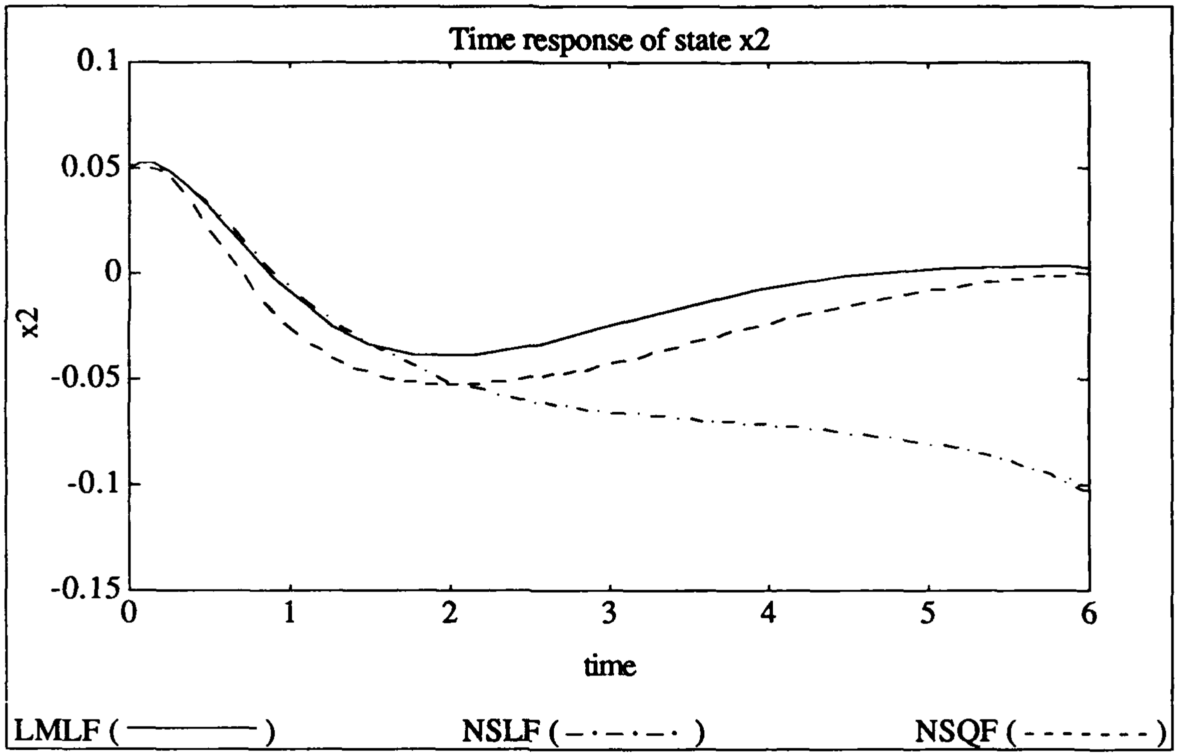


Fig. 5.2. Free response of state x_2 of Example 1 with nonzero initial conditions.

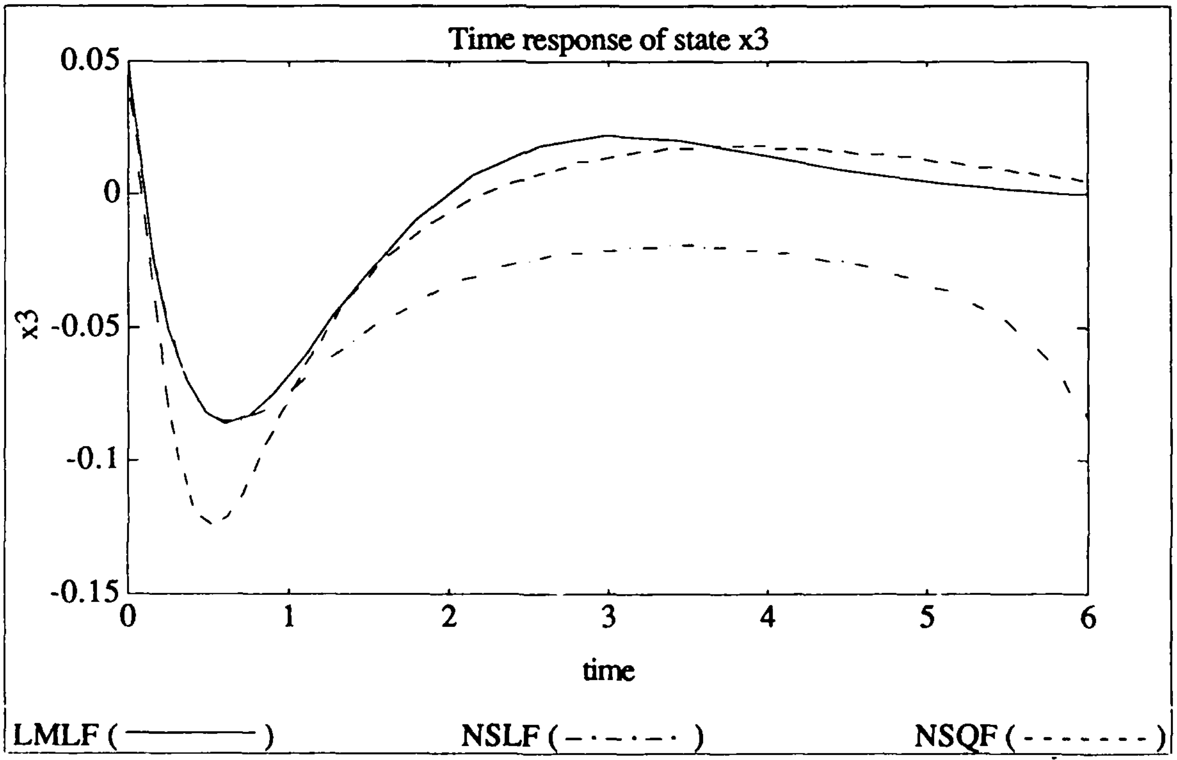


Fig. 5.3. Free response of state x_3 of Example 1 with nonzero initial conditions.

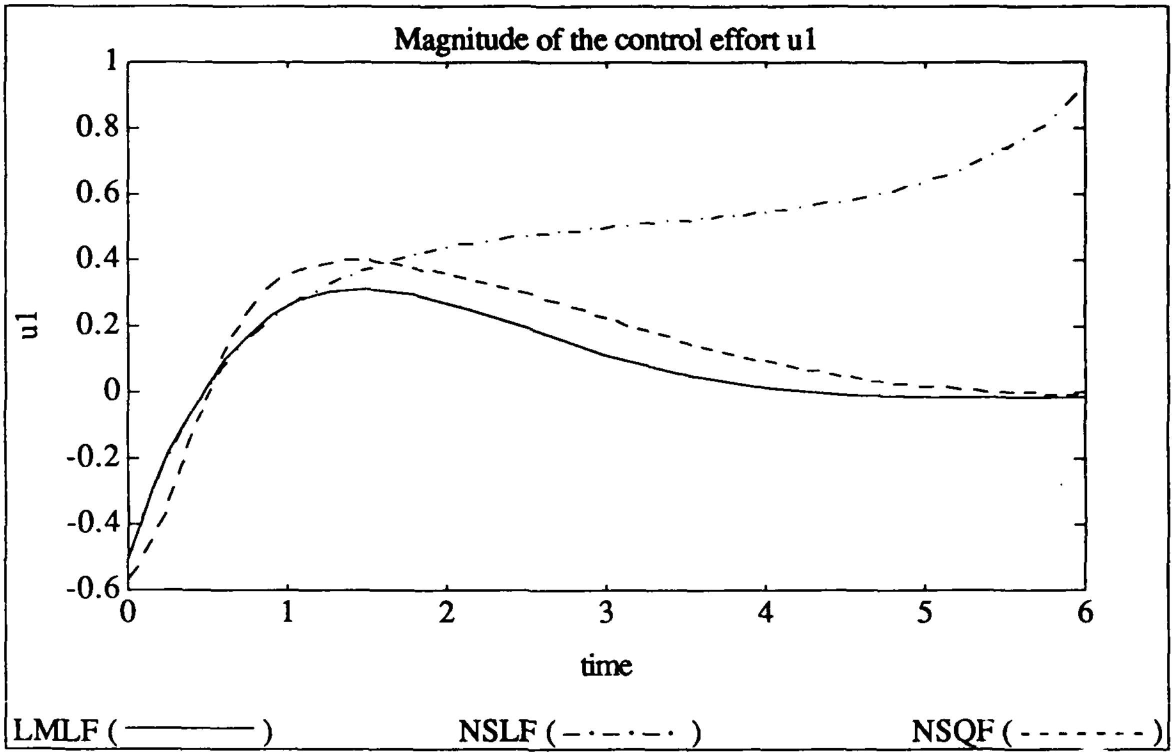


Fig. 5.4. Magnitudes of the inputs for Example 1 with nonzero initial conditions.

Figures 5.5, 5.6, and 5.7 show the phase portrait plots of the same simulation. We note that the plots shown are *projections* of the true three-dimensional phase portrait onto the respective planes shown. All models start at the same initial condition. While LMLF and NSQF decay towards the origin rapidly, the NSLF moves away from the origin, i.e. it clearly shows the tendency of instability.

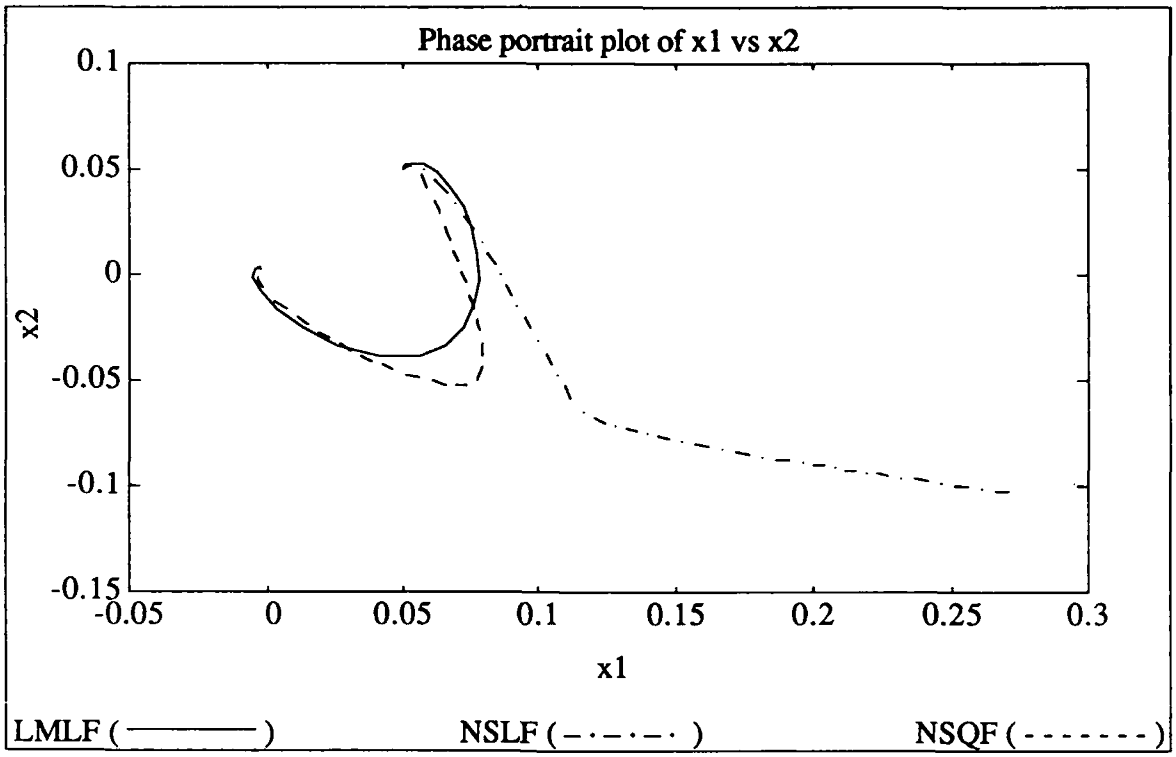


Fig. 5.5. Phase portrait of x_1 versus x_2 for Example 1 with nonzero initial conditions.

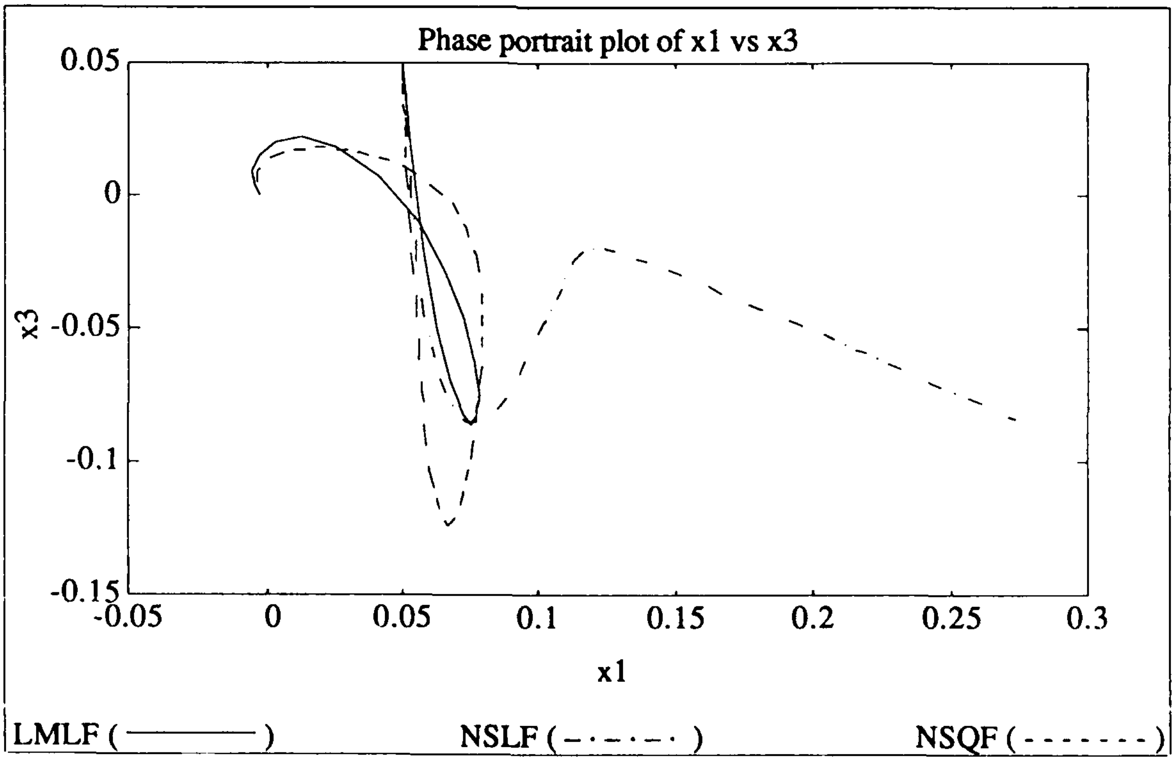


Fig. 5.6. Phase portrait of x_1 versus x_3 for Example 1 with nonzero initial conditions.

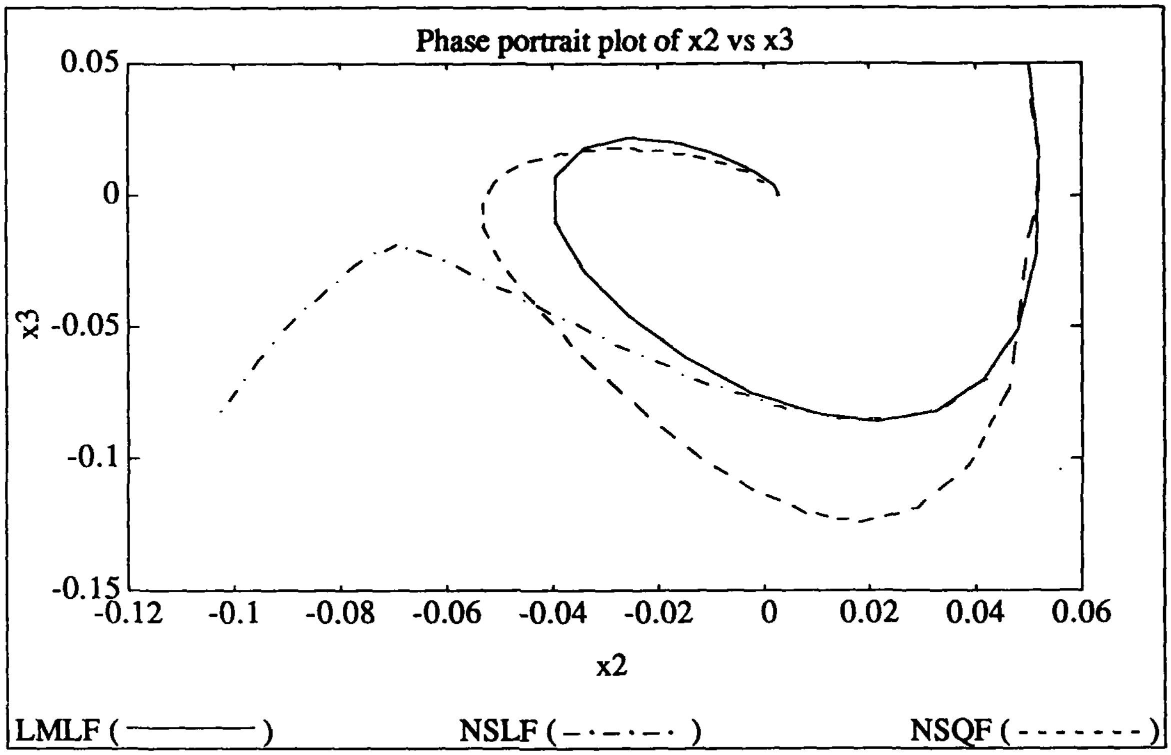


Fig. 5.7. Phase portrait of x_2 versus x_3 for Example 1 with nonzero initial conditions.

Figures 5.8 through 5.14 show the response of the system to an impulse input. All initial conditions have been set equal to zero. In this case, the nonlinear control causes the system to reach instability very rapidly. Note that the instability is more evident in the plot of control feedbacks. Even though the NSQF would achieve good tracking in the neighborhood of the nominal point, as seen in the earlier Figs. 5.1 through 5.7 for the simulation with nonzero initial conditions and no forcing, it has a tendency to decrease stability bounds for some systems. Stability properties of a nonlinear system are very difficult to analyze, and there is no rigorous theory that explains the stability behavior.

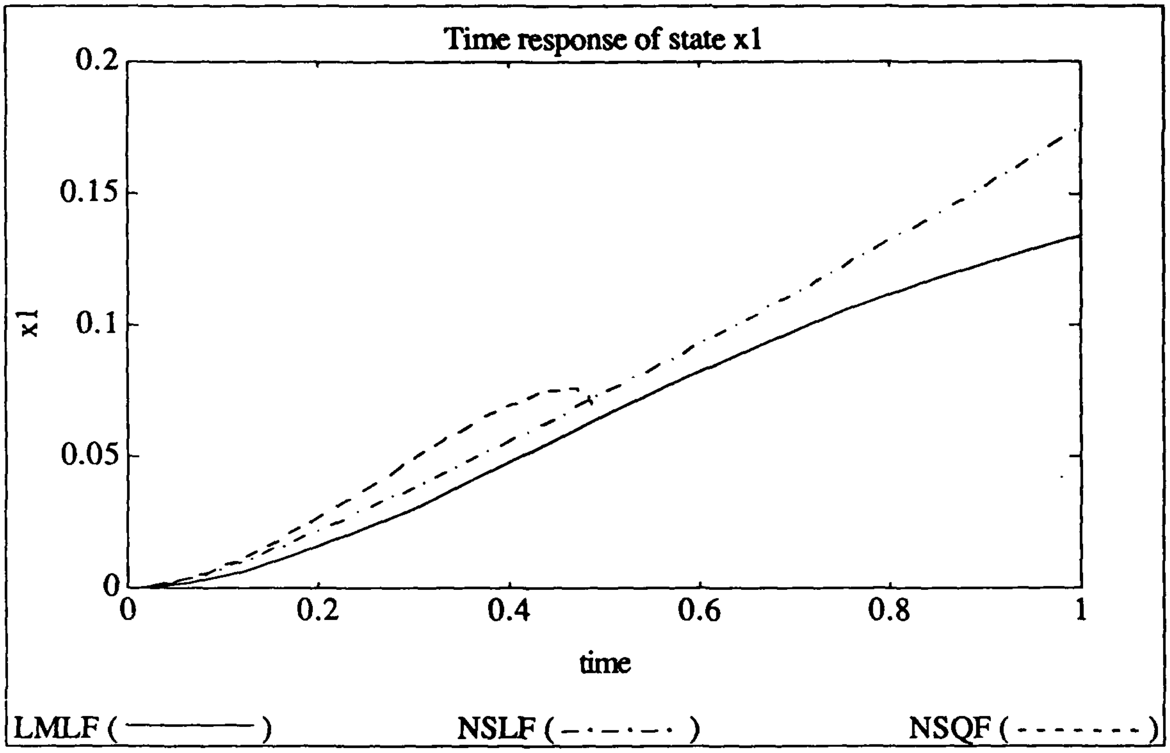


Fig. 5.8. Response of the state x_1 of Example 1 to an impulse input.

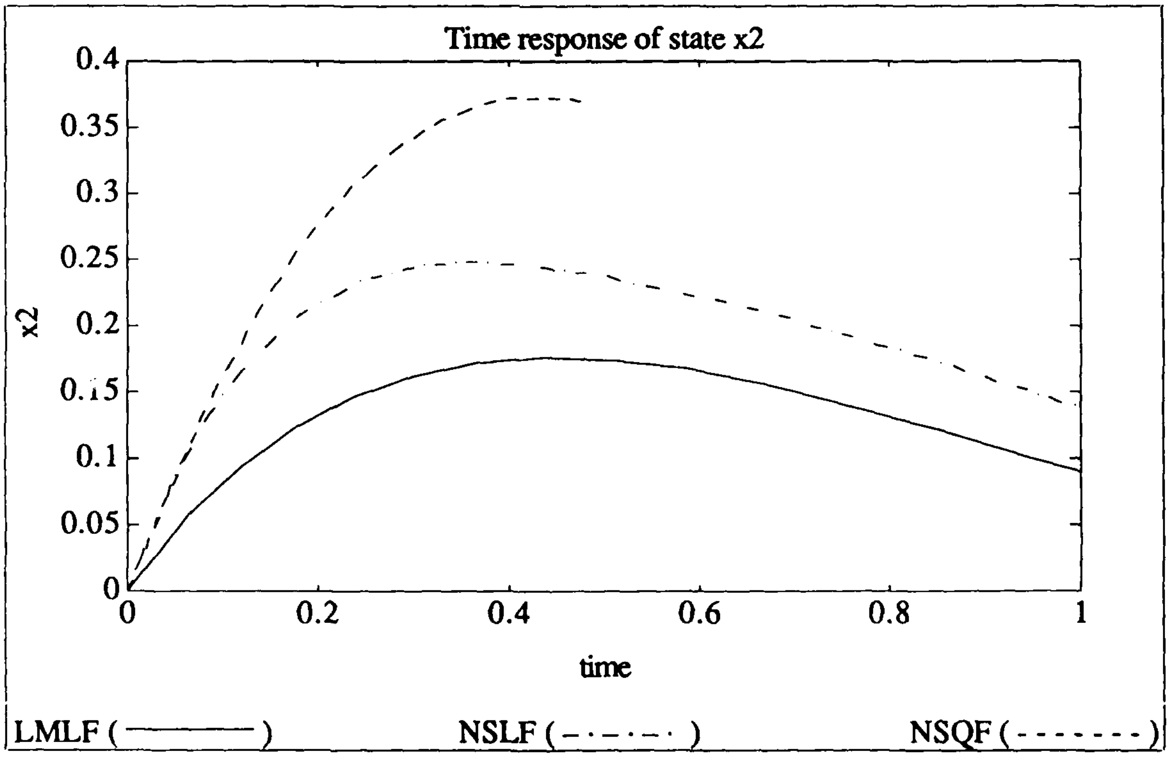


Fig. 5.9. Response of the state x_2 of Example 1 to an impulse input.

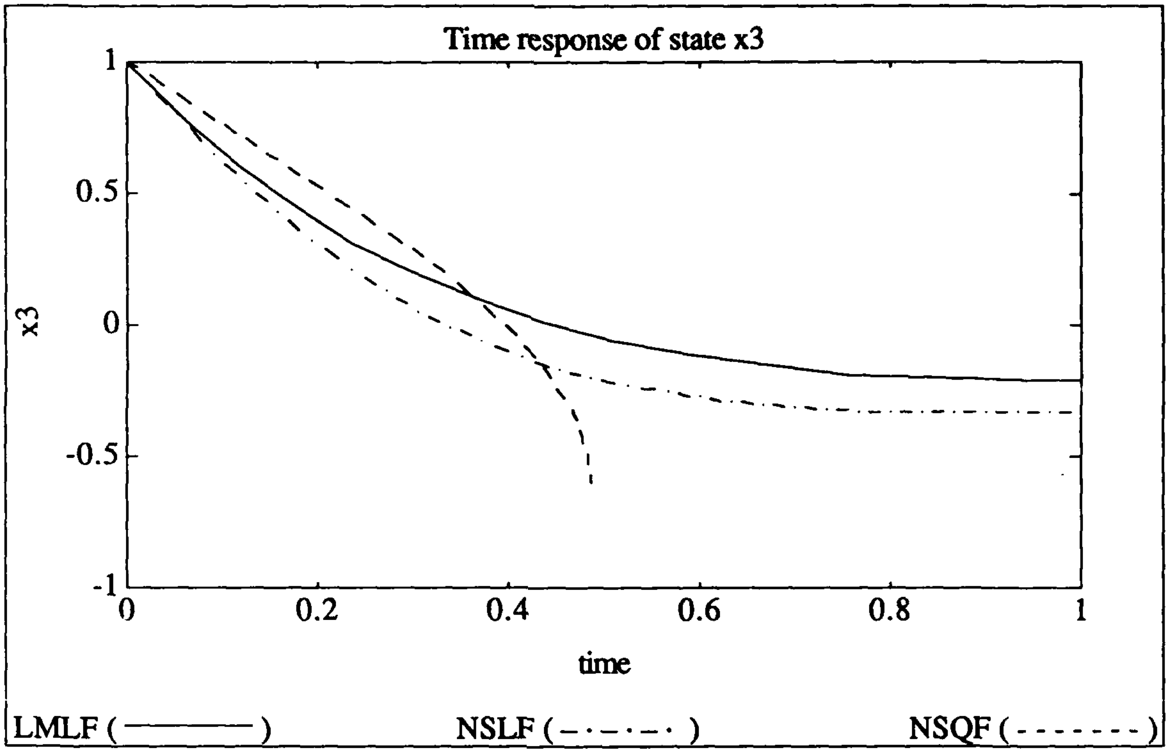


Fig. 5.10. Response of the state x_3 of Example 1 to an impulse input.

Fig. 5.11 for the magnitude of the linearizing control clearly shows the instability of the NSQF in this case. A possible cause of this behavior might be a loss of rank of the term $\left(I + \beta^{(1)}(x)\right)$ (a scalar in this example) which is inverted during calculating the nonlinear feedback (see Eqn. 5.5). Note that, as seen in Eqn. 5.13, this corresponds to a loss of controllability for the nonlinear system. The phase portrait plots for the impulse response are shown in Figs. 5.12 through 5.14.

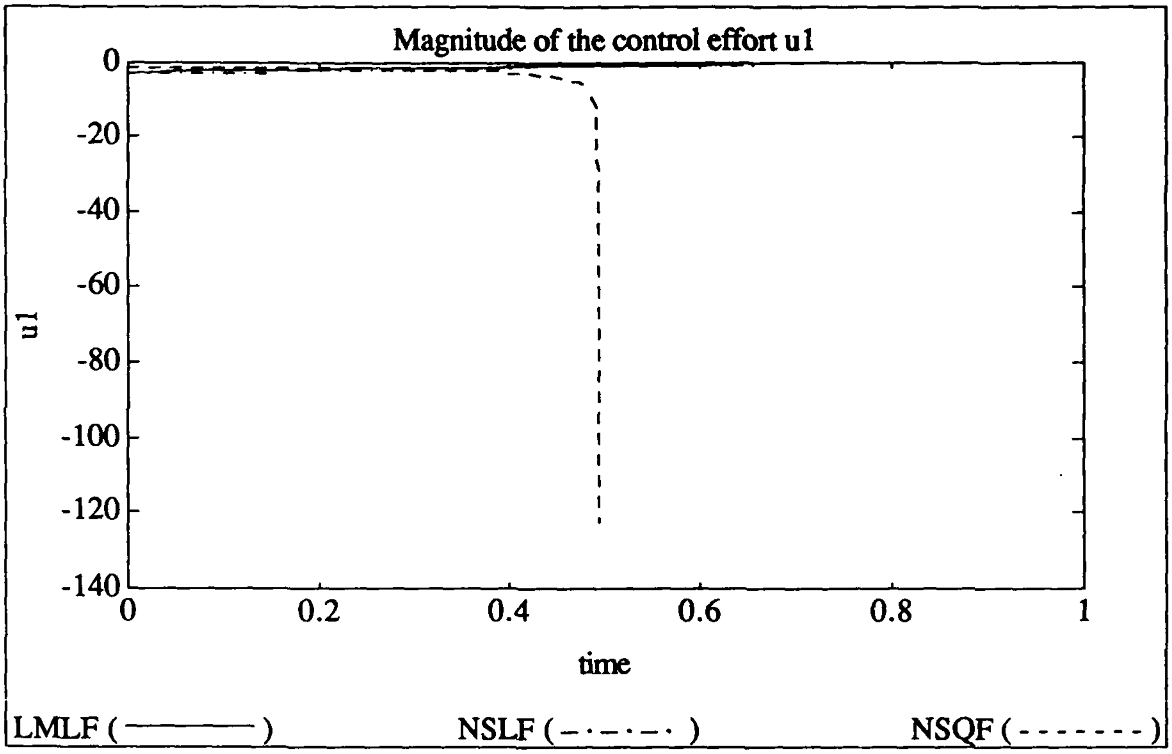


Fig. 5.11. Magnitude of the control inputs for Example 1 for an impulse input.

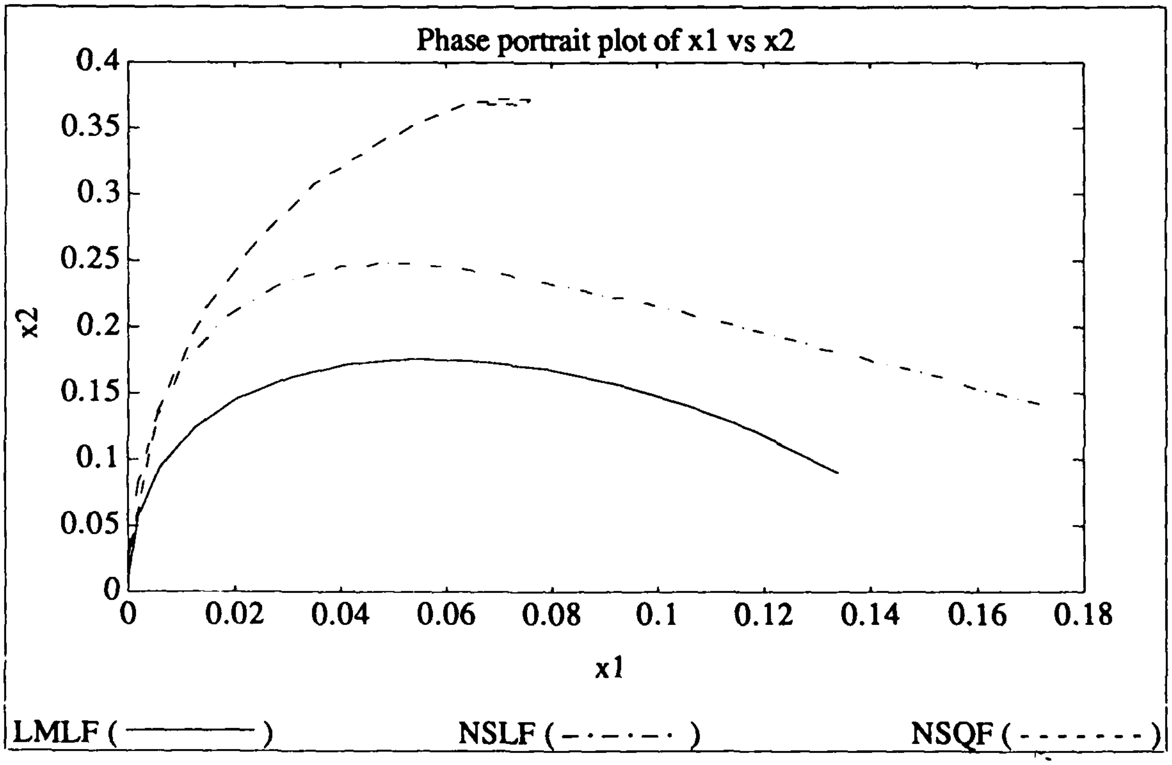


Fig. 5.12. Phase portrait of x_1 versus x_2 for Example 1 for an impulse input.

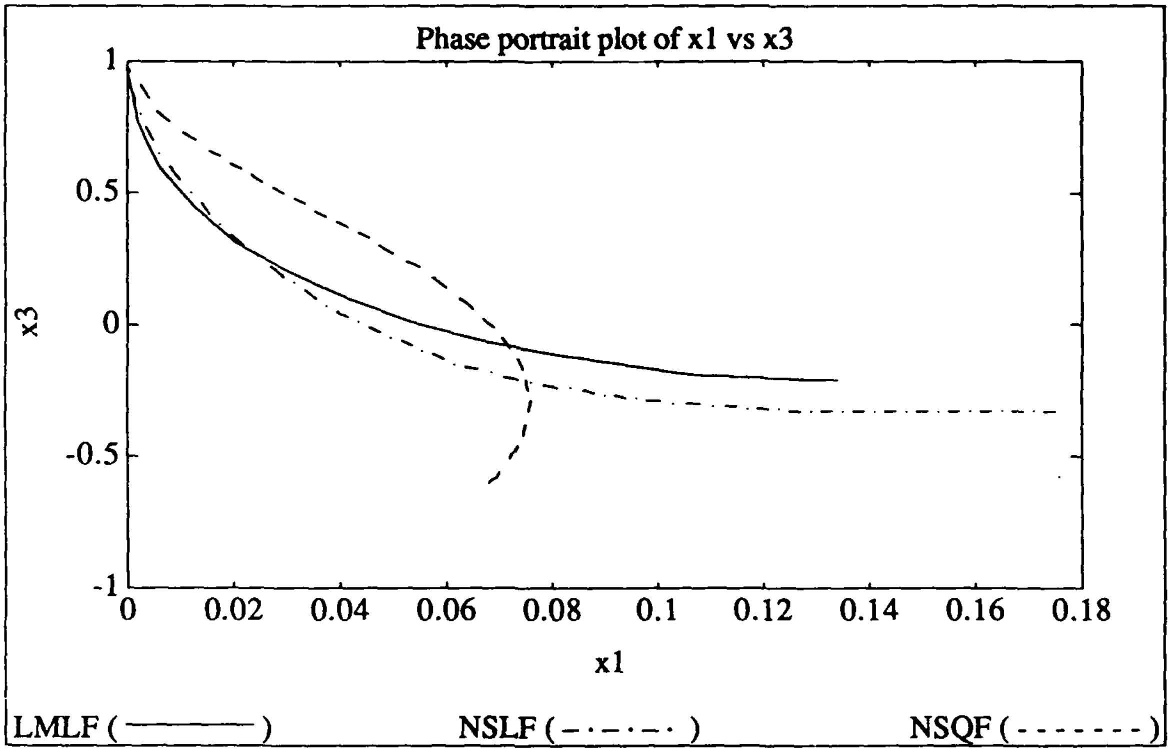


Fig. 5.13. Phase portrait of x_1 versus x_3 for Example 1 for an impulse input.

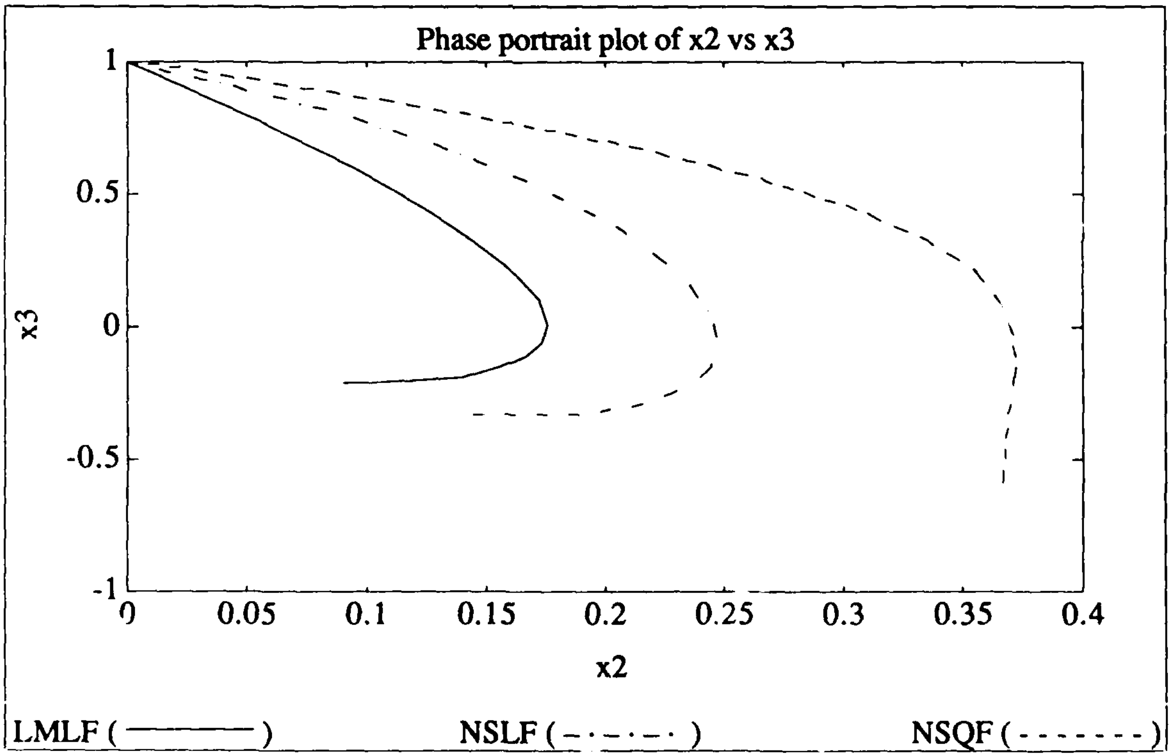


Fig. 5.14. Phase portrait of x_2 versus x_3 fnr Example 1 for an impulse input.

Figures 5.15 through 5.18 show the response of the system to a step input of magnitude 0.2. All initial conditions have been set to zero. In this case, the system NSLF reaches instability very rapidly. The NSQF is not unstable (even though for step inputs of higher amplitude, it would eventually exhibit unstable behavior), but it displays a constant steady state error in tracking the reference step input. Since the phase portrait plots don't offer much insight in this case, they were not plotted.

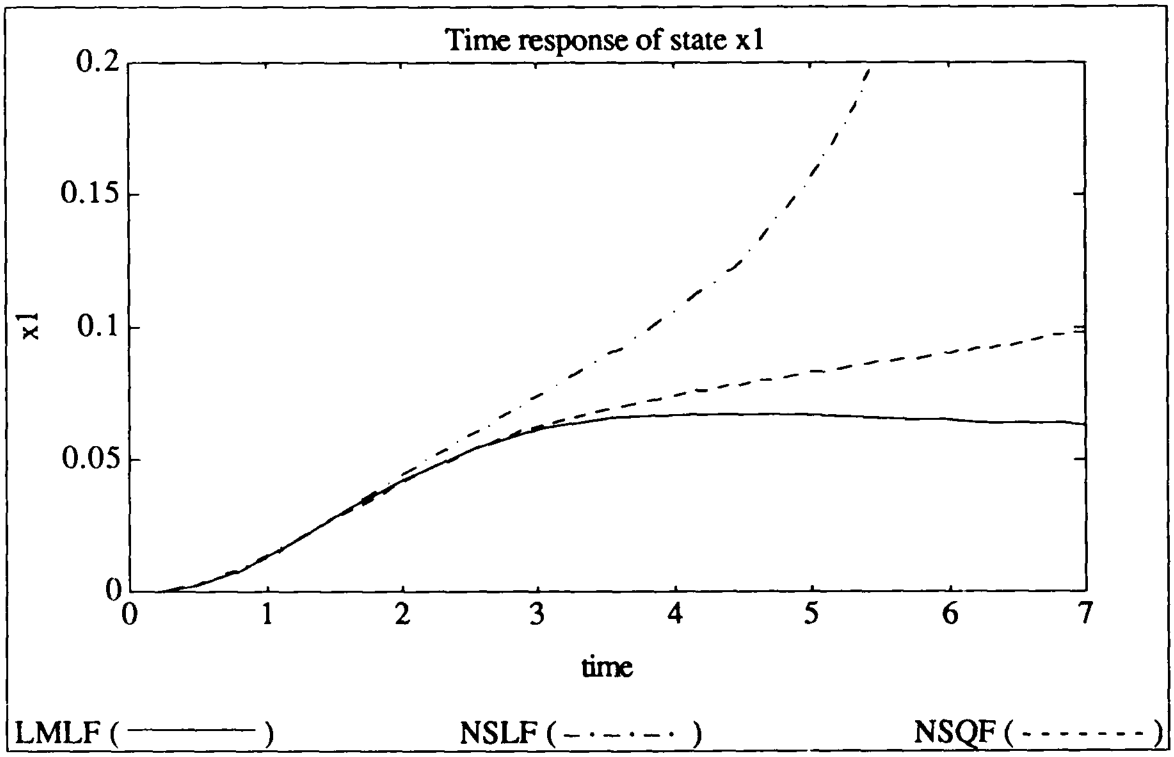


Fig. 5.15. Response of the state x_1 of Example 1 to a step input.

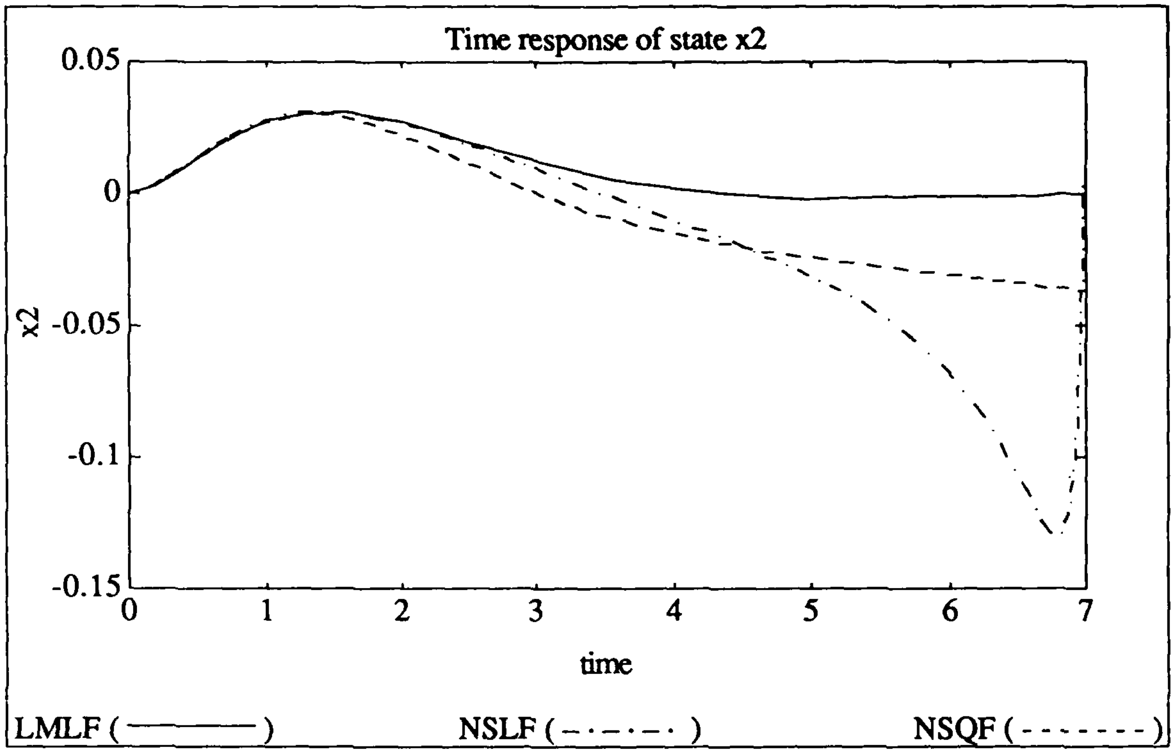


Fig. 5.16. Response of the state x_2 of Example 1 to a step input.

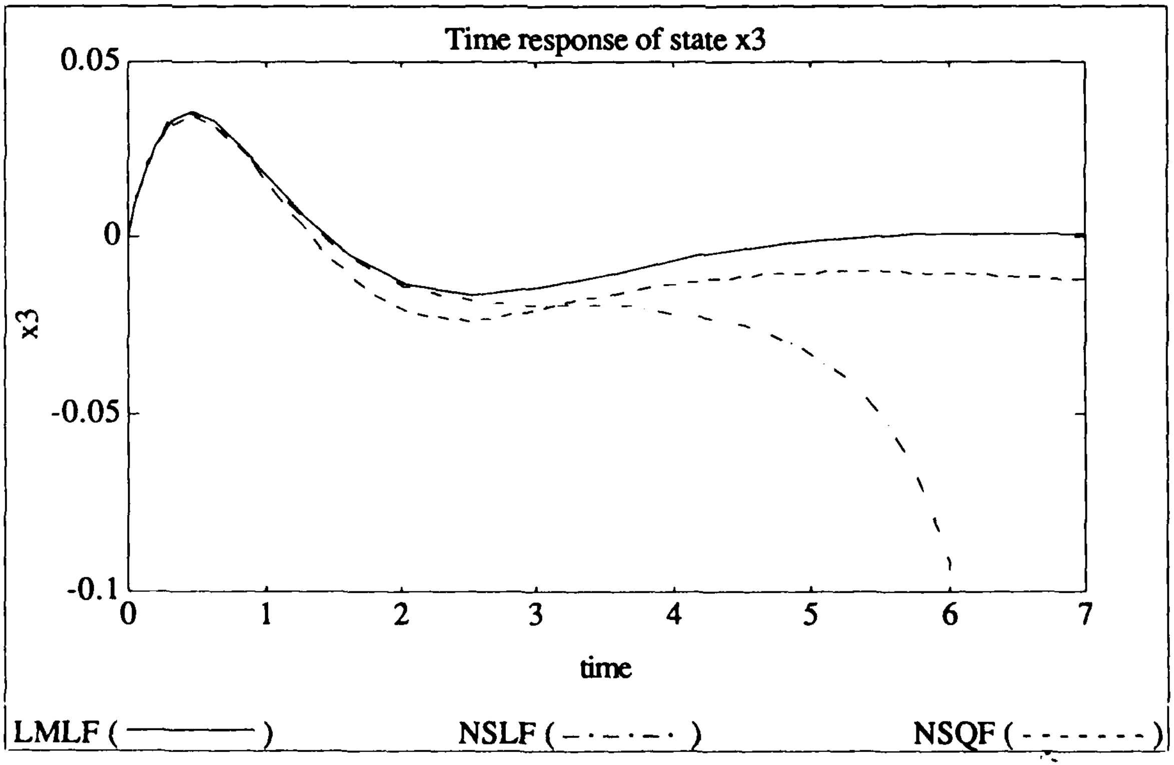


Fig. 5.17. Response of the state x_3 of Example 1 to a step input.

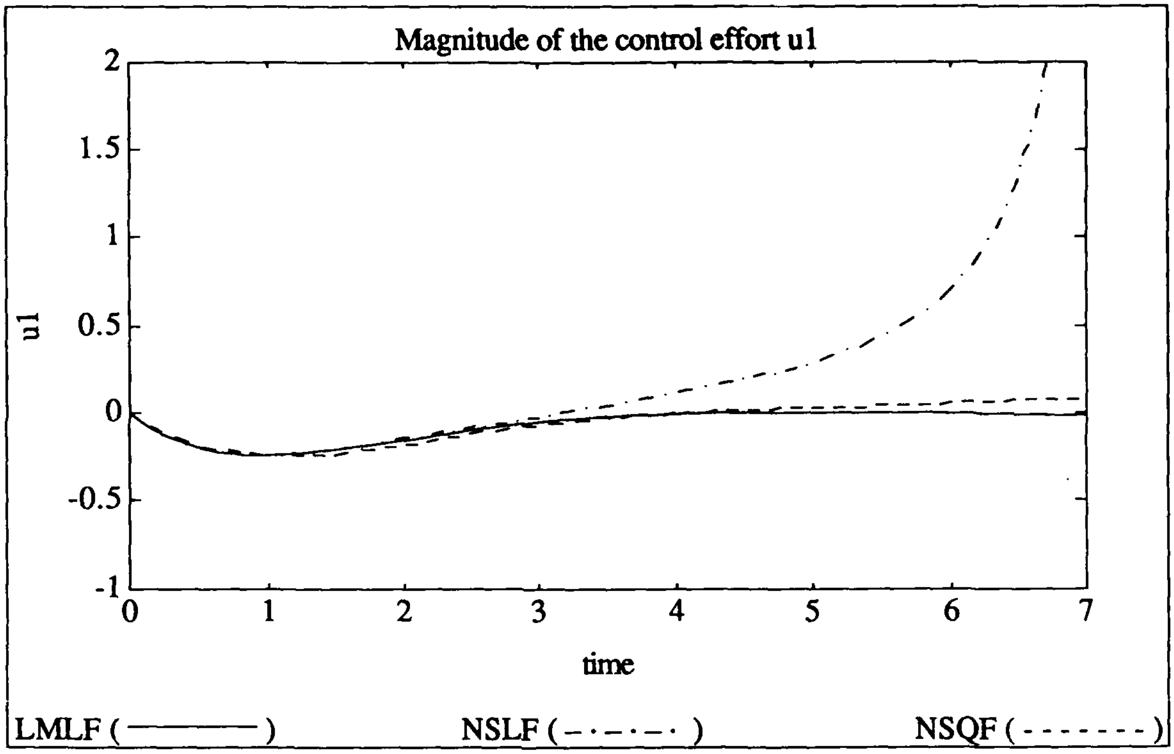


Fig. 5.18. Magnitudes of the control inputs for Example 1 for a step input.

The graphs shown in Figs. 5.19 through 5.25 present the response of the system to a sinusoidal input $u(t) = A\sin(\omega t)$ with the parameters $\omega = 6$, $A = 1$. This simulation is probably the most interesting case in displaying the advantage of the nonlinear feedback. The NSLF has a constant offset away from the equilibrium point. For NSQF, this average error rapidly goes to zero since in z coordinates, the system is exactly linear. After the initial transients die out, the LMLF and the NSQF oscillate around an equilibrium at the origin, i.e. their average is zero.

We should emphasize that the steady-state equilibrium around which each model oscillates is directly related to the average of the nonlinear terms in the vector fields over each period of the sinusoidal input. The strongest case we can make in favor of the NSQF is that this average is extremely close to zero, i.e. it is like that of an exactly linear model. Figs. 5.19, 5.20, and 5.21 show the transient and steady state time responses of all three models. The above argument is clearly obvious in these figures, especially in Fig. 5.19 since the state x_3 is farthest away from the input in terms of integration (note the Controller form of the linear part of the system in Eqn. 5.8).

One should also note that the very impressive improvement achieved by the quadratic control toward causing the system to track the linear model, as seen in Fig. 5.19, is due to the fact that this example is exactly linearizable. A nonlinear feedback and coordinate change that achieves exact linearization could as well be calculated using the Hunt-Su linearization method. However, in our approach, we also minimized the length of the vector formed by the coefficients of $\phi^{(2)}$, $\alpha^{(2)}$, and $\beta^{(1)}$. In contrast, the Hunt-Su theorem will not, in general, yield the minimum solution (in the sense that we have defined in Ch. 3) for the coordinate change and feedback.

For $n = 3$ the Hunt-Su method yields a one-parameter family of solutions to the approximate linearization problem for systems that are exactly linearizable up to degree 2, and the choice for the free parameter remains to be determined. The correct choice for this parameter in order to obtain the "smallest" coordinate change and feedback is precisely the solution found by our method. Therefore, the specific nonlinear coordinate change-nonlinear feedback pair calculated for this example is expected to yield a better performance during the transient response. In the steady state, the difference in performance between the coordinate change and feedback we have found and other solutions may not be appreciable if the trajectories are too close to the origin. The nature of our approach allows us to immediately extend the method to systems that are not exactly linearizable (see examples 2, 3, and 4).

Fig. 5.22 shows the plot of the feedback terms, where the feedback for NSQF has an offset from the equilibrium. This clearly indicates that the nonlinear input drives the system to a zero average by introducing a bias into the system. In Figs. 5.23, 5.24, and 5.25 we present the phase portrait plots of the three models. The steady state equilibrium points are again clearly seen, especially in the graph of x_1 vs x_2 where both the LMLF and the NSQF oscillate around the origin. The NSLF settles around a non-zero equilibrium.

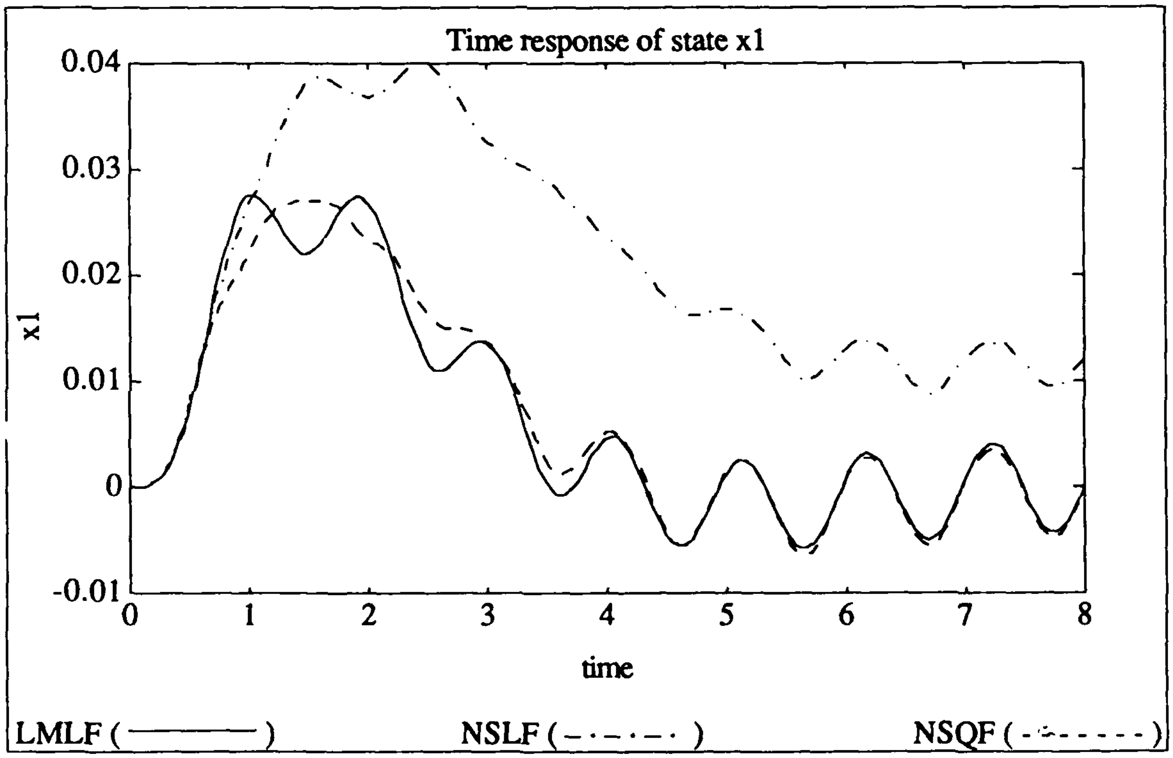


Fig. 5.19. Response of the state x_1 of Example 1 to a sinusoidal input.

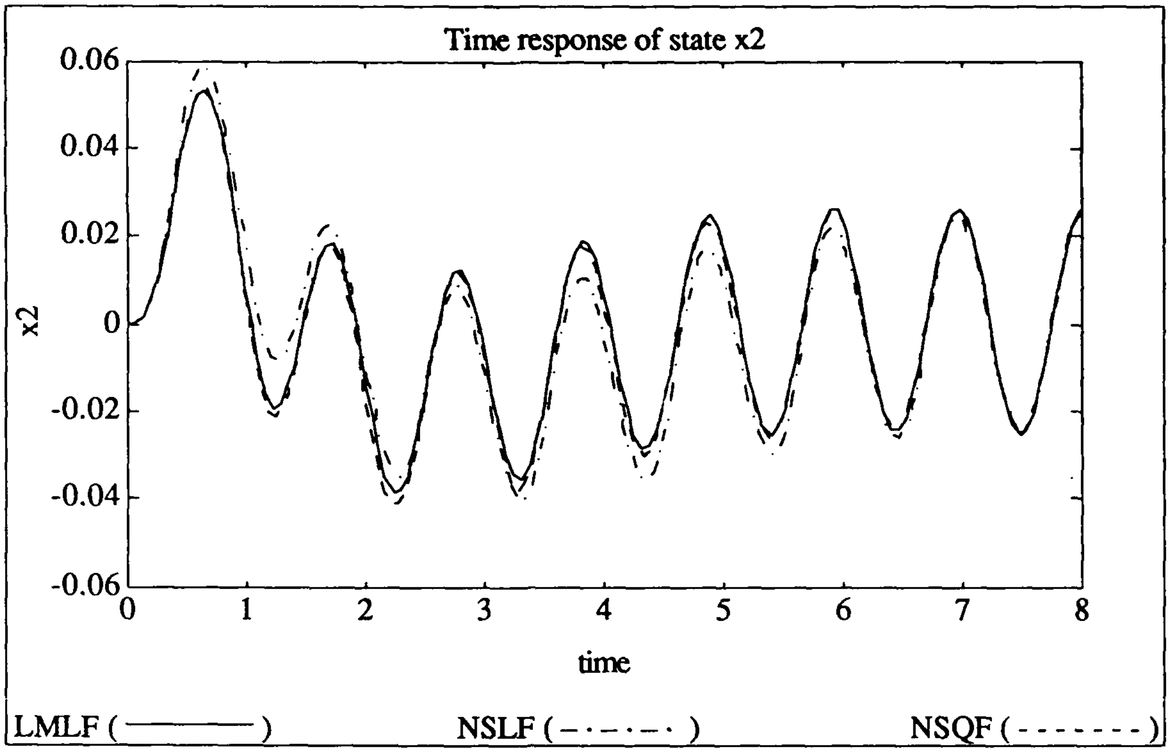


Fig. 5.20. Response of the state x_2 of Example 1 to a sinusoidal input.

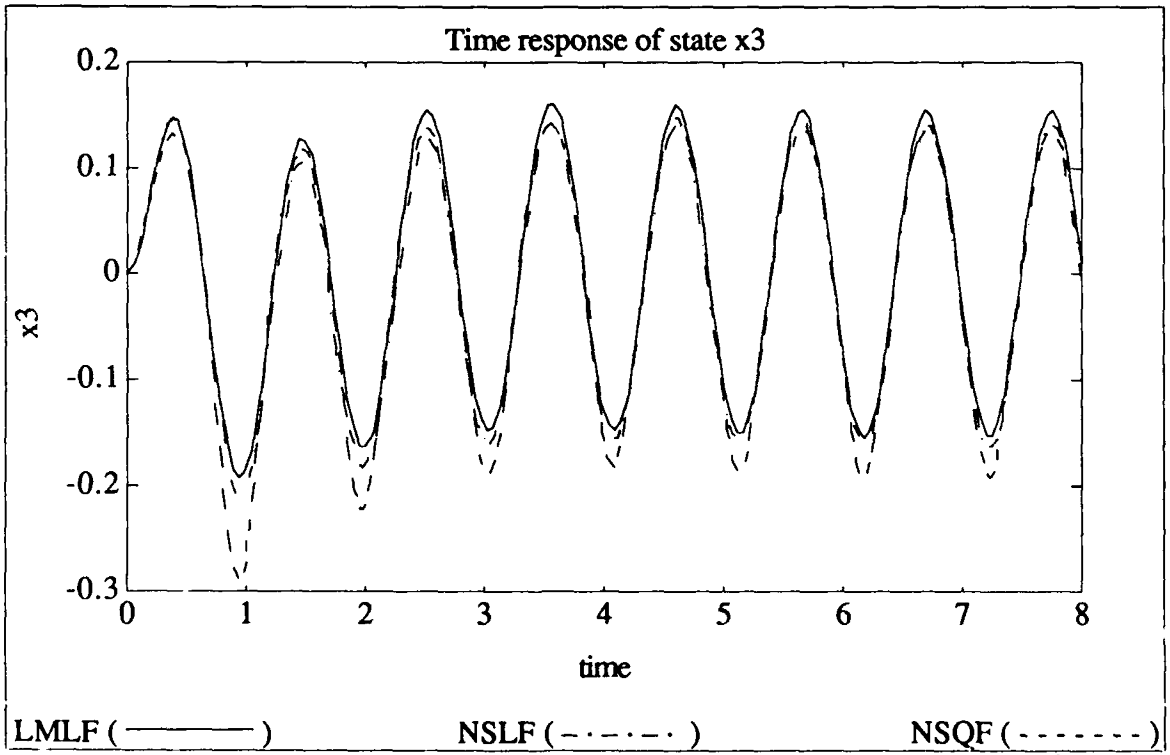


Fig. 5.21. Response of the state x_3 of Example 1 to a sinusoidal input.

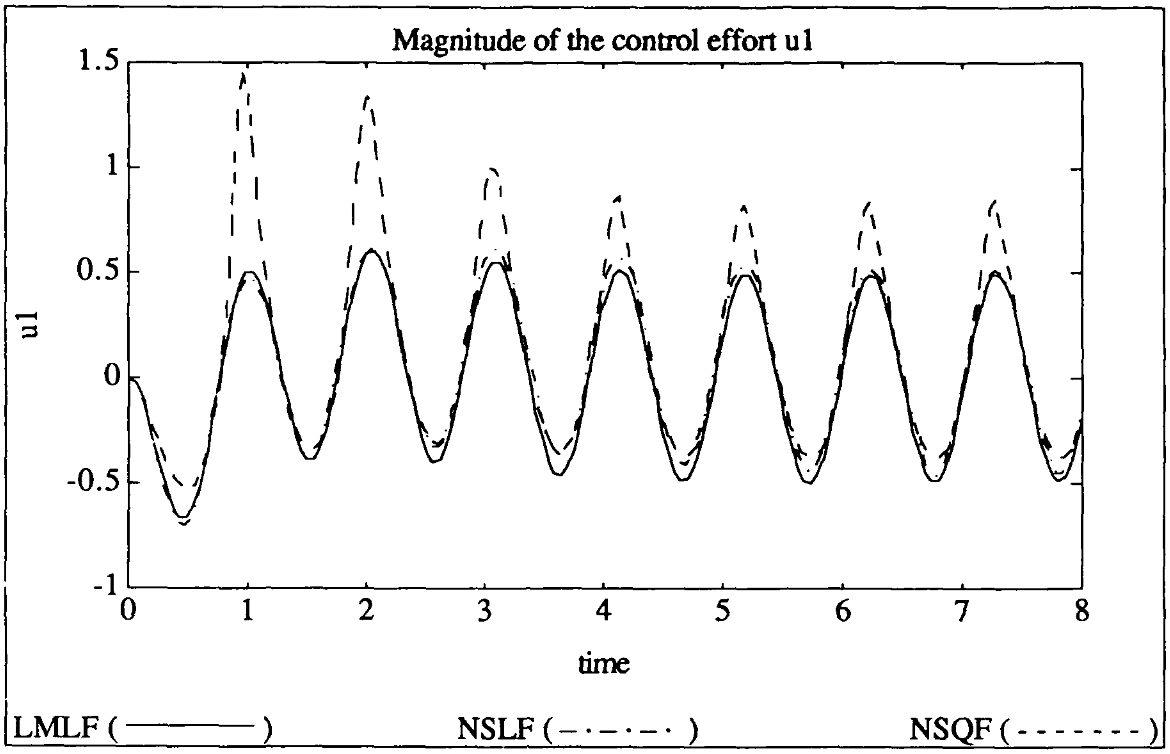


Fig. 5.22. Magnitude of the control efforts for Example 1 for a sinusoidal input.

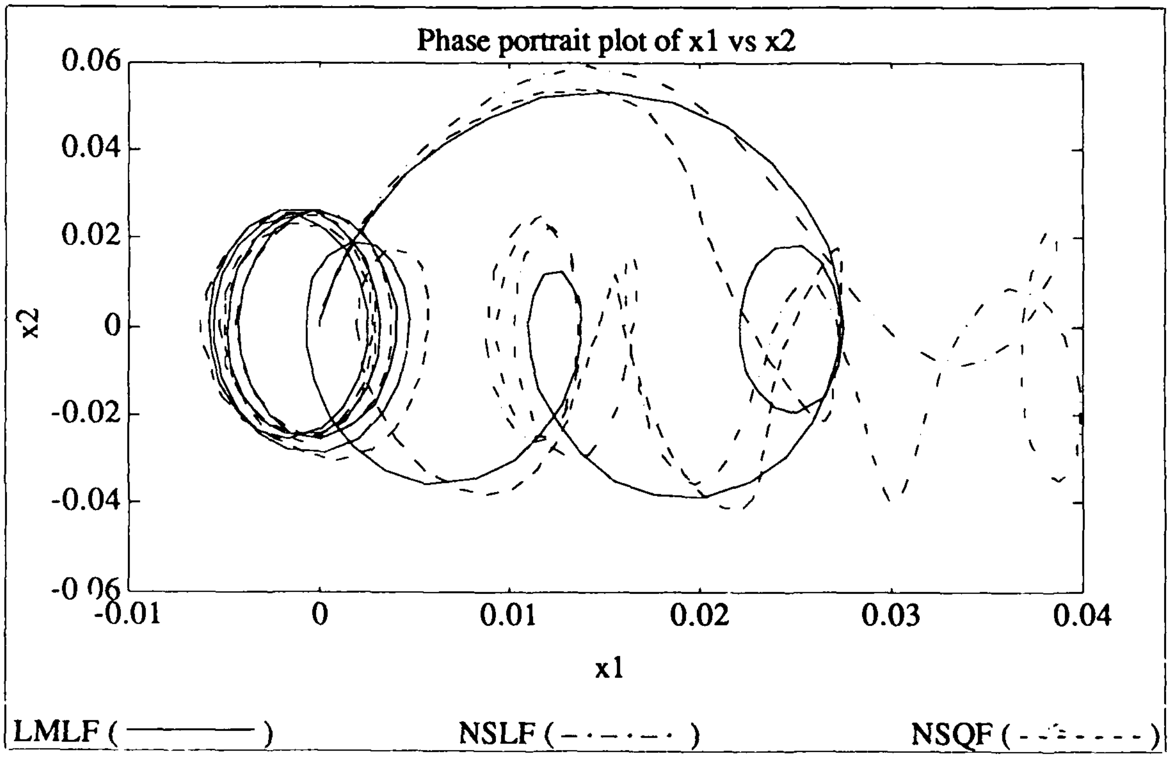


Fig. 5.23. Phase portrait of x_1 versus x_2 for Example 1 for a sinusoidal input.

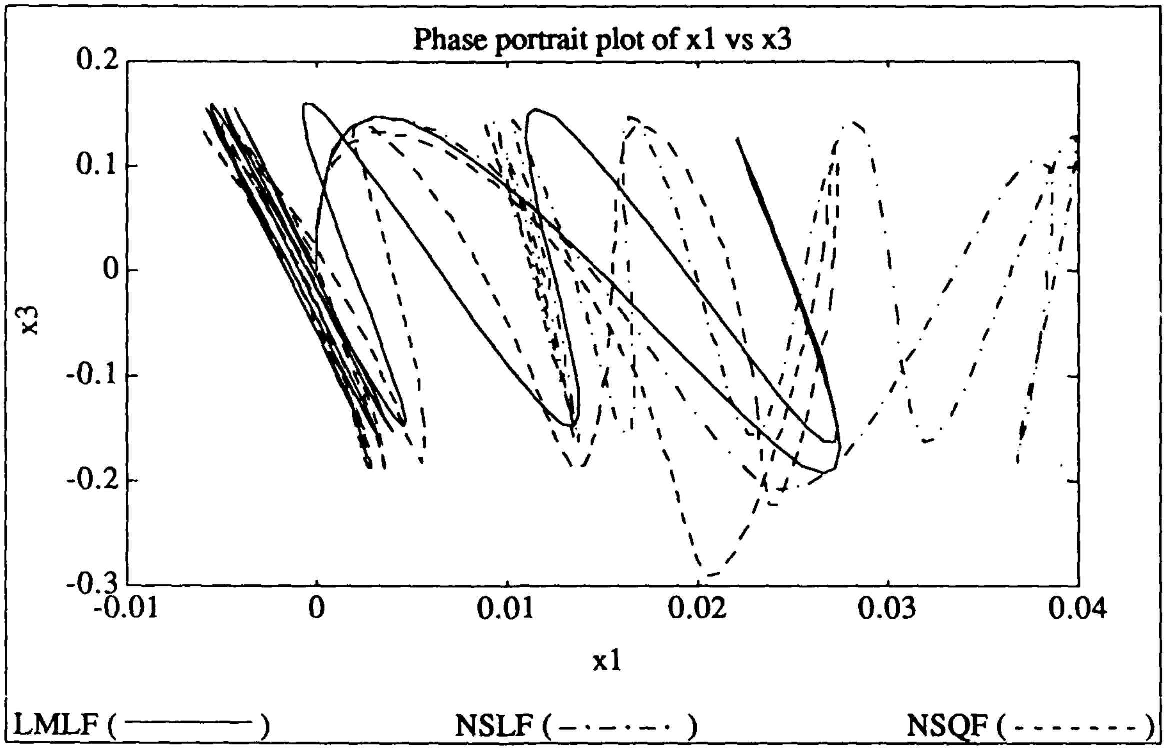


Fig. 5.24. Phase portrait of x_1 versus x_3 for Example 1 for a sinusoidal input.

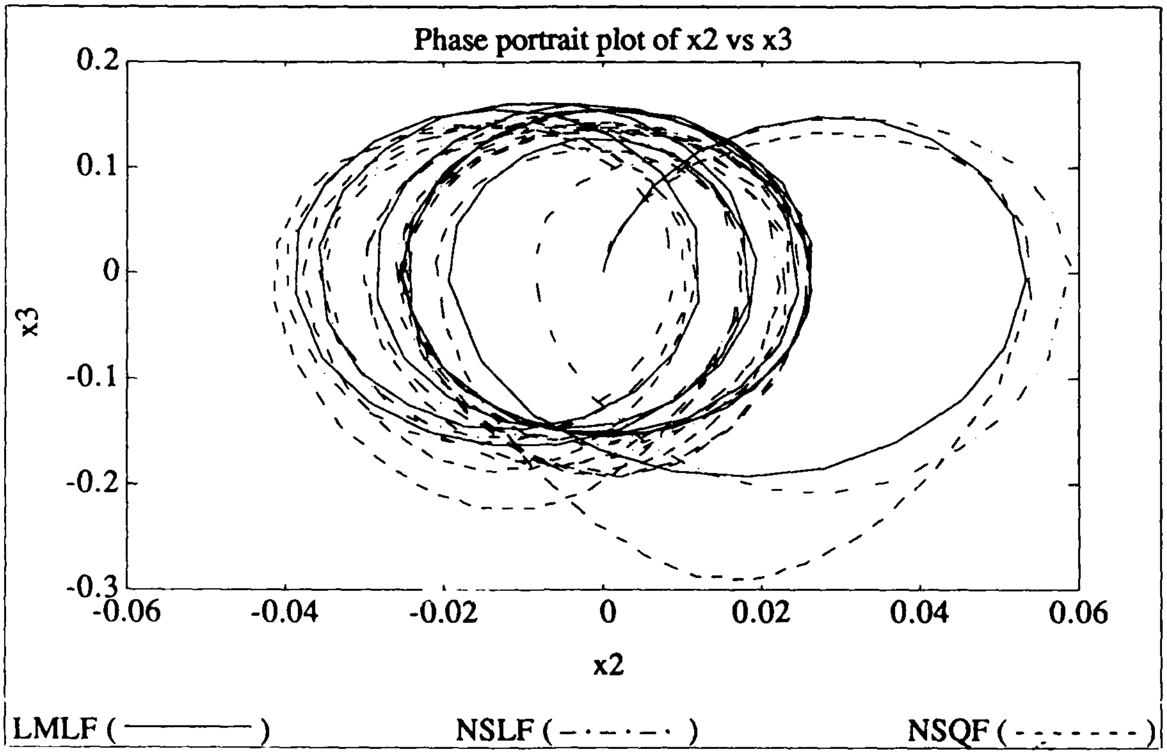


Fig. 5.25. Phase portrait of x_2 versus x_3 for Example 1 for a sinusoidal input.

5.3.2. Example 2:

As a second example we chose a nonlinear control system that is not exactly linearizable:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u + \begin{bmatrix} x_3{}^2 \\ x_1{}^2 + x_1 x_2 \\ 0 \end{bmatrix} \qquad (5.14)$$

The following coordinate transformation and feedback pair were found:

$$z_1 = x_1 - 0.47665x_1{}^2 + 0.11111x_1x_2 + 0.28221x_2{}^2 - 0.16667x_1x_3 \\ - 0.61997x_2x_3 + 0.30999x_3{}^2 \qquad (5.15a)$$

$$z_2 = x_2 - 0.95331x_1x_2 + 0.055556x_2{}^2 + 0.11111x_1x_3 + 0.34219x_2x_3 \\ + 0.32447x_3{}^2 \qquad (5.15b)$$

$$z_3 = x_3 + x_1{}^2 - x_1x_2 + 0.15781x_2{}^2 + 0.046694x_1x_3 - 0.028971x_2x_3 \\ + 0.48228x_3{}^2 \qquad (5.15c)$$

The terms in the nonlinear feedback were:

$$\alpha^{(2)}(x) = x_1x_3 + 0.30675x_2x_3 - 0.084526x_3{}^2 \qquad (5.16)$$

$$\beta^{(1)}(x) = -0.89775x_1 + 0.36997x_2 + 0.96336x_3 \qquad (5.17)$$

The above coordinate change and feedback found will *exactly* transform the following linearizable system (rather than (5.14)) into z coordinates:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u$$

$$+ \begin{bmatrix} 0.05555x_1+0.05555x_2 \\ -0.05555x_1-0.08333x_2-0.02777x_3 \\ 0.05555x_1+0.08333x_2+0.02777x_3 \end{bmatrix} u$$

$$+ \begin{bmatrix} -0.05555x_2^2-0.05555x_2x_3+0.9444x_3^2 \\ x_1^2+x_1x_2+0.05555x_2^2+0.05555x_2x_3+0.05555x_3^2 \\ -0.05555x_2^2-0.05555x_2x_3-0.05555x_3^2 \end{bmatrix} \tag{5.18}$$

i.e, (5.18) will transform with (5.15), (5.16) and (5.17) into z coordinates as

$$\dot{z} = Fz + Gv \tag{5.19}$$

with the input defined with

$$v = \alpha^{(2)}(x) + \left(\mathrm{I} + \beta^{(1)}(x)\right)u. \tag{5.20}$$

The next step in the computation process was designing a closed loop feedback for the linear part of the plant. The feedgack gains were calculated by the program with a quadratic optimal regulator design procedure. The matrix Q was taken to be the identity, and R was set to unity. The feedback gains were thus found to be:

$$K = [1 \quad 1.7321 \quad 1] \tag{5.21}$$

and these gains placed the eigenvalues of the linear part of the closed loop system at:

$\lambda_{1,2} = -0.866 \pm 0.5j$

$\lambda_3 = -1.$

Since we designed the closed loop feedback gains for the linearized system in z coordinates, we obtained the equivalent feedback u in x cordinates by using (5.1). This feedback was calculated as:

$$u = (I + \beta^{(1)}(x))^{-1}\{K(x - \phi^{(2)}(x)) - \alpha^{(2)}(x) + r\} \tag{5.7}$$

For the following graphs, similar to the first simulation, the curves are:

LMLF (————): Reference Linear Model with Linear Feedback.

NSLF (– · – · – ·): Nonlinear System with the Linear Feedback design.

NSQF (- - - - - - -): Nonlinear System with Quadratic Feedback design.

The plots in Figs. 5.26 through 5.29 show the response of the models to an initial condition of $x_1 = 0.35$; $x_2 = 0.35$; $x_3 = 0.35$ and a zero reference input. In Figs. 5.26, 5.27, and 5.28 the time responses of the states x_1, x_2, and x_3 are plotted, respectively. The NSQF shows some tendency toward instability until $t = 1$ (a simulation for slightly increased values for the initial conditions, not plotted, displayed a singularity in this neighborhood). After the NSQF recovers from this region, it tracks the LMLF extremely closely. The NSLF curve, on the other hand, settles down to the same value much later, and displays some amount of overshoot.

The improvement achieved by the NSQF in responding like a linear system, especially as seen in Fig. 5.26, makes a very strong case for our method. Despite the fact that the example is not an exactly linearizable system, there is almost an order of magnitude difference between the NSQF and the NSLF in percent deviation from the ideal linear model. The time that the NSQF settles at zero equilibrium is also very close to that of the LMLF, whereas the NSLF displays transient behavior for a much longer period of time.

The curves of the inputs of Fig. 5.29 shows the steep increase in the value of the control effort between $t = 0.5$ and $t = 1$, the region in which the NSQF has a tendency of instability. The inputs settle to zero afterwards.

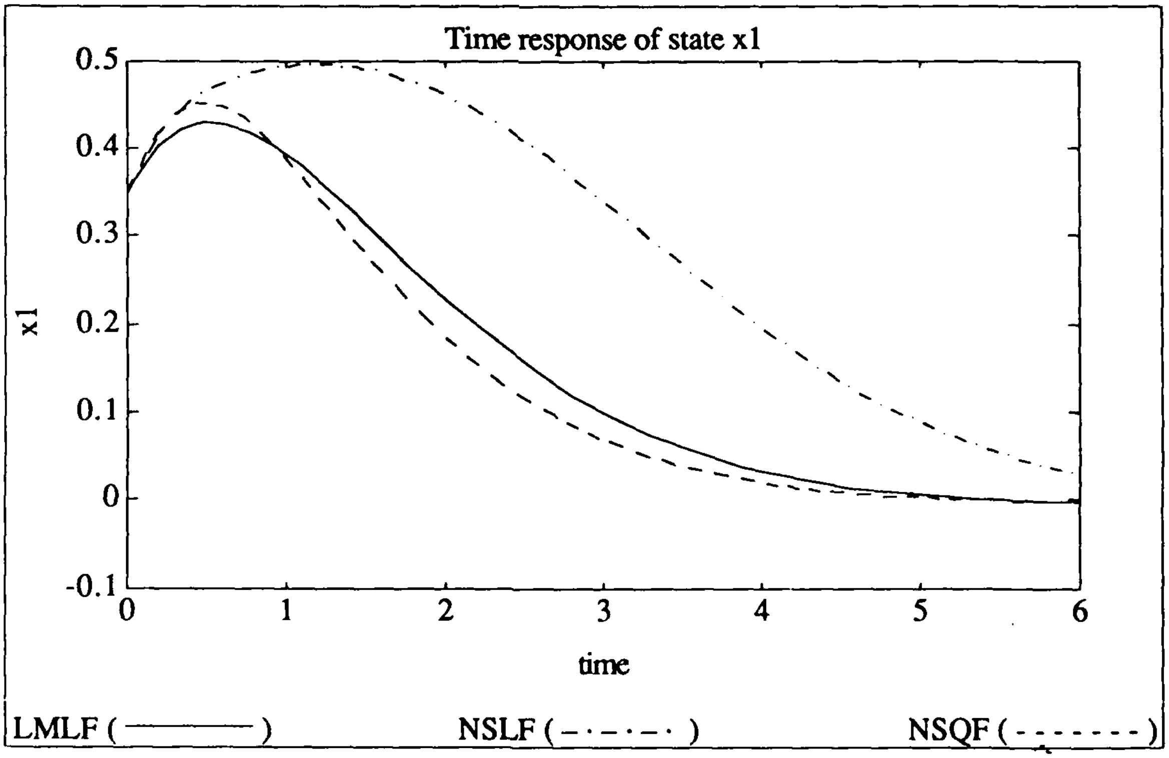


Fig. 5.26. Free response of state x_1 of Example 2 with nonzero initial conditions.

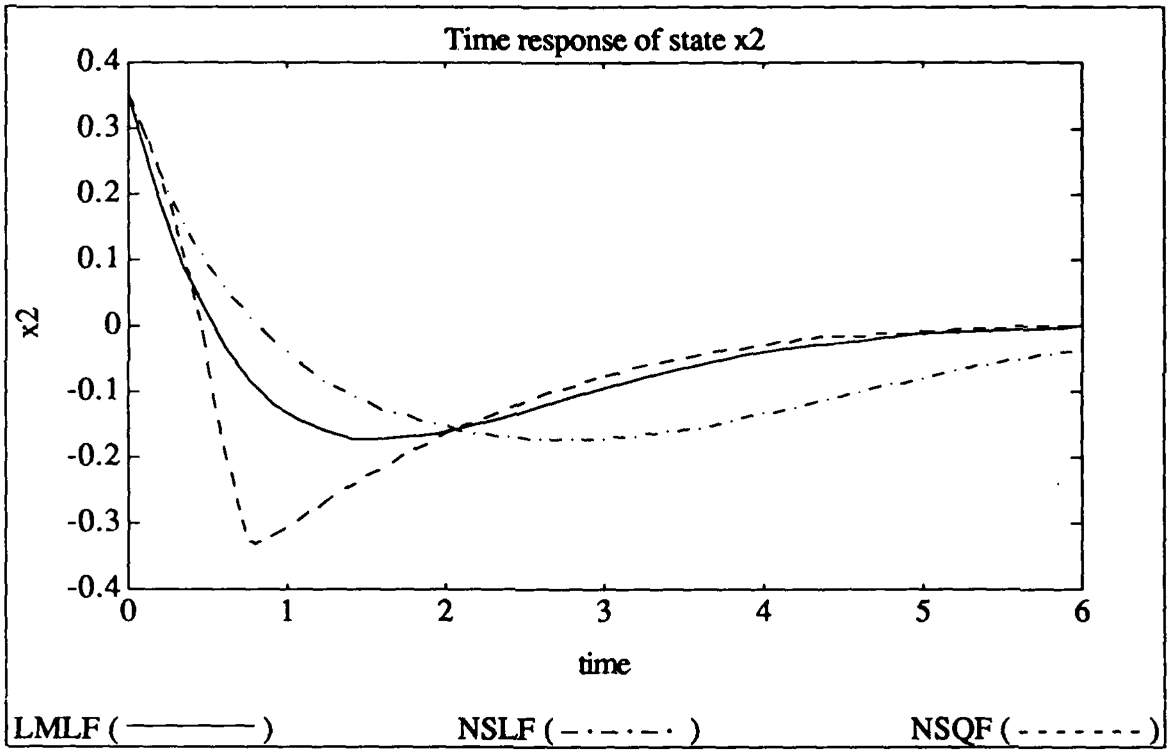


Fig. 5.27. Free response of state x_2 of Example 2 with nonzero initial conditions.

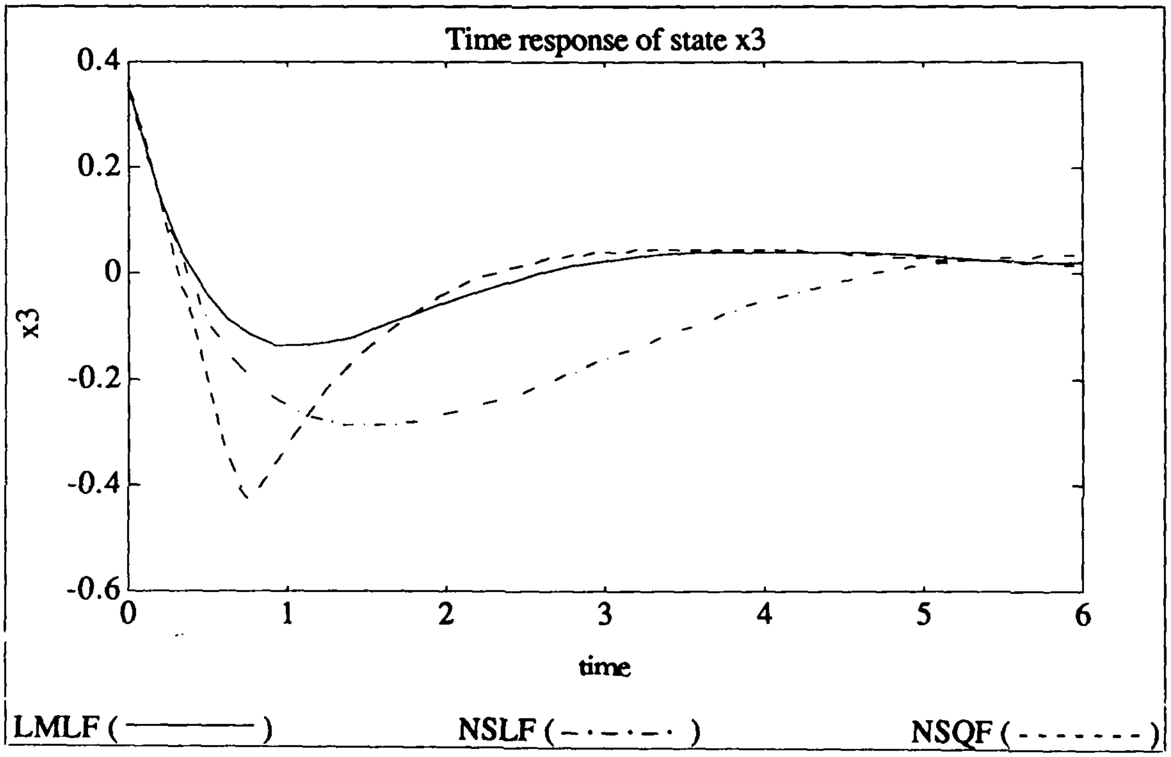


Fig. 5.28. Free response of state x_3 of Example 2 with nonzero initial conditions.

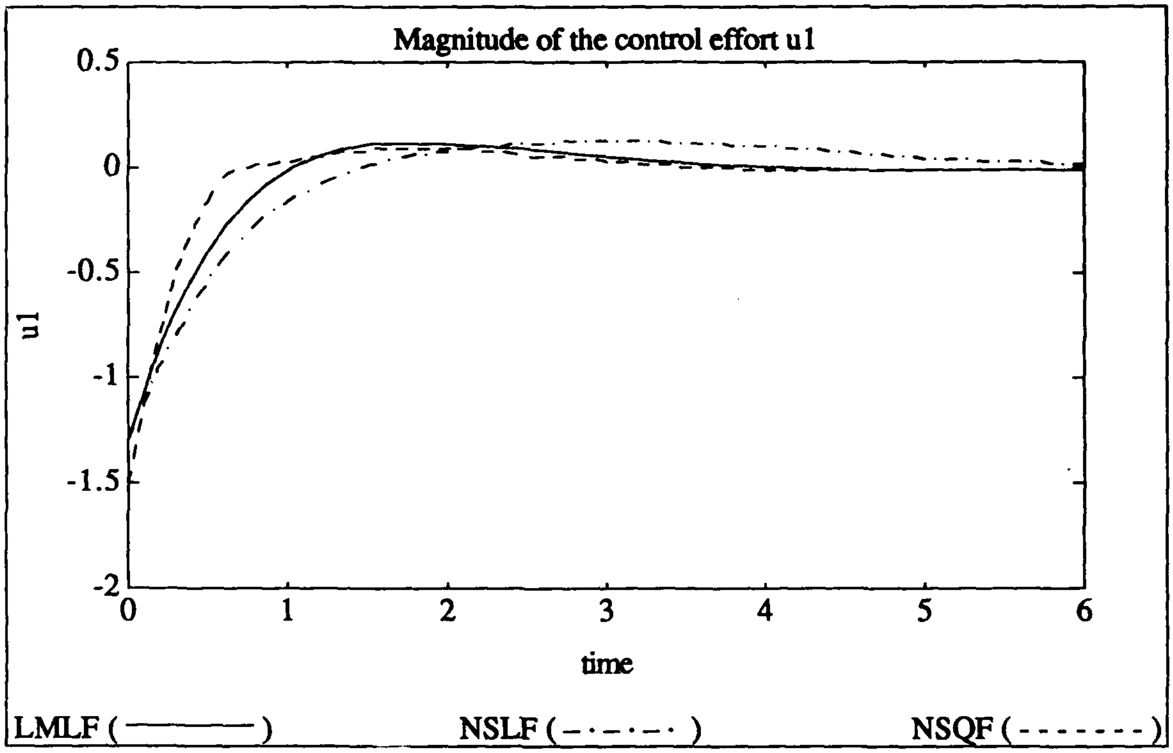


Fig. 5.29. Magnitudes of the inputs for Example 2 with nonzero initial conditions.

Figures 5.30 through 5.32 show the time response of the three models to an impulse input of unit magnitude. The NSQF displays unstable behavior, and its curve is not plotted through the entire time of the simulation because, due to instability, the integration algorithm could not carry out the computation any further. Since the stability properties of nonlinear systems are not well known, the cause of this instability is not entirely clear. Fig. 5.33 shows the magnitude of the inputs.

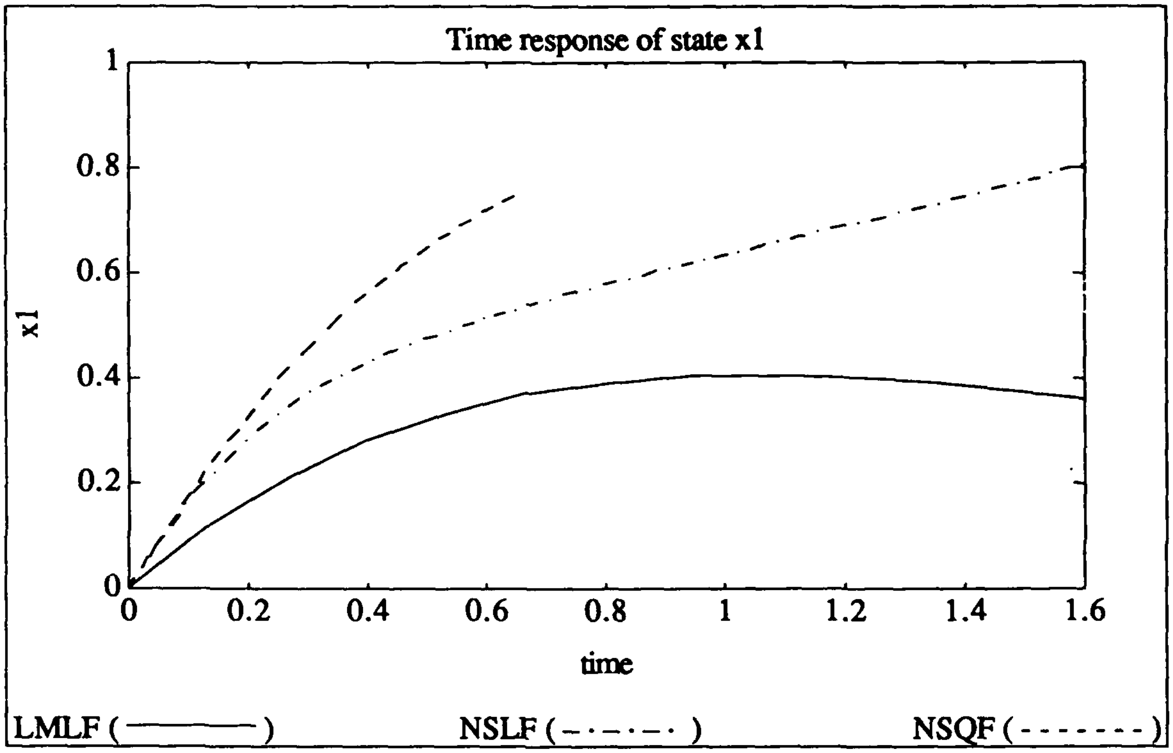


Fig. 5.30. Response of the state x_1 of Example 2 to an impulse input.

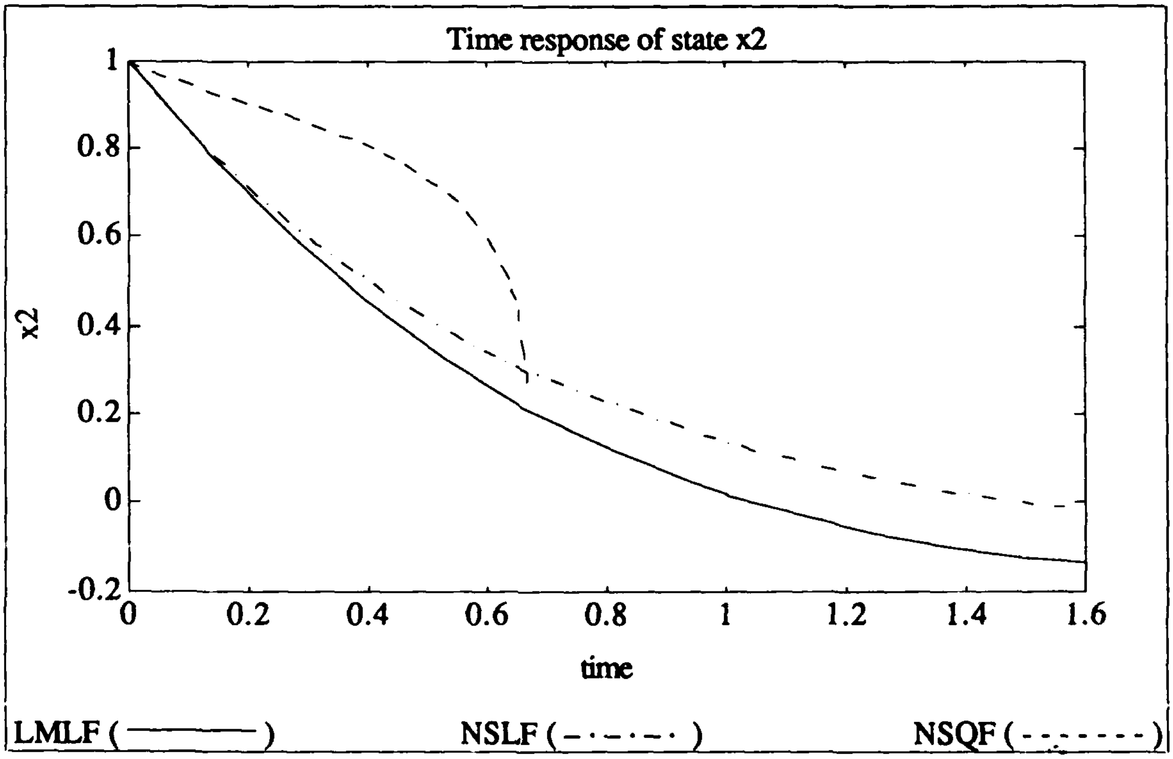


Fig. 5.31. Response of the state x_2 of Example 2 to an impulse input.

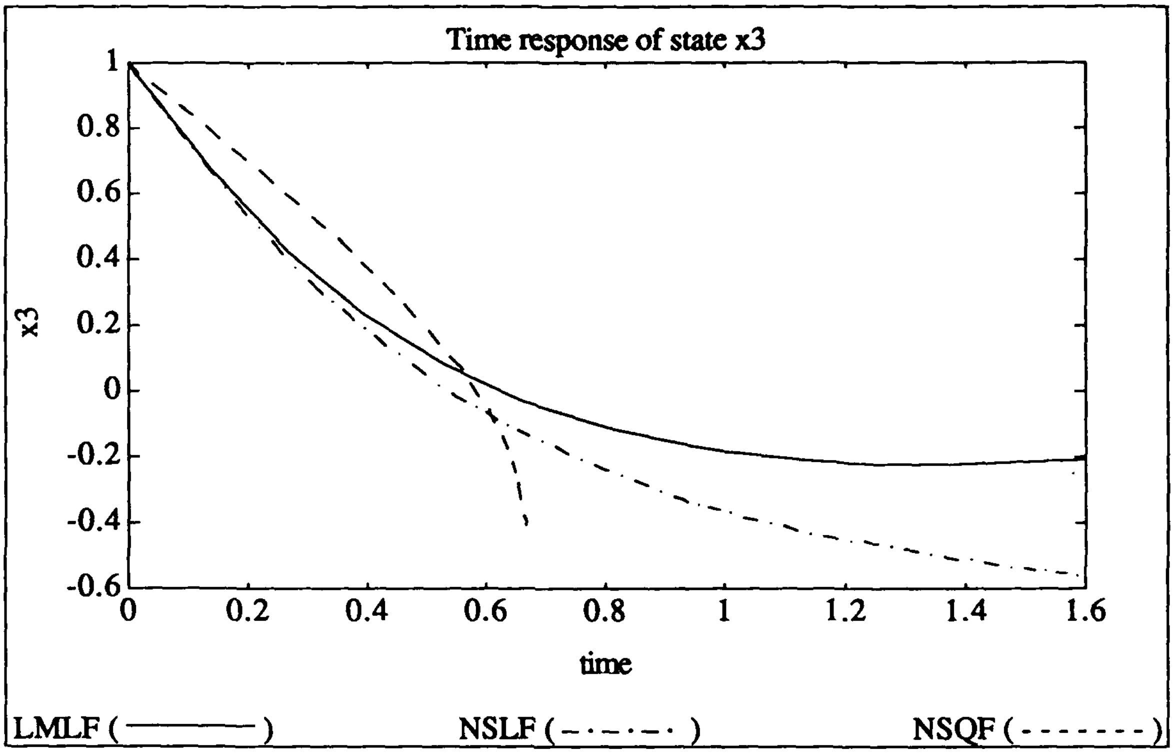


Fig. 5.32. Response of the state x_3 of Example 2 to an impulse input.

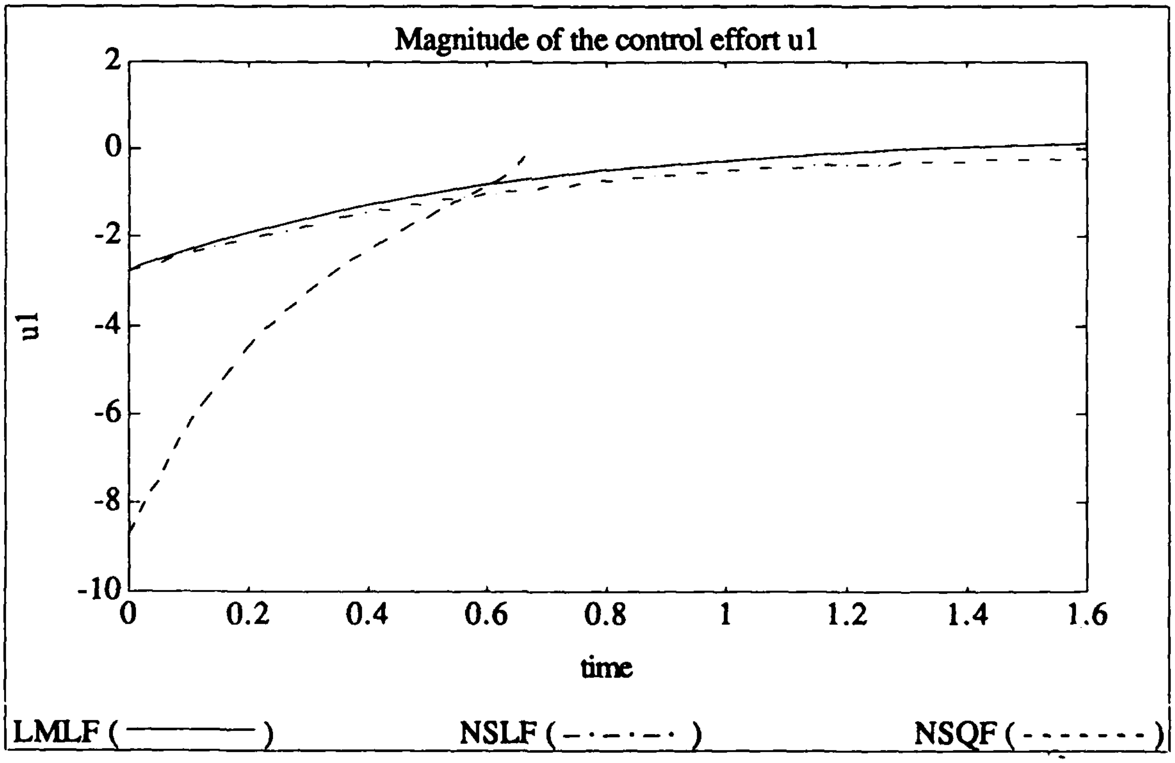


Fig. 5.33. Magnitude of the control inputs for Example 2 for an impulse input.

In Figs. 5.34 through 5.36 are plotted the time responses of the same models to a step input of amplitude 0.5. In these simulations, the NSQF exhibits stable behavior, and settles toward the optimally damped response curve of the LMLF within 5 time units. On the contrary, the NSLF quickly goes unstable and can not track the reference signal at all. One can choose a sufficiently small amplitude for the step input in order to obtain stable behavior for the NSLF, but the characteristics of the response curve in comparison with the NSQF would still display larger errors in tracking the LMLF. Fig. 5.37 shows the magnitude of the inputs in this case. The input values (except for NSLF, which goes unstable) settle at a steady state nonzero constant after some transient. This is expected because the reference signal is a step input.

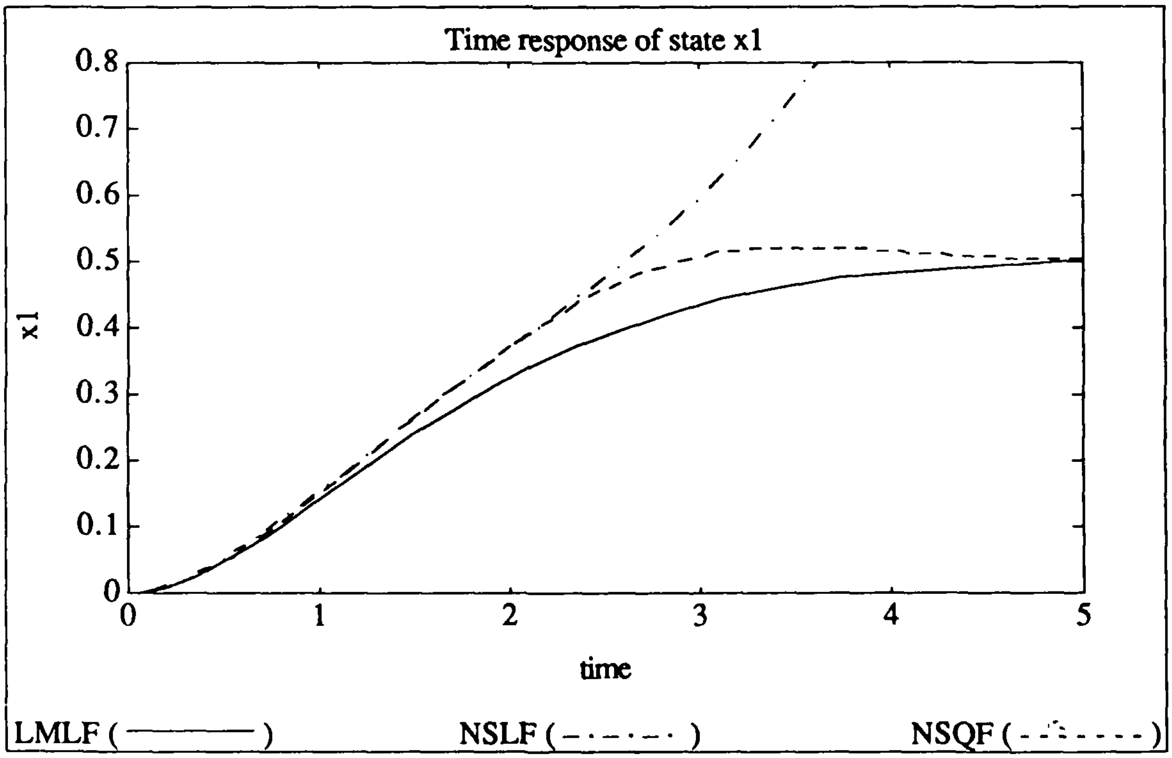


Fig. 5.34. Response of the state x_1 of Example 2 to a step input.

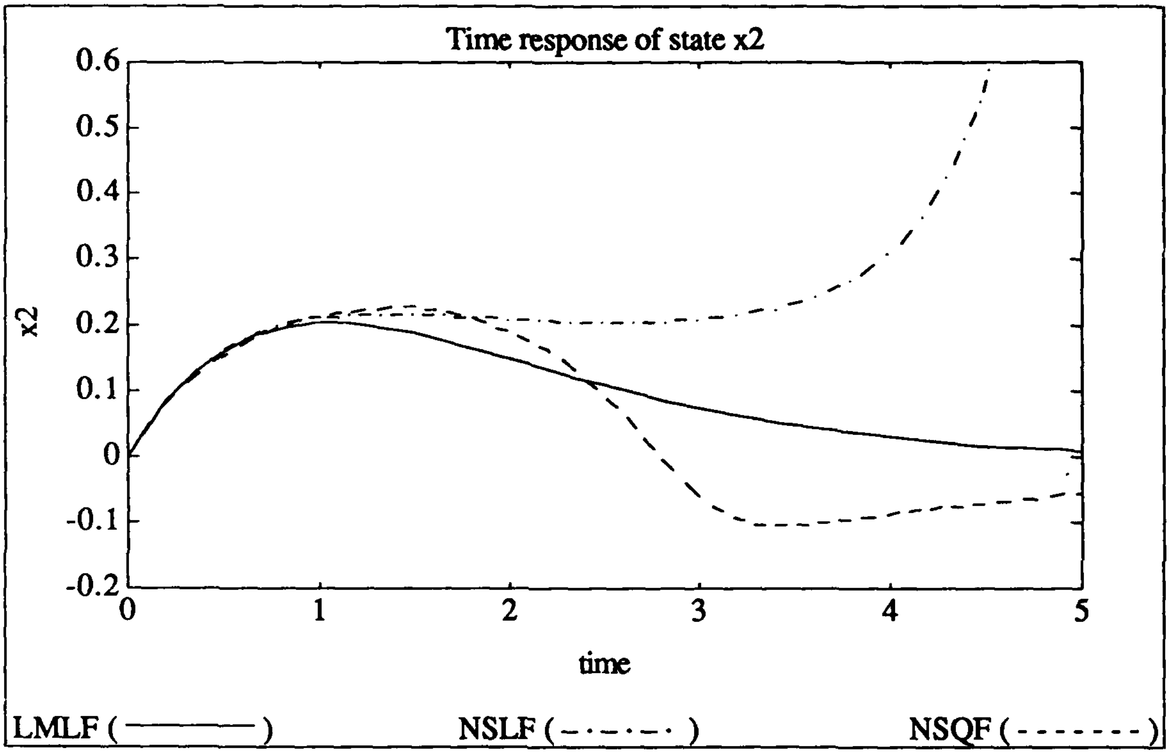


Fig. 5.35. Response of the state x_2 of Example 2 to a step input.

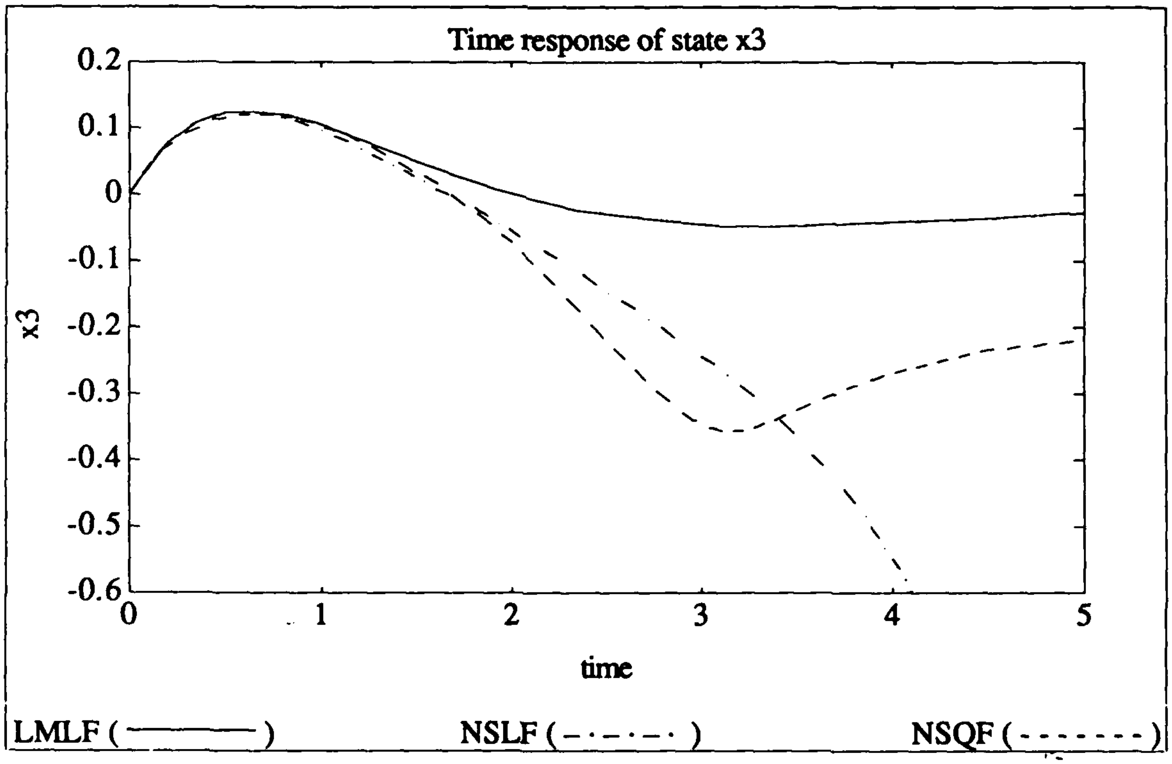


Fig. 5.36. Response of the state x_3 of Example 2 to a step input.

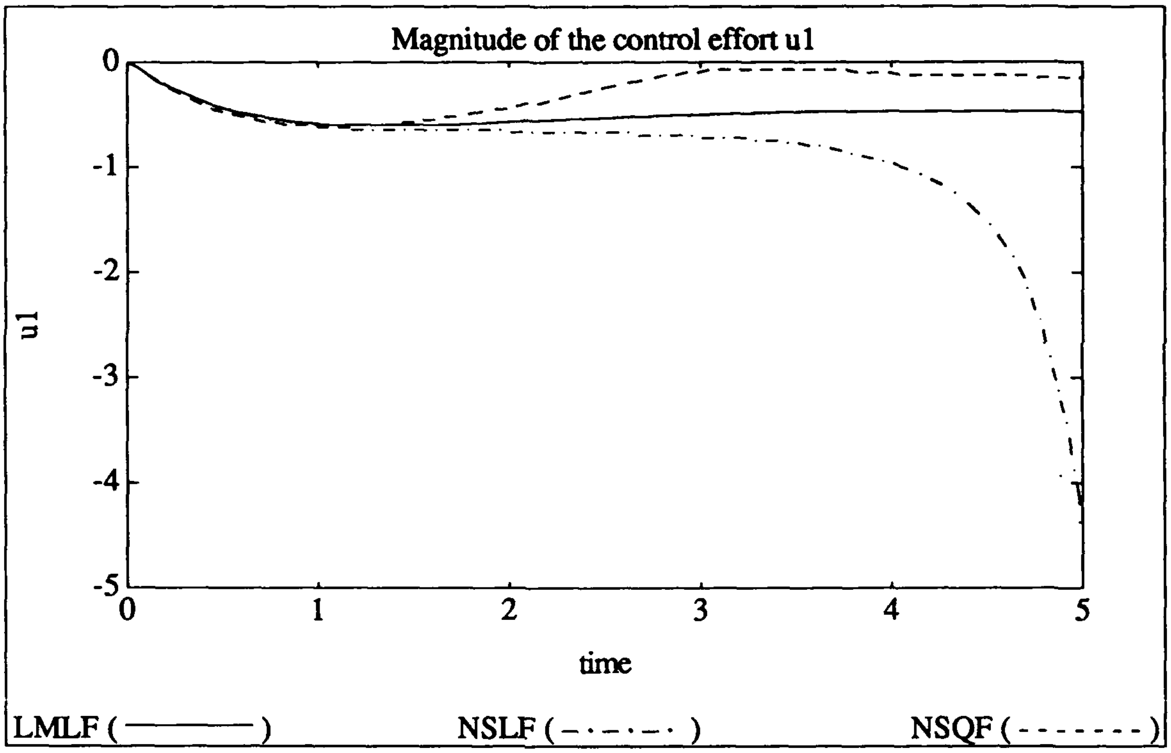


Fig. 5.37. Magnitudes of the control inputs for Example 2 for a step input.

Figs. 5.38, 5.39, and 5.40 show the transient and steady state time domain response of all three models of Example 2 to a sinusoidal input. The reference signal was $r(t) = A\sin(\omega t)$ with $\omega = 6$, $A = 1$. Similar comments as in Example 1 can be made for this set of responses. The NSQF clearly has an advantage over the NSLF in tracking the sinusoidal reference input. In this simulation, since the linearization is not exact, the NSQF does not follow the LMLF exactly. However, in comparison with the curve of NSLF this deviation is negligible. This simulation is again the most interesting case in displaying the advantage of the nonlinear feedback. The NSLF has a constant average offset from the equilibrium point. For NSQF, this offset is much smaller since in z coordinates, the system is closer to a linear model.

The above example verifies our claim in the approximate linearization procedure of minimizing the "distance" between the given system and the "closest linearizable" system. Of course the improvement achieved by the quadratic feedback is dependent on how far away the original nonlinear system is from the "closest linearizable" system before the linearization is done. After the initial transients die out, the LMLF settles into an equilibrium at the origin, i.e. its average is zero as expected from an exactly linear system. The NSQF displays oscillattions extremely close to the LMLF. Again we emphasize that the steady-state equilibrium around which each model settles is directly related to the average of the nonlinear terms in the vector fields over each period of the sinusoidal input. In the case of NSQF this average is very small, i.e. it is closer to an exactly linear model. The argument is more obvious in Fig. 5.38 since the state x_1 is farthest away from the input in terms of the number of integrations.

Fig. 5.41 shows the plot of the feedback inputs, each of which is periodic as expected. It is very interesting to observe that while the average values of the control inputs for LMLF and NSLF are zero (or almost zero), the average of the input for NSQF deviates from zero. This fact clearly has a connection to the above-mentioned argument for the averages of the responses for the three systems. It appears that the quadratic feedback introduces a bias into the nonlinear system which drives it towards the origin (on the average).

In Figs. 5.42, 5.43, and 5.44 we present the phase portrait plots of the three models. The steady state equilibrium points are again clearly seen, especially in the graphs of x_1 versus x_2 and x_1 versus x_3.

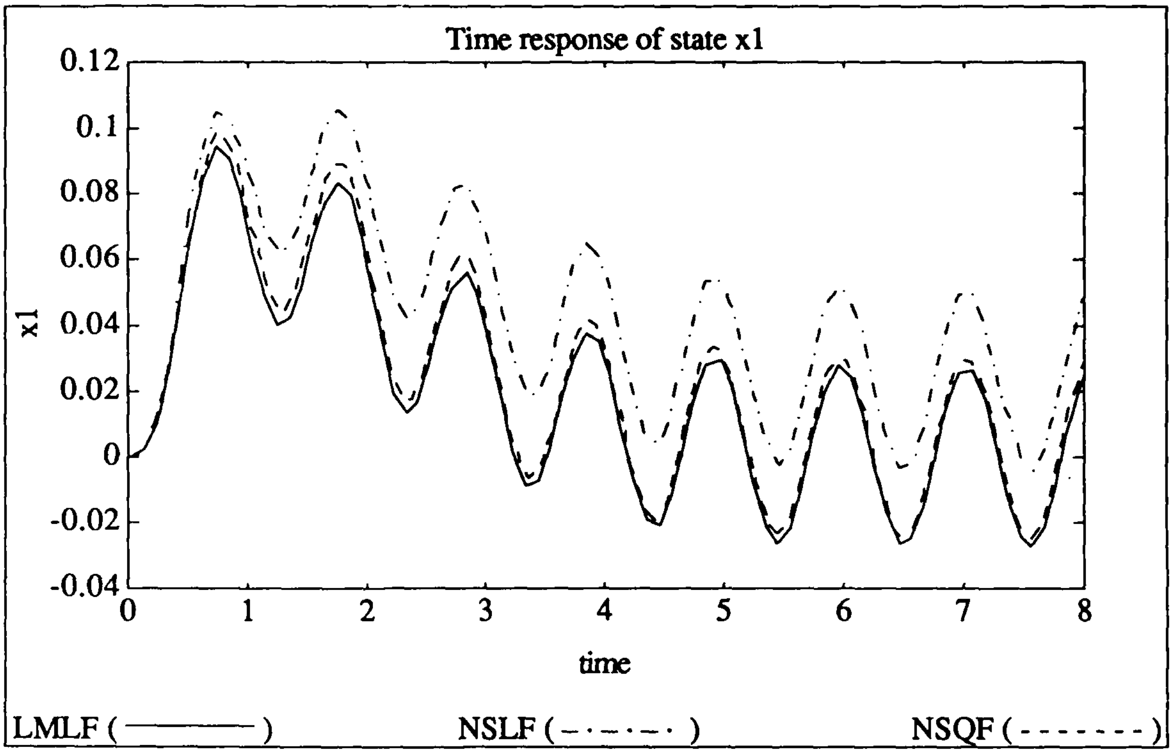


Fig. 5.38. Response of the state x_1 of Example 2 to a sinusoidal input.

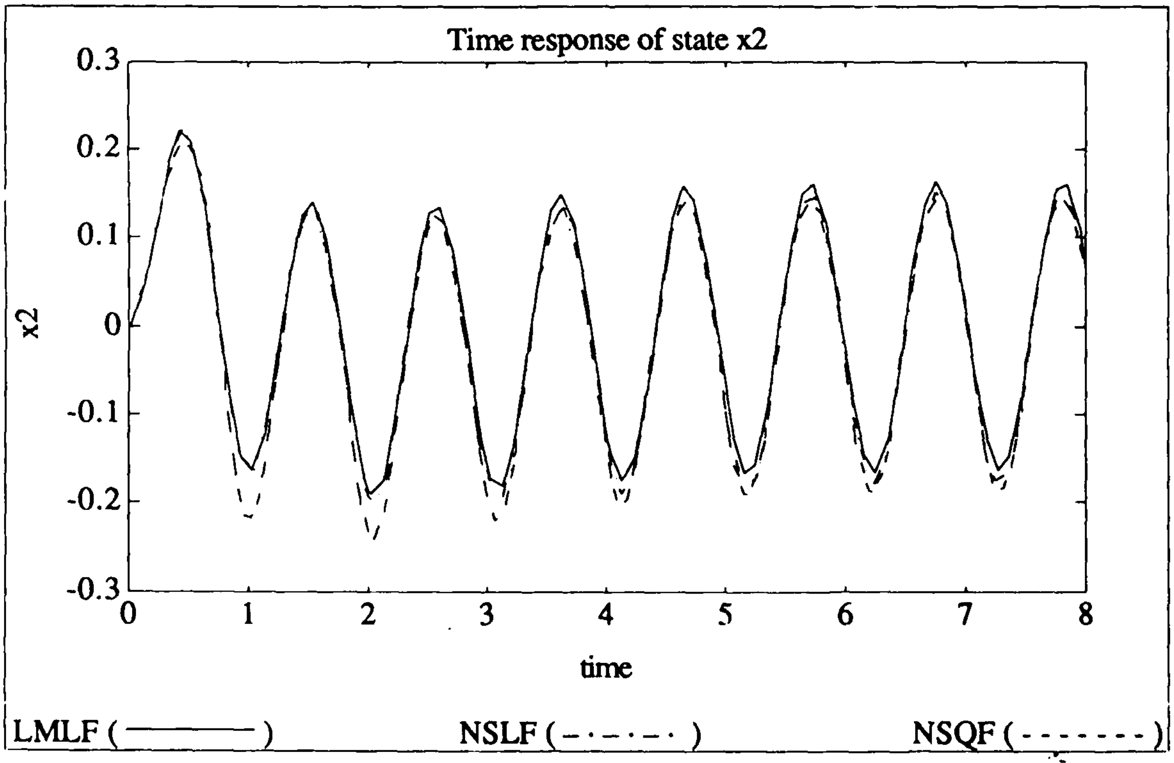


Fig. 5.39. Response of the state x_2 of Example 2 to a sinusoidal input.

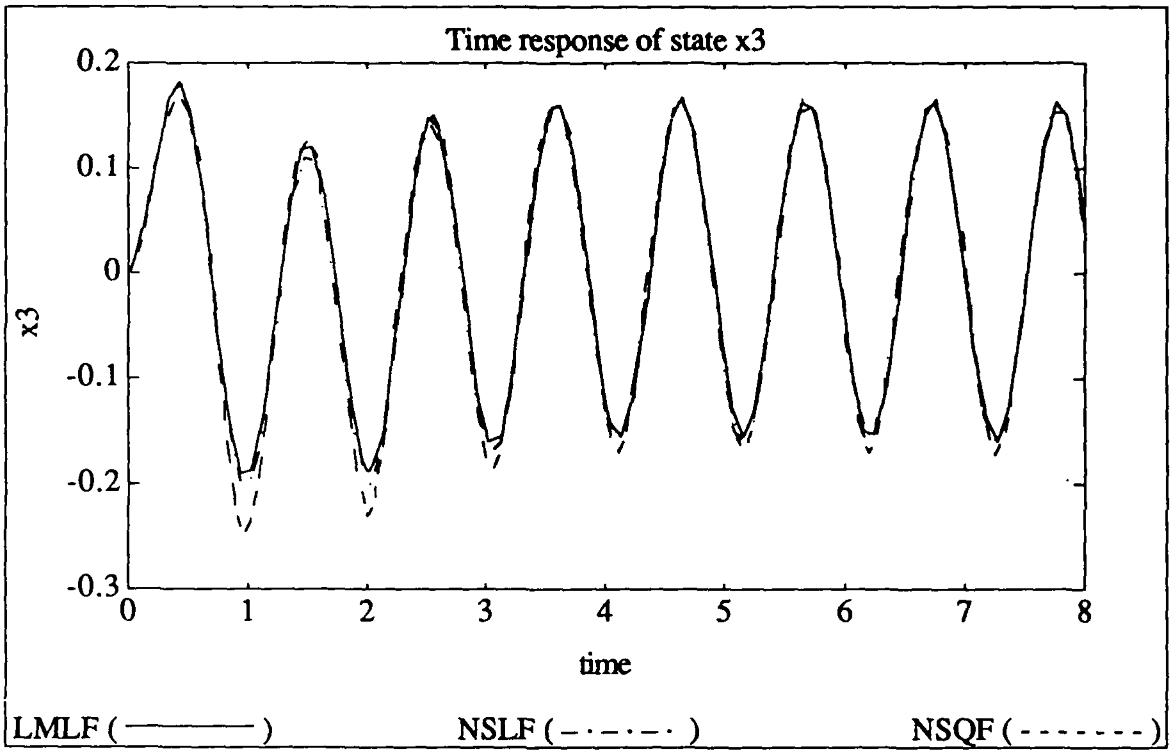


Fig. 5.40. Response of the state x_3 of Example 2 to a sinusoidal input.

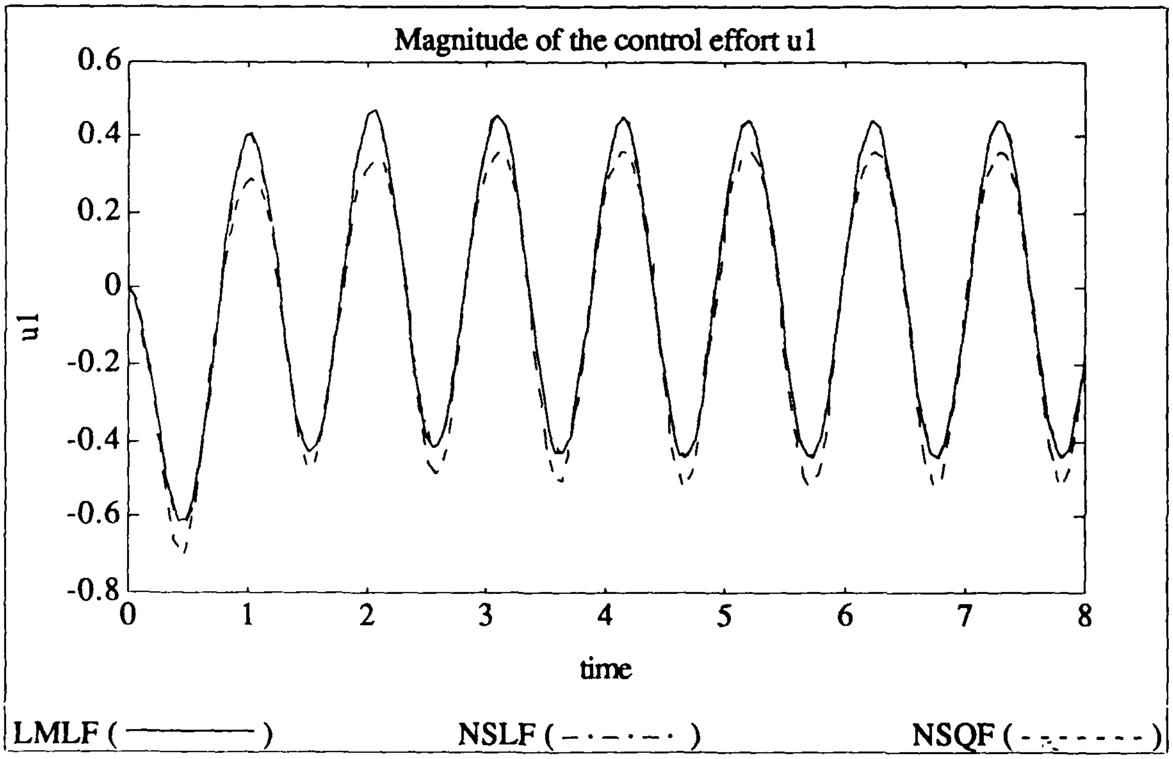


Fig. 5.41. Magnitude of the control efforts for Example 2 for a sinusoidal input.

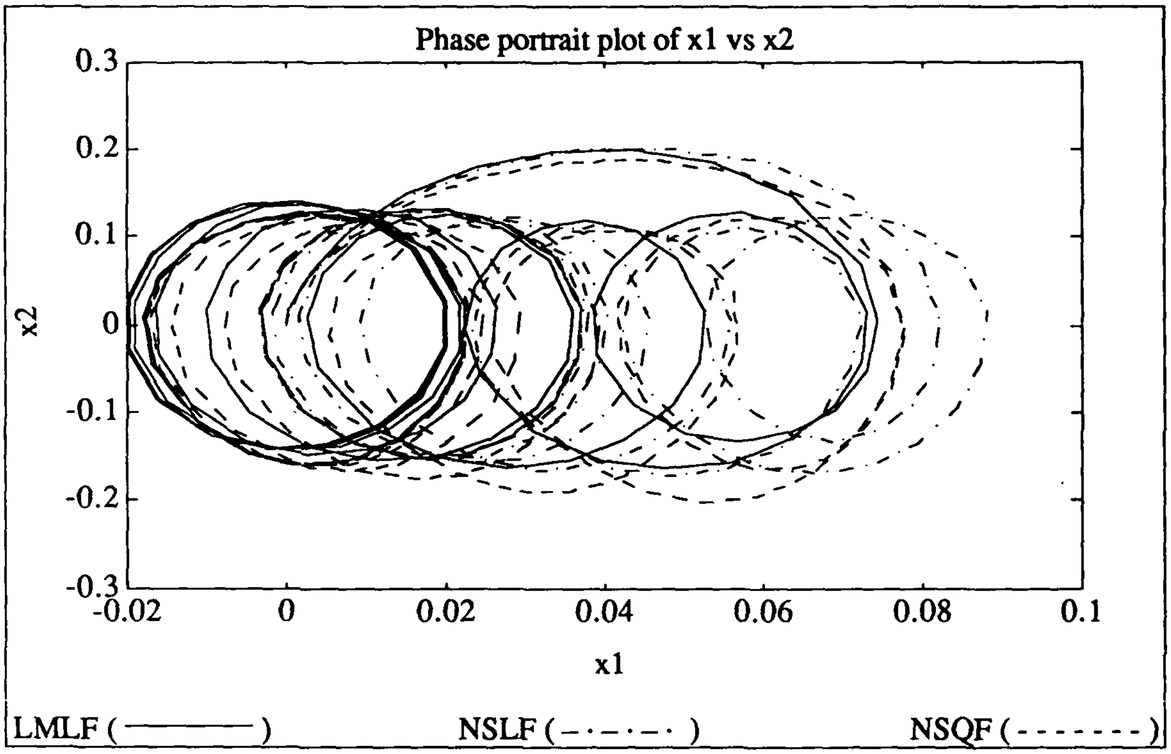


Fig. 5.42. Phase portrait of x_1 versus x_2 for Example 2 for a sinusoidal input.

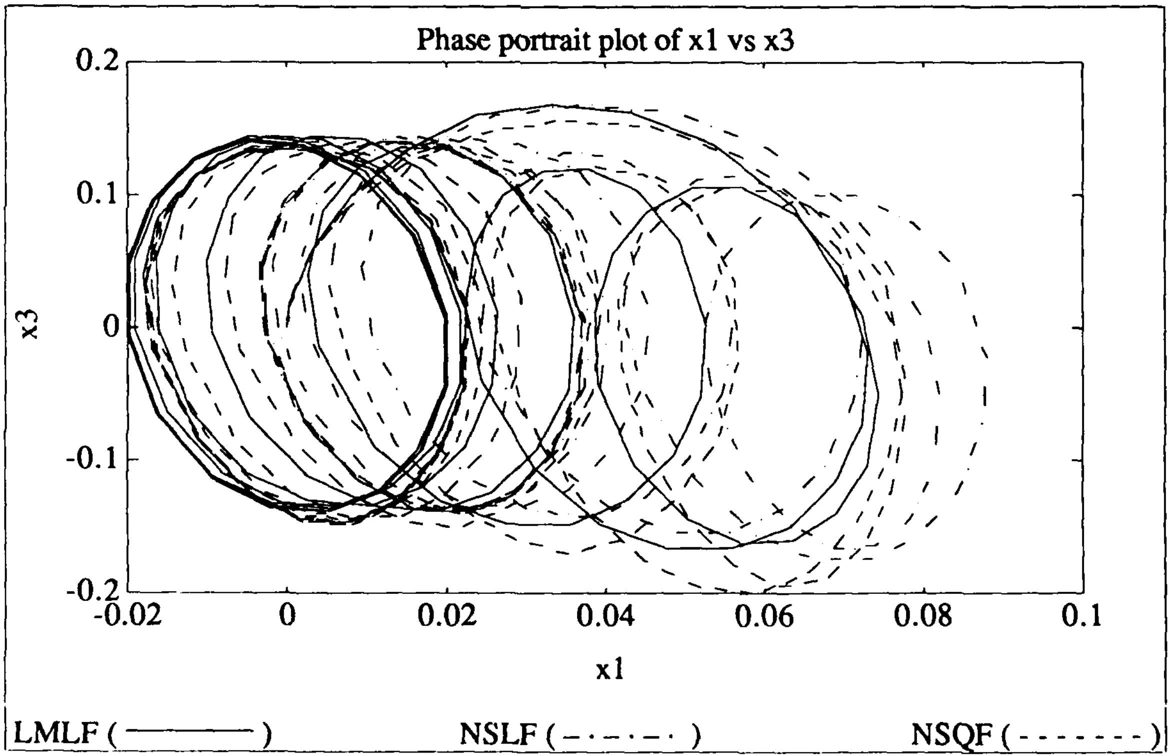


Fig. 5.43. Phase portrait of x_1 versus x_3 for Example 2 for a sinusoidal input.

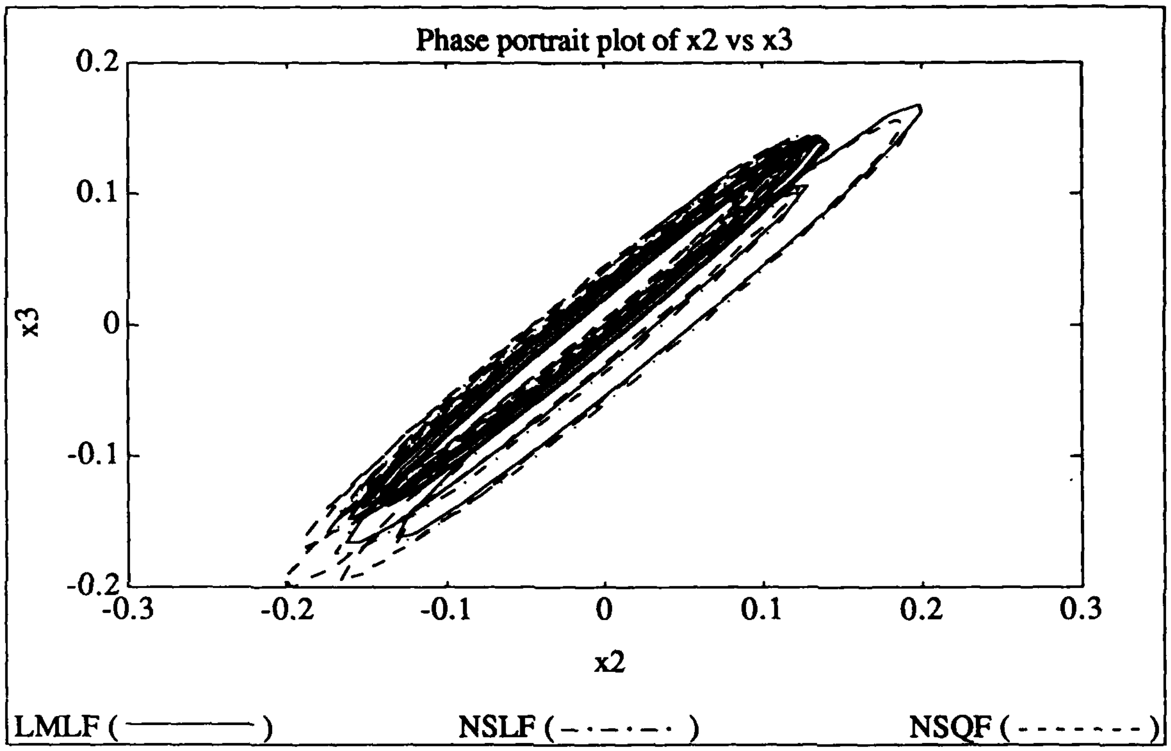


Fig. 5.44. Phase portrait of x_2 versus x_3 for Example 2 for a sinusoidal input.

5.4.3. Example 3:

The following third order system has been simulated as another example. This system is not exactly linearizable and it has open loop poles very close to the desired closed loop pole locations.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 1.259 & 0.0177 & -5.1128 \\ 0.5181 & -2.9646 & -12.2257 \\ 1.259 & 0.0177 & -4.1128 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} u$$

$$+ \begin{bmatrix} x_1^2 + 2x_1x_2 + x_1x_3 + x_2x_3 + x_3^2 \\ x_1^2 + 2x_1x_2 + 3x_1x_3 + x_2^2 + x_2x_3 + x_3^2 \\ x_1x_2 + x_1x_3 + 2x_2^2 + x_2x_3 \end{bmatrix} + \begin{bmatrix} x_1 + 2x_3 \\ x_2 + x_3 \\ 2x_3 \end{bmatrix} u \qquad (5.22)$$

The characteristic scales of each state and the input were assumed to be equal to unity. The following coordinate transformation and feedback pair were found:

$$z_1 = x_1 + 0.86986x_1^2 + 0.62932x_1x_2 - 2.8055x_1x_3 - 0.02927x_2^2$$
$$+ 0.21302x_2x_3 - 0.030389x_3^2 \qquad (5.23a)$$

$$z_2 = x_2 - 1.5358x_1^2 + 2.2273x_1x_2 + 1.0036x_1x_3 + 0.36424x_2^2 - 3.2297x_2x_3$$
$$+ 1.7879x_3^2 \qquad (5.23b)$$

$$z_3 = x_3 + 1.643x_1^2 - 0.062292x_1x_2 - 1.9677x_1x_3 + 0.26052x_2^2$$
$$- 0.25082x_2x_3 + 0.014877x_3^2 \qquad (5.23c)$$

The terms in the nonlinear feedback were:

$$\alpha^{(2)}(x) = 7.3167x_1^2 - 0.0043799x_1x_2 - 14.095x_1x_3 + 1.5513x_2^2 - 4.5535x_2x_3$$
$$+13.072x_3^2 \qquad (5.24)$$

$$\beta^{(1)}(x) = 1.1933x_1 + 0.72725x_2 - 0.43993x_3 \qquad (5.25)$$

The closest linearizable system had the following nonlinear terms:

$f^{(2)}(x) =$

$$\begin{bmatrix} 1.0001x_1^2+2.0002x_1x_2+1.0001x_1x_3+0.00044583x_2^2+1.0002x_2x_3+1.0001x_3^2 \\ x_1^2+2x_1x_2+3x_1x_3+x_2^2+x_2x_3+x_3^2 \\ 0.99981x_1x_2+x_1x_3+1.9996x_2^2+0.99981x_2x_3 \end{bmatrix}$$

$$g^{(1)}(x) = \begin{bmatrix} 1.0004x_1 + 0.00198412x_2 + 2.0003x_3 \\ 0.99986x_2 + x_3 \\ -0.000391772x_1 - 0.0017123x_2 + 1.9997x_3 \end{bmatrix} u$$

The system has open loop poles at -0.5227; $-2.6479 \pm 0.3841j$. A closed loop feedback for the linear part of the plant was calculated using a quadratic optimal regulator design procedure with the weighing on the states and the input equal to unity. The feedback gains were found as: $K = [-0.4730 \quad 0.0177 \quad 1.4224]$

and the above gains placed the eigenvalues of the linear part of the closed loop system at:

$\lambda_1 = -0.5056$

$\lambda_2 = -1.7647$

$\lambda_3 = -2.5633$

Since the open loop system already has stable roots, the linear feedback gains that are necessary to drive the system are small. In the simulation plots the response curves are defined with the following legend as before:

LMLF (————): Reference Linear Model with Linear Feedback.

NSLF (– · – · – ·): Nonlinear System with the Linear Feedback design.

NSQF (- - - - - - -): Nonlinear System with Quadratic Feedback design.

Figures 5.45 through 5.48 show the response of the system to an initial condition of x = [0.1 0 0]' (referred to as initial condition #1 in the figures) and the feedback input. The reference signal is set to zero. The response curves of Figs. 5.45, 5.46, and 5.47 show the vastly superior behavior of tne NSQF in terms of responding like a linear system, i.e. in tracking the LMLF extremely closely immediately after the transient response dies out.

We should note that in this specific example, the behavior of the NSQF in tracking the LMLF should be considered even more impressive because the system is not in controller canonical form. When a system is in controller form, the tracking errors appear to be much larger in those states that are further away from the input in terms of integration. This is because (considering a single input system) the signals are differentiated $n - 1$ times until the first state is reached, and all the errors are therefore amplified (Example 1 is a good demonstration of this fact). In this example, we observe excellent tracking behavior for NSQF on all the states, even without the benefit of the presence of controller form.

An inspection of the response characteristics of x_1 ,x_2 and x_3 in Figs. 5.45, 5.46, and 5.47, respectively, shows that in terms of settling times at the zero steady-state value, the NSLF is much superior to both the linear ideal model LMLF, and the NSQF. However, this is a fortuitous behavior of the system due to the specific nonlinearities in this example. The goal of the nonlinear control design is to track the linear system as close as possible, rather than to achieve the fastest response or the shortest settling time. One should also remember that the transient and the steady state responses of a linear system can be arbitrarily adjusted if the system is controllable and if there is no restriction in the magnitude of the requierd control effort. Therefore this should be seen as an isolated phenomenon which results from the particular nonlinearities present, as well as from the choice of eigenvalues of the linear part of the model.

During the transient part of the response the NSQF is not close to LMLF because the nonlinear terms are dominant. This is especially apparent in the curves of x_2 and x_3 (Figs. 5.46 and 5.47). After the first half of the simulation, this difference becomes virtually indistinguishable.

In the magnitude plot of the inputs (Fig. 5.48), the control effort for LMLF and NSLF are extremely small (almost zero for this simulation) and not visible on the scale of the plot. This is because the open loop system poles are very close to the closed loop values, and the control effort needed by a linear controller for driving the system to the origin is very small in both cases. In contrast, the NSQF needs some additional control effort to linearize the system. This magnitude approaches zero as the states approach the origin.

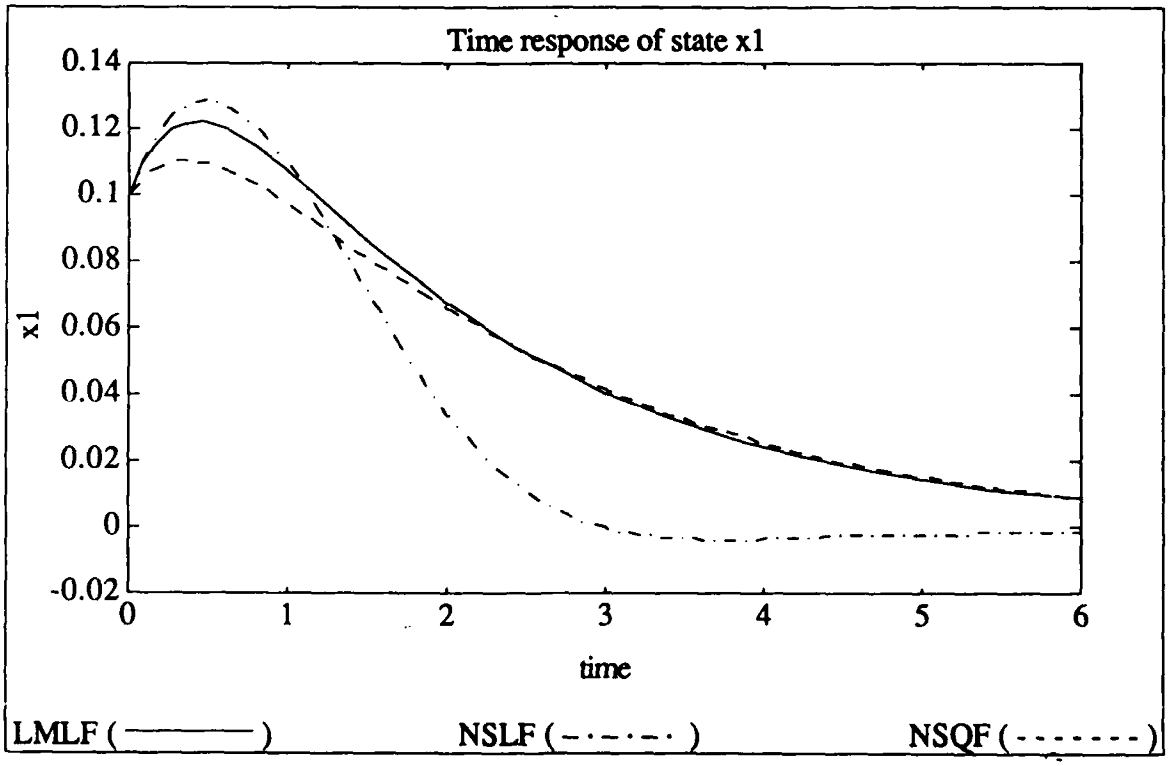


Fig. 5.45. Free response of state x_1 of Example 3 with nonzero initial condition #1.

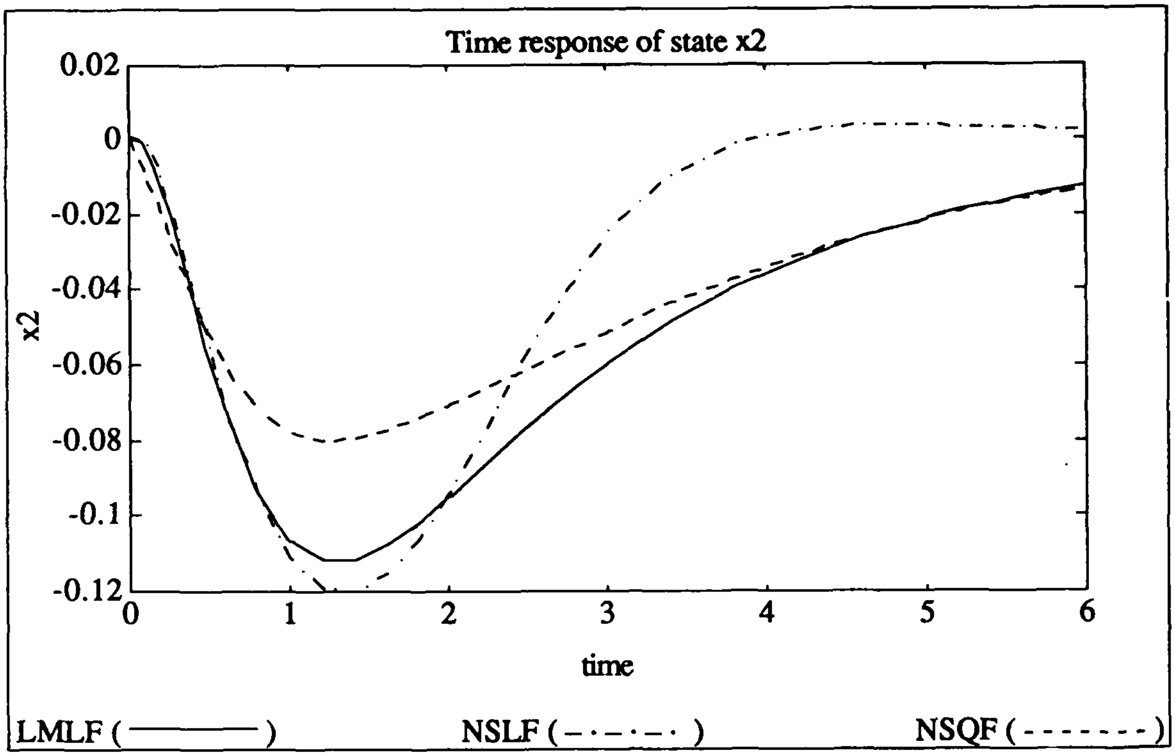


Fig. 5.46. Free response of state x_2 of Example 3 with nonzero initial condition #1.

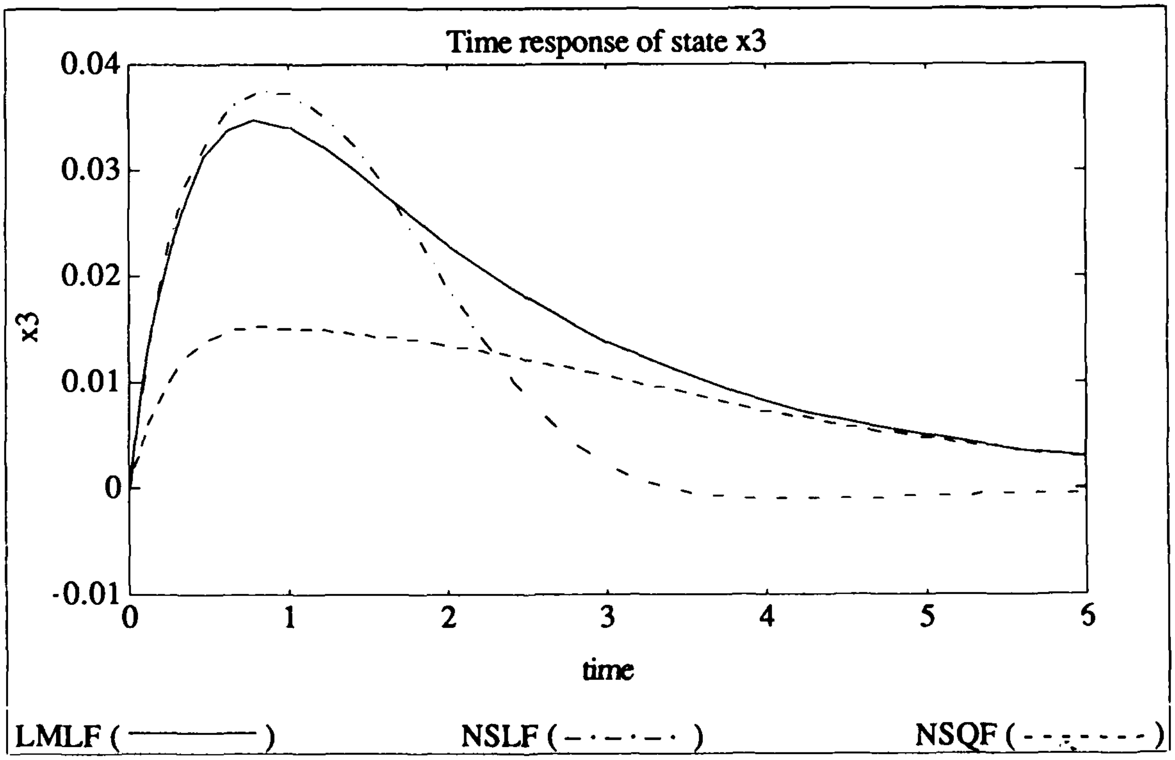


Fig. 5.47. Free response of state x_3 of Example 3 with nonzero initial condition #1.

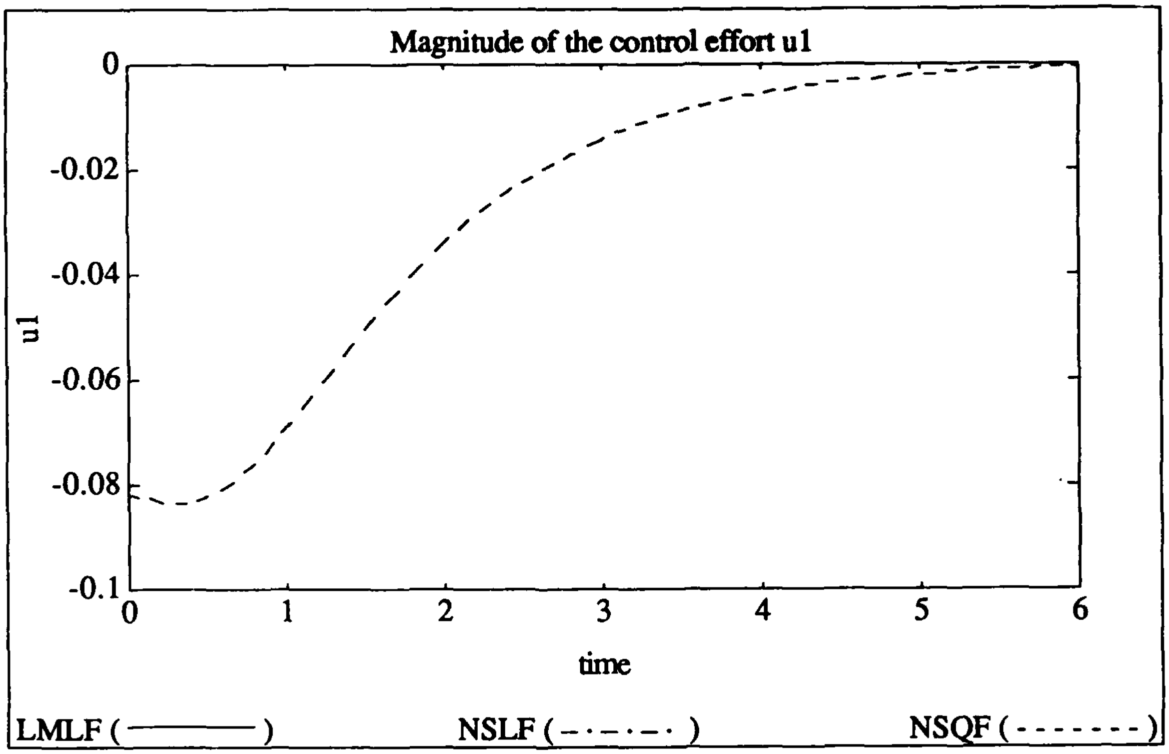


Fig. 5.48. Magnitudes of the inputs for Example 3 with nonzero initial condition #1.

In the following simulation, the initial condition of the previous case has been reversed in sign as $x = [-0.1\ 0\ 0]'$ (referred to as initial condition #2 in the captions of the graphs). The reason for choosing this set of initial conditions is to test the behavior of the nonlinear system for sensitivity to initial conditions. More specifically, we would like to check whether the NSQF will exhibit the same excellent performance for a variety of initial conditions as was seen in the simulation for the initial condition #1. The responses and the control efforts are shown in Figs. 5.49 through 5.52.

The NSLF is indeed initial condition sensitive (see the large deviations in Figs. 5.49, 5.50, 5.51 from the LMLF that it is supposed to track using the feedback law derived from the first order approximation). This is expected since the NSLF does not have the benefit of our nonlinear control scheme. The same plots show that the NSQF benefits greatly from the nonlinear control. Indeed, comparing the response curves of initial condition #1, the NSQF control displays an even more impressive response in tracking the LMLF, as seen from Figs. 5.49, 5.50, and 5.51 of the three states. To emphasize once more, the point here is not that the NSQF follows the LMLF better for this set of initial conditions over some others, but that it is *consistent* in this behavior for a variety of initial conditions.

The input effort is again very small for the LMLF and the NSLF compared to the NSQF. This is not regarded as a disadvantage for the quadratic control. The reason the NSQF seems to require large amounts of control effort to linearize the system is that the open loop eigenvalues of the linear part of the system are already near the desired closed loop locations and the linear feedback gains are therefore very small.

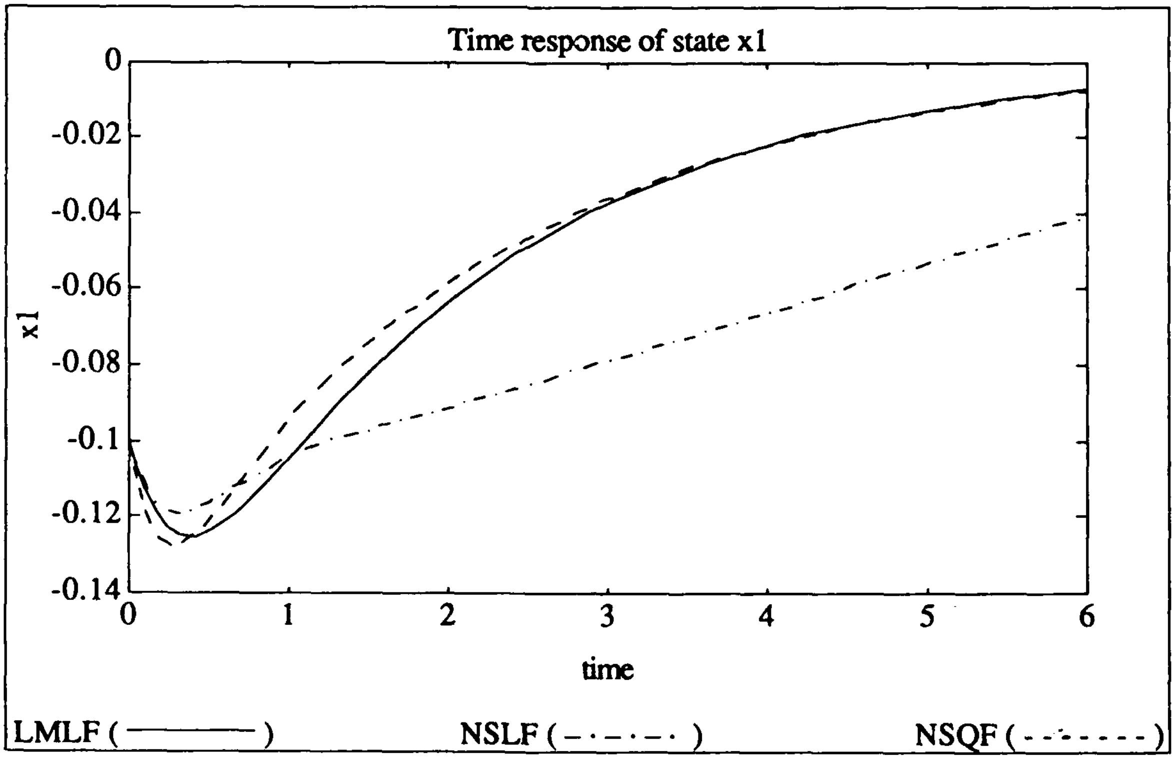


Fig. 5.49. Free response of state x_1 of Example 3 with nonzero initial condition #2.

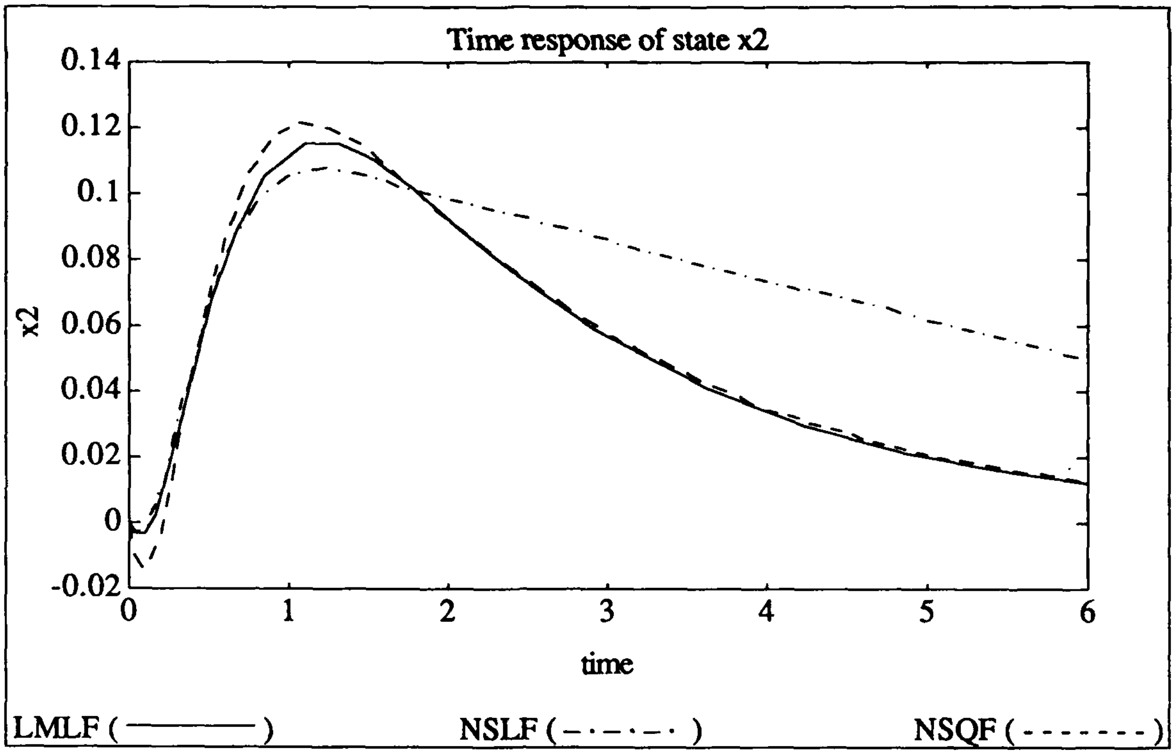


Fig. 5.50. Free response of state x_2 of Example 3 with nonzero initial condition #2.

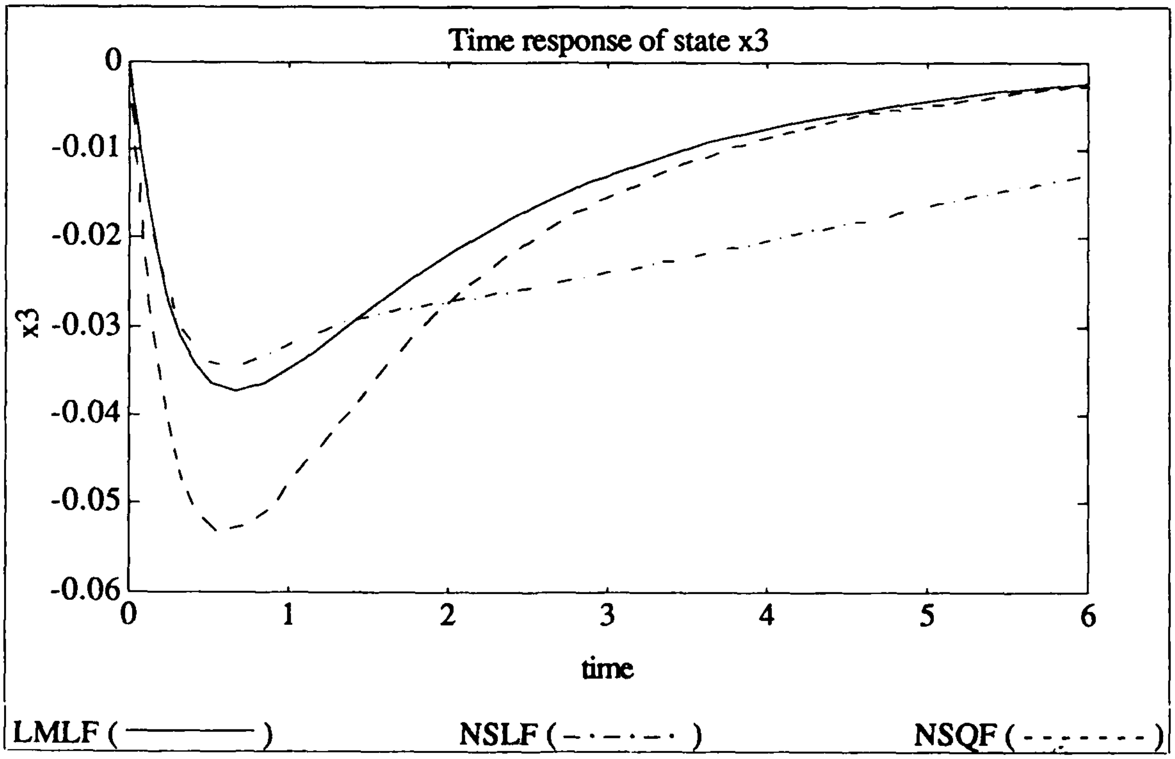


Fig. 5.51. Free response of state x_3 of Example 3 with nonzero initial condition #2.

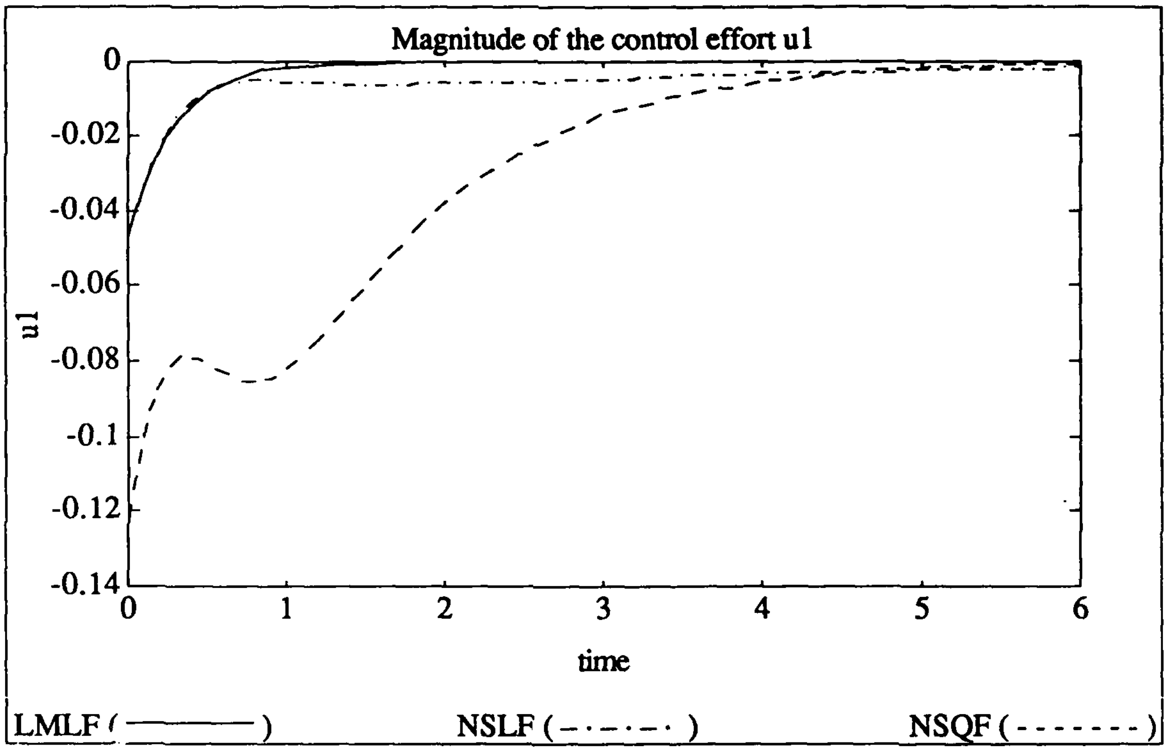


Fig. 5.52. Magnitudes of the inputs for Example 3 with nonzero initial condition #2.

We continue to further test the sensitivity of the system of Example 3 by applying nonzero initial conditions on other states. In the following, we present another set of simulations for the same system with nonzero initial conditions for x_3. The initial condition values are $x = [0\ 0\ 0.1]$ (initial condition #3) and $x = [0\ 0\ -0.1]$ (initial condition #4). Response curves and the control efforts for initial condition #3 are shown in Figs. 5.53 through 5.56. Figs. 5.57 through 5.60 present the responses and the input effort for initial condition #4.

Comparing the responses of state x_1 in Figs. 5.53 and 5.57 for initial conditions #3 and #4, we can state the same conclusions that were presented for initial conditions #1 and #2. The NSQF tracks the ideal linear model LMLF very closely as seen in both of these figures. Comparison of these two figures show that the sensitivity to changes in initial conditions of the NSQF is minimal. The tracking performance by the NSQF of the LMLF is excellent in both cases. Similar conclusions are reached when the responses for the set of initial conditions #3 and #4 of the state x_2 in Figs. 5.54 and 5.58, and of the state x_3 in Figs. 5.55 and 5.59 are compared.

The control efforts seen in Figs. 5.55 and 5.60 show that the NSQF requires larger input values compared to a linear design. As explained earlier, the reason that the LMLF and NSLF require such small input efforts is that the open loop and the closed loop eigenvalues (of the linear part of the example) are very close to each other.

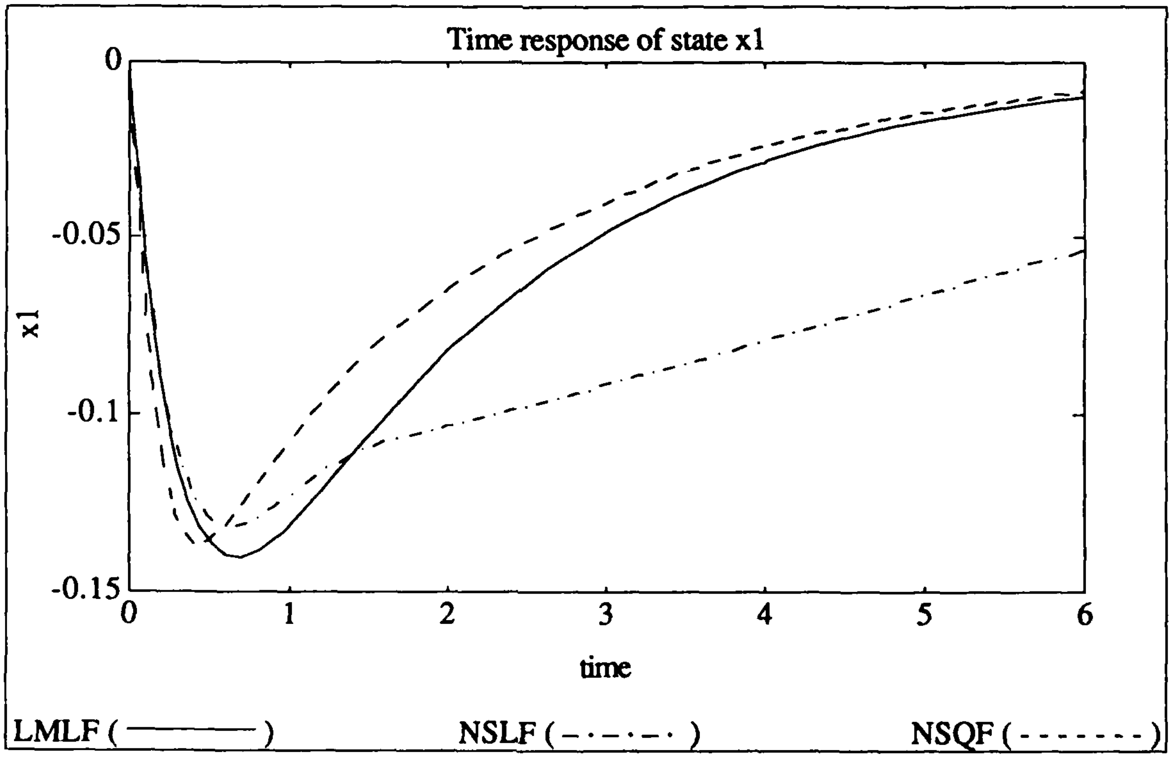


Fig. 5.53. Free response of state x_1 of Example 3 with nonzero initial condition #3.

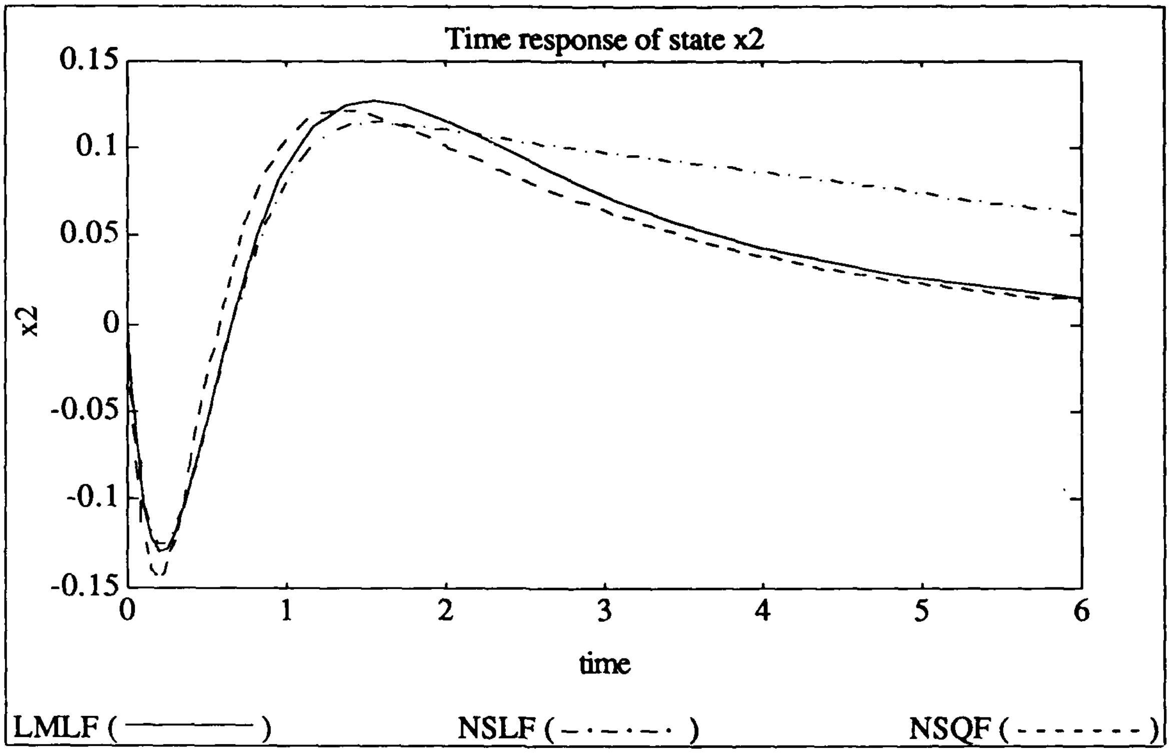


Fig. 5.54. Free response of state x_2 of Example 3 with nonzero initial condition #3.

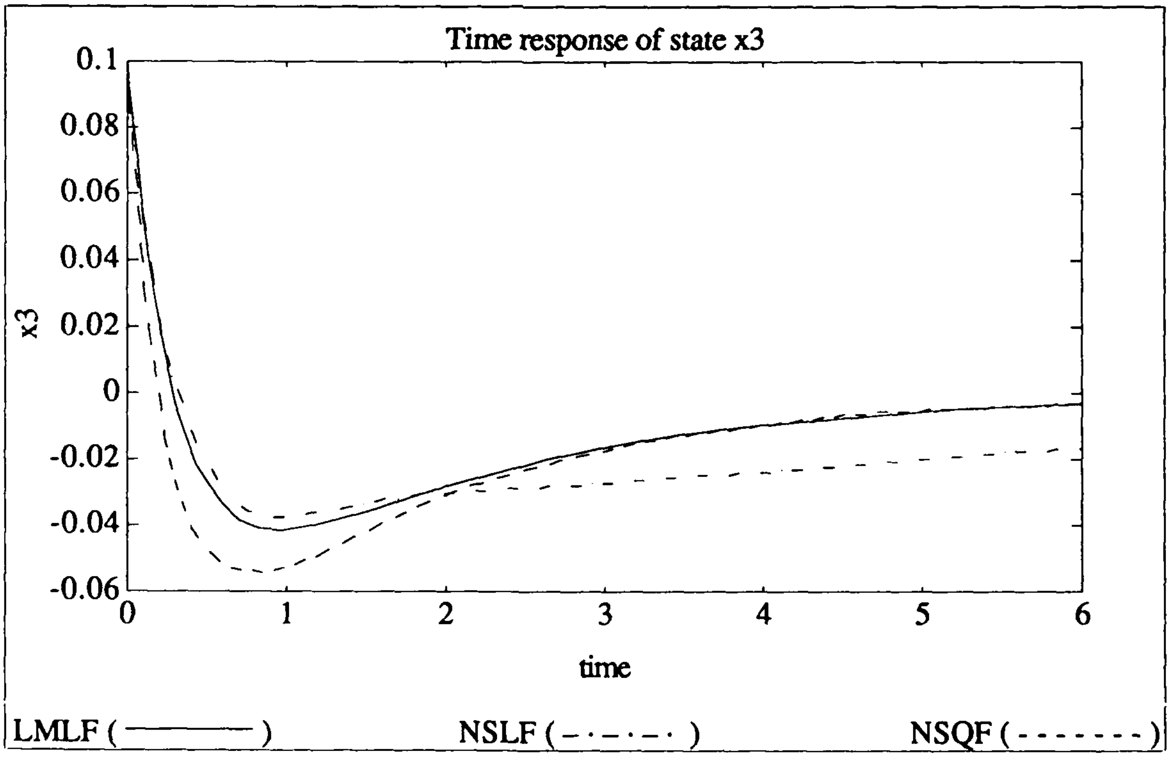


Fig. 5.55. Free response of state x_3 of Example 3 with nonzero initial condition #3:

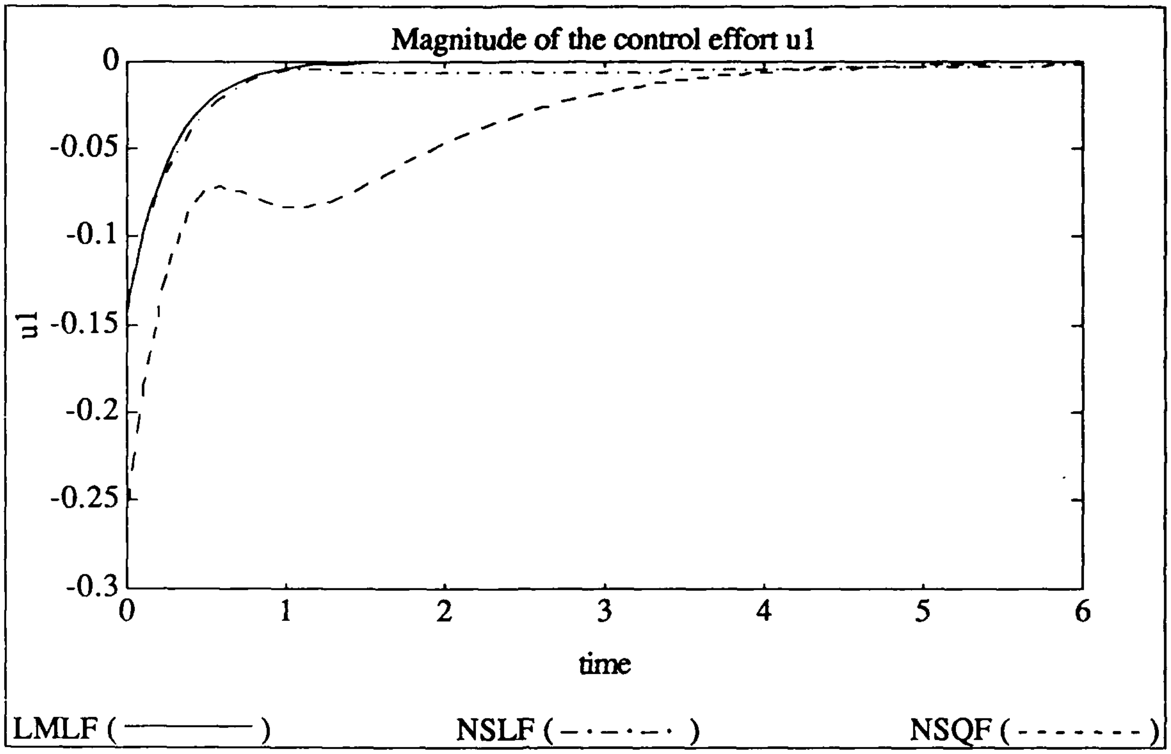


Fig. 5.56. Magnitudes of the inputs for Example 3 with nonzero initial condition #3.

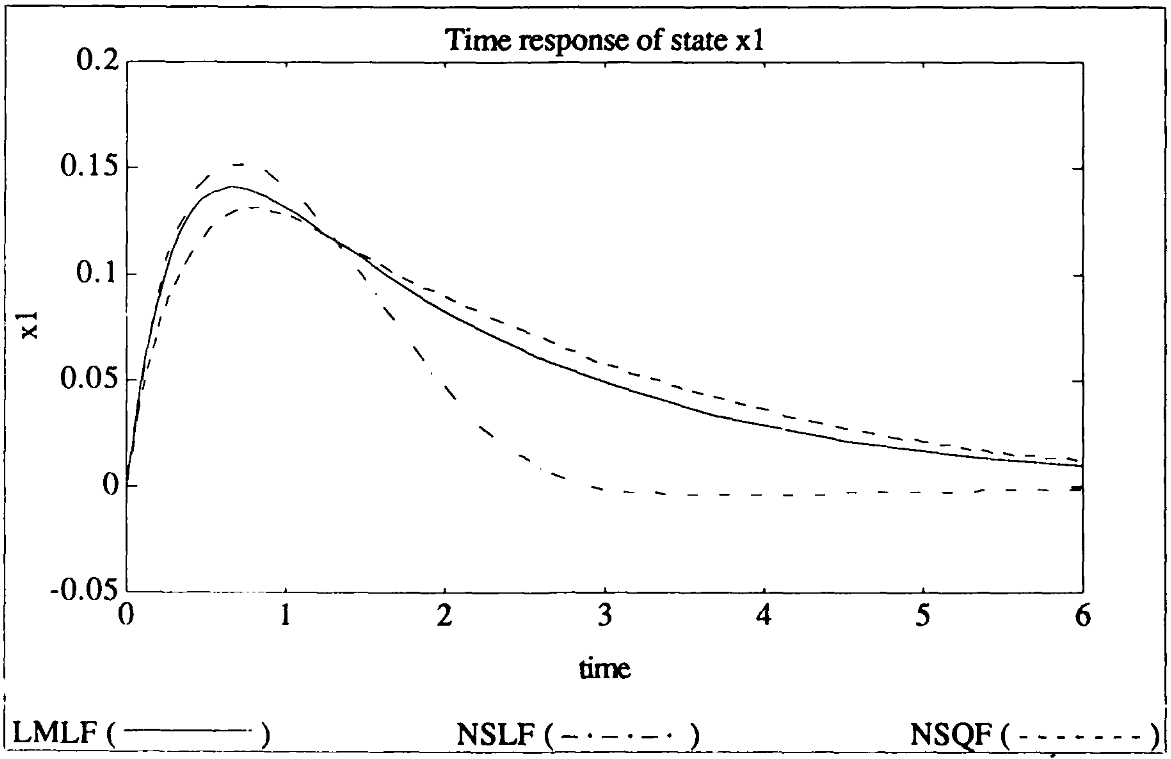


Fig. 5.57. Free response of state x_1 of Example 3 with nonzero initial condition #4.

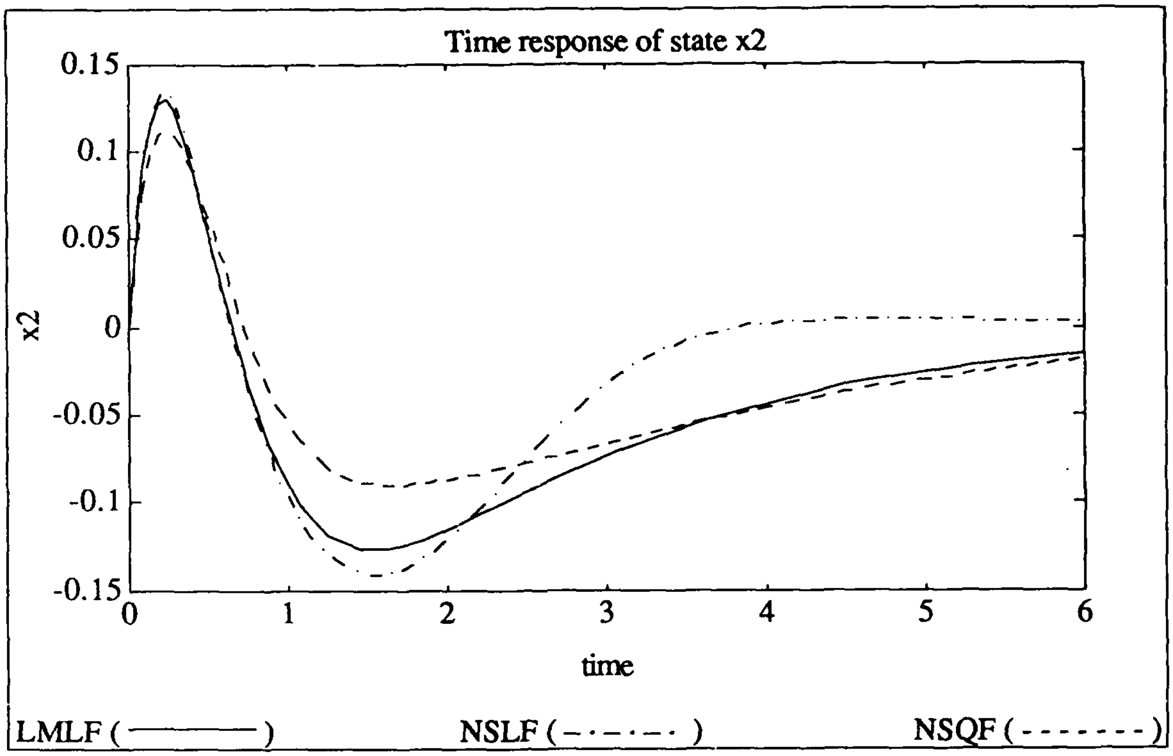


Fig. 5.58. Free response of state x_2 of Example 3 with nonzero initial condition #4.

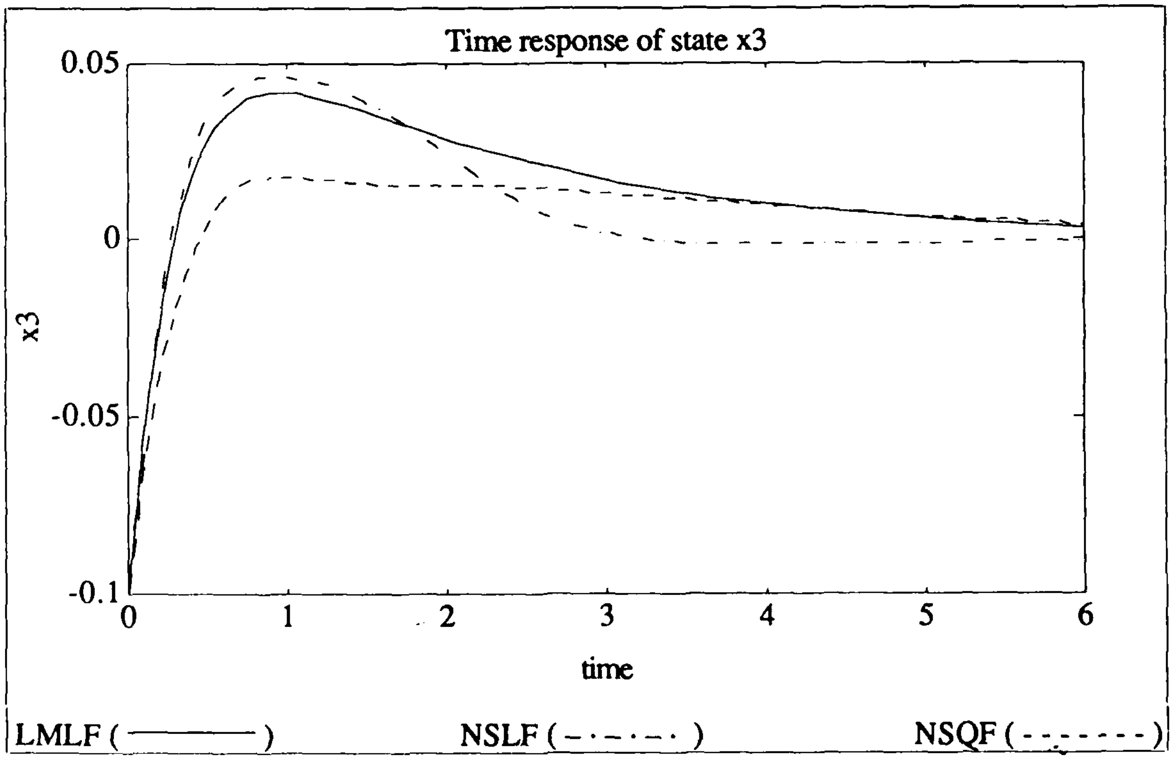


Fig. 5.59. Free response of state x_3 of Example 3 with nonzero initial condition #4.

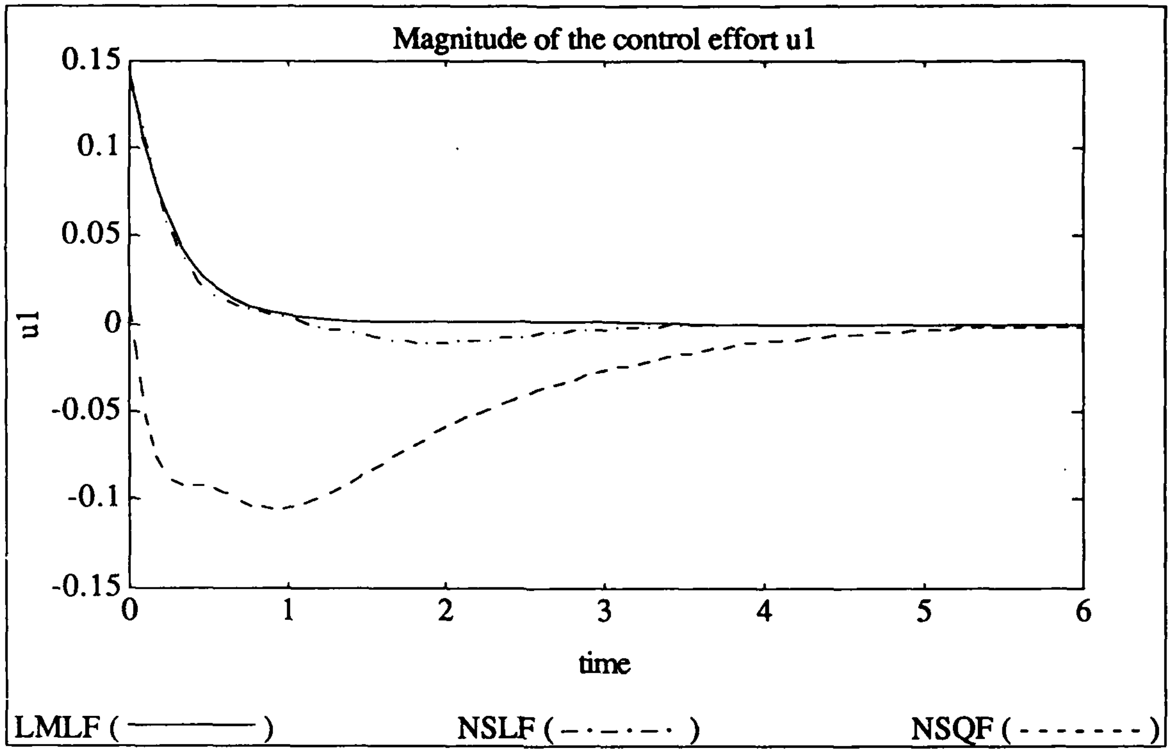


Fig. 5.60. Magnitudes of the inputs for Example 3 with nonzero initial condition #4.

As the final simulation of Example 3, we present the response of the system to a sinusoidal reference input of $r = \sin(6t)$. As in the Examples 1 and 2 presented earlier, the average value of the sinusoidal response gets closer to zero for the NSQF. This improvement is more apparent in the graphs of x_1 and x_3 in Figs. 5.61 and 5.63. For the response x_1 of the NSLF in Fig. 5.61, one can observe substantial nonzero average as well as a distorted non-sinusoidal shape in the response curve. The NSLF also displays a phase deviation away from the linear behavior, which is more apparent in the response of x_2 in Fig. 5.62. These three characteristics (namely nonzero average, non-sinusoidal response, and a phase shift away from the linear behavior) are typical characteristics of a nonlinear system, and they have been improved by the quadratic feedback of the NSQF.

In Fig. 5.63, the input for the NSQF is periodic and strongly nonlinear. The nonlinear shape of the input signal is not arbitrary; this specific nonlinearity in the input effectively cancels out the nonlinearities in the system (up to degree two). The magnitudes of the control inputs for LMLF and NSLF are too small to be seen in the given scale. Again this is because the linear feedback gains were small since the open loop poles were close to the closed loop values.

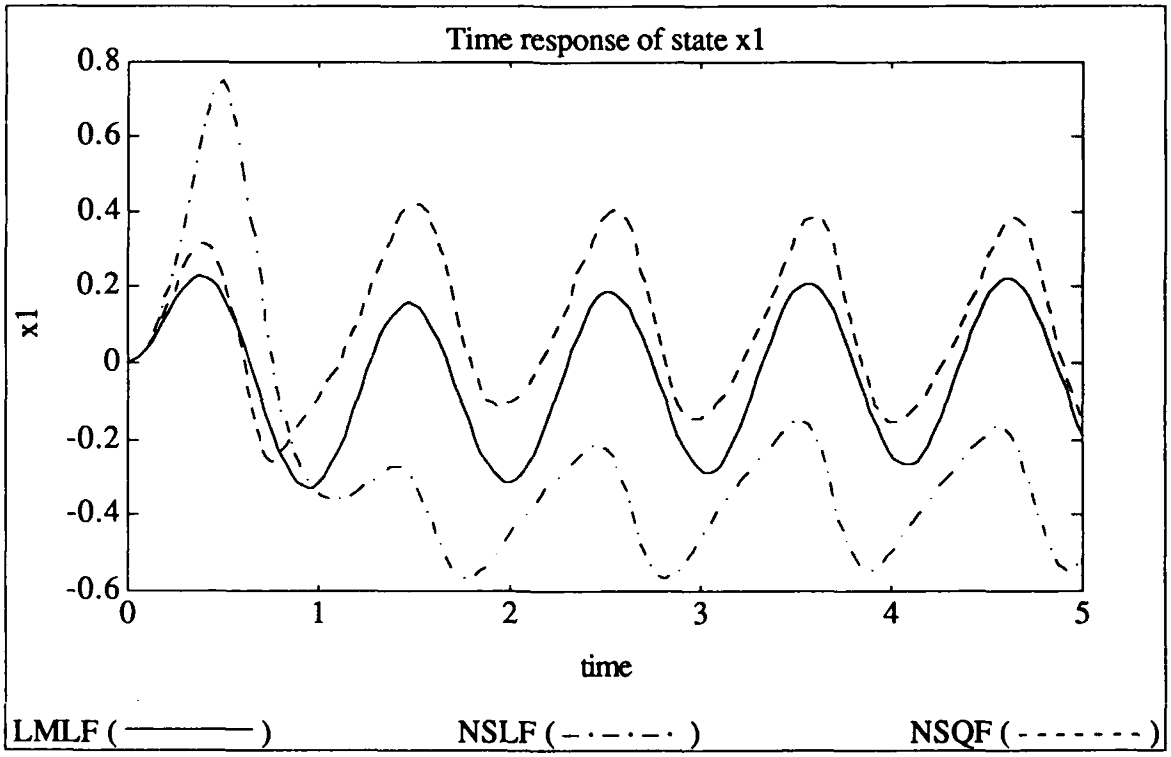


Fig. 5.61. Response of the state x_1 of Example 3 to a sinusoidal input.

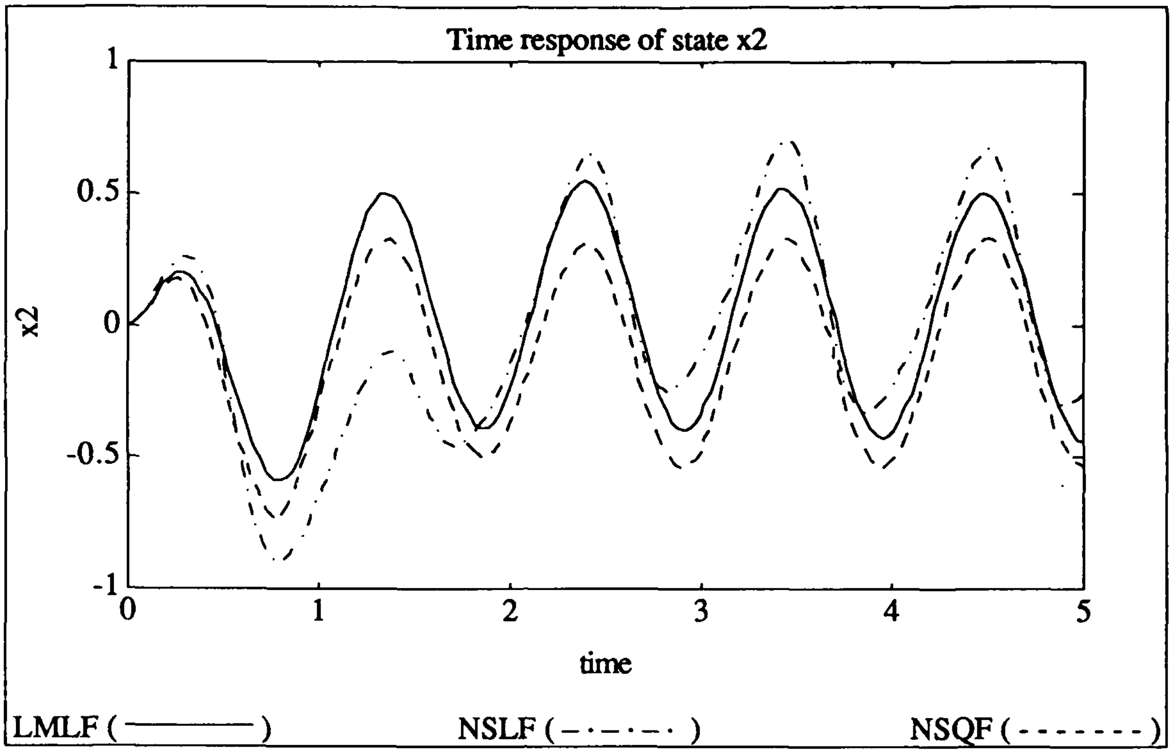


Fig. 5.62. Response of the state x_2 of Example 3 to a sinusoidal input.

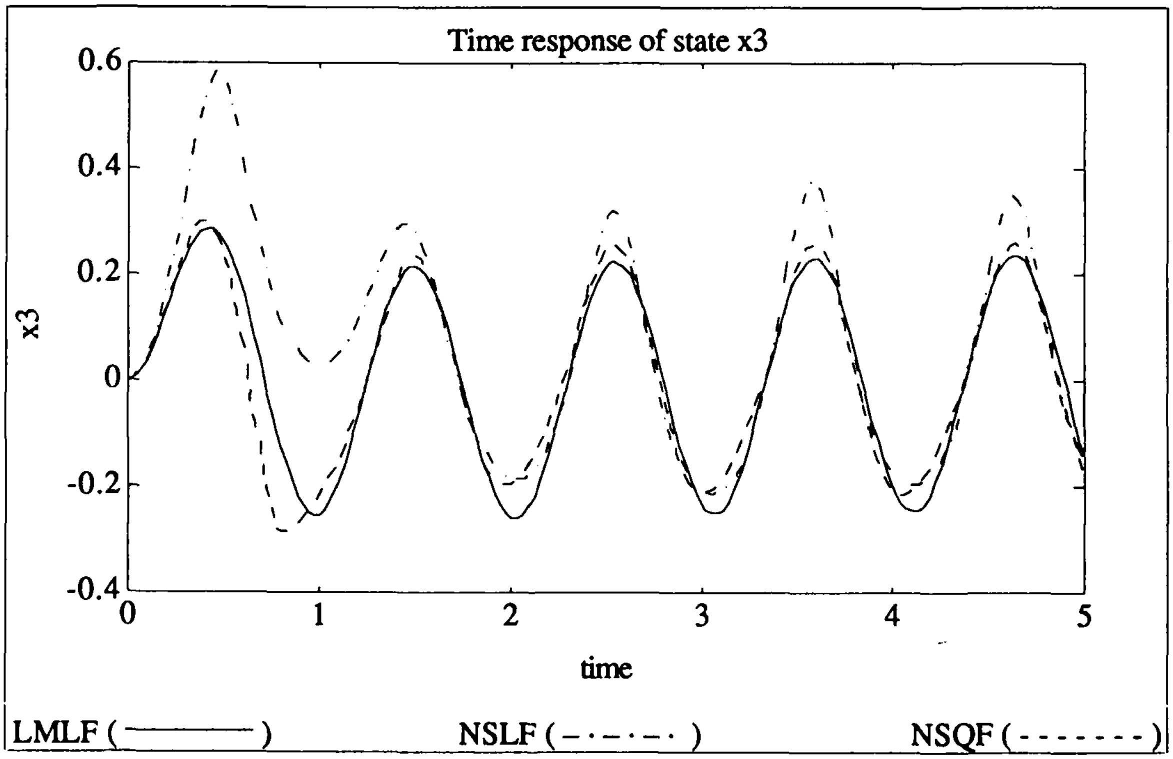


Fig. 5.63. Response of the state x_3 of Example 3 to a sinusoidal input.

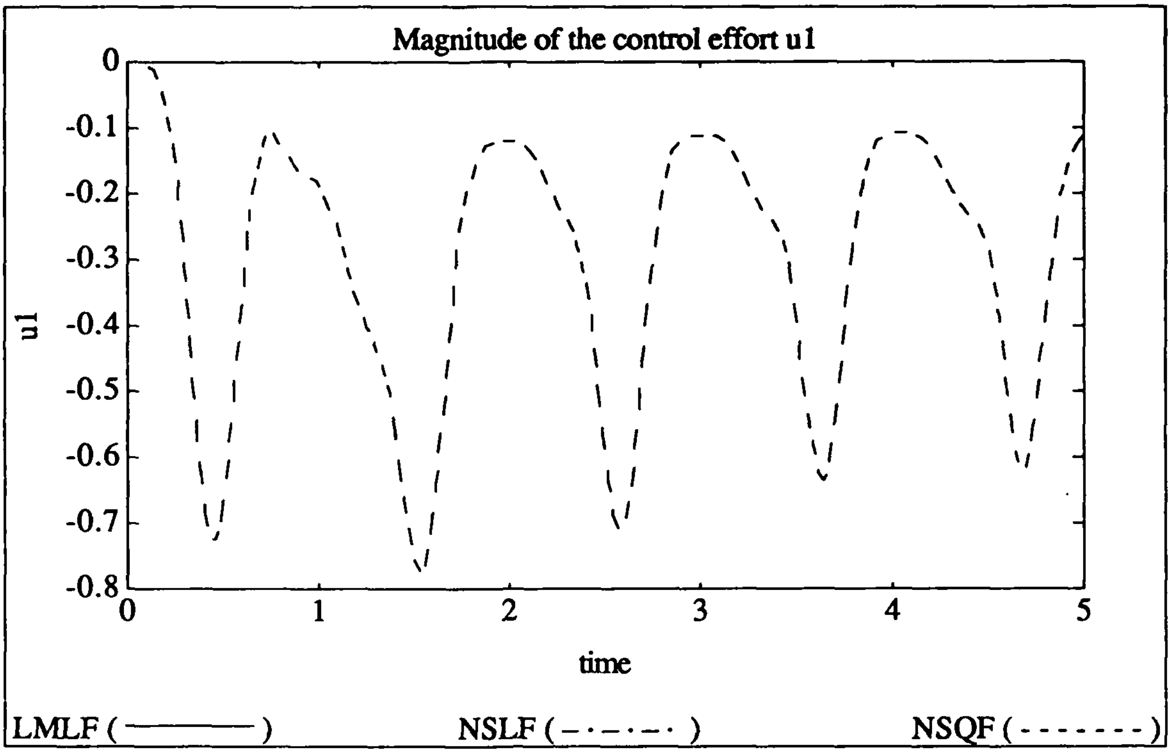


Fig. 5.64. Magnitude of the control efforts for Example 3 for a sinusoidal input.

5.3.4. Nearly Circular Satellite–Example 4:

As it was mentioned in the beginning of this section, in the simulations of Examples 1, 2, and 3, we have presented cookbook nonlinear systems. The first example was exactly linearizable; the second example was not exactly linearizable but in controller canonical form; and the third example had open loop poles very close to the desired closed loop poles.

The fourth example for the approximate linearization method that we present now, is a physical system. The fo'lowing problem has been adopted from Kailath [26]. We consider a satellite of mass *m* in planar earth orbit specified by its position and velocity in polar coordinates as

$$x \triangleq [r \ \ \dot{r} \ \theta \ \dot{\theta} \ ]' \tag{5.26}$$

where r is the orbital radius and θ is the angle of the radius vector with respect to a stationary, inertial reference line. The input thrusts or forces are written as

$$u = [u_r \ \ u_\theta \ ]' \tag{5.27}$$

which are inputs along the radial and tangential directions. These are applied by using small rocket engines. The equations of motion are:

$$\dot{x} = \begin{bmatrix} \dot{r} \\ r\dot{\theta}^2 - \frac{k}{r^2} + \frac{u_r}{m} \\ \dot{\theta} \\ \frac{-2\dot{r}\dot{\theta}}{r} + \frac{u_\theta}{mr} \end{bmatrix} \tag{5.28}$$

where m is the satellite mass and k is the orbital constant. When the equations of motion are linearized around a nominal orbit, the system is locally controllable with the above inputs. If either one of the controls is assumed to be lost, the linearized system is still controllable from only u_θ but not from only u_r.

In the following example we assume that $u_r \equiv 0$ and we control the satellite with u_θ only. The point of this assumption is that when both controls are present, all the nonlinear terms in the system can be cancelled exactly by appropriate feedback, and a nonlinear coordinate change becomes unnecessary. Furthermore, when both of the inputs are present, the system is exactly linearizable up to any desired degree. Thus the problem of feedback linearization using both of the inputs becomes an uninteresting exercise in algebra. With tangential thrust alone, however, the system becomes more challenging to control, and a nonlinear coordinate transformation is needed in addition to nonlinear feedback. With this assumption we present a stronger demonstration of the effectiveness of the approximate linearization method.

In the following, the system equations will be rewritten around the nominal states

$$x_{\text{nom}} = [r_0 \;\; 0 \;\; \omega_0 t \;\; \omega_0]$$

and the state vector x is redefined as perturbations from this nominal orbit:

$$x = [\; \delta r \;\; \delta \dot{r} \;\; \theta - \omega_0 t \;\; \delta\omega \;]' \tag{5.29}$$

Note that the notation in (5.29) has been changed from that of (5.28); the state vector x now represents the vector of perturbations.

When perturbed around the nominal orbit, the Taylor series expansion of the system equations including the second degree terms (i.e. truncated at the third degree in the series expansion) becomes

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega_0^2 & 0 & 0 & 2r_0\omega_0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{-2\omega_0}{r_0} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{mr_0} \end{bmatrix} u_\theta + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{-1}{mr_0^2} x_1 \end{bmatrix} u_\theta$$

$$+\begin{bmatrix} 0 \\ \frac{-3\omega_0^2}{r_0}x_1^2 + 2\omega_0 x_1 x_4 + r_0 x_4^2 \\ 0 \\ \frac{2\omega_0}{r_0^2}x_1 x_2 - \frac{2}{r_0}x_2 x_4 \end{bmatrix} \tag{5.30}$$

with the numerical values assigned as :

$\omega_0 = \pi/2$ rad/hr; $r_0 = 12.74$ Mm (Mm = 10^6 meters); $m = 1000$ kg.

These parameter values correspond to an earth satellite with a period of 4 hours and an orbit radius of 2 earth radii. The set of values represent an intermediate orbit between a geosynchronous satellite and a low altitude communications satellite.

Note that the satellite model in (5.30) is assumed to be controlled only by the tangential thrust u_θ. In order to apply the approximate linearization procedure, the state equations must be normalized according to the characteristic scales of each state. The following characteristic scales have been assumed for the states and the input:

x_{1c} = 12.74 Mm. = (2 earth radii)

x_{2c} = 1.58 Mm/hr.

$x_{3c} = \pi/2$ radians

$x_{4c} = \pi/2$ rad/hr.

$u_{\theta c}$ = 10,000 kg.Mm/hr.

The eigenvalues of the open-loop system are at $\lambda_{1,2} = 0$ and $\lambda_{3,4} = \pm 1.58j$, i.e. the system is critically stable. A linear quadratic optimal feedback design with unit weight on all the states and the input places the closed loop poles at:

$$\lambda_{1,2} = -2.9882 \pm 3.3655j$$

$$\lambda_{3,4} = -0.2429 \pm 0.1829j.$$

We note that the feedback is calculated in the scaled coordinates. Therefore unit weighting of the states and the input in the cost function of the quadratic optimal feedback design makes sense, unless dictated by other design criteria. The feedback gains that place the poles of the scaled system in the above locations were calculated as:

$$K = [-23.6995 \quad -1.0502 \quad 1.0000 \quad -13.0067]$$

For quadratic linearization, the following coordinate change and feedback was found (The computer program for approximate linearization presents the coordinate change and feedback in the unscaled coordinates so that an implementation of the feedback in the natural coordinates can be readily done):

$$z_1 = x_1 - 0.098137x_1^2 \tag{5.31a}$$

$$z_2 = x_2 - 0.19625x_1x_2 \tag{5.31b}$$

$$z_3 = x_3 - 0.0043866x_1x_2 + 0.0078554x_2x_4 \tag{5.31c}$$

$$z_4 = x_4 - 0.032852x_1^2 - 0.11776x_1x_4 - 0.004875x_2^2 + 0.31637x_4^2 \tag{5.31d}$$

$$\alpha^{(2)}(x) = -1767.1x_1x_2 - 10499x_2x_4 \tag{5.32}$$

$$\beta^{(1)}(x) = -0.19623x_1 + 0.63274x_4 \tag{5.33}$$

The second degree terms in the exactly linearizable system are:

$$f^{(2)} = \begin{bmatrix} 0 \\ -0.5878x_1^2 + 3.1597x_1x_4 + 12.736x_4^2 \\ 0.0014598x_2^2 \\ 0.01947x_1x_2 - 0.15701x_2x_4 \end{bmatrix}$$

$$g^{(1)}(x) = \begin{bmatrix} 0 \\ 0 \\ -6.1665\text{e-}07x_2 \\ -6.16\text{e-}06x_1 \end{bmatrix}$$

The system was simulated first with various nonzero initial conditions and no forcing input. Then a sinusoidal reference input was applied. In all of the following simulations, the curves are represented as:

LMLF (————): Reference Linear Model with Linear Feedback.

NSLF (– · – · – ·): Nonlinear System with the Linear Feedback design.

NSQF (- - - - - - -): Nonlinear System with Quadratic Feedback design.

In the first set of simulations, a displacement of +2 Mm along the radius from the nominal orbit was used as an initial condition, i.e. $x = [2 \;\; 0 \;\; 0 \;\; 0]'$ (initial condition #1). This value corresponds to a deviation of 2,000 kilometers from the nominal orbit. When a satellite is first put up into orbit, it is likely to encounter deviations of up to 20% from its nominal orbital radius. Therefore the initial condition chosen is a realistic value.

The response in Fig. 5.65 is for the radial displacement of the satellite from the nominal orbit. The NSQF follows the LMLF closely along the trajectory. In the radial velocity response, the NSQF is very close to the LMLF whereas the NSLF displays an overshoot first, and it does not get close to the linear system for a longer period of time. In Fig. 5.67 of the reference angle, the NSLF exhibits a large deviation from the linear ideal model LMLF. The NSQF tracks the linear ideal model extremely closely in this case. It appears from the large deviation exhibited by NSLF from the linear ideal model LMLF in Fig. 5.67 that the nonlinearities that affect the dynamics of the satellite along the tangential direction of the motion, i.e. those which influence the angular deviation, have a stronger effect on the motion. The nonlinear control is very effective in reducing these undesired nonlinear effects. This result is also due to the fact that the satellite is being controlled by the tangential thrust alone

The behavior of the NSQF in the plot of the angular velocity, Fig. 5.68, is very close to the LMLF. The large initial overshoot of the NSLF in the angular velocity is obviously the cause of the deviation in the angular displacement, since the displacement is the integral of the velocity.

The plot of the input efforts in Fig. 5.69 show that the control for the NSQF is closer to the LMLF. In the input plots, the initial value of the thrust was vey large for all three systems (about 40,000). Since there was not an appreciable difference between these curves for the first 1/4 hrs. of the simulation, part of the plot was removed to readjust the scale so that the difference between the inputs could be distinguished.

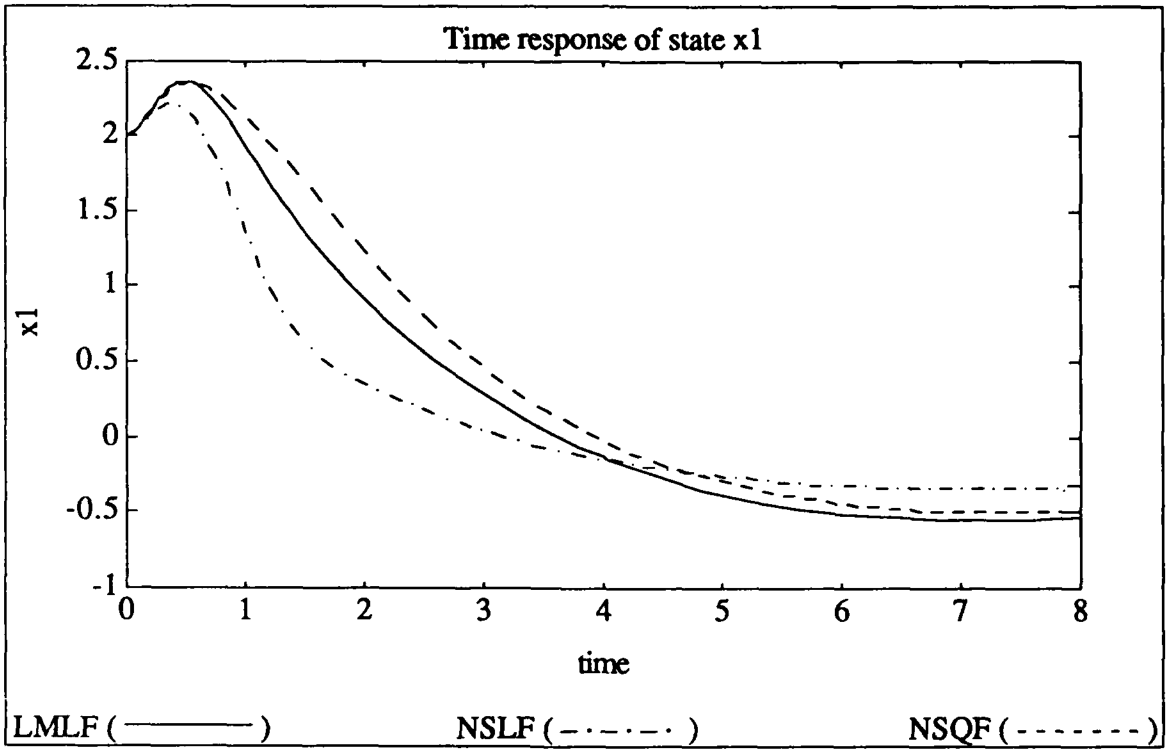


Fig. 5.65. Free response of state x_1 of Near Circular Satellite with nonzero initial condition #1.

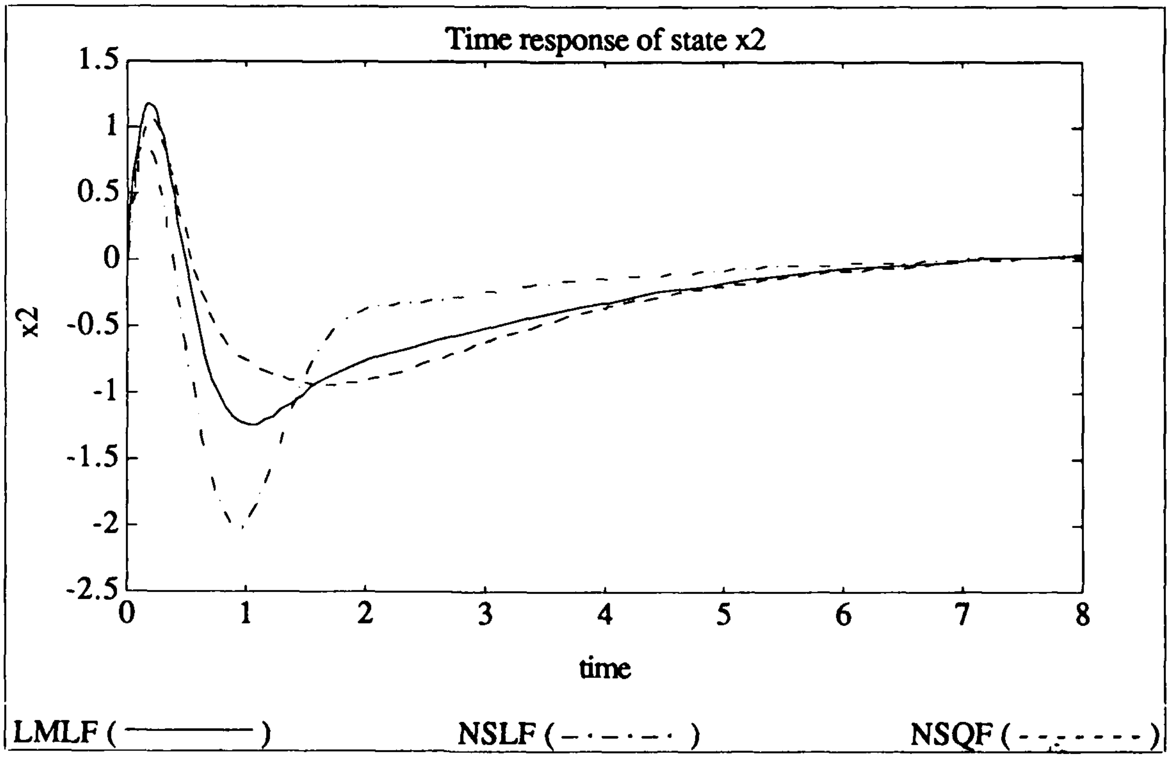


Fig. 5.66. Free response of state x_2 of Near Circular Satellite with nonzero initial condition #1.

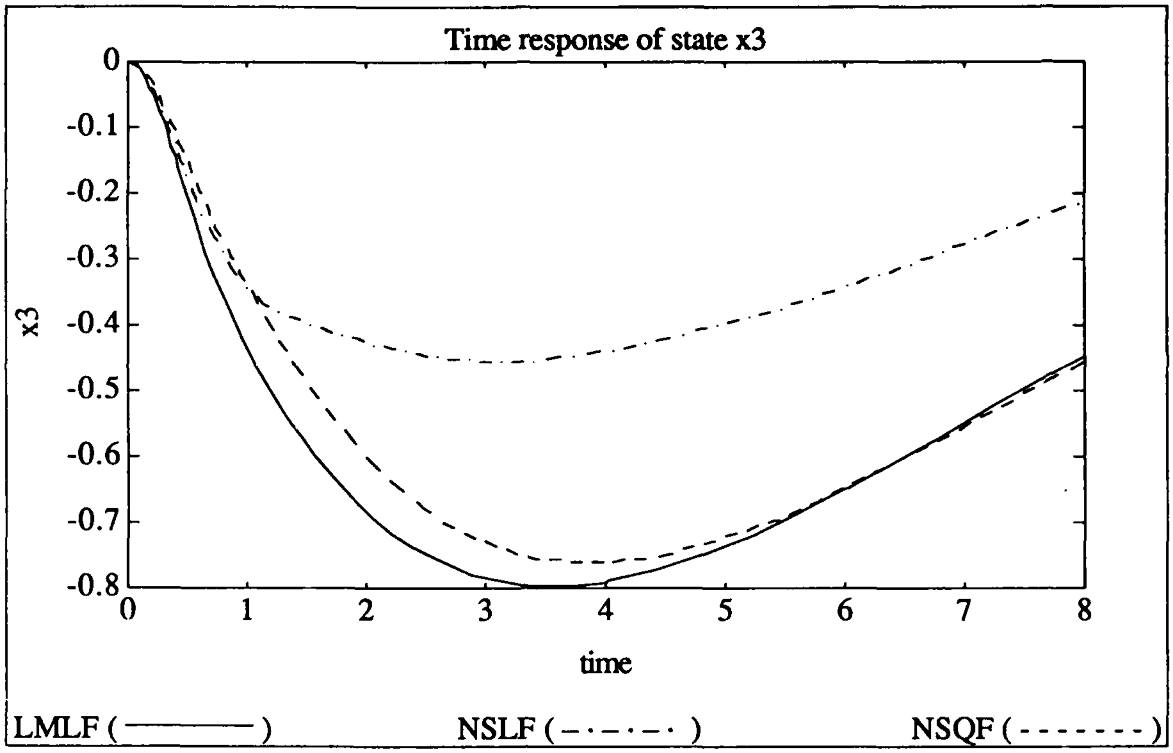


Fig. 5.67. Free response of state x_3 of Near Circular Satellite with nonzero initial condition #1.

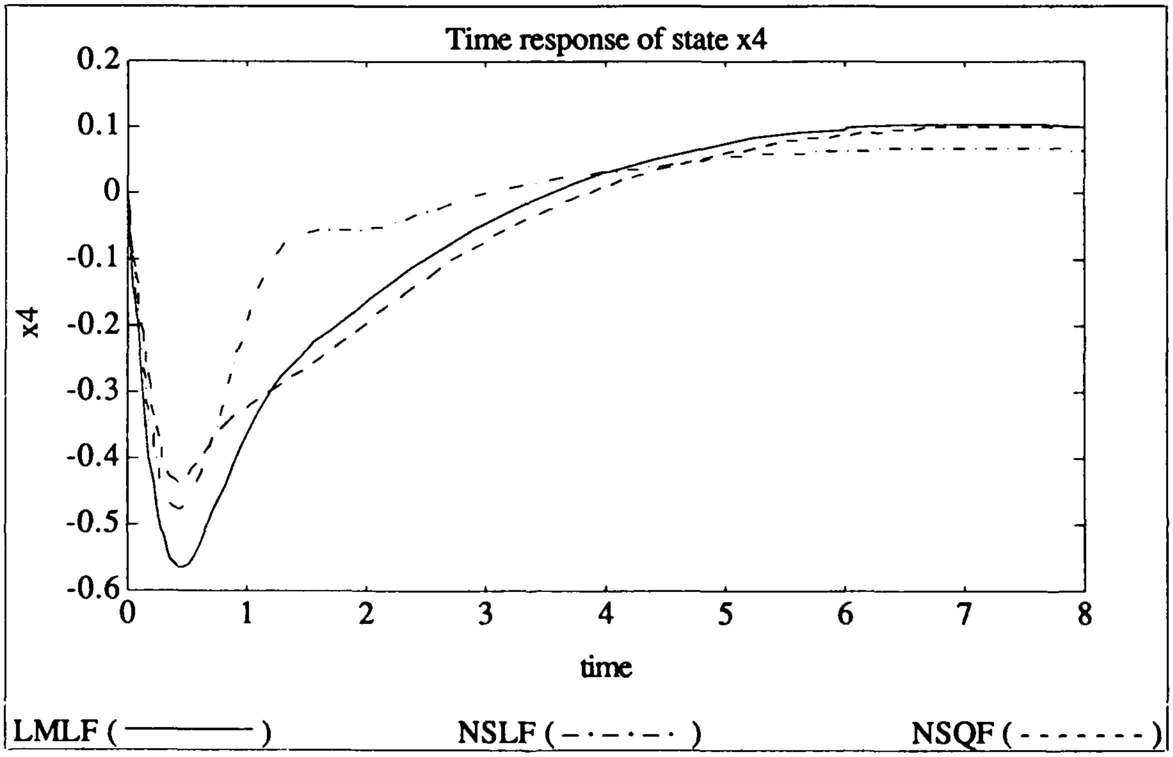


Fig. 5.68. Free response of state x_4 of Near Circular Satellite with nonzero initial condition #1.

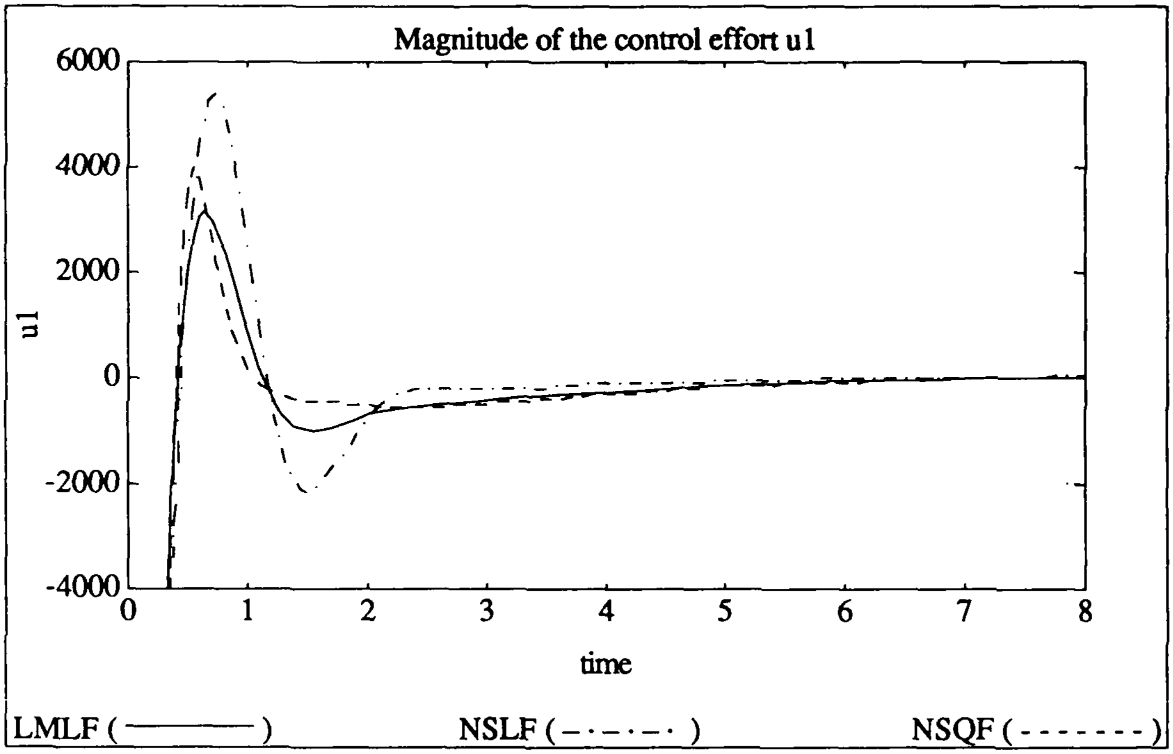


Fig. 5.69. Magnitudes of the inputs for Near Circular Satellite with nonzero initial condition #1.

In the Figures 5.70 to 5.74 the responses to an initial displacement of –2 Mm from the nominal orbit along the radius are shown (initial condition #2). This value is the negative of the initial condition #1, and it is intended as a test to check whether the response will display sensitivity to initial conditions, which was found to be absent in the NSQF of Example 3.

Comparing Fig. 5.65 for the initial condition #1 and Fig. 5.70 for the response of the radial deviation from the nominal orbit, we observe that the tracking by the NSQF of the LMLF is quite satisfactory. The NSQF displays a response that is only slightly sensitive to initial conditions. We note that when evaluating initial condition sensitivity for the nonlinear system, we are actually looking for the type of behavior typical of linear systems: Are the response curves symmetric with respect to the time axis when the sign of the initial conditions is changed? The two curves of NSFQ in Figs. 5.65 and 5.70 are not exactly symmetric, with a small overshoot displayed by the NSQF in Fig. 5.70 that is absent in the plot of Fig. 5.65. However, the difference is minor, and it does not affect the tracking of the LMLF.

Comparison of x_2 in Figs. 5.71 and 5.66, of x_3 in Figs. 5.72 and 5.67, and of x_4 in Figs. 5.73 and 5.68 between the two sets of initial condition responses #1 and #2 all show that the response of the NSQF is slightly initial condition dependent, but not as strongly as the curves of the NSLF. Immediately after the transient part of the response, the NSQF approaches extremely close to LMLF.

The plot for the magnitude of the inputs in fig. 5.74 has been rescaled to distinguish between the different inputs. Similar to the input plot of initial condition #1, in the section of the plot that is not shown, all the inputs were very close to each other. This rescaling helps us in distinguishing the following feature: For all the improvement achieved in the tracking of the linear ideal model by the NSQF, it is surprising that the input effort needed for this is actually smaller than those needed by the controls of LMLF and NSLF. This is a feature that was also present in the input plots of Fig. 5.69 of the initial condition #1.

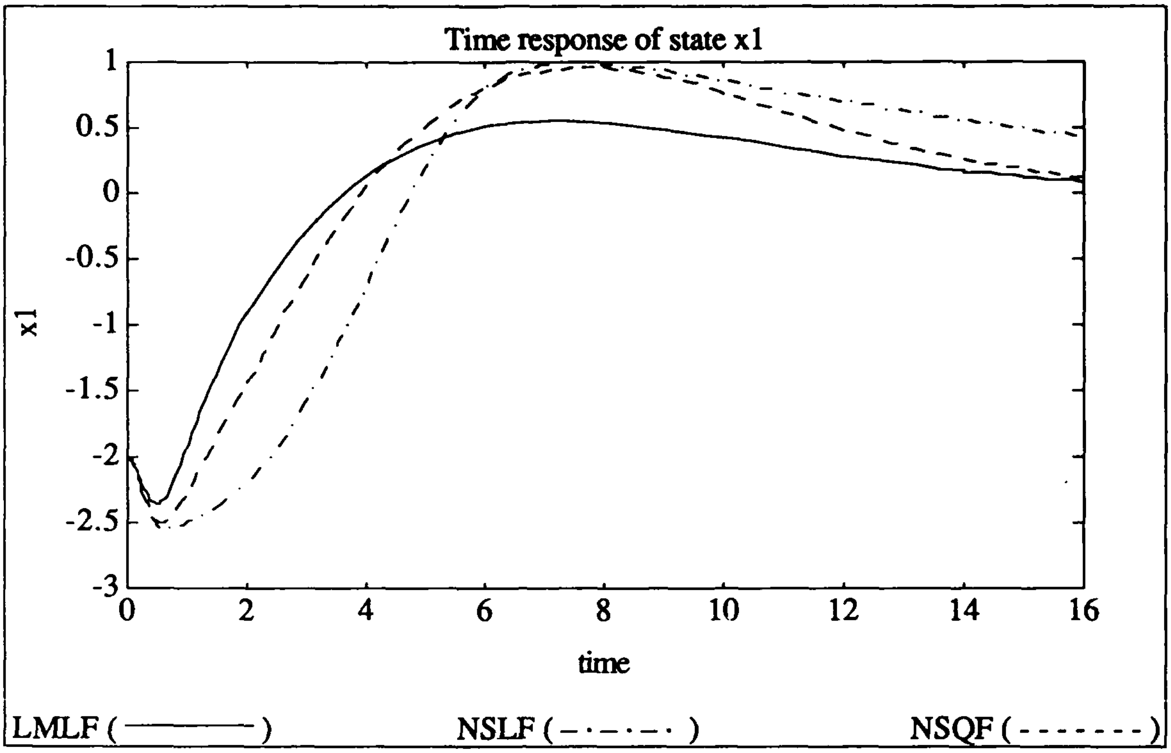


Fig. 5.70. Free response of state x_1 of Near Circular Satellite with nonzero initial condition #2.

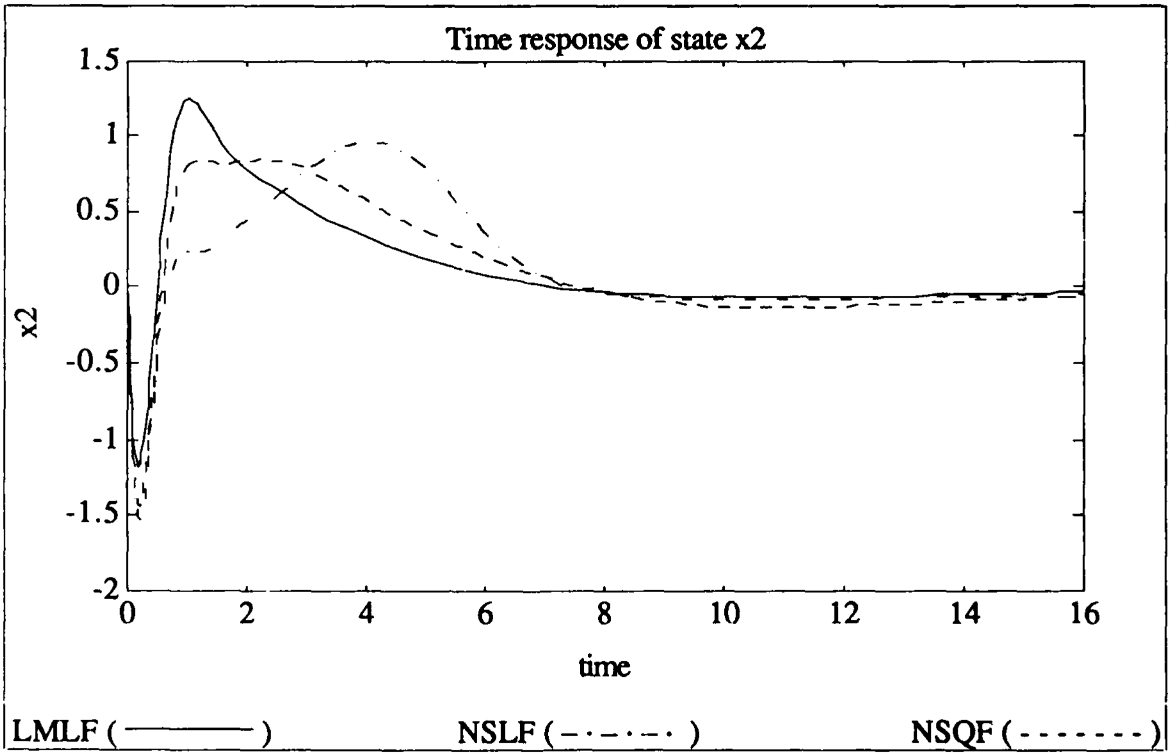


Fig. 5.71. Free response of state x_2 of Near Circular Satellite with nonzero initial condition #2.

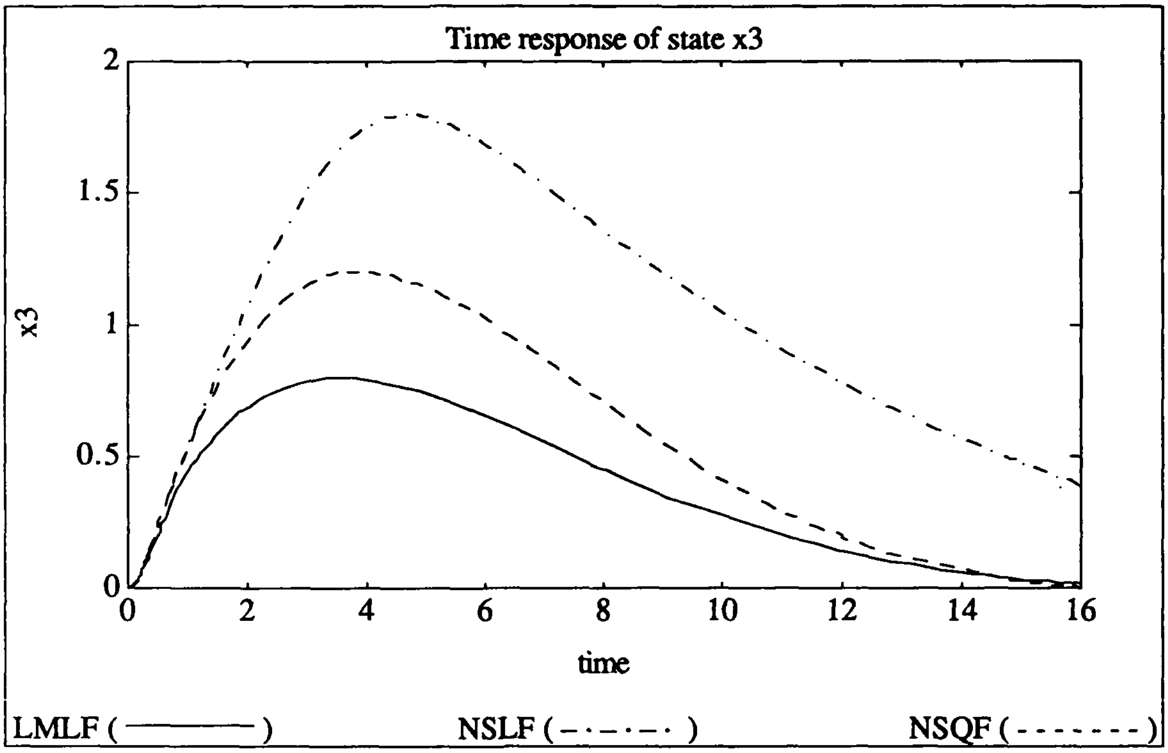


Fig. 5.72. Free response of state x_3 of Near Circular Satellite with nonzero initial condition #2.

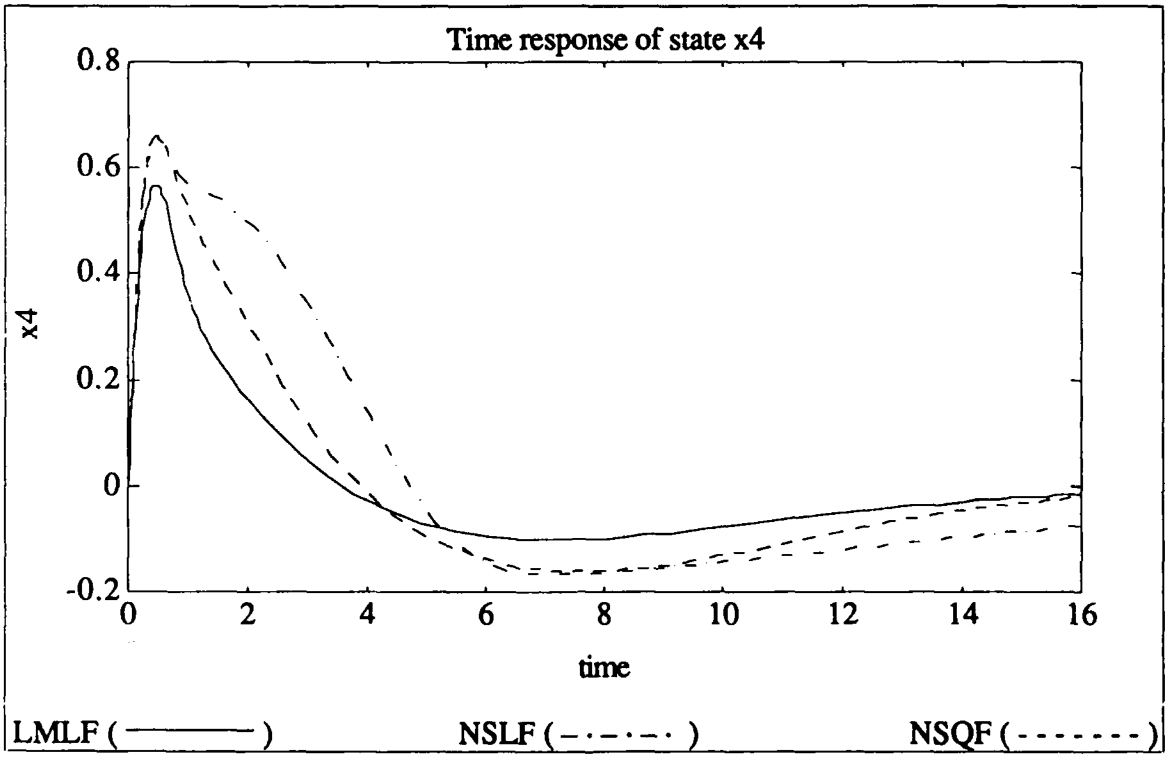


Fig. 5.73. Free response of state x_4 of Near Circular Satellite with nonzero initial condition #2.

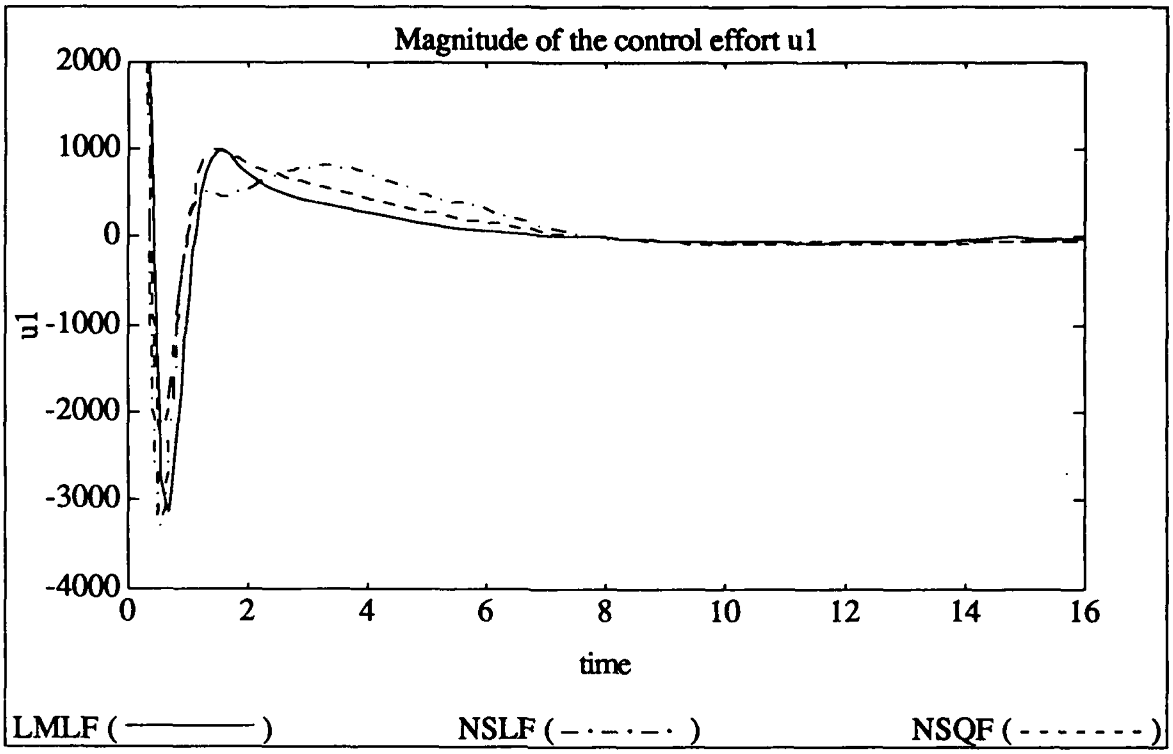


Fig. 5.74. Magnitudes of the inputs for Near Circular Satellite with nonzero initial condition #2.

An angular displacement of 1 radian from the nominal orbit has been applied for the following simulations (x = [0 0 1 0]'). The NSQF tracks the LMLF very closely in the radial displacement (Fig.5.75), radial velocity (Fig. 5.76), angular displacement (Fig. 5.77), and angular velocity (Fig. 5.78). These response curves show that the quadratic control is surprisingly effective in tracking the ideal linear model, especially when compared with the linear control of the nonlinear system, the NSLF. Please note that in the simulations of both the NSLF and the NSQF, we have integrated the *truncated* equations of motion presented in Eqn. (5.30).

The NSLF displays large transient errors and longer settling times in all the states. A reason why the NSQF is so effective in improving the system behavior in responding like the LMLF in the state of the angular position is because the system is being controlled by a tangential thrust, which directly affects the angular acceleration.

Because of the initially large magnitude of the input thrusts in Fig. 5.79, the scale of the plot was too large to allow the different inputs to be clearly distinguished, and we truncated a small portion of the plot by rescaling the axes. This is why the input curves are missing for the first 1/3 hours of the simulation. In this truncated part, the input values for all systems were as high as 3000, and they were very close to each other. Since we are running fictitious simulations, we can afford such large input forces. In reality, however, such thrust values may not be available for being able to recover from the large initial deviations we have imposed on the system, and longer settling times would be physically more realistic.

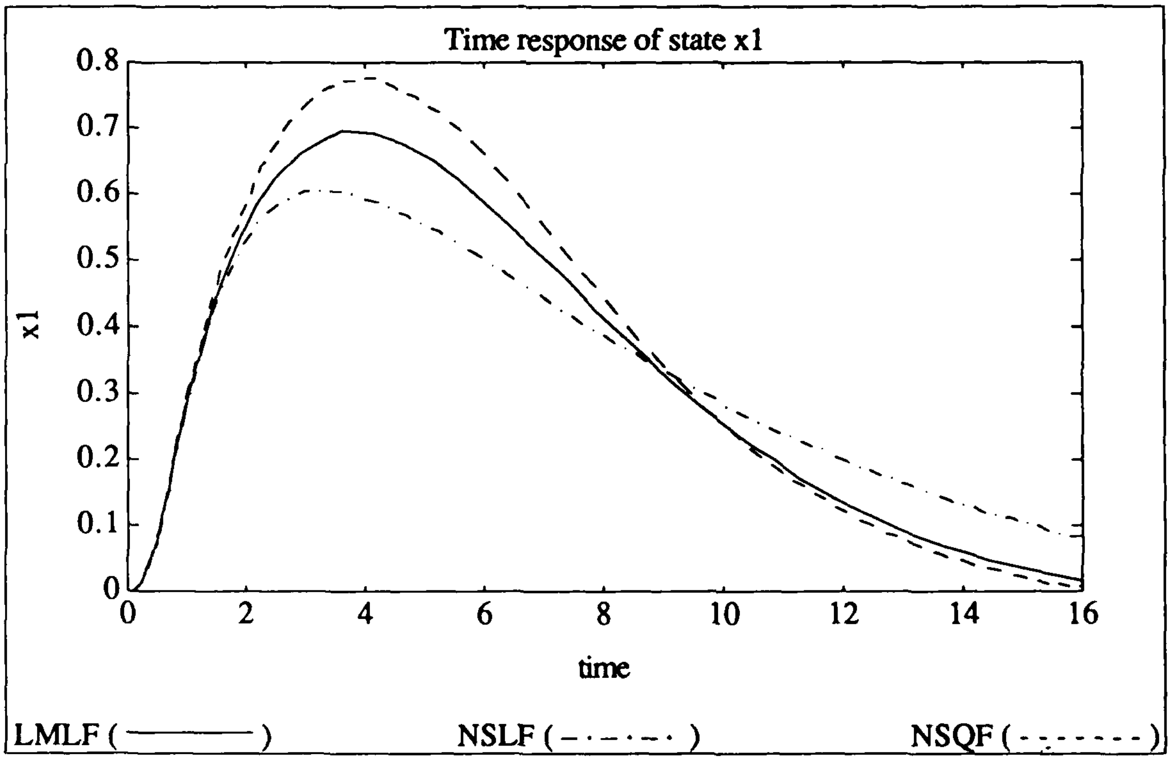


Fig. 5.75. Free response of state x_1 of Near Circular Satellite with nonzero initial condition #3.

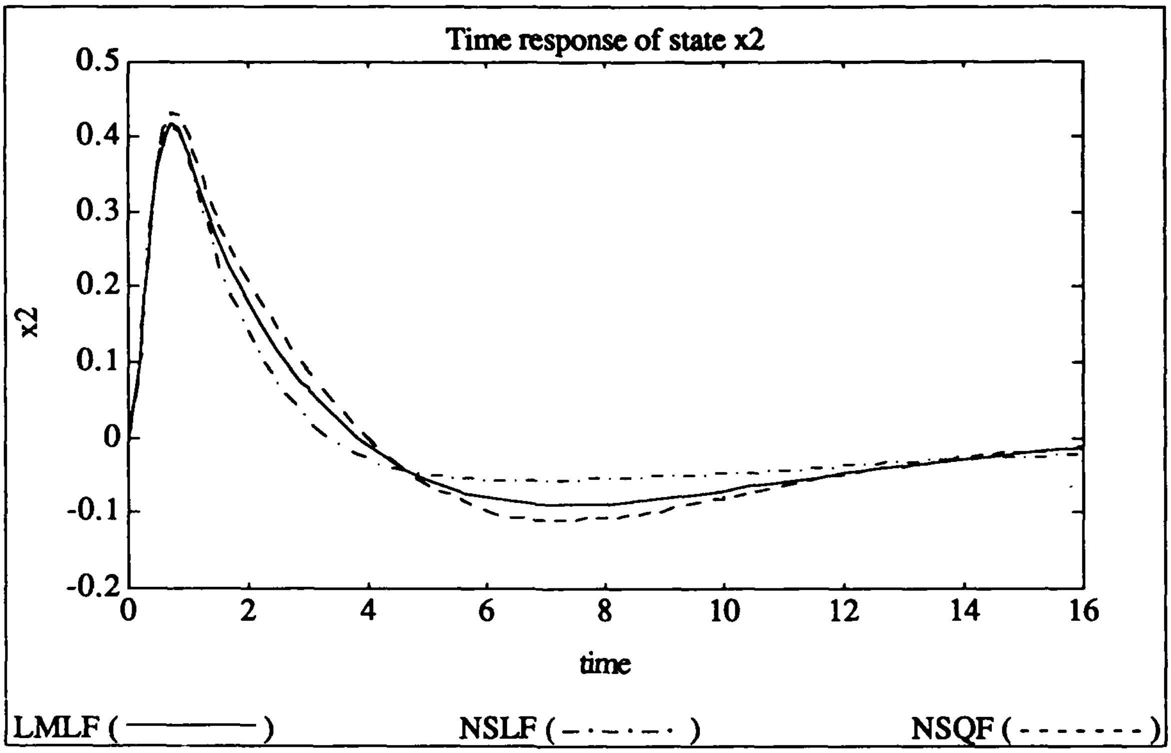


Fig. 5.76. Free response of state x_2 of Near Circular Satellite with nonzero initial condition #3.

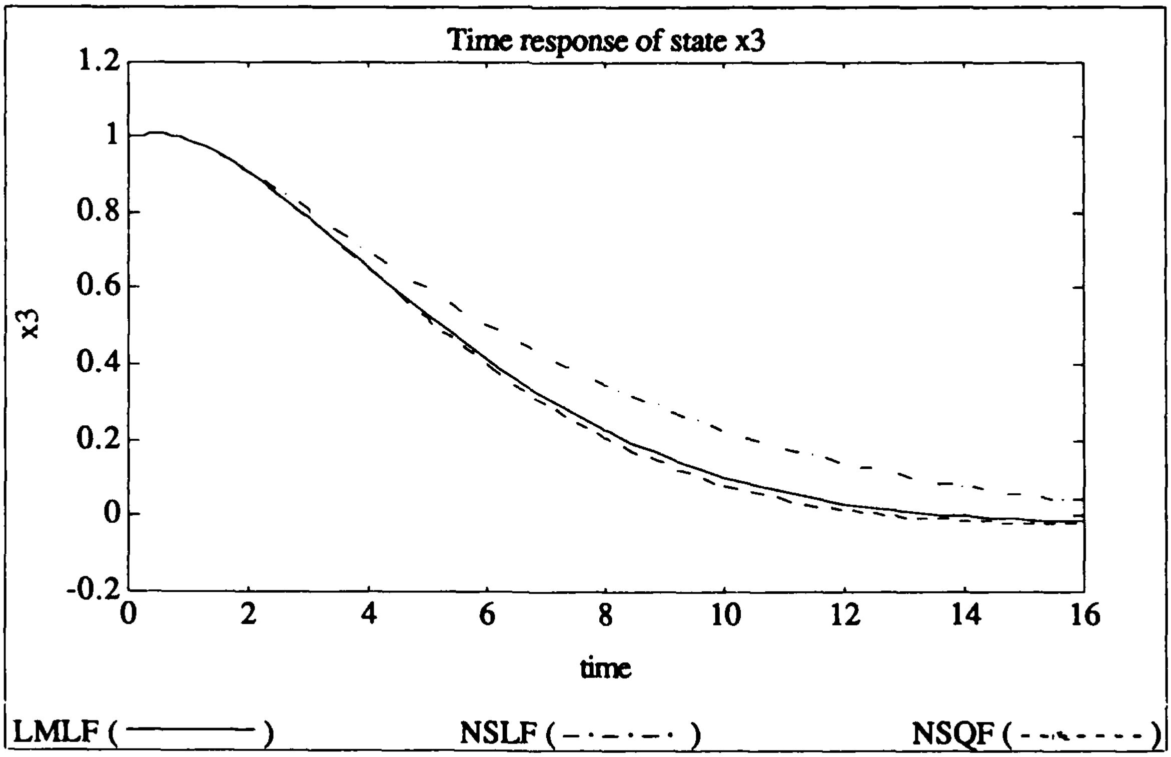


Fig. 5.77. Free response of state x_3 of Near Circular Satellite with nonzero initial condition #3.

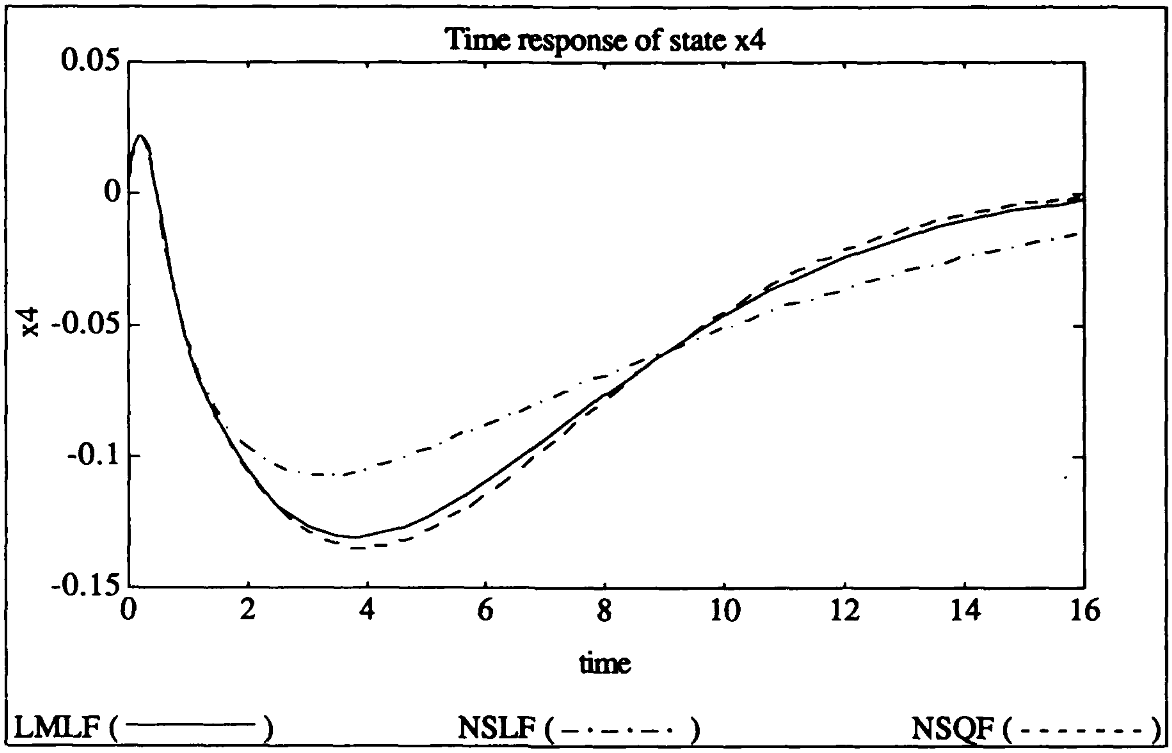


Fig. 5.78. Free response of state x_4 of Near Circular Satellite with nonzero initial condition #3.

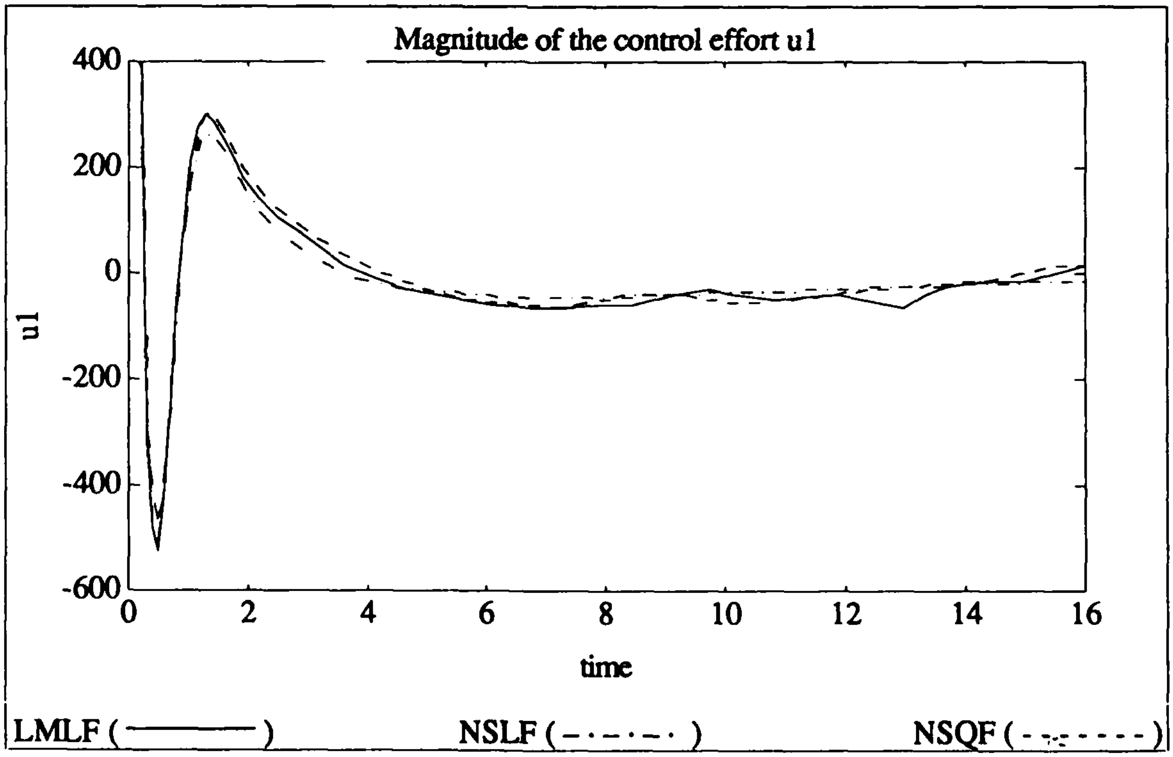


Fig. 5.79. Magnitudes of the inputs for Near Circular Satellite with nonzero initial condition #3.

Figures 5.80 through 5.84 show simulation results for an angular displacement of –1 radian from the nominal orbit, an initial condition with a sign opposite of the previous case. In the captions of these plots, this case is referred to as initial condition #4. The discussion for the responses in this case is nearly identical to that of the initial condition #3. The NSQF is extremely succesful in the tracking of the LMLF in all the states in the Figs. 5.80, 5.81, 5.82 and 5.83. The sensitivity properties to a change in initial conditions is excellent, and in this sense the NSQF behaves almost exactly like a linear system.

As in the earlier simulation, the input effort is initially very large, and the difference between various controls could be clearly seen only when a small initial portion of the graph was truncated.

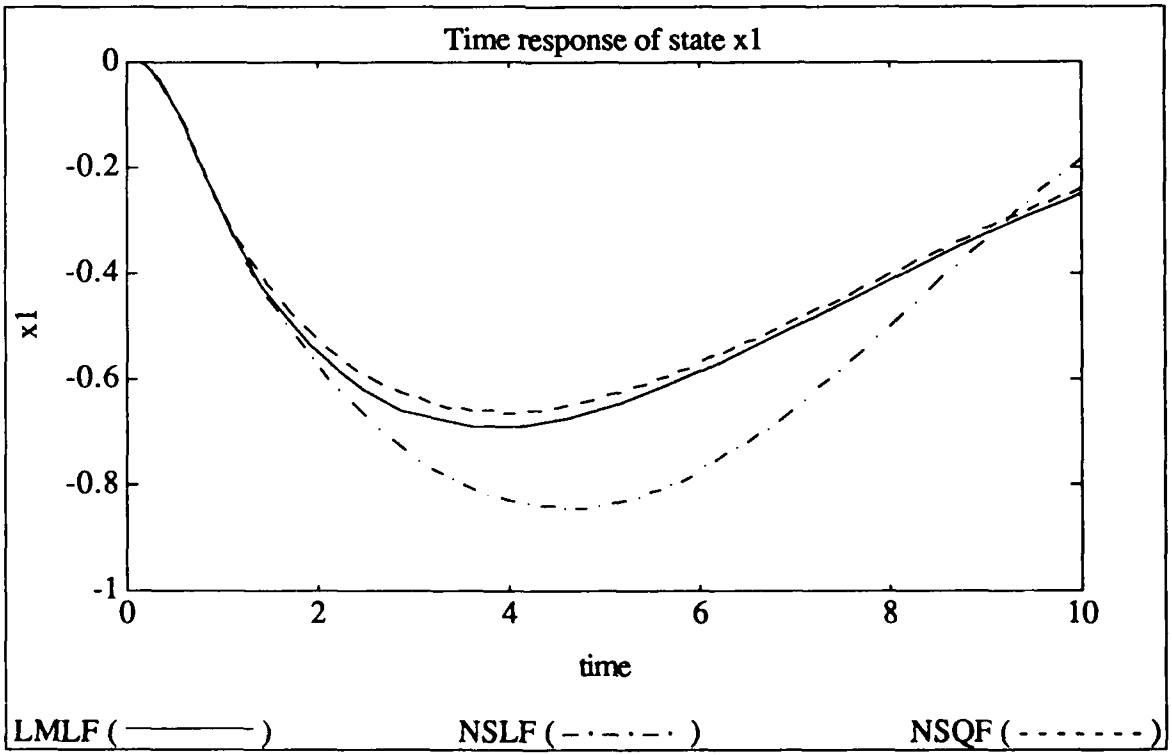


Fig. 5.80. Free response of state x_1 of Near Circular Satellite with nonzero initial condition #4.

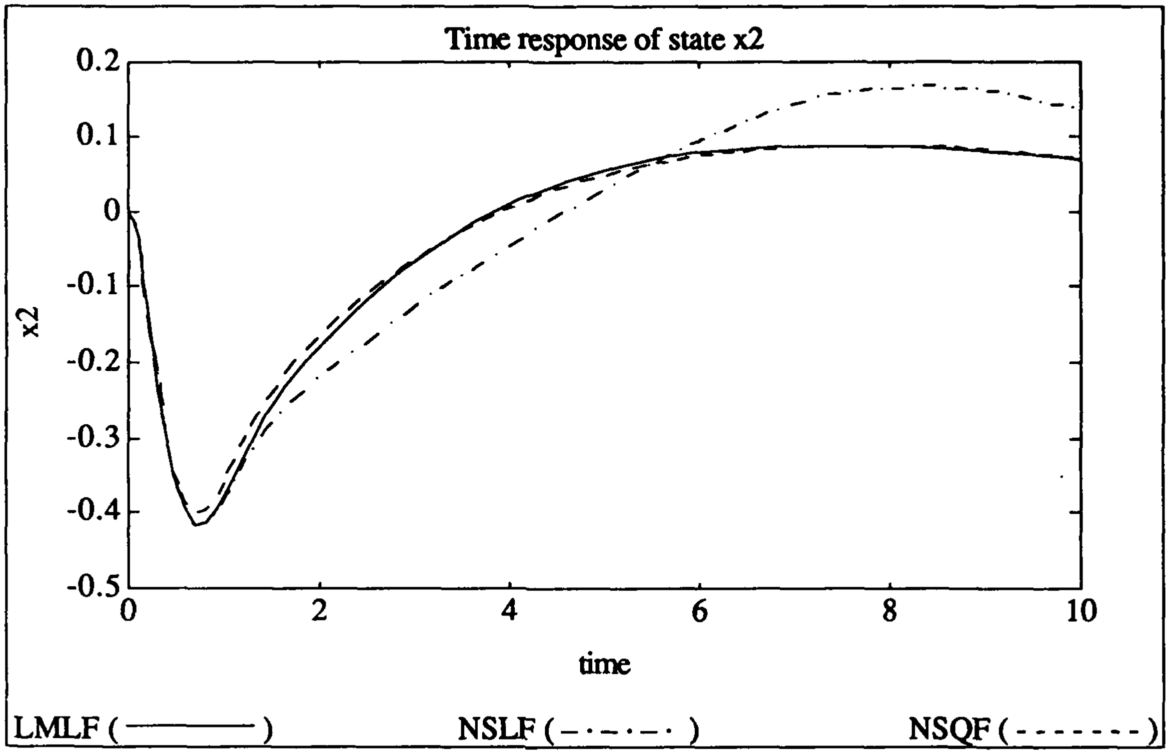


Fig. 5.81. Free response of state x_2 of Near Circular Satellite with nonzero initial condition #4.

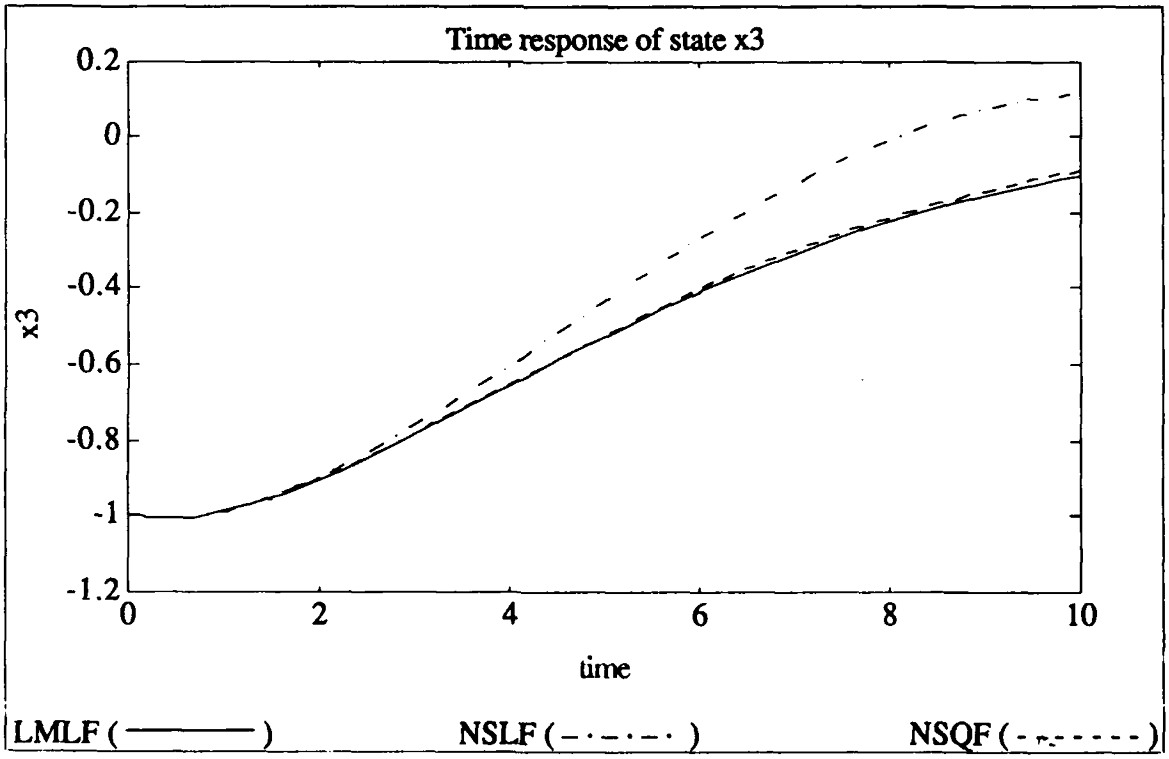


Fig. 5.82. Free response of state x_3 of Near Circular Satellite with nonzero initial condition #4.

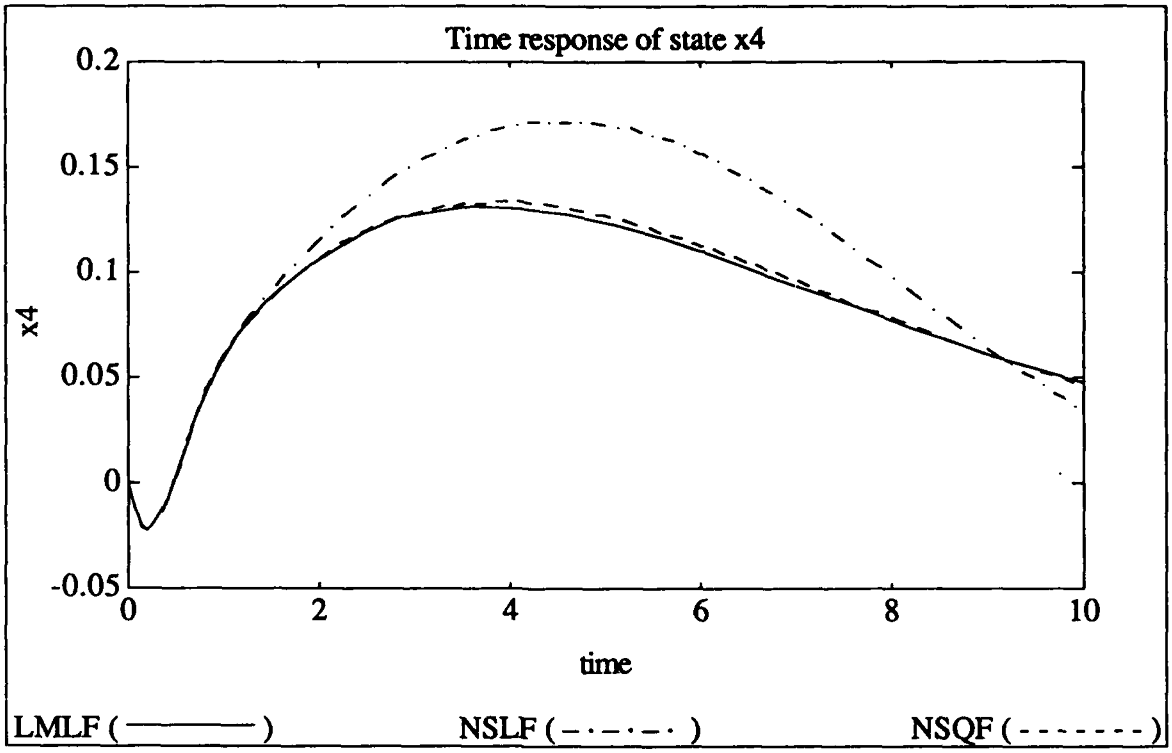


Fig. 5.83. Free response of state x_4 of Near Circular Satellite with nonzero initial condition #4.

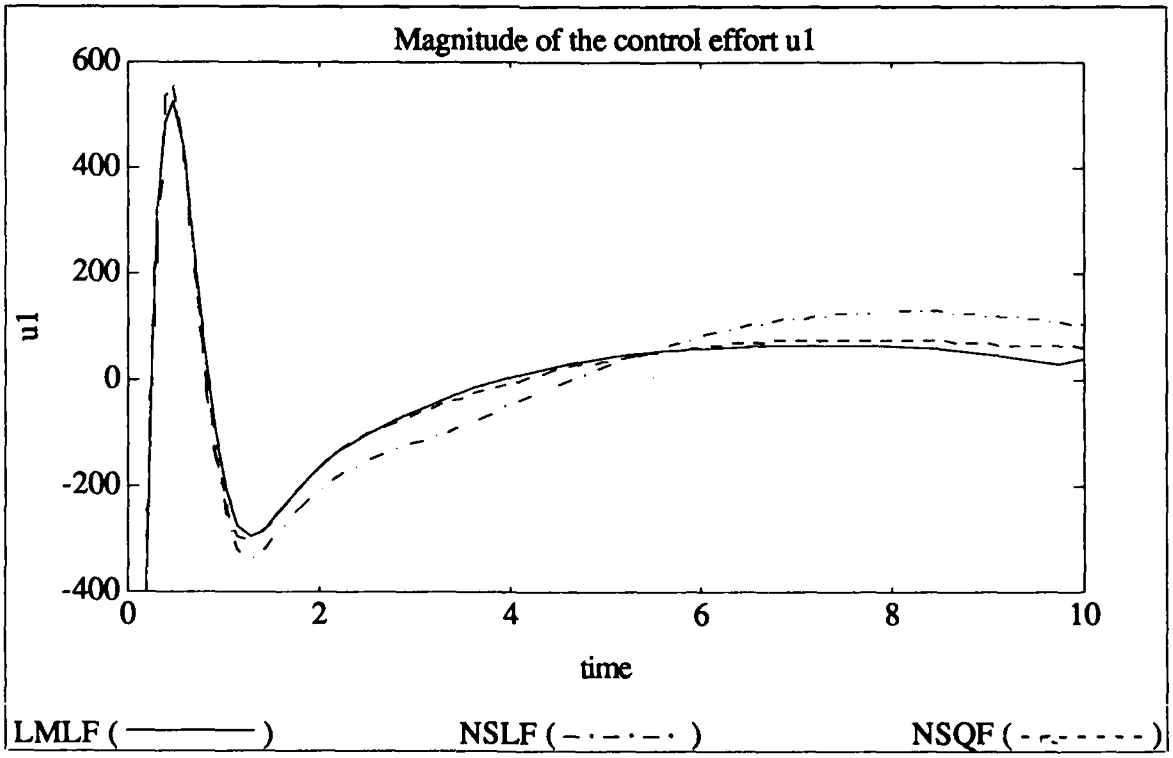


Fig. 5.84. Magnitudes of the inputs for Near Circular Satellite with nonzero initial condition #4.

Finally, for the last simulation, a sinusoidal forcing of 30,000sin(6*t*) has been applied to the satellite. Note that this reference signal has the same units as the input. The amplitude of the sinusoidal forcing magnitude may seem very large, but we note that as seen in Eqn. (5.28), the thrust is divided by the mass and the orbital radius to yield the angular acceleration. This value has been chosen large so as to be able to effectively distinguish between the behavior of the NSLF and the NSQF. Such a large reference forcing is not physically realizable, and this portion of the example is simulated only for the sake of demonstrating the effectiveness of our method. The period of the disturbance was chosen to be close to the orbital period of the satellite. This type of periodic disturbances are typical for a satellite, and they might represent external effects such as solar pressure or gravitational forces due to other celestial objects.

In the radial displacement response of Fig. 5.85 we observe the same phenomenon seen in the simulations of Examples 1, 2, and 3, i.e. the quadratic feedback causes the average value of the sinusoidal response to approach zero. Similar behavior is displayed with a larger deviation in the angular displacement response of Fig. 5.87. The average drift in all the models away from the equilibrium in this fugure was very interesting, and the simulation was extended until $t = 20$ to observe the steady state behavior, which is shown in Fig. 5.88. The NSLF displays a large drift from the reference angle. The LMLF shows a sinusoidal response with zero average. The NSQF is much closer to the LMLF in average than the NSLF. However it is obvious From Fig. 5.88 that the NSQF will also exhibit a steady state average error. This is because the nonlinear system is not linearizable, and the error is due to the residual second degree terms that could not be removed by coordinate change and feedback. Note that the magnitudes of the nonlinearities are significant for these values of the states.

Figs. 5.86 and 5.89 show the radial and angular velocities, respectively. In these curves, the linear and nonlinear control responses are not readily distinguishable. The difference becomes noticeable when the velocities are integrated to obtain the displacements, as discussed above. In the the angular velocity response x_4 of the satellite, we observe a large overshoot in the NSQF. In trying to correct the nonlinearities, the quadratic controller introduces very large deviations.

Fig. 5.90 shows that the input thrust for the NSQF also assumes very large values during the transient part of the response. This may be due to a tendency of the quadratic control to drive the system unstable for large deviations from the equilibrium. However, whether this large input force is really necessary to effectively cancel the nonlinear terms in the system (i.e. are the nonlinear terms really so large to need such a strong force to cancel them), or the higher degree approximation of our approach starts failing for these large values of the states is not clear.

Large values for the input force was also observed for Example 1 presented earlier in this chapter. In order to improve on the performance of the nonlinear controller for some systems that display similar tendencies of instability, certain modifications in the design procedure may have to be made. It is not entirely clear at this moment what these modifications should be, and this is certainly one of the open research questions that extends from this work.

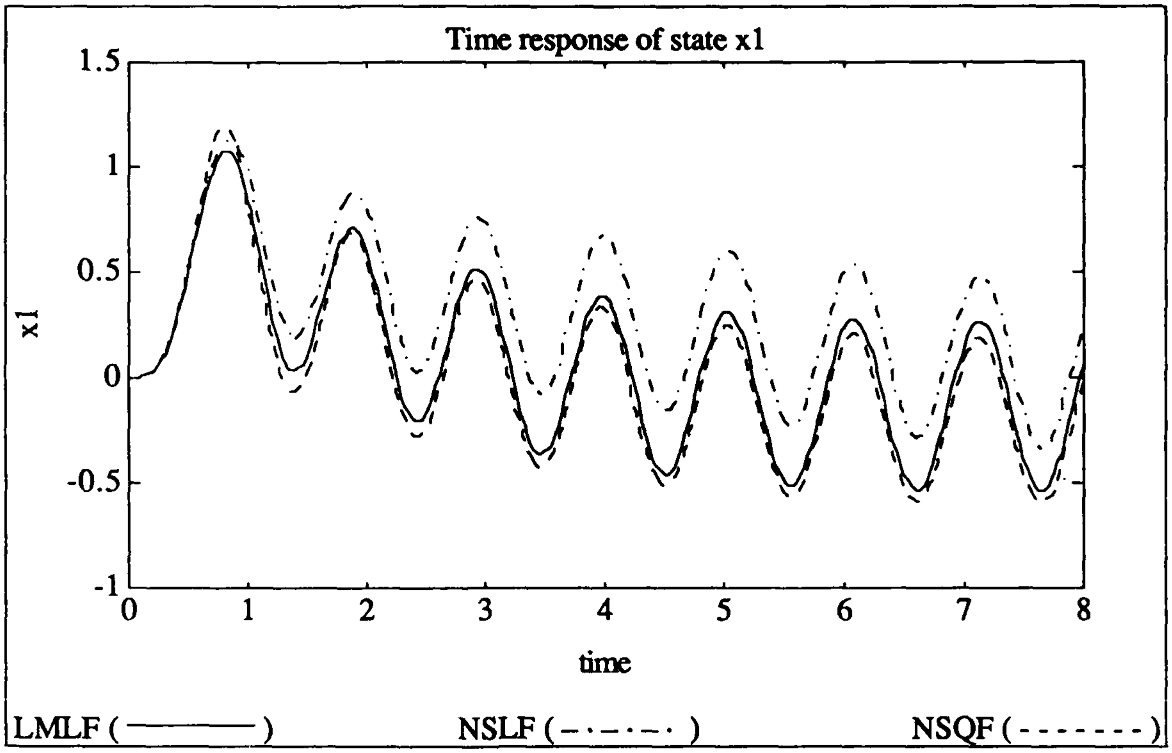


Fig. 5.85. Response of the state x_1 of Near Circular Satellite to a sinusoidal input.

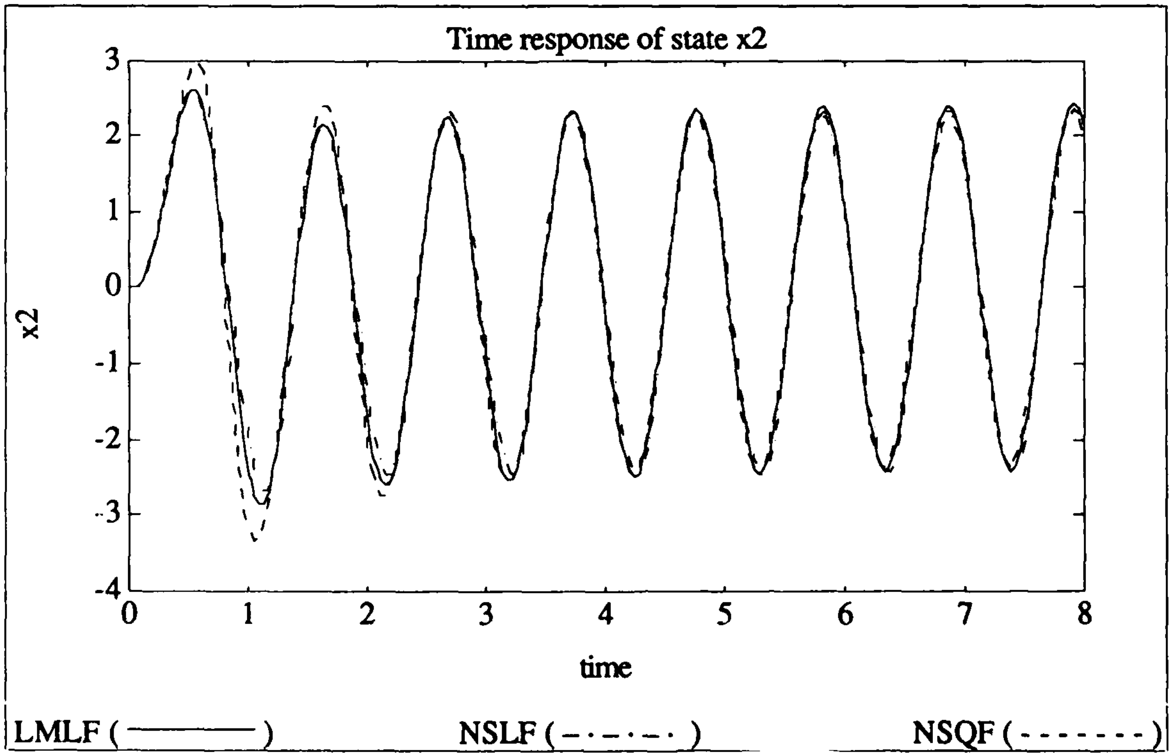


Fig. 5.86. Response of the state x_2 of Near Circular Satellite to a sinusoidal input.

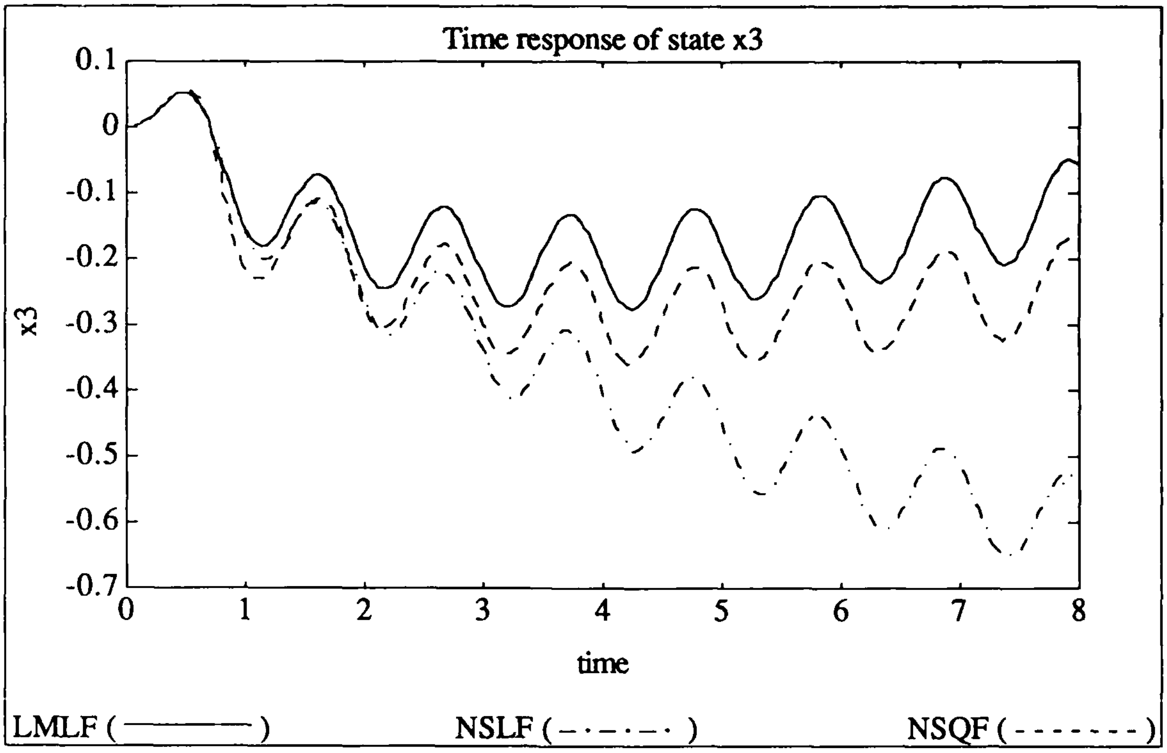


Fig. 5.87. Response of the state x_3 of Near Circular Satellite to a sinusoidal input.

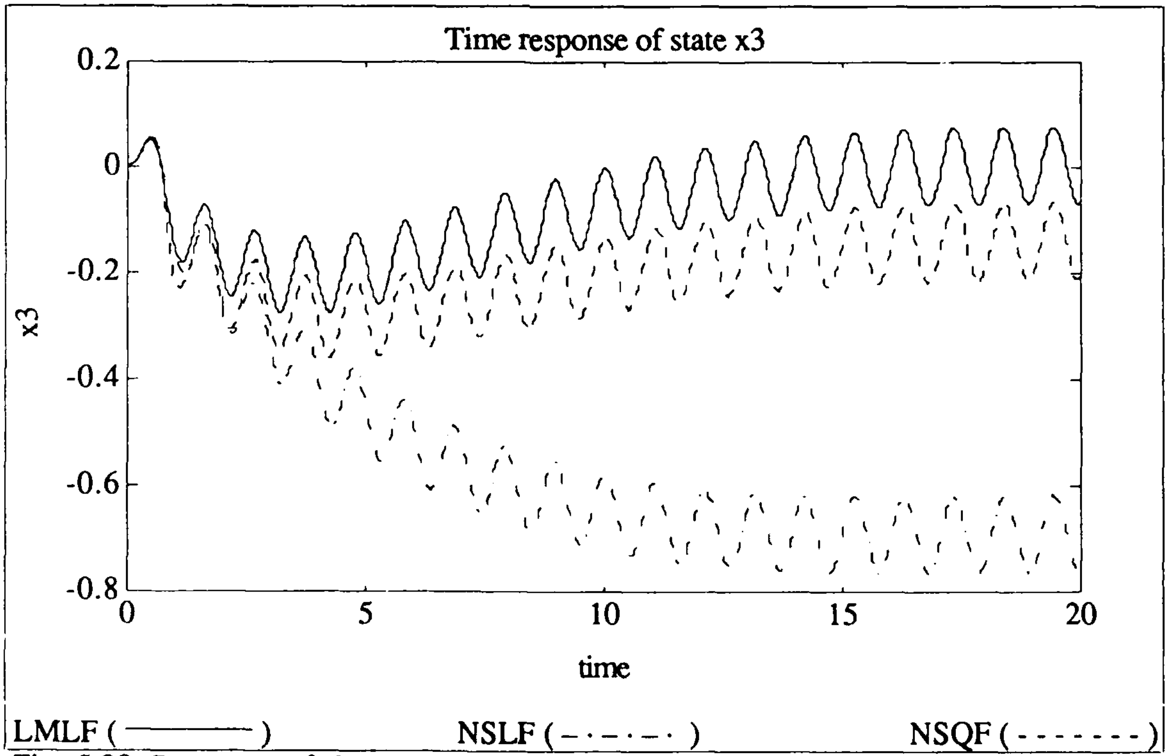


Fig. 5.88. Response of the state x_3 of Near Circular Satellite to a sinusoidal ihput for extended time.

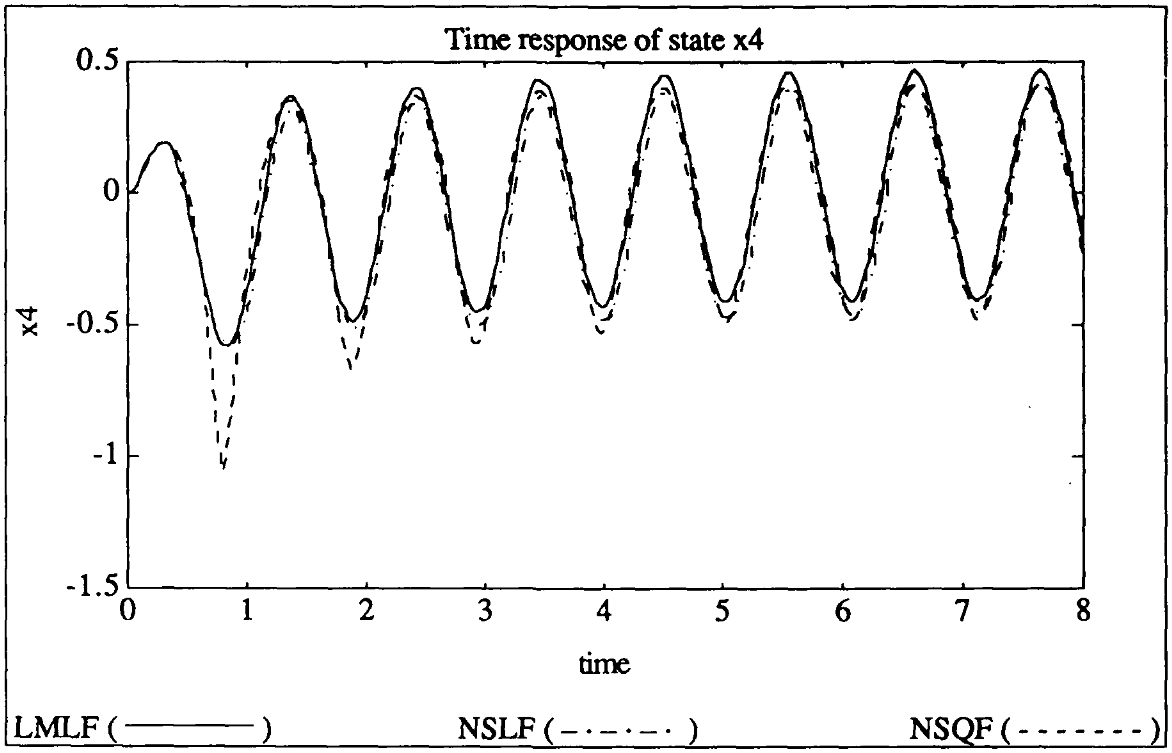


Fig. 5.89. Response of the state x_4 of Near Circular Satellite to a sinusoidal input.

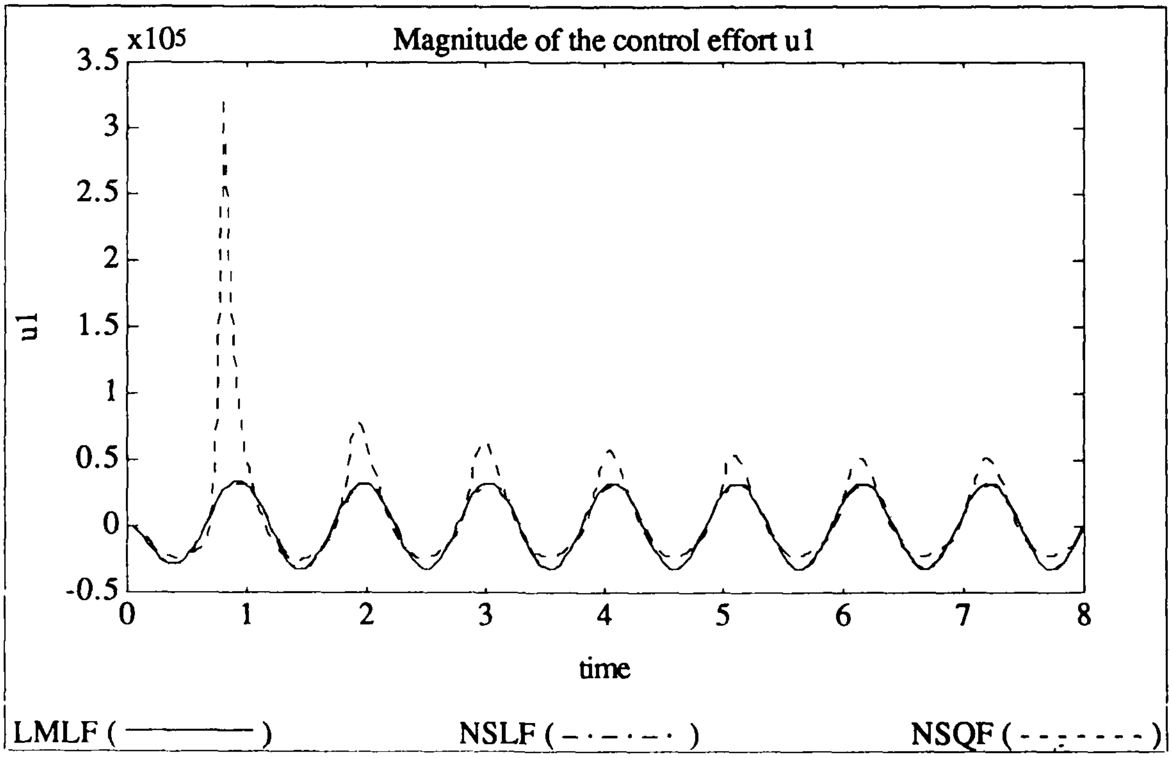


Fig. 5.90. Magnitude of the control efforts for Near Circular Satellite for a sinusoidal input.

6. SUMMARY, CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

6.1. Summary

In this report, we have developed a method for approximate linearization of nonlinear control systems using coordinate transformation and feedback pairs. First, a series expansion of a given nonlinear system at some nominal operating point was obtained. Based on the Hunt-Su theorem [17,18] and the approximate linearization results in [29] we formulated the problem of linearizing a nonlinear control system by nonlinear coordinate transformation and nonlinear feedback. The form of the coordinate change and feedback were chosen such that in the vicinity of the nominal operating point the transformation approaches the identity map. The differential equations of the system were then evaluated in the new set of coordinates using the given coordinate change and feedback.

With the goal of eliminating the nonlinear terms of a given degree in the series expansion, we obtained the Generalized Homological Equations. These equations were evaluated after choosing a suitable basis in which the ρth degree vector valued monomials were expressed. When the coefficients of the monomials of similar powers were set equal to each other, we obtained a linear system of equations in the unknown coefficients of the coordinate change and feedback, i.e. in the coefficients of $\underline{\phi}^{(2)}(x)$, $\alpha^{(2)}(x)$ and $\beta^{(1)}(x)$. Since these terms are coefficients of monomials of degree ρ, the resulting system of linear equations is approximately of order $n^{\rho + 1}$ where n is the dimension of the nonlinear vector field.

The calculation procedures to solve this set of equations were then implemented in the MATLAB computer program. MATLAB is an application language with a rich collection of matrix based computational algorithms. There are toolbox packages available for control systems, system identification, and signal processing. It is also a programming environment with a convenient user interface. By using the built-in subroutines available in the package and writing additional subroutines, functions and procedures one can create programs that automate the solutions of mathematical problems. This package proved to be a very suitable tool for writing the computer program that implements the results of this report.

The computer program prompts the user for the input parameters of the system to be linearized such as order of the system and number of inputs, characteristic scales for each state and input, linear part of the plant and the input coefficient, and nonlinear terms of second degree in the states and the inputs. After obtaining the parameters of the system, the program first calculates the linear and nonlinear parts of the system in the scaled coordinates. A linear set of equations in the unknown coefficients of the coordinate change and feedback is then obtained. This large system of equations is solved by a singular value decomposition subroutine. The solution obtained is presented to the user as the result of the first part of the program. At this point, having found the coordinate change and feedback needed, one can terminate the program, or continue to design a feedback law for the linear part of the system.

In the program, the nonlinear feedback needed to achieve the second degree linearization is calculated in the scaled natural coordinates. Prior to running simulations, symbolic expressions for the differential equations are generated to speed up the integrations. Three different models are simulated using the above feedback: 1) An exactly linear model (which agrees with the linear part of the system) to serve as a benchmark, 2) The nonlinear system with a linear feedback design based on a first degree approximation, 3) The nonlinear system that is linearized up to second degree with a quadratic feedback and with the additional linear feedback as in 1) and 2).

Various forms of initial conditions and reference input signals (such as impulses, steps and sinusoidal inputs) can be applied to all three systems. They are integrated for a specified length of time. At the end of the simulations, the resulting state trajectories are converted back to natural unscaled coordinates and plotted for comparison. The feedback inputs for all three simulations may also be plotted. Phase portraits of the states can also be made. Once the plots are completed, one can extend the simulations for a longer period of time, apply different inputs for a different set of simulations, design different forms of feedback for the linear part of the system or terminate the program.

Using the computer program as a tool, we found nonlinear feedback and coordinate transformation pairs for various example nonlinear systems. We tested the resulting closed loop systems with different initial conditions and impulse, step or sinusoidal forcing inputs. The improvement of the second degree approximation over the first degree design was superior in almost all the cases that were tested. The systems that are exactly linearizable showed the best improvement, and when the errors due to the second degree terms in the coordinate change became small, the behavior of the linearized systems were almost exactly linear.

For the type of systems that are not exactly linearizable, there was a noticable improvement, especiallly in the steady state responses to a disturbance. The most impressive results were obtained when the systems were forced by a sinusoidal forcing function. Unlike linear systems, the response of nonlinear systems to a sinusoidal forcing do not usually have zero mean. Depending on the nonlinear termsthat are present, there may be a nonzero bias value which is approximately the average of the nonlinearities over a period of the forcing function. In this case, it was observed from the simulations that the quadratic feedback decreased this bias term, and for exactly linearizable systems the bias was exactly zero in the steady state. The type of disturbances a control system encounters in the real world are not impulse or step inputs but noisy signals in general. The linearized systems have not been tested with noisy inputs. However, the response behavior for the sinusoidal inputs may be an indication that the improvements will be satisfactory.

6.2. Conclusions and Future Research Directions

One of the main contributions of this report is obtaining a numerical solution to the approximate linearization problem for nonlinear control systems. The exact solvability of the homological equations and the linearizability of a system according to the Hunt Su theorem (up to the specified degree) are equivalent conditions. Therefore whenever a control system is not exactly linearizable, it is impossible to find an exact solution to this system of linear equations. When a solution exists, the coordinate change and feedback that are found transform the nonlinear system into an exactly linear system (up to the specified degree) in a new set of coordinates.

When there is no exact solution, we have solved a least squares problem by finding an approximate solution to the system of equations using the singular value decomposition. Before finding an approximate solution, the state variables were normalized according to their characteristic scales. By defining the least squares problem in a consistent way, we minimized the norm of the error between an approximate solution and a "nearby" exact solution. As shown in Section 3, this is equivalent to solving the associated singular value decomposition problem for the linear system of equations in the scaled coordinates.

The Hunt-Su theorem is severe in its restriction that a transformation and feedback that linearizes a nonlinear system has to achieve the transformation *exactly*, i.e. up to an infinite degree in the series expansion of the nonlinear terms. When it exists, one can calculate the coordinate change-feedback pair using the Hunt-Su theorem. If the Hunt-Su test fails, there is nothing more one can do to linearize the system within that framework. One of the contributions of this report is to offer the above described approximate solution to the linearization problem in the case when the Hunt-Su theorem fails.

The improvement achieved by our method depends on whether or not before the linearization a given nonlinear system was far away from the “minimum” solution yielded by the singular value decomposition method. Obviously, another important factor in the performance of the linearized system is the relative magnitude of the nonlinear terms in the system compared to the linear part. If the nonlinearities are very small, the linear part of the system will dominate the behavior, and the improvements in such a case may not be significant.

The numerical calculations needed for the transformation and feedback for a second degree linearization are computationally expensive. Further improvements in the calculation speeds of the transformation and feedback may be possible (perhaps with the goal of a real-time implementation) by removing some of the redundancies in the linear system of equations that are obtained from the homological equations.

A next step in this research may be the formulation of the third degree linearization problem. It is expected that a third degree linearization will improve on the linear behavior of a nonlinear system in a larger neighborhood of the operating point. However, the numerical calculations necessary in this case are even more burdensome. The question that arises is whether a third degree linearization for a single operating point contributes a justifiable improvement for a broader range in the state space, or one should perform successive second degree linearizations at more than one nominal point. The answer to this question is not clear and may depend on the severity of the nonlinearities, or on the individual application.

Since we have not solved the third degree linearization problem in this report, a performance comparison with the second degree linearization is not possible at this time. Provided a given system is exactly linearizable, a third degree linearization may be a worthwhile pursuit. However, if the system is not exactly linearizable, since there are some second degree terms that can not be removed, the benefits of a third degree linearization in the presence of these second degree terms may not be crucial. This argument also depends on the relative magnitudes of the third degree terms.

It should be pointed out that the formulation of the third and even higher degree linearizations are similar to the second degree linearization, as shown in Chapter 3. Aside from some additional technical difficulty of dealing with a higher dimensional system of equations (the sizes of which are directly related to the presence of third –or higher– degree monomials) the numerical solution is also the same.

During the testing of the computer program with various examples, it was observed that the stability characteristics of the example systems were sometimes adversely affected. The nonlinear terms that are present in a system may occasionally augment system stability in certain regions in the state space, which may depend on their magnitudes and signs at any given point in the trajectory. Attempting to remove these nonlinear terms may sometimes cause a system to have a smaller basin of stability after linearization.

Our central focus in this report has been the improvement on the response of a nonlinear system in favor of a linear one, and the stability issue has been considered only locally, i.e. for the linear part of the system. Clearly, in a reasonably close neighborhood of the origin, the stability properties will be dominated by the linear part of the vector field. Since linear systems can be made globally asymptotically stable, it was initially expected that a system that has been approximately linearized would always have a larger basin of stability compared to the same system before the approximate linearization. This intuitive generalization is not correct in many cases.

The issue of nonlinear stability is a difficult problem that has not been completely solved. The stability of a system with nonlinear feedback of the type we are proposing may depend on important properties of the overall system such as nonlinear controllability, loss of convergence in the series expansion of the vector field, a loss of rank in the inversion of the coordinate change and feedback, etc. These issues are beyond the scope of this report, but the investigation of stability properties of the method is a natural extension of the work.

The results presented in this report and their implementation in a computer program that yields the solution to approximate linearization problems, help us to analyze various control systems and to gain further insight into the nonlinear behavior of control systems. Valuable experience was obtained from the various numerical experiments that have been performed. Future research directions will certainly be influenced by these observations as well.

REFERENCES:

[1] Abed, E.H., and J.H. Fu, "Local feedback stabilization and bifurcation control, I. Hopf Bifurcation", *Systems and Control Letters*, vol. 7, pp. 11–17, 1986.

[2] Aeyels, D., "Stabilization of a class of nonlinear systems by a smooth feedback control", *Systems and Control Letters*, vol. 5, pp. 289-294, 1985.

[3] Akhrif, O., and G.L. Blankenship, "Computer algebra for analysis and design of nonlinear control systems," *Proceedings of the 1987 American Control Conference*, pp. 547- 554, Minneapolis, Minn., Jun. 10-12, 1987.

[4] Arnold, V. I., *Geometric Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[5] Hedrick, J.K, and H.M. Paynter, eds., *Nonlinear System Analysis and Synthesis: Volume 1 - Fundamental Principles*, ASME, 1978.

[6] Ramnath, R.V., J.K. Hedrick, and H.M. Paynter, eds., *Nonlinear System Analysis and Synthesis: Volume 2 - Techniques and Applications*, ASME, 1980.

[7] Banks, S.P., *Mathematical Theories of Nonlinear Systems*, Prentice Hall International Series in Systems and Control Engineering, M.J. Grimble, ed., Prentice Hall International (UK), 1988.

[8] Bestle, D. and M Zeitz, "Canonical form observer design for non-linear time-variable systems," *Int. J. Control*, vol. 38, no. 2, pp. 419-431, 1983.

[9] Brunovsky, P., "A classification of linear controllable systems," Kybernetica cislo, 3, pp. 173-188, 1970.

[10] Claude, D., "Everything you always wanted to know about linearization," in *Algebraic and Geometric Methods in Nonlinear Control Theory*, M. Fliess and M. Hazewinkel, eds., D. Reidel Publishing Company, pp. 181-226, 1986.

[11] Chen, Y., *Nonlinear Feedback and Computer Control of Robot Arms,* Ph.D. Thesis, Sever Institute of Technology, Washington University, St. Louis, Missouri, 1984.

[12] Frezza, R., S. Karahan, A.J. Krener, M. Hubbard, "Application of an efficient nonlinear filter," in *Analysis and Control of Nonlinear Systems,* C.I. Byrnes, C.F. Martin, and R.e. Saeks (editors), pp. 223-237, North-Holland, Amsterdam, 1988.

[13] Freund, E., "Fast nonlinear control with arbitrary pole placement for industrial robots and manipulators," in *Robot Motion,* eds. Brady et al, MIT Press, pp. 147-167, 1982.

[14] Guckenheimer, J. and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields,* Springer-Verlag, New York, 1983.

[15] Hermann, R., "On the accessibility problem in control theory," in *Int. Symp. on Nonlinear Differential Equations and Nonlinear Mechanics,* New York, Academic Press, pp 325-332, 1963.

[16] Hermann, R., and A.J. Krener, "Nonlinear controllability and observability," *IEEE Trans. Automat. Contr.,* vol. AC-22, no. 5, pp.728-740, 1977.

[17] Hunt, L.R., and R. Su, "Linear equivalents of nonlinear time-varying system," *International Symposium on the Mathematical Theory of Netwoks and Systems,* Santa Monica, pp. 119-123, 1981.

[18] Hunt, L.R., R. Su, and G. Meyer, "Global transformations of nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. AC 28, pp. 24-31, 1983.

[19] Hunt, L.R., R. Su, and G. Meyer, "Design for multi-input nonlinear systems," in *Differential Geometric Control Theory,* R. W. Brockett, R. S. Millman and H. J. Sussmann, eds., Birkhauser, Boston, pp. 268-298, 1983.

[20] Iliĉ-Spong, M., R. Marino, S.M. Peresada, and D.G. Taylor, "Feedback linearizing control of switched reluctance motors," *IEEE Trans. Automat. Contr.,* vol.AC-32, 1987.

[21] Isidori, A., *Nonlinear Control Systems: An Introduction,* vol. 72, Lecture Notes in Control and Information Sciences, M. Thoma, ed., Springer-Verlag Berlin, 1985.

[22] Isidori, A., *Lectures on Nonlinear Control,* Notes prepared for a course at the Carl Cranz Gesselschaft (3-6 August), 1987.

[23] Isidori, A. and A.J. Krener, "On the feedback equivalence of nonlinear systems," *Systems and Control Letters,* vol. 2, pp.118-121, 1982.

[24] Isidori, A. and A. Ruberti, "On the synthesis of linear input-output responses for nonlinear systems," *Systems and Control Letters,* vol. 4, pp 17-22, 1984.

[25] Jakubczyk, B. and W. Respondek, "On linearization of control systems," *Bull. Acad. Polon. Sci., Ser. Sci. Math. Astronom. Phys.,* 28, pp.517-522, 1980.

[26] Kailath, T., *Linear Systems,* Prentice Hall Information and System Science Series, T. Kailath, ed., 1980.

[27] Krener, A.J., "On the equivalence of control systems and the linearization of nonlinear systems," *SIAM J. Control,* vol. 11, pp. 670-676, 1973.

[28] Krener, A.J., "A decomposition theory for differentiable systems," *SIAM J. Contr. Opt.,* v. 15, pp. 670-676, 1977.

[29] Krener, A.J., "Approximate linearization by state feedback and coordinate change," *Systems and Control Letters,* vol. 5, pp.181-185, 1984.

[30] Krener, A.J., "New approaches to the design of nonlinear compensators," *Proceedings of Berkeley Ames Conf. on Nonlinear Problems, Aerodynamics and Flight Control,* Math. Sci. Press, 1984.

[31] Krener, A.J., "The intrinsic geometry of dynamic observations," *Algebraic and Geometric Methods in Nonlinear Control Theory,* M. Fliess and M. Hazewinkel, eds., D. Reidel Publishing Company, pp 77-87, 1986.

[32] Krener, A.J., "Normal forms for linear and nonlinear systems", *Contemporary Mathematics,* vol. 68, pp. 157-189, 1987.

[33] Krener, A.J., and Isidori, A. "Linearization by output injection and nonlinear observers," *Systems and Control Letters,* vol. 3, pp. 47-52, 1983.

[34] Krener, A.J., S. Karahan, and M. Hubbard, "Approximate normal norms for nonlinear control systems," *Proceedings of 27th IEEE Conference on Decision and Control,* vol. 2, pp. 1223-1229, Austin, TX, December 7-9, 1988.

[35] Krener, A.J., S. Karahan, M. Hubbard,and R. Frezza, "Higher order linear approximations to nonlinear systems," *Proceedings of 26th IEEE Conference on Decision and Control*, vol. 1, pp. 519-523, Los Angeles, CA, December 9-11, 1987.

[36] Krener, A.J., and W. Respondek, "Nonlinear observers with linearizable error dynamics," *SIAM J. Contr., Opt.*, vol. 23, pp. 197-216, 1985.

[37] Lesiak, J., and A.J. Krener, "The existence and uniqueness of Volterra series," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 1090-1095, 1978.

[38] Meyer, G., R.L. Hunt, and R. Su, "Design of an autopilot by means of linearizing transformations," NASA Technical Memo 84295, 1984.

[39] Meyer, G., R. Su and R.L. Hunt, "Application of nonlinear transformations to automatic flight control," *Automatica*, vol. 20, pp. 103-107, 1984.

[40] Morse, A.S., "Structural invariants of linear multivariable systems," *SIAM J. Contr.*, vol. 11, pp.446-465, 1973.

[41] Nicosia, S., P. Tomei, and A. Tornambé, "A nonlinear observer for elastic robots," *IEEE Journal of Robotics and Automation*, vol. 4, no. 1, 1988.

[42] Poincaré, H., *Oeuvres*, Tome 1, Gauthier-Villars, Paris, 1928.

[43] Su, R., "On the linear equivalents of nonlinear systems," *Systems and Control Letters*, vol. 2, pp 48-52, 1982.

[44] Sussmann, H.J., "Lie Brackets, real analyticity and geometric control," in *Differential Geometric Control Theory*, R. W. Brockett, R. S. Millman and H. J. Sussmann, eds., Birkhauser, Boston, pp. 1-116, 1983.

[45] Tarn, T.J., A.K. Bejczy, A. Isidori, Y. Chen, "Nonlinear feedback in robot arm control," *Proc. 23rd Conference on Decision and Control*, pp.736-751, Las Vegas, NV, 1984.

[46] Wonham, W.M. *Linear Multivariable Control: A Geometric Approach (2nd edition)*, Vol. 10, Application of Mathematics Series, Springer-Verlag, New York, 1979.

APPENDIX A:

AN EXAMPLE SESSION FOR THE COMPUTER PROGRAM

The following example session has been recorded during a sample run of the Approximate Linearization program. The MATLAB command "diary on" starts recording the screen output. The command "diary off" stops the recording of the session. Note that the graphical output and anything typed by the user can not be recorded. The user responses have been added later and they are emphasized as boldface type.

```
mainmenu

      ••••••••••••••••••••••••••••••••••
      •   SECOND DEGREE APPROXIMATION   •
      •   TO NONLINEAR CONTROL SYSTEMS  •
      •                                 •
      •       Controller  Version       •
      •                                 •
      •               by                •
      •                                 •
      •    S i n a n   K a r a h a n    •
      ••••••••••••••••••••••••••••••••••
Welcome to the Second Degree Linearization Program
Press a key to start

----- M A I N   M E N U -----

      1) Help on program and functions
      2) Enter new nonlinear system parameters
      3) Solve the second degree linearization problem
      4) Feedback design for the models
      5) Simulate the system and compare performances
      6) Exit


 Select a menu number: 1
%
%This help routine presents a *very* brief introduction to
%the second degree approximation program.  For more detailed
%information about the program and the background theory
```

```
%please consult to the accompanying "Second Degree
%Approximation to Nonlinear Control Systems using MATLAB:
%Users' Manual".
echo off
Press any key
%
% The program solves for second order approximations of
% nonlinear control systems.  Your control system should be
% in the following format:
%
%  dx              (2)      (1)           3
%  -- = Fx + Gu + f   (x) + g   (x)u + O(x,u)
%  dt
%
% where F is the n*n plant matrix, G is the n*1 input vector,
%  (2)                                                  (1)
% f    is the second degree part of the vector field, g
% is the first degree part of the input coefficients. All of
% these terms have to be calculated ;, the user from the
% series expansionof the control system at the nominal
% operating point 0. For further details on the individual
% subroutines and function programs called by the routines,
% use the help utility in MATLAB.
echo off;
Press any key

----- M A I N   M E N U -----

      1) Help on program and functions
      2) Enter new nonlinear system parameters
      3) Solve the second degree linearization problem
      4) Feedback design for the models
      5) Simulate the system and compare performances
      6) Exit


 Select a menu number: 2
----- Input one of the following: -----

      1) Enter data manually
      2) Enter a filename to retrieve data


 Select a menu number:
----- For entry of data, choose -----

      1) Novice mode
      2) Expert mode


 Select a menu number:
 enter order of the system, n: 3

 enter dimension of the control, m: 1
```

```
Enter a characteristic scale for each state (used for
normalizing)

 scale of x1 = 1
 scale of x2 = 1
 scale of x3 = 1

Enter a characteristic scale for each input

 scale of u1 = 1
 Enter 3x3 plant matrix, F: [0 1 0;0 0 1;-3 2 -1]
     0     1     0
     0     0     1
    -3     2    -1

 Enter 3x1 control matrix, G: [0;0;1]
     0
     0
     1
Enter the coefficients of the second degree terms f2 in the
series expansion of the control system in the following
order:
x1x1  x1x2  x1x3  x2x2  x2x3  x3x3
Note the dimensions of f2:
3 rows [for each state equation from 1 to 3] by 6 columns
[for the
coefficients of above terms].

 f2 = [1 0 -1 0 0 0;0 -1 0 1 0 1;0 0 0 0 1 0]
     1     0    -1     0     0     0
     0    -1     0     1     0     1
     0     0     0     0     1     0

Enter coefficients of the first degree terms in g1 in the
form of a 3 x 3 matrix, where
 [1]
g  (x)*u = g11*x*u1

each g1i is a 3 by 3 matrix, and g1 is formed of row blocks
of g1i.

 g1 = [1 0 0;0 0 0;0 1 0]
     1     0     0
     0     0     0
     0     1     0


 Enter a file name to store this data: sessiondat
Note: The variables are being saved in the following order:
 n  m  xscale  uscale  f  g  f2  g1

----- Enter one of the following -----
```

```
  1) Display the variables n  m  xscale  uscale  f  g  f2  g1
  2) Return to main menu


 Select a menu number: 1
 Your system is as follows:
~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Dimension of the system:
 n
~~~
     3
Dimension of the control
 m
~~~
     1

 Scale factors of the states, x1 through xn :
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
     1     1     1
 Scale factor(s) of the input(s), u1 through um :
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
     1

Press any key

Linear part of the plant:
 F
~~~
     0     1     0
     0     0     1
    -3     2    -1

Constant part of the input vector:
 G
~~~
     0
     0
     1

Press any key

Second degree part of the plant:
  (2)
 f   (x)
~~~~~~~~~
f2(1) = +x(1)*x(1)-x(1)*x(3)
f2(2) = -x(1)*x(2)+x(2)*x(2)+x(3)*x(3)
f2(3) = +x(2)*x(3)

Press any key

First degree coefficient of the input(s):
(Each g1i  multiplies i'th input ui)
```

```
 g11
~~~~~~~
+x(1)
0
+x(2)

Corrections should be done by choosing #2 from the main menu.
Press any key

----- M A I N   M E N U -----

      1) Help on program and functions
      2) Enter new nonlinear system parameters
      3) Solve the second degree linearization problem
      4) Feedback design for the models
      5) Simulate the system and compare performances
      6) Exit


 Select a menu number: 3
• Calculating the scaled variables...

•
• We seek a quadratic change of coordinates
•          (2)
• z = x - phi   (x)
• and feedback
•         (2)              (1)
• v = alpha   (x) + (I + beta   (x)).u
• which transforms our system into the form
•
•   dz                    2         3
•  ---- = Fz + Gv + R(z,v)   + O(z,v)
•   dt
•                                                   2
• where, if the system is exactly linearizable, R(z,v)
• (residual) is zero.
•


•
• We use the homological equations
•
•             (2)             (2)      (2)
•  [ F.x, phi   (x) ] + alpha   (x) = f   (x)
•
•             (2)               (1)      (1)
•  [ G.u, phi   (x) ] + G.beta   (x) = g   (x).u
•

• Calculating the system of equations...

•                                                      2
• We construct a large linear system LX = B in the O(x,u)
```

```
• coefficients of the homological equations.  This system has
•
•                2            2
•  ROWS:      n  (n+1)/2 + mn
•
•                2                       2
•  COLUMNS:   n  (n+1)/2 + mn(n+1)/2 + m n
•
• In general, the column rank is deficient and the solution
• is overdetermined.
• We use the SVD algorithm to get the "nearest" possible
• solution.

• Solving for the coordinate change and feedback...
• SVD may take a while, please wait.
• Calculations done.
Press any key

----- Please choose a menu item -----

      1) Display coordinate change and feedback
      2) Show the closest linearizable system
      3) Exit to main menu


 Select a menu number: 1
press any key
Phi2 in the second degree coordinate change z=x-phi2(x):
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
phi2(1) = +0.19697*x(1)*x(1)+x(1)*x(3)-0.5*x(2)*x(2)
phi2(2) = -4*x(1)*x(1)+2.3939*x(1)*x(2)
phi2(3) = -7*x(1)*x(2)+2.3939*x(1)*x(3)+1.3939*x(2)*x(2)-
x(3)*x(3)
Press any key
Second degree part of feedback: Alpha2
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
alpha2(1) = -1.4091*x(1)*x(1)+7*x(1)*x(2)-
2*x(1)*x(3)+7.1061*x(2)*x(2)-0.18182*x(2)*x(3)-x(3)*x(3)
Press any key
First degree part in the feedback: Beta1
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
beta1(1,1) = -2.3939*x(1)+x(2)+2*x(3)

press any key

----- Please choose a menu item -----

      1) Display coordinate change and feedback
      2) Show the closest linearizable system
      3) Exit to main menu


 Select a menu number: 2
The system is exactly linearizable up to degree 2.
```

```
----- Please choose a menu item -----

      1) Display coordinate change and feedback
      2) Show the closest linearizable system
      3) Exit to main menu


 Select a menu number: 3
----- M A I N   M E N U -----

      1) Help on program and functions
      2) Enter new nonlinear system parameters
      3) Solve the second degree linearization problem
      4) Feedback design for the models
      5) Simulate the system and compare performances
      6) Exit


 Select a menu number: 4
Closed loop feedback design for the linear part of the plant:

----- Input one of the following: -----

      1) Specify closed loop eigenvalues
      2) Design Linear Quadratic Optimal feedback


 Select a menu number: 2
Quadratic regulator will minimize: integral(x'Qx + u'Ru)dt
Enter 3 by 3 <positive semi-definite> matrix Q

 Q = [1 0 0;0 1 0;0 0 1]
Enter 1 by 1 <positive definite> matrix r

 R = 1
The gains found are:
kfeed =
    0.1623    6.9158    2.9789
The resulting closed loop eigenvalues are:
cleig =
  -0.7245 + 0.8515i
  -0.7245 - 0.8515i
  -2.5299
Press a key

----- M A I N   M E N U -----

      1) Help on program and functions
      2) Enter new nonlinear system parameters
      3) Solve the second degree linearization problem
      4) Feedback design for the models
      5) Simulate the system and compare performances
      6) Exit
 Select a menu number: 5

•The following three systems will be simulated in scaled
coordinates:
•
[1]    dx/dt = (F+G.K).x + r
•
•                             (2)       (1)              (1)
[2]    dx/dt = (F+G.K)x + f   (x) + g   (x)Kx + (G + g   (x))r
•
•                        (2)           (1)
[3]    dx/dt = Fx + f   (x) + (G + g   (x))u(x)
•
• where r is a reference input signal.
• The nonlinear feedback u(x) in [3] is :
•                          -1
• u(x) = (I + beta1(x))   {K(x - phi2(x)) - alpha2(x) + r}.
•
• [2] is the nonlinear system with a linear feedback design.
• [3] is the same system with linear + quadratic feedback.
• [1] is a linear reference model. It is expected that as
• x approaches zero, [3] will agree with [1] closer than [2].
Calculating symbolic expressions for [1], [2] and [3]...

Press any key

----- Simulation Menu -----

      1) Free response-nonzero initial conditions
      2) Impulse input-zero or nonzero initial conditions
      3) Step input-zero or nonzero initial conditions
      4) Sinusoid input-zero or nonzero initial conditions
      5) Exit to main menu

 Select a menu number: 1
Enter initial conditions (column vector): x0 =[0.1;0.1;0.1]
 Simulation will start from initial time t0 = 0.

 Enter final time (*scaled* time): tf = 5
• Simulation of the linear model done
• Simulation of the nonlinear system with linear control done
• Simulation of the nonlinear system with quadratic control
done

----- Plot Menu -----

      1) Plot time response
      2) Plot phase portrait
      3) Extend simulation for some more time
      4) Return to simulation menu
 Select a menu number: 1
Legend for plots (colors seen in color monitors only):
#1: Linear reference model (___ Solid; red)
#2: Nonlinear system with linear feedback (._. Dashdot;
green)
#3: Nonlinear system with quadratic feedback (---Dashed;
blue)

Enter simulations to see: "1" for #1, "12" for #1&#2, etc
Please enter the number(s) in increasing order

 Enter curve number(s) : 123
Press a key to see each plot
```

(The plots are viewed on the graphics screen of MATLAB)

```
 Do you want to see the control input magnitudes?(y/n): y
• Calculating the magnitudes of the control inputs...
```

(The plot is viewed on the graphics screen of MATLAB)

```
----- Plot Menu -----

      1) Plot time response
      2) Plot phase portrait
      3) Extend simulation for some more time
      4) Return to simulation menu

 Select a menu number: 4
----- Simulation Menu -----

      1) Free response-nonzero initial conditions
      2) Impulse input-zero or nonzero initial conditions
      3) Step input-zero or nonzero initial conditions
      4) Sinusoid input-zero or nonzero initial conditions
      5) Exit to main menu


 Select a menu number:  5
----- M A I N   M E N U -----

      1) Help on program and functions
      2) Enter new nonlinear system parameters
      3) Solve the second degree linearization problem
      4) Feedback design for the models
      5) Simulate the system and compare performances
      6) Exit

 Select a menu number: 6

 ŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸ
 ŸŸ Have a good day ! ŸŸ
 ŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸŸ
```

APPENDIX B:

ADDITIONAL PUBLICATIONS

The following publications are supported by the grant and they are included in this appendix:

PUBLISHED:

1. 1986 Krener, A. J., "Normal forms for linear and nonlinear systems. In Differential Geometry," *The Interface between Pure and Applied Mathematics*, M. Luksik, C. Martin and W. Shadwick, eds. Contempary Mathematics V.68, American Mathematical Society, Providence, 157-189.
2. 1987 Krener, A. J., S. Karahan, M. Hubbard and R. Frezza, "Higher order linear approximations to nonlinear control systems," *Proceedings, IEEE Conf. on Decision and Control*, Los Angeles, 519-523.
3. 1987 Karahan, S., "Determining Torque and Velocity Limits on Joint Actuators for Robot Arms with Coupled Joint Motion," *Proceedings of the Ninth IASTED International Symposium on Robotics and Automation*, Santa Barbara, CA, 96-99.
4. 1988 Frezza, R., S. Karahan, A. J. Krener and M. Hubbard, "Application of an efficient nonlinear filter," in *Analysis and Control of Nonlinear Systems*, C.I. Byrnes, C.F. Martin and R.E. Saeks, eds. North Holland, Amsterdam, 223-238.
5. 1988 Phelps, A.R. and A.J. Krener, "Computation of observer normal form using MACSYMA," in *Analysis and Control of Nonlinear Systems*, C.I. Byrnes, C.F. Martin and R.E. Saeks, eds. North Holland, Amsterdam, 475-482.
6. 1988 Krener, A.J., "Reciprocal processes, second order stochastic differential equations and PDE's of conservation and balance," in *Analysis and Control of Nonlinear Systems*, C.I. Byrnes, C.F. Martin and R.E. Saeks, eds. North Holland, Amsterdam, 579-590.
7. 1988 Krener, A.J. and H. Schaettler, "The structure of small-time reachable sets in low dimensions," *SIAM Journal on Control and Optimization*, 27:120-147.
8. 1988 Krener, A.J., "Reciprocal diffusions and stochastic differential equations of second order," *Stochastics* 24:393-422.

9. 1988 Krener, A.J., S. Karahan and M. Hubbard, "Approximate normal forms of nonlinear systems," *Proceedings, IEEE Conf. on Decision and Control*, San Antonio,1223-1229.

IN PRESS:

1. Krener, A. J., "Nonlinear controller design via approximate normal forms," *Proceedings of the IMA Summer Institute on Signal Processing*, University of Minnesota, 1988.
2. Clark, J.M.C., "A local characterization of reciprocal diffusions," *Stochastics*, to appear.

SUBMITTED:

1. Krener, A.J. and Y. Zhu, "The fractional representation of a class of nonlinear systems."
2. Kang, W. and A.J. Krener, "Observation of a rigid body from measurements of a principle axis."

Contemporary Mathematics
Volume 68, 1987

Normal Forms for Linear and Nonlinear Systems*

Arthur J. Krener[†]

1 Introduction

It is well-known that a state space description of a controllable linear system can be transformed to controllable or controller form by a linear change of state variables. A state space description of an observable linear system can be transformed to observable or observer form by a linear change of state variables. Moreover the former are closely related to right matrix fractional descriptions (RMFD) of the transfer function and the latter are closely related to left matrix fractional descriptions (LMFD). These facts can be found in many texts such as Wolovich [1] or Kailath [2]. (The reader should be warned that the controllable/controller and observable/observer terminologies are not standard, we follow that of [2]). Unfortunately there is no one treatment of this material which is suitable for our purposes so we devote Sections 2 and 3 to a review. This is by way of preparation for our discussion of the existence and uniqueness of normal forms for nonlinear systems in Sections 4 and 5. Our treatment generalizes Zeitz [22] who discussed similar forms for scalar input and scalar output nonlinear systems.

2 Linear Normal Forms

Throughout this paper we shall use the following notation. The *indices* $\ell_1, \dots, \ell_q$ are positive integers summing to n. A *prime triple* (A,B,C) with indices $\ell_1, \dots, \ell_q$ is a triple of block diagonal matrices of dimension $n \times n$, $n \times m$ and $p \times n$ of the

*This paper is in final form and no version of it is submitted elsewhere.

[†]Research supported in part by the Air Force Office of Scientific Research under 85NM215.

form

$$A = \text{BlockDiag.} \begin{bmatrix} 0 & 1 & \cdots & 0 \\ & & \ddots & 1 \\ 0 & \cdots & & 0 \end{bmatrix}^{\ell_i \times \ell_i} \tag{1}$$

$$B = \text{BlockDiag.} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}^{\ell_i \times 1} \tag{2}$$

$$C = \text{BlockDiag.} \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^{1 \times \ell_i} \tag{3}$$

The "prime" terminology was introduced by Morse [3].

Consider the linear state space description

$$\begin{aligned} \dot{\xi} &= F\xi + Gu && (4) \\ y &= H\xi && (5) \end{aligned}$$

where $F^{n\times n}$, $G^{n\times m}$ and $H^{p\times n}$. The system is said to be *controllable if*

$$\text{rank}\ \{F^{r-1}G^j :\ j = 1,\ldots,m;\ r = 1,\ldots,n\} = n \tag{6}$$

(Note: F^r denotes the r^{th} power of F, G^j denotes the j^{th} column of G and H_i is the i^{th} row of H.)

Every controllable linear system has *controllability indices* $\ell_1,\ldots,\ell_m \geq 0$ characterized by $\ell_1 + \ldots + \ell_m = n$ and

$$\begin{aligned} &\text{rank}\{F^{r-1}G^j :\ j = 1,\ldots,m;\ r = 1,\ldots,\ell\} = \\ &\text{rank}\{F^{r-1}G^j :\ j = 1,\ldots,m;\ r = 1,\ldots,\ell \wedge \ell_j\} \end{aligned} \tag{7}$$

for $\ell = 1,\ldots,n$. The minimum of ℓ and ℓ_j is denoted by $\ell_i \wedge \ell_j$. The set of controllability indices is uniquely determined by F and H and does not change under linear state feedback. There can be some freedom, m in the ordering of the controllability indices even when the ordering of the inputs remains fixed. This is because there may be several orderings which satisfy (7). Of course a change of variables in the input space or a reordering of the inputs can change the order of the controllability indices. We could reorder the inputs so that $\ell_1 \leq \ldots \ell_m$ or the reverse but we shall not do so. To simplify notation we shall restrict our attention to systems where the controllability indices are positive, $\ell_1,\ldots,\ell_m \geq 1$. A general system can be made to satisfy this condition by deleting dependent columns of G.

An alternative characterization of property (7) of the controllability indices is that

$$F^{\ell_j}G^j = 0 \tag{8}$$

$$\text{mod } \{F^{r-1}G^i : i = 1, \ldots, m;\ r = 1, \ldots, (\ell_j + 1) \wedge \ell_i\}.$$

The controllabilities indices of (4), (5) are said to be *strict* if (8) holds mod $\{F^{r-1}G^i : i = 1, \ldots, m;\ r = 1, \ldots, \ell_j \wedge \ell_i\}$. The controllability indices are strict iff there is only one ordering of the controllability indices satisfying (7).

It is always possible to make a linear change of input coordinates $\tilde{u} = \beta u$ that makes the controllability indices strict for the new pair $(\tilde{F}, \tilde{G}) = (F, G\beta^{-1})$ without changing their order. One way of accomplishing this to define $1 \times n$ vectors $K_1, \ldots, K_m$ by

$$K_i F^{r-1} G^j = \begin{cases} 0 & 1 < r < \ell_j \\ \delta_i^j & r = \ell_j \end{cases} \tag{9}$$

and let β be the $m \times m$ non-singular matrix whose $i - j$ entry is

$$\beta_i^j = K_i F^{\ell_i - 1} G^j. \tag{10}$$

In this case β satisfies

$$\beta_i^j = \delta_i^j \quad \ell_i \leq \ell_j. \tag{11}$$

Moreover β is the only such matrix which makes the controllability indices strict and leave the order invariant. A change of input coordinates $\bar{u} = \lambda \tilde{u}$ preserves the strictness of the controllability indices while leaving the order invariant iff $\lambda_i^j = 0$ for $\ell_i > \ell_j$.

The system (4), (5) is said to be observable if

$$\text{rank } \{H_i F^{r-1} : i = 1, \ldots, p;\ r = 1, \ldots, n\} = n. \tag{12}$$

Every observable linear system has *observability indices* $\ell_1, \ldots, \ell_p \geq 0$ characterized by $\ell_1 + \cdots + \ell_p = n$ and

$$\text{rank } \{H_i F^{r-1} : i = 1, \ldots, p;\ r = 1, \ldots, \ell\} =$$

$$\text{rank } \{H_i F^{r-1} : i = 1, \ldots, p;\ r = 1, \ldots, \ell \wedge \ell_i\} \tag{13}$$

for $\ell = 1, \ldots, n$. The set of indices is uniquely determined by H and F and does the change under linear change of coordinates in the state and output spaces and linear output injection. There can be freedom in the ordering of the observability indices even when the order of the outputs remains fixed. This is because there may be several orderings which satisfy (13). Of course a change of output variables or a reordering of the outputs can change the order of the observability indices. We could reorder the outputs so that $\ell_1 \leq \ldots \leq \ell_p$ or the reverse but we shall not do so. We shall restrict our discussion to systems where all the observability indices are positive.

Similarly an alternative characterization of property (13) of the observability indices is that

$$H_i F^{\ell_i} = 0 \tag{14}$$

$$\text{mod } \{H_j F^{r-1} : \ j = 1, \ldots, p; \ r = 1, \ldots, (\ell_i + 1) \wedge \ell_j\}.$$

The observability indices of (4), (5) are said to be *strict* if (14) holds mod $\{H - jF^{r-1} : \ j = 1, \ldots, p; \ r = 1, \ldots, \ell_i \wedge \ell_j\}$. The observability indices are strict iff there is only one ordering of them satisfying (13). It is always possible to make a linear change of output coordinates $y = \gamma \tilde{y}$ that makes the observability indices strict for the new pair $(\tilde{H}, \tilde{F}) = (\gamma^{-1} H, F)$ while not changing their order. One way of accomplishing this is to define $n \times 1$ vectors $Q^1, \ldots, Q^p$ by

$$H_i F^{r-1} Q^j = \begin{cases} 0 & 1 \leq r \leq \ell_i \\ \delta_i^j & \quad r = \ell_i \end{cases} \tag{15}$$

and let γ be the $p \times p$ non-singular matrix whose $i - j^{th}$ entry is

$$\gamma_i^j = H_i F^{\ell_j - 1} Q^j. \tag{16}$$

In this case γ satisfies

$$\gamma_i^j = \delta_i^j \quad \ell_i \geq \ell_j. \tag{17}$$

Moreover γ is the only such matrix which makes the observability indices strict and leaves the order invariant because a change of output coordinates $\tilde{y} = \mu \bar{y}$ preserves the strictness and order of the observability indices iff $\mu_i^j = 0$ for $\ell_i < \ell_j$.

The *controllable form* of a linear system is

$$\dot{x} = Ax - \alpha Cx + Bu \tag{18}$$

$$y = \gamma x \tag{19}$$

where (A, B, C) is a prime triple with indices $\ell_1, \ldots, \ell_m$ and α and γ are arbitrary matrices of dimensions $n \times m$ and $p \times n$.

The following facts are well-known and/or can be easily proved. A system in controllable form is controllable with controllability indices $\ell_1, \ldots, \ell_m$. A system (4), (5) can be transformed into controllable form (18), (19) by a linear change of state coordinates $\xi = Tx$ iff it is controllable. If (4), (5) is controllable with controllability indices $\ell_1, \ldots, \ell_m$ then the x coordinates of (18), (19) are defined by taking as a basis the columns of the matrix T

$$T = [F^{\ell_1 - 1} G^1, \ldots, G^1, \ldots, F^{\ell_m - 1} G^m, \ldots, G^m] \tag{20}$$

Let $x = T^{-1}\xi$ have components

$$x^* = (x_{11}, \ldots, s_{1\ell_1}, \ldots, x_{m1}, \ldots, x_{m\ell_m}), \tag{21}$$

where $*$ denotes transpose, then $F^{\ell_j - r}G^j$ in ξ coordinates becomes the unit vector in the x_{jr} direction in x coordinates. The j^{th} column of α and the matrix γ are given by

$$\begin{aligned} \alpha^j &= -T^{-1}F^{\ell_j}G^j && (22)\\ \gamma &= HT. && (23) \end{aligned}$$

It can be shown that $\alpha^j_{ir} = 0$ if $\ell_i - r > \ell_j$. The controllability indices are strict iff $\alpha^j_{ir} = 0$ for $\ell_i - r \geq \ell_j$. The controllable form (18), (19) of the linear system (4), (5) and the associated x coordinates (20), (21), (22), (23) are unique up to reordering of the controllability indices. The *observable form* of a linear system is

$$\begin{aligned} \dot{x} &= Ax - B\alpha x + \beta u && (24)\\ y &= Cx && (25) \end{aligned}$$

where (A, B, C) is a prime triple with indices $\ell_1, \ldots, \ell_p$ and α and β are arbitrary matrices of dimensions $p \times n$ and $n \times m$.

A system in observable form is observable with observability indices $\ell_1, \ldots, \ell_p$. A system (4), (5) can be transformed into observable form (24), (25) by a linear change of state coordinates $\xi = Tx$ iff it is observable. If (4), (5) is observable with observability indices $\ell_1, \ldots, \ell_p$ then the x coordinates of (18), (19) are of the form

$$x^* = (x_{11}, \ldots, x_{1\ell_1}, \ldots, x_{p1}, \ldots, x_{p\ell_p}) \tag{26}$$

where T^{-1} is defined by

$$x_{ir} = H_i F^{r-1}\xi. \tag{27}$$

The i^{th} row of α and the matrix β are given by

$$\begin{aligned} \alpha_i &= -H_i F^{\ell_i} T && (28)\\ \beta &= T^{-1}G. && (29) \end{aligned}$$

It can be shown that $\alpha_i^{jr} = 0$ if $r > \ell_i + 1$. The observability indices are strict iff $\alpha_i^{jr} = 0$ for $r > \ell_i$. The observable form (24), (25) of the linear system (4), (5) and the associated x coordinates (26), (27), (28), (29) are unique up to a reordering of observability indices.

The *controller form* of linear system is

$$\dot{x} = Ax - B\alpha x + B\beta u \tag{30}$$

$$y = \gamma x \tag{31}$$

where (A, B, C) is a prime triple with indices $\ell_1, \ldots, \ell_m$ and α, β, γ are matrices of dimensions $m \times n, m \times m, p \times n$. These matrices are arbitrary except β must be non-singular.

A system in controller form is controllable with controllability indices $\ell_1, \ldots, \ell_m$ and the controllability indices are strict relative to the input $\tilde{u} = \beta u$. A system (4),(5) can be transformed into controller form (30), (31) by a linear change of state coordinates $\xi = Tx$ iff it is controllable. If (4), (5) is controllable with controllability indices $\ell_1, \ldots, \ell_m$, then let β be defined by (10). One can define a pseudo-output for (4), (5).

$$\psi = K\xi \tag{32}$$

where K is the $m \times n$ matrix defined by (9). The square system (4) and (32) is observable with strict observability indices $\ell_1, \ldots, \ell_m$. The observable form realization of (4) and (32) is a controller form realization of (4), (5). The x coordinates of (30), (31) are of the form (20) and

$$x_{jr} = K_j F^{r-1} \xi. \tag{33}$$

The matrix γ is given by (23) and the i^{th} row of α is given by

$$\alpha_i = -K_i F^{\ell_i} T. \tag{34}$$

Since the observability indices of (4) and (32) are strict, we have

$$\alpha_i^{jr} = 0 \quad r > \ell_i. \tag{35}$$

In general controller form realizations are not unique. However the controller form realization which satisfies (11) and (35) is unique up to reordering of the controllability indices.

The *observer form* of a linear system is

$$\dot{x} = Ax - \alpha C x + \beta u \tag{36}$$

$$y = \gamma C x \tag{37}$$

where (A, B, C) is a prime triple with indices $\ell_1, \ldots, \ell_p$ and α, β, γ are indices of dimensions $n \times p, n \times m, p \times p$. These matrices are arbitrary except that γ must be non-singular.

A system in observer form is observable with observability indices $\ell_1, \dots, \ell_p$ and the observability indices are strict relative to the output $\tilde{y} = \gamma^{-1}y$. A system (4), (5) can be transformed into observer form (36), (37) by a linear change of state coordinates $\xi = Tx$ iff it is observable. If (4), (5) is observable with observability indices $\ell_1, \dots, \ell_p$, let γ be given by (16). One can define a pseudo-input μ

$$\dot{\xi} = F\xi + Q\mu \tag{38}$$

where Q is the $n \times p$ matrix defined by (15). The square system (38) and (5) is controllable with strict controllability indices $\ell_1, \dots, \ell_p$. The controllable form realization of (38) and (5) is an observer form realization of (4), (5). The x coordinates of (36), (37) are of the form (26) and defined by $\xi = Tx$ where

$$T = [F^{\ell_1 - 1}Q^1, \dots, Q^1, \dots, F^{\ell_p - 1}A^p, \dots, Q^p]. \tag{39}$$

The matrix β is given by (29) and the j^{th} column of α is given by

$$\alpha^j = -T^{-1}F^{\ell_j}Q^j.] \tag{40}$$

Since the controllability indices of (38) and (5) are strict, we have

$$\alpha^j_{ir} = 0 \quad 1 \leq r \leq \ell_i - \ell_j. \tag{41}$$

In general observer form realizations are not unique. However, the observer form realization which satisfies (17) and (41) is unique up to a reordering of the observability indices.

Remark 2.1 *The controller form (30), (31) of a system is very useful in designing a linear state variable feedback to stabilize the system. The observer form (36), (37) is very useful in designing asymptotic observers. Together they can be used to stabilize a system by dynamic output feedback (also called observer based compensation). See [1] or [2] for details.*

Remark 2.2 *Controllable and observable forms are easier to compute and are useful for finding the observer and controller forms of related systems.*

3 MFD's

The purpose of this section is to emphasize the very close relationship between the normal forms of a linear system described above and the so-called polynomial matrix fractional descriptions of its transfer function. For linear systems it is only a matter of personal preference which representation we choose to work with. This is not the case for the nonlinear systems because they don't have nice frequency domain descriptions. Our treatment is similar to that of [1] and [2].

Throughout we shall use the following notation. Given the indices $\ell_1, \dots, \ell_q$ where $n = \ell_1 + \cdots + \ell_q$ then $\Delta(s)$, $\Phi(s)$ and $\Psi(s)$ are block diagonal polynomial matrices of dimensions $q \times q$, $q \times n$ and $n \times q$ of the form

$$\Delta(s) = \text{BlockDiag} \left[\; s^{\ell_i} \; \right]^{1\times 1} \tag{42}$$

$$\Phi(s) = \text{BlockDiag} \left[\; s^{\ell_i - 1} \quad \cdots \quad 1 \; \right]^{1\times \ell_i} \tag{43}$$

$$\Psi(s) = \text{BlockDiag} \begin{bmatrix} 1 \\ \vdots \\ s^{\ell_i - 1} \end{bmatrix}^{\ell_1 + 1} \tag{44}$$

The linear state space description

$$\dot{x} = Ax + v \tag{45}$$

$$z = x \tag{46}$$

with input v, state x, output z, all of dimension n, has the following polynomial matrix description in the transform domain

$$\Delta(s)\xi(s) = \Phi(s)v(s) \tag{47}$$

$$z(s) = \Psi(s)\xi(s) \tag{48}$$

where the so called "partial state" $\xi(s)$ is defined by

$$\xi(s) = Cx(s) \tag{49}$$

or equivalently

$$x(s) = \Psi(s)\xi(s) \tag{50}$$

(Here $x(s)$ denotes the Laplace transform of $x(t)$, etc.). The (A, B, C) of the above are a prime triple with indices $\ell_1, \dots, \ell_q$ so that

$$C\Psi(s) = \Phi(s)B = I^{q\times q}. \tag{51}$$

From this we quickly obtain MFD's corresponding to the 4 normal forms of the last section. For a system in controllable form (18), (19) we use the relations

$$v(s) = -\alpha Cx(s) + Bu(s) \tag{52}$$

$$y(s) = \gamma x(s). \tag{53}$$

Let $q = m$, then from (47), (48) and (49), (50), (51) we obtain

$$\begin{aligned} u(s) &= (\Delta(s) + \Phi(s)\alpha)\xi(s) && (54)\\ y(s) &= \gamma\Psi(s)\xi(s) && (55) \end{aligned}$$

which is a RMFD of the form

$$y(s) = N(s)D^{-1}(s)u(s) \qquad (56)$$

where

$$\begin{aligned} D(s) &= (\Delta(s) + \Phi(s)\alpha) && (57)\\ N(s) &= \gamma\Psi(s). && (58) \end{aligned}$$

Given a RMFD (56) we can always obtain a controllable form realization. Recall that a polynomial matrix is *unimodular* if it has an inverse which is a polynomial matrix. If we multiply $N(s)$ and $D(s)$ on the right by a unimodular matrix we don't change the transfer function. In this way we can insure that the matrix of highest column coefficients of D is invertible and even more equals the identity.

Let $\ell_1, \ldots, \ell_m$ be the column degrees of D, then $D(s)$ and $N(s)$ can be written as (57), (58) thus defining α and γ. This yields a controllable form realization of (56).

For a system in controller form (30), (31) we use the relations

$$\begin{aligned} v(s) &= -B\alpha x(s) + B\beta u(s) && (59)\\ y(s) &= \gamma z(s) && (60) \end{aligned}$$

and so we obtain the RMFD (56) where

$$\begin{aligned} D(s) &= \beta^{-1}(\Delta(s) + \alpha\Psi(s)) && (61)\\ N(s) &= \gamma\Psi(s). && (62) \end{aligned}$$

Of course we can go backwards. Given the RMFD (56) we multiply $N(s)$ and $D(s)$ on the right by a unimodular matrix so that the matrix of highest column coefficients of $D(s)$ is nonsingular. The decomposition (61), (62) defines α, β and γ of a controller realization of the transfer function.

For a system in observable form (24), (25) we use the relations

$$\begin{aligned} v(s) &= -B\alpha x(s) + \beta u(s) && (63)\\ y(s) &= Cz(s) && (64) \end{aligned}$$

We obtain the LMFD

$$y(s) = D^{-1}(s)N(s)u(s) \tag{65}$$

where

$$\begin{aligned} D(s) &= \Delta(s) + \alpha\Psi(s) && (66) \\ N(s) &= \Phi(s)\beta. && (67) \end{aligned}$$

On the other hand given a LMFD (65) we can multiply $D(s)$ and $N(s)$ on the left by a unimodular matrix to obtain the decomposition (66), (67). This defines α and β of an observable form realization.

For a system in observer form (36), (37) we use the relations

$$\begin{aligned} v(s) &= -\alpha Cx(s) + \beta u(s) && (68) \\ y(s) &= \gamma Cz(s) && (69) \end{aligned}$$

which lead to a LMFD (65) where

$$\begin{aligned} D(s) &= (\Delta(s) + \Phi(s)\alpha)\gamma^{-1} && (70) \\ N(s) &= \Phi(s)\beta. && (71) \end{aligned}$$

Given the LMFD (65) the decomposition (70), (71) defines α, β and γ of an observer form realization.

4 Nonlinear Observable and Controller Forms

Henceforth we focus our attention on the nonlinear system

$$\begin{aligned} \dot{xi} &= f(\xi) + g(\xi)u && (72) \\ y &= h(\xi) && (73) \end{aligned}$$

where $\xi \in \mathbf{R}^n$, $u \in \mathbf{R}^m$, $y \in \mathbf{R}^p$ and f, g, h are smooth (C^∞) functions. We are interested in (72), (73) in some open connected subset M of the state space containing the nominal operating point ξ^0.

We introduce some terminology and notation. The *Lie derivatives* of the function $h_i(\xi)$ by the vector fields $f(\xi)$ and $g^j(\xi)$ are functions defined by

$$\begin{aligned} L_f(h_i)(\xi) &= \frac{\partial h_i}{\partial \xi}(\xi)f(\xi) && (74) \\ L_{g^j}(h_i)(\xi) &= \frac{\partial h_i}{\partial \xi}(\xi)g^j(\xi). && (75) \end{aligned}$$

Of course these operations can be iterated,

$$L_f^r(h_i) = L_f(L_f^{r-1}(h_i)). \tag{76}$$

The *differential* dh_i of a function h_i is a one form defined by

$$dh_i(\xi) = \frac{\partial h_i}{\partial \xi}(\xi). \tag{77}$$

A *one form* ω is a row vector field or more precisely a C^∞ linear combination of differentials.

$$\omega(\xi) = (\omega^1(\xi), \ldots, \omega^n(\xi)) = \sum k_i(\xi) dh_i(\xi) \tag{78}$$

where $k_i(\xi)$ and $h_i(\xi)$ are smooth functions. A one form can be paired with a vector field (all vector fields are columns unless otherwise stated) to obtain a function

$$\langle \omega, f \rangle(\xi) = \omega(\xi) f(\xi) = \sum_{i=1}^{n} \omega^i(\xi) f_i(\xi). \tag{79}$$

A vector field can also Lie differentiate a form to obtain another one form

$$L_f(\omega) = \omega \frac{\partial f}{\partial \xi} + (\frac{\partial \omega^*}{\partial \xi} f)^* \tag{80}$$

where $*$ denotes transpose. In particular

$$L_f(dh_i) = d(L_f(h_i)). \tag{81}$$

A vector field can also Lie differentiate another vector field to yield a third vector field.

$$ad(f)g^j = [f, g^j] = \frac{\partial g^j}{\partial \xi}(\xi) f(\xi) - \frac{\partial f}{\partial \xi}(\xi) g^j(\xi). \tag{82}$$

This can be iterated,

$$ad^r(f)g^j = [f, ad^{r-1}(f)g^j]. \tag{83}$$

The operation (82) is also called the *Lie bracket* (82) of the vector fields and can be thought of both as a multiplication and as a differentiation. This is evidenced by the following Liebnitz-type formula called the Jacobi identity

$$[f, [g^i, g^j]] = [[f, g^i], g^j] + [g^i, [f, g^j]]. \tag{84}$$

Moreover the pairing (79) satisfies a Liebnitz formula with respect to Lie differentiation

$$L_f(\langle \omega, g^j \rangle) = \langle L_f(\omega), g^j \rangle + \langle \omega, [f, g^j] \rangle \tag{85}$$

For the readers unfamiliar with these concepts we suggest the calculation of the above definitions and formulas in the linear case (4), (5) where

$$
\begin{aligned}
f(\xi) &= F\xi && (86)\\
g^j(\xi) &= G^j && (87)\\
h_i(\xi) &= H_i\xi. && (88)
\end{aligned}
$$

We define

$$\mathcal{E}^\ell = C^\infty\{L_f^{r-1}(dh_i) : i = 1,\dots,p;\ r = 1,\dots,\ell\} \qquad (89)$$

where $C^\infty\{\cdot\}$ means the linear span over C^∞ coefficients. Such a collection of one forms which is closed under addition and multiplication by C^∞ functions is called a *codistribution.* We denote by $\mathcal{E}^\ell(\xi)$ the linear space of $1 \times n$ vectors obtained by evaluating the one forms of $\mathcal{E}^\ell$ at the point ξ.

Given indices $\ell_1,\dots,\ell_p$ we define

$$\mathcal{E}^\ell_{\ell_1,\dots,\ell_p} = C^\infty\{L_f^{r-1}(dh_i) : i = 1,\dots,p;\ r = 1,\dots,\ell \wedge \ell_i\} \qquad (90)$$

and $\mathcal{E}^\ell_{\ell_1,\dots,\ell_p}(\xi)$ the vector space obtained by evaluation of these forms at ξ.

The system is (72), (73) has *observability indices* $\ell_1,\dots,\ell_p$ around ξ^0 if $\ell_1 + \cdots + \ell_p = n$, and

$$\text{dimension } \mathcal{E}^n(\xi) = n \qquad (91)$$

and

$$\mathcal{E}^\ell(\xi) = \mathcal{E}^\ell_{\ell_1,\dots,\ell_p}(\xi) \qquad (92)$$

for $\ell = 1,\dots,n$ and all ξ in some neighborhood of ξ^0. The reader who has done the suggested calculations recognizes (91) as a generalization of (12) and (92) a generalization of (13). The *observability indices* are strict if

$$L_f^{\ell_i}(dh_i) \in \mathcal{E}^{\ell_i} \qquad (93)$$

for $i = 1,\dots,p$. This generalizes the linear definition.

The set of observability indices of (72), (73) is uniquely determined by h and f and is invariant under changes of coordinates in the state and output spaces. There can be some freedom in the ordering of the indices even when the ordering of the outputs remains fixed. The observability indices are strict iff there is only one ordering satisfying (92). To simplify notation we restrict our attention to systems where all the observability indices are positive.

Condition (91) could be called *zero input observability.* It means that the state $\xi(t)$ of (72), (73) can be distinguished from its neighbors by the output $y(t)$ and its first $n-1$ time derivatives along the trajectories near ξ^0 corresponding to

$u(t) = 0$. Unlike the linear case, (91) does not imply the existence of observability indices satisfying (92) around every ξ^0 but only for a generic, (i.e. an open and dense) set of ξ^0's. The latter condition implies that the functions

$$x_{ir} = L_f^{r-1}(h_i)(\xi) \tag{94}$$

for $i = 1, \ldots, p;\ r = 1, \ldots, \ell_i$ are valid local coordinates on the state space. When (72),(73) has observability indices around a point ξ^0 which is a critical point of f, $(f(\xi^0) = 0)$ then they agree with the observability indices of the linear approximating system to (72), (73) at ξ^0.

The *observable form* of a nonlinear system is

$$\dot{x} = Ax - B\alpha(x) + \beta(x)u \tag{95}$$

$$y = Cx \tag{96}$$

where (A, B, C) is a prime triple with indices $\ell_1, \ldots, \ell_p$ and α, β are smooth $m \times 1$, $n \times m$ matrix valued functions of x.

Proposition 4.1 . *A nonlinear system in observable form (95), (96) has observability indices $\ell_1, \ldots, \ell_p$. A nonlinear system (72), (73) can be transformed into observable form (95), (96) by a change of local coordinates around ξ^0 iff the system has observability indices around ξ^0. If (72), (73) has observability indices $\ell_1, \ldots, \ell_p$ around ξ^0 then the x coordinates of the form (26) given by (94) transform it to observable form. The observable form of a nonlinear system, the associated x coordinates and the nominal x-operating point $x^0 = T^{-1}(\xi^0)$ are unique up to a reordering of the observability indices. The functions α and β of the observable form (95), (96) are given by*

$$\alpha_i = L_f^{\ell_i}(h_i) \tag{97}$$

$$\beta_{ir}^j = L_{g^j} L_f^{r-1}(h_i). \tag{98}$$

The observability index assumption (92) implies that

$$d\alpha_i = L_f^{\ell_i}(dh_i) \in \mathcal{E}^{\ell_i+1} \tag{99}$$

which means that α_i does not depend on x_{jr} if $r > \ell_i + 1$. The observability indices are strict (93) iff α_i does not depend on a x_{jr} if $r \geq \ell_i + 1$, in other words

$$d\alpha_i = L_f^{\ell_i}(dh_i) \in \mathcal{E}^{\ell_i}. \tag{100}$$

The proof of this result is relatively straightforward, for example see [4], section 2.

We now turn to controllability properties of (72), (73). We define

$$\mathcal{D}^\ell = C^\infty\{\mathrm{ad}^{r-1}(-f)g^j : j = 1,\dots,m;\ r = 1,\dots,\ell\}. \tag{101}$$

This is a collection of vector fields closed under addition and multiplication by C^∞ functions; such an object is called a *distribution*. Given indices $\ell_1,\dots,\ell_m$, let

$$\mathcal{D}^\ell_{\ell_1,\dots,\ell_m} = \{\mathrm{ad}^{r-1}(-f)g^j : j = 1,\dots,m;\ r = 1,\dots,\ell \wedge \ell_j\}. \tag{102}$$

The system (72), (73) has *controllability indices* $\ell_1,\dots,\ell_m$ around ξ^0 if $\ell_1 + \cdots + \ell_m = n$ and

$$\text{dimension } \mathcal{D}^n(\xi) = n \tag{103}$$

and

$$\mathcal{D}^\ell(\xi) = \mathcal{D}^\ell_{\ell_1,\dots,\ell_m}(\xi) \tag{104}$$

for $\ell = 1,\dots,n$ and all ξ in some neighborhood of ξ^0. Of course (103) is a generalization of (6) and (104) is a generalization of (7). The *controllability indices are strict* if

$$\mathrm{ad}^{\ell_j}(-f)g^j \in \mathcal{D}^{\ell_j}_{\ell_1,\dots,\ell_m} \tag{105}$$

for $i = 1,\dots,n$. This generalizes the linear definition.

The set of controllability indices of (72), (73) is uniquely determined by f and g and is invariant under change of coordinates in the state space and nonlinear state feedback, i.e. $u = \alpha(x) + \beta(x)v$ where $\beta(x)$ is $m \times m$ invertible. There can be some freedom in the ordering of the indices even when the output is fixed. The controllability indices are strict iff there is only one ordering satisfying (104). For notational convenience, we restrict our attention to systems where the controllability indices are positive.

Condition (103) could be called *local linear controllability* for if ξ^0 is a critical point of f, $(f(\xi^0) = 0)$ then the linear approximation to (72), (73) at ξ^0 is controllable iff (103) holds. Once again (103) does not imply the existence of controllability indices satisfying (104) around every ξ^0, only for an open, dense set of ξ^0's. When (72), (73) has controllability indices around a critical point ξ^0, they agree with the controllability indices of the linear approximating system to (72), (73) at ξ^0.

The *controller form* of a nonlinear system is

$$\begin{aligned} \dot{x} &= Ax - B\alpha(x) + B\beta(x)u & (106)\\ y &= \gamma(x) & (107) \end{aligned}$$

where (A, B, C) is a prime triple with indices $\ell_1,\dots,\ell_m$ and α, β, γ are smooth $m \times 1$, $m \times m$, $p \times 1$ matrix valued functions of x which are arbitrary except that $\beta(x)$ must be nonsingular. The question of when a nonlinear control system can be transformed to controller form has been independently solved by several authors [5,6,7,8,9,22]. Some only considered special cases like $m = 1$ or $\beta(x)$ =constant. Our treatment follows Hunt and Su [8].

Recall that a distribution $\mathcal{D}$ is *involutive* if it is closed with respect to Lie bracket, i.e. $[q^1, q^2] \in \mathcal{D}$ whenever $q^1, q^2 \in \mathcal{D}$. Given a distribution we can consider the under-determined systems of partial differential equations.

$$\langle dk, q\rangle = 0 \;\text{ forall}\; q \in \mathcal{D} \tag{108}$$

for the unknown function $k(\xi)$. The question of existence and uniqueness of local solutions to (73) is addressed by the following.

Frobenius Theorem Suppose $\mathcal{D}$ is of constant codimension d. $\mathcal{D}$ is involutive iff locally there exists d independent solutions $k_1, \ldots, k_d$ to (73). Any other solution $k(\xi)$ is a function of $k_1(\xi), \ldots, k_d(\xi)$.

Proposition 4.2 *([8], see also [5,6,7,8,9,22]). A nonlinear system in controller form (106), (107) has controllability indices $\ell_1, \ldots, \ell_m$ which are strict relative to the input $\tilde{u} = \beta u$. A nonlinear system (72), (73) can be transformed into controller form (106), (107) by a local change of coordinates around ξ^0 iff it has controllability indices $\ell_1, \ldots, \ell_m$ and $\mathcal{D}^{\ell_j - 1}$ is involutive for $j = 1, \ldots, m$.*

Proof The proof of the first statement is a straightforward verification.

As for the second suppose (72), (73) can be transformed to controller form by $\xi = T(x)$. Using the C matrix of the prime triple we define a pseudo-output.

$$\psi = k(\xi) = CT^{-1}(\xi) \tag{109}$$

then the function k satisfies

$$\begin{aligned} L_f^{\ell_i}(k_i) &= \alpha_i && (110)\\ L_{g^j} L_f^{r-1}(k_i) &= \begin{cases} 0 & 1 \le r < \ell_i \\ \beta_i^j & \quad r = \ell_i \end{cases} && (111) \end{aligned}$$

Using the Liebnitz formula (85) and induction we can show that (111) is equivalent to

$$\langle L_f^s(dk_i), \mathrm{ad}^{r-s-1}(-f)g^j\rangle = \begin{cases} 0 & 1 \le r < \ell_i, \quad 0 \le s < r \\ \beta_i^j & \quad r = \ell_i, \quad\;\; 0 \le s < r \end{cases} \tag{112}$$

From this it follows that for every $q \in \mathcal{D}^\ell$

$$\langle L_f^{r-1}(dk_i), q\rangle = 0 \tag{113}$$

for $i = 1,\dots,n$ and $r = 1,\dots,\ell_i - \ell$. Moreover from the invertibility of β it follows that the functions $\{L_f^{r-1}(k_i)\ i = 1,\dots,m$ and $r = 1,\dots,\ell_i - \ell\}$ are independent. There are as many such functions as the codimension of $\mathcal{D}^\ell$ so by the Frobenius theorem $\mathcal{D}^\ell$ is involutive for all ℓ and in particular for $\ell = \ell_j - 1$; $j = 1,\dots,m$.

On the other hand if $\mathcal{D}^{\ell_j - 1}$ is involutive for $j = 1,\dots,m$ then by repeated application of the Frobenius theorem one can find independent functions $k_1,\dots,k_m$ satisfying (112) where β is some invertible $m \times m$ matrix valued function. If we define x coordinates by

$$x_{jr} = L_f^{r-1}(k_j)(\xi) \tag{114}$$

for $j = 1,\dots,m$; $r = 1,\dots,\ell_j$ then these coordinates transform the nonlinear system to controller form (106), (107). The functions α and β are given by (110), (111).

When it exists, the controller form of a nonlinear system is not unique. From the proof of the above we see that the controller form is completely determined by the choice of the pseudo-output $k(\xi)$ satisfying (111) for some invertible $\beta(\xi)$. If $\tilde{k}(\xi)$ is another solution of (111) then (112) and (113) imply that $\tilde{k}_i(\xi)$ is a function of $x_{jr} = L_f^{r-1}(k_j)$ for $\ell_j \geq \ell_i$ and $r = 1,\dots,\ell_j - \ell_i$.

Notice that the nominal operating point $x^0 = T^{-1}(\xi^0)$ of the controller form is determined by the choice of $k(\xi)$. In particular there exists k such that $x^0 = 0$ iff there exists u^0 such that $f(\xi^0) + g(\xi^0)u^0 = 0$.

Another point worth mentioning is that the system (72) with pseudo-output $\psi = k(\xi)$ does not necessarily have observability indices $\ell_1,\dots,\ell_m$. This would be the case iff in addition to (112), $k(\xi)$ satisfies

$$\langle L_f^{\ell_i}(dk_i), \mathrm{ad}^{r-1}(-f)g^j\rangle = 0$$

for $r = 1,\dots,\ell_j - \ell_i - 1$.

We might try to obtain a unique controller form by requiring that α and β also satisfy the nonlinear generalizations of (11) and (35), namely

$$\beta_i^j(\xi) = \delta_i^j \quad \ell_i \leq \ell_j \tag{115}$$

$$\frac{\partial \alpha_i}{\partial x_{jr}} = 0 \quad r > \ell_i \tag{116}$$

But this would reduce the number of nonlinear systems that admit a controller form. The conditions (115), (116) imply that

$$\langle dk_i, \mathrm{ad}^{r-1}(-f)g^j\rangle = \begin{cases} 0 & 1 \leq r < \ell_j \\ \beta_i^j & r = \ell_j \end{cases} \tag{117}$$

This is a system of first order partial differential equations for the unknown functions $k_1, \dots, k_m$. The solvability of such a system is addressed by the following. **Integrability Theorem** Let $q^1(\xi), \dots, q^n(\xi)$ be an n linearly independent n dimensional vector fields. There exists a solution $k = (k_1, \dots, k_m)$ to the system

$$\langle dk_i, q^j \rangle = \begin{cases} \delta_i^j & j = 1, \dots, m \\ 0 & j = m+1, \dots, n \end{cases}$$

iff

$$[q^i, q^j] \in \mathcal{D} \quad i, j = 1, \dots, n$$

where $\mathcal{D}$ is the distribution spanned by $\{q^{m+1}, \dots, q^n\}$. The solution is unique up to a choice of $k(\xi^0)$.

From this theorem we see that there exists a solution to (117) iff

$$[\mathrm{ad}^{r-1}(-f)g^i, \ \mathrm{ad}^{s-1}(-f)g^j] \in \mathcal{D} \tag{118}$$

for $i, j = 1, \dots, n$; $r = 1, \dots, \ell_i$, $s = 1, \dots, \ell_j$ where $\mathcal{D}$ is the distribution given by

$$\mathcal{D} = C^\infty\{\mathrm{ad}^{r-1}(-f)g^j : j = 1, \dots, n; \ r = 1, \dots, \ell_j - 1\}. \tag{119}$$

Condition (118),(119) is considerably more stringent then $\mathcal{D}^{\ell_j - 1}$ being involutive for $j = 1, \dots, m$. In particular suppose we consider a generic nonlinear system (72), (73) with $n = 2$ and $m = 1$. Around a generic point ξ^0, the vector fields g^1 and $\mathrm{ad}(-f)g^1$ are linear independent hence such a system has a controllability index $\ell_1 = 2$. The distribution $\mathcal{D}^1 = C^\infty\{g^1\}$ is trivially involutive so such a system has a controller form. However condition (118), (119) which in this case is

$$[g^1, \ \mathrm{ad}(-f)g^1] \in C^\infty\{g^1\}$$

is not generically satisfied.

Suppose (72), (73) has controllability indices $\ell_1, \dots, \ell_m$ around ξ^0. Regardless of whether or not it admits a controller form around ξ^0, it is always possible to make the controllability indices strict by a change of input coordinates $\tilde{u} = \beta(\xi)u$ when $\beta_i^j(\xi) = \delta_i^j$ for $\ell_i \leq \ell_j$ as in the linear case. We define one forms $\omega_1(\xi), \dots, \omega_p(\xi)$ by

$$\langle \omega_i, \mathrm{ad}^{r-1}(-f)g^j \rangle = \begin{cases} 0 & 1 \leq r < \ell_i \\ \delta_i^j & r = \ell_i \end{cases} \tag{120}$$

From this and the controllability index assumptions (104) it follows that

$$\langle \omega_i, \mathrm{ad}^{r-1}(-f)g^j \rangle = 0 \quad \ell_j < r < \ell_i. \tag{121}$$

Moreover by repeated use of the Liebnitz formula (85) we see that (120), (121) is equivalent to

$$\langle L_f^{r-1}(\omega_i), g^j \rangle = \begin{cases} 0 & 1 \leq r < \ell_j \text{ or } \ell_j < r < \ell_i \\ \delta_i^j & r = \ell_j \end{cases} . \tag{122}$$

We define β by

$$\beta_i^j = \langle L_f^{\ell_i - 1}(\omega_i), g^j \rangle. \tag{123}$$

Immediately we see that $\beta_i^j = \delta_i^j$ for $\ell_i \leq \ell_j$ so β is invertible. It is not hard to show that the system defined by $(\tilde{f}, \tilde{g}) = (f, g\beta^{-1})$ has strict controllability indices $\ell_1, \ldots, \ell_m$. Notice that if (117) is solvable then $\omega_i = dk_i$.

5 Nonlinear Controllable and Observer Forms

The *controllable form* of a nonlinear system is

$$\dot{x} = Ax + \alpha(Cx) + Bu \tag{124}$$
$$y = \gamma(x) \tag{125}$$

where (A, B, C) is a prime triple with indices ℓ_1, ℓ_m and α, γ are smooth matrix valued functions of dimensions $n \times 1$, $p \times 1$. Notice that α is a function of the pseudo-output $\psi = Cx$ while γ is a function of x.

Notice that if $\alpha(\psi)$ is a linear function of ψ then the dynamics (124) of the nonlinear controllable form agree with the dynamics (18) of the linear controllable form. Hence the question of the existence of a nonlinear controllable form is closely related to the question of linearizing the dynamics (124) by a change of state coordinates. This latter question has a long history going back to Poincare [16]. For more recent work see [17,18,19,20,21].

For the most part the controllable forms of nonlinear systems have not appeared explicitly in the literature. But as one might expect they have arisen implicitly in some of the work on observer form [4,10], and on linearization [21]. The following is a reformulation of similar results from [10], [21] and [22].

Proposition 5.1 *A nonlinear system in controllable form has controllability indices $\ell_1, \ldots, \ell_m$. A nonlinear system (72), (73) can be transformed into controllable form (124), (125) by a change of local coordinates around ξ^0 iff it has controllability indices $\ell_1, \ldots, \ell_m$ and*

$$[\mathrm{ad}^{r-1}(-f)g^i,\ \mathrm{ad}^{s-1}(-f)g^j] = 0 \tag{126}$$

for $i, j = 1, \ldots, m$ and $r = 1, \ldots, \ell_i$, $s = 1, \ldots, \ell_j$ around ξ^0. Controllable form and the associated x coordinates are unique up to a choice of the nominal x-operating point $x^0 = T^{-1}(\xi^0)$ and up to reordering of the controllability indices.

The dynamics (72) of a nonlinear system can be linearized, or equivalently, can be transformed to the dynamics of linear controllable form (18) by a change of state coordinates around ξ^0 iff (72) has controllability indices $\ell_1, \ldots, \ell_m$ and (126) holds for $i, j = 1, \ldots, m$ and $r = 1, \ldots, \ell_i + 1$; $s = 1, \ldots, \ell_j + 1$.

Proof Consider the nonlinear system in controllable form (124), (125). It is a straightforward calculation to show that

$$\mathrm{ad}^{r-1}(-Ax + \alpha(\psi))B^j = \begin{cases} A^{r-1}\,B^j & 1 \le r \le \ell_j \\ \frac{\partial \alpha}{\partial \psi_j} & r = \ell_j + 1 \end{cases} . \tag{127}$$

Hence the controllable form (124), (125) has controllability indices $\ell_1, \ldots, \ell_m$. Moreover if (72), (73) can be transformed to (124), (125) by a change of state coordinates then clearly (72), (73) must have the same controllability indices and (126) must hold.

On the other hand suppose (72), (73) has controllability indices $\ell_1, \ldots, \ell_m$ and (126) holds. By the integrability theorem of Section 4 with $m = n$, we can choose coordinate functions $x_{ir}(\xi)$, $i = 1, \ldots, m$; $r = 1, \ldots, \ell_i$ such that

$$\langle dx_{ir},\ \mathrm{ad}^{\ell_j - s}(-f)g^j \rangle = \delta_i^j \delta_r^s \tag{128}$$

for $i, j = 1, \ldots, m$ and $r = 1, \ldots, \ell_i$, $s = 1, \ldots, \ell_j$.

In the x coordinates, $\mathrm{ad}^{\ell_j - s}(-f)g^j$ becomes the unit vector in the direction x_{js} or in other words

$$\frac{\partial x}{\partial \xi}\mathrm{ad}^{\ell_j - s}(-f)g^j = A^{\ell_j - s}B^j \tag{129}$$

for $j = 1, \ldots, m$ and $s = 1, \ldots, \ell_j$, where A, B are from the prime triple with indices $\ell_1, \ldots, \ell_m$. Let $\tilde{f}(x)$ be the transform of $f(\xi)$ into x coordinates.

$$\tilde{f}(x) = \frac{\partial x}{\partial \xi}(\xi(x)) f(\xi(x)) \tag{130}$$

Then from (129), (130) we have if $s > 1$

$$\begin{aligned} \frac{\partial}{\partial x_{js}} \tilde{f}(x) &= [A^{\ell_j - s}B^j,\ \tilde{f}(x)] \\ &= [\frac{\partial x}{\partial \xi}\mathrm{ad}^{\ell_j - s}(-f)g^j,\ \frac{\partial x}{\partial \xi} f] \\ &= \frac{\partial x}{\partial \xi}\mathrm{ad}^{\ell_j - s + 1}(-f)g^j \end{aligned}$$

$$= A^{\ell_j - s+1} B^j \tag{131}$$

From this we conclude that

$$\tilde{f}(x) = Ax + \alpha(Cx)$$

where $C_i x = x_{i1}$, $i = 1, \ldots, m$.

We now prove the last part of the theorem. If the nonlinear dynamics (72) can be transformed to the dynamics (18) of linear controllable form then by the above it must have controllability indices $\ell_1, \ldots, \ell_m$ and (126) must hold for $r = 1, \ldots, \ell_i$ and $s = 1, \ldots, \ell_j$. Moreover we see from (127) that $\mathrm{ad}^{\ell_j}(-f)g^j$ must transform to a constant vector field in x coordinates so (126) must hold for $r = 1, \ldots, \ell_i + 1$ and $s = 1, \ldots, \ell_j + 1$.

On the other hand if (126) holds for $r = 1, \ldots, \ell_i + 1$ and $s = 1, \ldots, \ell_j + 1$ then $\mathrm{ad}^{\ell_k}(-f)g^k$ must be a constant linear combination of the frame of vector fields $\{\mathrm{ad}^{r-1}(-f)g^i,\ i = 1, \ldots, m;\ r = 1, \ldots, \ell_i\}$. To see this, suppose for some functions $\lambda^k_{ir}(\xi)$

$$\mathrm{ad}^{\ell_k}(-f)g^k = \sum_{i=1}^{m} \sum_{r=1}^{\ell_i} \mathrm{ad}^{r-1}(-f)g^i \lambda^k_{ir}.$$

Bracketing with $\mathrm{ad}^{s-1}(-f)g^j$ yields

$$0 = \sum_{i=1}^{m} \sum_{r=1}^{\ell_i} \mathrm{ad}^{r-1}(-f)g^i L_{\mathrm{ad}^{s-1}(-f)g^j}(\lambda^k_{ir}).$$

The linear independence of the vector fields of the frame implies that for $j = 1, \ldots, m$; $s = 1, \ldots, \ell_j$

$$0 = L_{\mathrm{ad}^{s-1}(-f)g^j}(\lambda^k_{ir})$$

hence λ^k_{ir} is a constant. By (127) this implies that

$$\alpha_{ir}(\psi) = \sum_{k=1}^{m} \lambda^k_{ir}\psi_k \quad \text{QED.}$$

Notice that it is more difficult for a nonlinear system (72), (73) to have a controllable form than to have a controller form. Clearly conditions (126) implies that $\mathcal{D}^{\ell_j - 1}$ is involutive for $j = 1, \ldots, m$. This extra difficulty is partially explained by the extra freedom afforded by $\beta(x)$ in the controller form which is lacking in the controllable form. Zeitz defines controllable form with $\beta(x)$ present [22]. There is also more freedom in the α of the controller form than the α of the controllable form. The former is an $\mathbf{R}^m$ valued function of $\mathbf{R}^n$ while the latter is a $\mathbf{R}^n$ valued function of $\mathbf{R}^m$. The linear terms of the Taylor series expansion have the same number of degrees of freedom, $n\ m$, but there are more degrees of freedom in the higher order terms of the controller form that the controllable form. In particular for the terms of order 2, there are $mn(n+1)/2$ degrees of freedom in the former and $nm(m+1)/2$ in the latter.

The *observer form* of a nonlinear system is

$$\dot{x} = Ax - \alpha(Cx) + \beta(Cx)u \tag{132}$$

$$y = \gamma(Cx) \tag{133}$$

where (A, B, C) is a prime triple with indices $\ell_1, \ldots, \ell_p$ and α, β and γ are smooth matrix valued functions of dimensions $n \times 1$, $m \times m$ and $p \times 1$. They are arbitrary except that γ must be a local diffeomorphism. We let $\bar{y} = Cx$.

Observer form is useful in the construction of asymptotic observers

$$\dot{\hat{x}} = A\hat{x} + \alpha(\bar{y}) + \beta(\bar{y})u + M(\bar{y} - C\hat{x}) \tag{134}$$

with linear error dynamics

$$\dot{\tilde{x}} = (A - MC)\tilde{x}. \tag{135}$$

The question of when a nonlinear system can be transformed to observer form has been considered by several authors [4], [10,11,12,13,14], [22]. Most treated only special cases like $p = 1$ or γ =identity. The general solution can be found in [4]. The approach taken in [4] is similar to the approach described above for the linear case.

Suppose the nonlinear system (72), (73) can be transformed into observer form (132), (133) by a local change of coordinates around ξ^0. Using the B matrix of the prime triple we add a pseudo-input μ to (132)

$$\dot{x} = Ax - \alpha(Cx) + \beta(Cx)u + B\mu. \tag{136}$$

When u is held constant at 0, (136) can be viewed as the controllable form relative to the pseudo-input μ.

We transform (136) back to ξ-coordinates

$$\dot{\xi} = f(\xi) + g(\xi)u + \bar{q}(\xi)\mu$$

which defines the vector fields $\bar{q} = \bar{q}^1, \ldots, \bar{q}^p$. These vector fields satisfy

$$\langle L_f^{r-1}(d\bar{y}_i),\ \bar{q}^j\rangle = \begin{cases} 0 & 1 \leq r < \ell_i \\ \delta_i^j & r = \ell_i \end{cases} . \tag{137}$$

If $\bar{q}$ is known then we can recover the observer form by choosing local coordinates x_{ir} to satisfy

$$\langle dx_{ir},\ \mathrm{ad}^{\ell_j - s}(-f)\bar{q}^j\rangle = \delta_i^j \delta_r^s \tag{138}$$

for $i, j = 1, \ldots, p$, $r = 1, \ldots, \ell_i$ and $s = 1, \ldots, \ell_j$. Such coordinates exist iff

$$[\mathrm{ad}^{\ell_i - r}(-f)\bar{q}^i,\ \mathrm{ad}^{\ell_j - s}(-f)\bar{q}^j] = 0 \tag{139}$$

for $i, j = 1, \ldots, p$; $r = 1, \ldots, \ell_i$; $s = 1, \ldots, \ell_j$ and

$$[\mathrm{ad}^{\ell_i - r}(-f)\bar{q}^i, g^j] = 0 \tag{140}$$

for $i = 1, \ldots, p$; $j = 1, \ldots, n$ and $r = 2, \ldots, \ell_i$.

Summarizing the discussion, an observer form (132,133) of (72), (73) exists iff there exists a change of coordinates $y = \gamma(\bar{y})$ on the output space and vector fields $\bar{q}^1, \ldots, \bar{q}^p$ determined by γ via (137) such that (139),(140) holds. In effect (139),(140) constitute an overdetermined system of partial differential equations for the change of coordinates $y = \gamma(\bar{y})$ on the output space. To analyze such equations we must introduce the geometric concept of a Koszul connection on the output space. Let $\phi^i(y), i = 1, 2, \ldots$ denote vector fields on the p dimensional output space. A *Koszul connection* on y-space is a mapping Δ from pairs of such vector fields to vector fields.

$$\Delta : (\phi^i, \phi^j) \mapsto \Delta_{\phi^i}(\phi^j) \tag{141}$$

This mapping is linear over C^∞ functions in the first argument and satisfies a Liebnitz formula in the second argument. In other words if $\lambda_i(y)$ and $\mu_j(y)$ are smooth functions then

$$\Delta_{\sum_i \lambda_i \phi^i}\Big(\sum_j \mu_j \phi^j\Big) = \sum_{i,j}\big(\lambda_i \mu_j \Delta_{\phi^i}(\phi^j) + \lambda_i L_{\phi^i}(\mu_j)\phi^j\big). \tag{142}$$

If $\phi^1(y), \ldots, \phi^p(y)$ is a local frame of vector fields then Δ is completely determined by its *Christoffel symbols* $\Gamma_k^{ij}(y)$ relative to this frame. These are defined by the expressions

$$\Delta_{\phi^i}(\phi^j) = \sum_k \Gamma_k^{ij} \phi^k. \tag{143}$$

Consider a second frame $\bar{\phi}^1, \ldots, \bar{\phi}^p$ related to the first by

$$\bar{\phi}^i = \sum_{p=1}^{p} \phi^\rho \mu_\rho^i \tag{144}$$

and

$$\phi^r = \sum_{k-1}^{\rho} \bar{\phi}^k \lambda_k^r \tag{145}$$

where $\Lambda = (\lambda_i^j)$ is a $p \times p$ nonsingular matrix valued function of y and $\Lambda^{-1} = (\mu_i^j)$. It follows from (141,142) (143) and (144, 145) that

$$\Gamma_k^{ij} = \sum_{\rho,\sigma,\tau} \mu_\rho^i \mu_\sigma^j \lambda_k^\tau \Gamma_\tau^{\rho\sigma} + \sum_{\rho,\tau} \mu_\rho^i \lambda_k^\tau L_{\phi^\rho}(\mu_\tau^j). \tag{146}$$

A Koszul connection Δ has *zero curvature* if there exists a frame where the Christoffel symbols are zero. From equation (146) we obtain the partial differential equation that such a change of frame must satisfy. It is more conveniently written in matrix notation where Γ^ρ denotes the $p \times p$ matrix $(\Gamma_\tau^{\rho\sigma})$ with row index τ and column index σ,

$$0 = \Gamma^\rho \Lambda^{-1} + L_{\phi^\rho}(\Lambda^{-1}) \tag{147}$$

or equivalently

$$L_{\phi^\rho}(\Lambda) = \Lambda \Gamma^\rho. \tag{148}$$

The integrability condition for this is

$$L_{\phi^\rho} L_{\phi^\sigma}(\Lambda) - L_{\phi^\sigma} L_{\phi^\rho}(\Lambda) - L_{[\phi^\sigma,\phi^\rho]}(\Lambda) = 0 \tag{149}$$

or equivalently

$$\Lambda(\Gamma^\rho \Gamma^\sigma - \Gamma^\sigma \Gamma^\rho + L_{\phi^\rho}(\Gamma^\sigma) - L_{\phi^\rho}(\Gamma^\rho) - \sum_\tau c_\tau^{\sigma\rho} \Gamma^\tau) = 0 \tag{150}$$

where $C_\tau^{\sigma\rho}$ are the *structural coefficients* of the frame

$$[\phi^\sigma, \phi^\rho] = \sum_\tau C_\tau^{\sigma\rho} \phi^\tau. \tag{151}$$

The coefficient of Λ in (150) is the *curvature* of Δ.

It is convenient to work with frames of vector fields arising from coordinates on the output space. Suppose y and $\bar{y}$ are two different coordinate systems and ϕ and $\bar{\phi}$ are the associated frames, i.e.

$$L_{\phi^i} = \frac{\partial}{\partial y_i} \quad L_{\bar{\phi}^i} = \frac{\partial}{\partial \bar{y}_i}. \tag{152}$$

These frames are related by the chain rule

$$\phi = \bar{\phi} \frac{\partial \bar{y}}{\partial y} \tag{153}$$

so

$$\Lambda = \partial \bar{y} / \partial y. \tag{154}$$

A Koszul connection Δ is *flat* if there exists a coordinate frame for which the Christoffel symbols are zero. Such coordinates are said to be *flat* relative to the connection. Suppose $\Gamma_{\tau}^{\phi\sigma}$ are the Christoffel symbols relative to coordinates y. Clearly we can find new coordinates $\bar{y}$ where Christoffel symbols are zero iff we can solve the pair of partial differential equations (147,148) and (154).

We rewrite these as

$$\frac{\partial}{\partial y_i}\Lambda = \Lambda\Gamma^i \tag{155}$$

$$\frac{\partial \bar{y}}{\partial y} = \Lambda. \tag{156}$$

The integrability condition for the first is the zero curvature condition (150) which can be rewritten as

$$\Gamma^j\Gamma^i - \Gamma^i\Gamma^j + \frac{\partial\Gamma^i}{\partial y_j} - \frac{\partial\Gamma^j}{\partial y_i} = 0 \tag{157}$$

for $i, j = 1, \ldots, p$. The integrability condition for the second is

$$\Gamma_k^{ij} - \Gamma_k^{ji} = 0. \tag{158}$$

The left side of (158) is called the torsion of the connection Δ. In summary a Koszul connection is flat (i.e. has Christoffel symbols zero relative to some coordinate frame) iff it has zero curvature (157) and zero torsion (158).

Suppose $\bar{y}$ are flat coordinates for a flat connection Δ. It follows from (155,156) that another set of coordinates y is flat iff y and $\bar{y}$ are affinely related, i.e. for some constant invertible matrix Λ

$$\bar{y} = \Lambda y + \bar{y}^0.$$

The relevance of the above for the problem of transforming a system to observer form is explained by the following lemmas.

Lemma 5.1 *Suppose the nonlinear system (72,73) has one distinct observability index $\ell = \ell_1 = \ldots = \ell_p$ of multiplicity p. Define vector fields $q^1, \ldots, q^p$ by*

$$\langle L_f^{r-1}(dy_i), q^j\rangle = \begin{cases} 0 & 1 \le r \le \ell_i \\ \delta_i^j & r = \ell_j \end{cases} \tag{159}$$

Define p^3 functions $\Gamma_k^{ij}(\xi)$ by

$$\Gamma_k^{ij} = \frac{1}{\ell}\langle L_f(dy_k), [\mathrm{ad}^{\ell-1}(-f)q^i, \mathrm{ad}^{\ell-2}(-f)q^j]\rangle. \tag{160}$$

Let $\bar{y} = \bar{y}(y)$ be a change of output coordinates and $\bar{q} = \bar{q}^1, \ldots, \bar{q}^p$ be vector fields defined by (137). Define another p^3 function $\bar{\Gamma}_k^{ij}(\xi)$ by

$$\bar{\Gamma}_k^{ij} = \frac{1}{\ell}\langle L_f(d\bar{y}_k), [\mathrm{ad}^{\ell-1}(-f)\bar{g}^i, \mathrm{ad}^{\ell-2}(-f)\bar{g}^j]\rangle. \tag{161}$$

Then Γ_k^{ij} and $\bar{\Gamma}_k^{ij}$ are related by

$$\bar{\Gamma}_k^{ij} = \sum_{\rho,\sigma,\tau} \frac{\partial y_\rho}{\partial \bar{y}_i}\frac{\partial y_\sigma}{\partial \bar{y}_j}\frac{\partial \bar{y}_k}{\partial y_\tau}\Gamma_\tau^{\rho\sigma} + \sum_{\rho,\tau}\frac{\partial y_\rho}{\partial \bar{y}_i}\frac{\partial \bar{y}_k}{\partial y_\tau}\frac{\partial}{\partial y_\rho}(\frac{\partial y_\tau}{\partial \bar{y}_j}). \tag{162}$$

The proof of this lemma can be found in [15]. Notice that the lemma asserts that Γ_k^{ij} transform like the Christoffel symbols of a connection on the output space, not that they are Christoffel symbols. If $\Gamma_k^{ij}(\xi)$ are actually only functions of y then they define a connection on the output space and this connection is independent of the choice of output coordinates.

Lemma 5.2 *Suppose the nonlinear system (72,73) has one distinct observability index $\ell = \ell_1 = \ldots = \ell_p$ and can be transformed to observer form (132,133) then $\Gamma_k^{ij}(\xi)$ defined by (160) are functions only of y and define a flat connection on the output space.*

Proof We compute the symbols $\bar{\Gamma}_k^{ij}$ given by (161) where $\bar{y} = Cx$ are the transformed output coordinates of the observer form. The vector fields $\bar{q}(\xi)$ defined by (143) transform to B in x coordinates. By induction we obtain

$$\mathrm{ad}^{r-1}(-Ax + \alpha(\bar{y}))B^j = \begin{cases} A^{r-1}B^j & 1 \leq r \leq \ell \\ \frac{-\partial\alpha}{\partial \bar{y}^j} & r = \ell + 1 \end{cases} \tag{163}$$

From

$$\begin{aligned}[]
[\mathrm{ad}^{\ell-1}(-Ax + \alpha(\bar{y}))B^i, \mathrm{ad}^{\ell-2}(-Ax + \alpha(\bar{y}))B^j] &= \\
[A^{\ell-1}B^i, A^{\ell-2}B^j] = 0. &
\end{aligned} \tag{164}$$

It follows that

$$\bar{\Gamma}_k^{ij} = 0. \tag{165}$$

If Γ_k^{ij} are defined by (160) then (162) and (165) shows that they are functions of y alone and can be transformed to zero by a change of output coordinates. Hence they define a flat connection on the output space. **QED**.

From these lemmas we immediately obtain the following theorem [15].

Theorem 1 *Suppose the nonlinear system (72,73) has one distinct observability index $\ell = \ell_1 = \ldots = \ell_p$ around ξ^0. It can be transformed to observer form around ξ^0 iff* *the $\Gamma_k^{ij}(\xi)$ defined by (160) are functions only of y, hence define a Koszul connection on the output space*

this connection is flat

for any flat coordinates $\bar{y}$ on the output space the vector fields defined by (137) satisfy the commutative conditions (139,140).

Consider a system with one distinct observability index $\ell = \ell_1 = \ldots = \ell_p$ which is in nonlinear observable form, i.e.

$$\dot{\rho}_{ir} = \begin{cases} \xi_{ir+1} & 1 \leq r < \ell \\ f_i(\xi) & r = \ell \end{cases} \tag{166}$$

for $i = 1, \ldots, p$ and $r = 1, \ldots, \ell$.

The vector yields $q^1, \ldots, q^p$ defined by (159) are just the unit vectors in the directions $\xi_{1\ell}, \ldots, \xi_{p\ell}$. The Γ_k^{ij} defined by (160) are given by

$$\Gamma_k^{ij} = \frac{-1}{\ell} \frac{\partial}{\partial \xi_{j2}} \frac{\partial}{\partial \xi_{i\ell}} f_k(\xi). \tag{167}$$

The change of output coordinates $\bar{y} = \gamma^{-1}(y)$ must satisfy the partial differential equations (155,156) or

$$\frac{\partial}{\partial y_i}\left(\frac{\partial \bar{y}_\rho}{\partial y_j}\right) = \frac{-1}{\ell} \sum_k \frac{\partial \bar{y}_\rho}{\partial y_k} \frac{\partial^2 f_k}{\partial \xi_{j2} \partial \xi_{i\ell}}. \tag{168}$$

The integrability conditions for this are the zero curvature condition (157) or

$$\sum_\rho \frac{\partial^2 f_\sigma}{\partial \xi_{\rho 2} \partial \xi_{i\ell}} \frac{\partial^2 f_\rho}{\partial \xi_{r2} \partial \xi_{j\ell}} - \frac{\partial^2 f_\sigma}{\partial \xi_{\rho 2} \partial \xi_{j\ell}} \frac{\partial^2 f_\rho}{\partial \xi_{r2} \partial \xi_{i\ell}} =$$

$$\ell\left(\frac{\partial^3 f_\sigma}{\partial \xi_{j1} \partial \xi_{r2} \partial \xi_{i\ell}} - \frac{\partial^3 f_\sigma}{\partial \xi_{i1} \partial \xi_{r2} \partial \xi_{j\ell}}\right) \tag{169}$$

and the zero torsion condition (158) or

$$\frac{\partial^2 f_k}{\partial \xi_{j2} \partial \xi_{i\ell}} = \frac{\partial^2 f_k}{\partial \xi_{i2} \partial \xi_{j\ell}}. \tag{170}$$

If these are satisfied then we can solve (168). If (139,140) are satisfied then we can solve (138) to find the x coordinates of observer form.

Needless to say this is a very tedious process. There is a necessary condition that a system in nonlinear observable form (166) must satisfy to be transformable to observer form. We define the *degree* of the variable ξ_{ir} to $r - 1$ and the degree of a product of such variables to be the sum of the degrees of its factors. If (166) can be transformed into observer form then $f_i(\xi)$ must be a polynomial of degree at most ℓ. We refer the reader to [4] for a proof of this. In particular if $\ell = 2$ then this degree condition, the zero curvature condition (169) and (139, 140) are necessary and sufficient. The torsion free condition (170) is trivially satisfied. It follows from (169) that (139) need only be checked for $r = s = 1$ and $i \neq j$.

If $p = 1$ then trivially the curvature and torsion are zero and (168) reduces to a first order linear ordinary differential equation for the quantity $d\bar{y}/dy$. It is solvable if the degree condition on f_1 is satisfied. In particular when $p = 1$ and $\ell = 2$ the degree condition and (140) are necessary and sufficient for the existence of observer form.

We now discuss the case where there are several distinct observability indices. The general approach is as before. To find the observer form of (72,73) if it exists we seek an appropriate change of output coordinates $\bar{y} = \gamma^{-1}(y)$ which allows us to define vector fields $\bar{q}$ via (137). If (139,140) are satisfied then we can solve (138) for the x coordinates of the observer form.

The presence of several distinct observability indices complicates the search for $\bar{y}$ and forces us to proceed in stages. Notice that for a system in observer form the observability indices are strict for the output $\bar{y} = Cx$. This is because

$$L^{r-1}_{(Ax-\alpha(y))}(d\bar{y}) = CA^{r-1} \bmod \mathcal{E}^{r-1}$$

and the output indices are strict for the pair C, A. So any nonlinear system that admits an observer form must admit a change of output coordinates which make the output indices strict. Moreover, the problem of transforming a nonlinear system with strict observability indices into observer form is greatly simplified by the following fact.

A change of output coordinates $\bar{y} = \gamma^{-1}(y)$ preserves the order and strictness of the observability indices iff

$$\frac{\partial \bar{y}_i}{\partial y_j} = 0 \text{ for } \ell_i < \ell_j. \tag{171}$$

To find a change of output coordinates which make the observability indices of (72,73) strict we start by defining vector fields $q^1, \ldots, q^p$ via (159).

It follows by the standard induction argument using the Liebnitz formula (85) that (143) implies

$$\langle dy_i, \mathrm{ad}^{r-1}(-f)q^j \rangle = \begin{cases} 0 & 1 \leq r < \ell_i \\ \beta_i^j & r = \ell_j \\ 0 & \ell_i < r < \ell_j \end{cases} \tag{172}$$

Moreover the vector fields $\mathrm{ad}^{r-1}(-f)q^j$ $j = 1,\ldots,p$; $r = 1,\ldots,\ell_j$ form a frame of n independent vector fields. These characterize $\mathcal{E}^\ell$ as

$$\mathcal{E}^\ell = \{\mathrm{ad}^{r-1}(-f)q^j :\ j = 1,\ldots,p;\ r = 1,\ldots,\ell_j - \ell\}\perp$$

$$\mathcal{E}^\ell = \{\text{oneforms } \omega :\ \langle \omega, \mathrm{ad}^{r-1}(-f)q^j\rangle = 0\ j = 1,\ldots,p;\ r = 1,\ldots,\ell_j - \ell_i\}. \tag{173}$$

Suppose $\bar{y} = \gamma^{-1}(y)$ is a change of output coordinates which preserves the ordering of the observability indices. The observability indices are strict relative to the $\bar{y}$ output iff

$$L_f^{\ell_i}(d\bar{y}_i) \in \mathcal{E}^{\ell_i} \tag{174}$$

or equivalently by (173)

$$\langle L_f^{\ell_i}(d\bar{y}_i)\mathrm{ad}^{r-1}(-f)q^j\rangle = 0 \tag{175}$$

for $r = 1,\ldots,\ell_j - \ell_i$. By induction and the Liebnitz formula this is equivalent to

$$\langle d\bar{y}_i, \mathrm{ad}^{r-1}(-f)q^j\rangle = 0 \tag{176}$$

for $r = \ell_{i+1},\ldots,\ell_j$. Since $d\bar{y} = \partial\bar{y}/\partial y\ dy$, (172) implies that (176) must hold for $r = 1,\ldots,\ell_j - 1$ also. We have shown that the observability indices are strict relative to the output $\bar{y}$ iff (176) holds for $r = 1,\ldots,\ell_j - 1$ when $\ell_j \le \ell_i$ and for $r = 1,\ldots,\ell_j$ when $\ell_j > \ell_i$.

We define p distributions

$$\mathcal{Y}^i = C^\infty\{\mathrm{ad}^{r-1}(-f)q^j :\ r = 1,\ldots,\ell_j - 1 \text{ if } \ell_j \le \ell_i \text{ and}$$
$$r = 1,\ldots,\ell_j \text{ if } \ell_j > \ell_i. \tag{177}$$

As we have just seen a change of coordinates $\bar{y} = \gamma^{-1}(y)$ preserves the ordering of the observability indices and makes them strict iff

$$d\bar{y}_i \perp \mathcal{Y}^i\ i = 1,\ldots,p. \tag{178}$$

This is an underdetermined system of first order PDE's for $\bar{y}$. By employing the Frobenius Theorem, we obtain the following reformulation of Theorem 4.2 in [4].

Proposition 5.2 *Suppose the nonlinear system (72,73) has observability indices $\ell_1,\ldots,\ell_p$ around ξ^0. There exists a local change of output coordinates $\bar{y} = \gamma^{-1}(y)$ which preserves the ordering of the observability indices and makes them strict iff the distributions $\mathcal{Y}^1,\ldots,\mathcal{Y}^p$ are involutive.*

Lemma 5.3 *Suppose the nonlinear system has strict observability indices $\ell_1,\ldots,\ell_p$ and $\ell = \min\{\ell_1,\ldots,\ell_p\}$. Define vector fields q by (159) and symbols Γ_k^{ij} by (160) then*

$$\Gamma_k^{ij} = 0 \text{ if } \ell_i > \ell \text{ or } \ell_k > \ell.$$

Proof Equation (160) can be rewritten as

$$\ell\Gamma^{ij}_k = L_{\mathrm{ad}^{\ell-1}(-f)g^i} L_{\mathrm{ad}^{\ell-2}(-f)g^j} L_f(y_k)$$
$$- L_{\mathrm{ad}^{\ell-2}(-f)q^j} L_{\mathrm{ad}^{\ell-1}(-f)q^i} L_f(y_k). \tag{179}$$

By (159)

$$L_{\mathrm{ad}^{\ell-2}(-f)q^j} L_f(y_k) = L_{q^j} L_f^{\ell-1}(y_k) = \begin{cases} 0 & \ell_k > \ell \\ \delta^j_k & \ell_k = \ell \end{cases}$$

so the first term on the right of (179) is always zero. If $\ell_k > \ell$ then

$$L_{\mathrm{ad}^{\ell-1}(-f)g^i} L_f(Y_k) = L_{q^i} L_f^{\ell}(y_k) = \begin{cases} 0 & \ell_k > \ell+1 \\ \delta^j_k & \ell_k = \ell+1 \end{cases}$$

so $\Gamma^{ij}_k = 0$.

Suppose $\ell_k = \ell$ and $\ell_i > \ell$. Then $q^i \in \mathcal{E}^{\ell\perp}$ so by the strictness assumption (174,175) it follows

$$L_{\mathrm{ad}^{\ell-1}(-f)q^i} L_f(y_k) = L_{q^i} L_f^{\ell}(y_k) = 0$$

so $\Gamma^{ij}_k = 0$.

Lemma 5.4 *Suppose the nonlinear system has strict observability indices $\ell_1, \ldots, \ell_p$ and $\ell = \min\{\ell_1, \ldots, \ell_p\}$. Define vector fields q by (159) and symbols Γ^{ij}_k by (160). Let $\bar{y} = \bar{y}(y)$ be a change of coordinates among those outputs of lowest observability index, i.e.*

$$\frac{\partial \bar{y}_i}{\partial y_j} = \delta^j_i \quad \text{if } \ell_i \text{ or } \ell_j > \ell. \tag{180}$$

Define $\bar{q}$ by (137) and $\bar{\Gamma}^{ij}_k$ by (161) then Γ^{ij}_k and $\bar{\Gamma}^{ij}_k$ are related as Christoffel symbols (162).

The proof of this is similar to that of Lemma 5.2, see [15].

Lemma 5.5 *Suppose the nonlinear system (72,73) has strict observability indices $\ell_1, \ldots, \ell_p$ and $\ell = \min\{\ell_1, \ldots, \ell_p\}$. If (72,73) admits an observer form (132,133) then the $\Gamma^{ij}_k(\xi)$ defined by (160) are functions only of y and define a flat connection on the output space.*

Proof By Lemma 5.6 we know that $\Gamma^{ij}_k = 0$ if $\ell_i > \ell$ or $\ell_k > \ell$. So all we need to show is the existence of an observer form for (72,73) implies the existence of a change of coordinates among those outputs of lowest observability index (180) which to transform the Γ_k^{ij} to zero for $\ell_i = \ell_j = \ell_k = \ell$. But it is clear from the proof of Lemma 5.3 that if we were to compute the Γ_k^{ij} defined by (160) for a system in observer form then they are zero.

By Lemma 5.7 the Γ_k^{ij} for $\ell_i = \ell_j = \ell_k = \ell$ transform like Christoffel symbols under a change of coordinates among those outputs of lowest observability index. By (180) the change of output coordinates to observer form $\bar{y} = \gamma^{-1}(y)$ transform the outputs of lowest observability index among themselves and can be used to take Γ_k^{ij} to zero for $\ell_i = \ell_j = \ell_k = \ell$.

If $\Gamma_k^{ij}(\xi)$ defined by (160) are the Christoffel symbols of a flat connection on the output space then we can solve the partial differential equations (155,156) to find flat coordinates $\bar{y}$. These coordinates are not necessarily the $\bar{y}$ of the observer form if it exists. But at least those of lowest observability index are because of (171). We change notation and denote the flat coordinates by y.

The next stage is to find the next smallest distinct observability index $\ell' = \min\{\ell_i > \ell\}$. We define new symbols

$$\Gamma_k^{ij} = \frac{1}{\ell'}\langle L(dy_k), [\mathrm{ad}^{\ell'_i-1}(-f)q^i, \mathrm{ad}^{\ell'_j-1}(-f)q^j]\rangle \tag{181}$$

where $\ell'_i = \ell' \wedge \ell'$.

It is not hard to see by an argument similar to Lemma 5.7 that $\Gamma_k^{ij} = 0$ if ℓ_i or $\ell_j > \ell'$. Moreover if $\ell_i = \ell_j = \ell_k = \ell$ then the Γ_k^{ij} of (181) are just ℓ/ℓ' times the Γ_k^{ij} of (160). The latter are zero by our choice of flat output coordinates.

For reasons explained below, if the system admits an observer form then Γ_k^{ij} defined by (181) define a flat connection on the output space. If this is so then we solve (155,156) for new flat output coordinates $\bar{y}$. Because of the above remarks the change of coordinates will satisfy

$$\frac{\partial \bar{y}_i}{\partial y_j} = \delta_i^j \text{ if } \ell_i = \ell' \text{ or } \ell_j > \ell'.$$

We continue on in this fashion until we have exhausted the list of observability indices or found symbols which do not define a flat connection. If the latter does not happen then the last flat coordinates $\bar{y}$ are the desired output coordinates of the observer form. The observer form will exist if (139) is satisfied for $r = 1, \ldots, \ell_i$; $s = 1, \ldots, \ell_j$ and (140) holds for $r = 2, \ldots, \ell_i$; $j = 1, \ldots, m$.

To see why this approach is valid consider a system (72), (73) which can be transformed into observer form. Using Lemmas 5.7 and 5.8 we can assume that $\bar{y}_i = y_i$ for those outputs of lowest observability index $\ell_i = \ell$. Assuming that

$u = 0$ we have

$$\begin{array}{cc} y_i = \xi_{i1} & y_i = x_{i1} \\ \dot{\xi}_{i1} = \xi_{i2} & \dot{x}_{i1} = x_{i2} - \alpha_{i\ell} \\ \vdots & \vdots \\ \dot{\xi}_{i\ell} = f_i(\xi) & \dot{x}_{i\ell} = -\alpha_{i\ell} \end{array} \tag{182}$$

By comparing these we arrive at

$$\xi_{ir} = x_{ir} - \sum_{s=1}^{r-1} (\frac{d}{dt})^{r-s-1} \alpha_{is} \tag{183}$$

and

$$f_i(\xi) = -\sum_{s=1}^{\ell} (\frac{d}{dt})^{\ell-s} \alpha_s. \tag{184}$$

We add dummy state variables ξ_{ir}, x_{ir} for $r = \ell + 1, \ldots, \ell'$ to (182) as follows

$$\begin{array}{cc} y_i = \xi_{i1} & y_i = x_{i1} \\ \dot{\xi}_{i1} = \xi_{i2} & x_{i1} = x_{i2} - \alpha_{i\ell} \\ \vdots & \vdots \\ \dot{\xi}_{i\ell} = \xi_{i\ \ell+1} + f_i(\xi) & \dot{x}_{i\ell} = x_{i\ \ell+1} - \alpha_{i\ell} \\ \dot{\xi}_{i\ \ell+1} = \xi_{i\ \ell+2} & \dot{x}_{i\ \ell+1} = x_{i\ \ell+2} \\ \vdots & \vdots \\ \dot{\xi}_{i\ell'} = 0 & \dot{x}_{i\ell'} = 0 \end{array} \tag{185}$$

It is not hard to see using (184) that these are transforms of each other under (183) and

$$\xi_{ir} = x_{ir} \quad \ell < r \leq \ell'. \tag{186}$$

Hence if the original system (182) can be transformed to observer form and $y_i = \bar{y}_i$ then so can the modified system (185). Moreover for the modified system the smallest observability index is now ℓ' rather that ℓ so we can apply Lemmas 5.7 and 5.8. It is a straightforward calculation to show that the symbols of the modified system defined by (160) with ℓ replaced by ℓ' are the same as those given by (181).

References

[1] W.A. Wolovich, *Linear Multivariable Systems*, Springer Verlag, N.Y., 1974.

[2] T. Kailath, *Linear Systems*, Prentice Hall, Englewood Cliffs, 1980.

[3] A.S. Morse, *Structural invariants of linear multivariable systems*, SIAM J. Control 11, 1973, pp.446–465.

[4] A.J. Krener and W. Respondek, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control and Opt. 23, 1985, pp.197–216.

[5] R.W. Brockett, *Feedback invariants for nonlinear systems*, IFAC, Helsinki, 2, 1978, pp.115–1120.

[6] B. Jakubczyk and W. Respondek, *On the linearization of control systems*, Bull. Acad. Polon. Sci., Ser. Math. Astro. Phys. 28, 1980, pp.517–522.

[7] R. Sommer, *Control design for multivariable non-linear time-varying systems*, Int. J. Control 31, 1980, pp.883–891.

[8] L.R. Hunt and R. Su, *Linear equivalents of nonlinear time varying systems*, Int. Symp. Math. Theory, Network and Systems, Santa Monica, 1981, pp.119–123.

[9] A.A.Zhevnin and A.P. Krishchenko, *Controllability of nonlinear systems and synthesis of control algorithms*, Soc. Phys. Dokl. 26–6, 1981, pp.559–561.

[10] A.J. Krener and A. Isidori, *Linearization by output injection and nonlinear observers*, Sys. and Control Letter, 3, 1983, pp.47–52.

[11] D. Bestle and M. Zeitz, *Canonical form observer design for nonlinear time variable system*, Int. J. Control, 38, 1983, pp.419–431.

[12] H. Keller, *Nonlinear observer design via two canonical forms*, Univ. Karlsruhe, preprint.

[13] H. Keller, *Nonlinear observer design by transformation into a generalized observer canonical form*, Univ. Karlsruhe, preprint.

[14] H. Fritz and H. Keller, *Design of nonlinear observers by a two-step transformation*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel (eds.). D. Reidel, Dordrehct, to appear.

[15] A.J. Krener, *The intrinsic geometry of nonlinear observations*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel (eds.). D. Reidel, Dordrehct, to appear.

[16] H. Poincare, Oeuvres, Tome 1 (Gauthier-Villars, Paris 1928).

[17] R. Hermann, *The formal linearization of a semisimple Lie algebra of vector fields about a singular point*, Trans, Amer. Math Soc. 130, 1968, pp.105–109.

[18] V.W. Guillemin and S. Sternberg. Remarks on a paper of Hermann, Trans. Amer. Math Soc. 130, 1968, pp.110–116.

[19] J.L. Sedwick and D.L. Elliott, *Linearization of analytic vector fields in the transitive case*, J. Diff. Eq. 26, 1977, pp. 370–390.

[20] A.J. Krener, *On the equivalence of control systems and the linearization of nonlinear systems*, SIAM J. Control 11, 1973, pp. 670–676.

[21] W. Respondek, *Linearization, feedback and Lie brackets*, Proc. of Conf. on the Geometric Theory of Nonlinear Control Systems, Wroclaw Technical University Press, Wroclaw, 1985.

[22] M. Zeitz, *Canonical forms for nonlinear systems*, Proc. of Conf. on the Geometric Theory of Nonlinear Control Systems, Wroclaw Technical University Press, Wroclaw, 1985.

Received Spring of 1986
University of California, Davis

HIGHER ORDER LINEAR APPROXIMATIONS TO NONLINEAR CONTROL SYSTEMS*

Arthur J. KRENER[†], Sinan KARAHAN[††], Mont HUBBARD[††], Ruggero FREZZA[†]

[†] Department of Mathematics

[††] Department of Mechanical Engineering

University of California
Davis, CA 95616

Abstract: One traditional approach in the analysis and design of nonlinear control systems is a first order approximation by a linear system. A new approach is to use nonlinear change of coordinates and feedback to construct linear approximations that are accurate to second and higher orders. However, the algebraic calculations required to obtain these aproximations are somewhat lengthy. In this paper, the theoretical framework for finding such change of coordinates for a nonlinear system are described. A software package that symbolically solves these transformations is currently being prepared.

1. Introduction

There is no general method for dealing with all nonlinear systems because nonlinear differential equations are virtually devoid of a general method of attack. A well-known and straightforward way to analyze nonlinear control systems is to obtain a linear approximation of the plant dynamics around a nominal operating point and design a feedback law for the resulting linear system. If the nonlinearities are strong, this approximation is valid for only a limited range of the operating regime, and performance degradation or loss of stability of the control system may occur as the system moves away from the nominal point. Then it may be necessary to repeat the linearization and design a new controller for the updated linear representation. This process is repeated as often as necessary, as dictated by the nonlinearities in the plant.

Another approach is to feed some nonlinear correction terms into the linearized plant model to compensate for the inaccuracies involved in the approximations. However, it is usually not straightforward to find such correction terms. Poincaré's theory of normal forms produces a fruitful technique for transforming a nonlinear vector field to a simpler form in the neighborhood of an equilibrium point. Another method employed in robotics is the cancellation of all the nonlinear terms by feedback. Alternatively, with the method of linearizing transformations one seeks a change of coordinates and state feedback to transform the nonlinear system into a linear one. Various forms of this question have been addressed by Brockett [8], Hunt and Su [3], Jakubczyk and Respondek[4], Sommer [9], and Krener [1,5]. The concepts on which the present paper is based, and the necessary and sufficient conditions for the existence of a solution, have been treated by Krener in [2].

The method proposed is to find a nonlinear change of coordinates for a nonlinear system to construct a linear approximation of the plant dynamics accurate to second or higher order. Based on these more accurate approximations one should be able to design controllers that give improved performance over a wider range of operating conditions. The computations required to calculate these transformations are somewhat complicated. As suggested in [2], this difficulty may be overcome with the aid of a symbolic algebraic computation package. The goal of this paper is to describe the theoretical framework for finding the required transformations.

2. Linearizing Transformations

Let us consider a nonlinear system in which the control u enters the dynamics in a linear fashion:

$$\dot{x} = f(x) + g(x)u \tag{1a}$$

$$x(0) = x^{*}. \tag{1b}$$

where $x \in \Re^n$ and $u \in \Re^m$. The system is assumed to be at rest at the nominal operating point (x^*; $u^* = 0$). For brevity of the expressions we will assume $x^* = 0$. The calculations can be easily extended to the case $x^* \neq 0$. First, consider the linearization of (1) at x^*:

$$\dot{x} = Ax + Bu \tag{2a}$$

$$A = \frac{\partial f}{\partial x}(0), \quad B = g(0). \tag{2b}$$

We will seek a coordinate change for (1) of the form identity plus

* Research supported by the Air Force Office of Scientific Research under Grant No. 85-0267.

higher order terms, such that the resulting linear plant will agree with (1) up to an error of order $O(x,u)^{p+1}$ (i.e. terms of $O(x)^{p+1}$ and $O(x)^{p}u$) where p is the degree of approximation. Obviously, Eqn. (2) results when p = 1. In the following, the case for p = 2 will be derived and the results will be generalized to any arbitrary order p by induction.

We assume a transformation of the form:

$$x = z + \phi^{(2)}(z) \tag{3}$$

where z denotes the new set of coordinates. $\phi^{(2)}$ is a polynomial of degree 2, the monomial coefficients of which are to be found.

The time derivative of (3) yields:

$$\dot{x} = \dot{z} + \frac{\partial \phi^{(2)}(z)}{\partial z}\dot{z} \tag{4}$$

We solve for the differential equation in the new coordinates z:

$$\dot{z} = (\, I + \frac{\partial \phi^{(2)}(z)}{\partial z})^{-1}\,\dot{x} \tag{5}$$

To evaluate (5), the functions f(x) and g(x) are expanded in a Taylor series, and (3) is introduced:

$$\begin{aligned} f(x) &= f^{(1)}(x) + f^{(2)}(x) + O(x)^3 \\ &= f^{(1)}(z + \phi^{(2)}(z)) + f^{(2)}(z) + O(z)^3 \\ &= Az + A\phi^{(2)}(z) + f^{(2)}(z) + O(z)^3 \end{aligned} \tag{6}$$

$$\begin{aligned} g(x) &= g^{(0)}(x) + g^{(1)}(x) + O(x)^2 \\ &= B + g^{(1)}(z) + O(z)^2 \end{aligned} \tag{7}$$

The term $(\, I + \frac{\partial \phi^{(2)}}{\partial z})^{-1}$ in (5) is expanded in a series around z = 0 as:

$$(\, I + \frac{\partial \phi^{(2)}}{\partial z})^{-1} = (\, I - \frac{\partial \phi^{(2)}}{\partial z} + (\frac{\partial \phi^{(2)}}{\partial z})^2 - \ldots). \tag{8}$$

Then, combining (6), (7), and (8) in (5) and expanding we get:

$$\dot{z} = Az + Bu + A\phi^{(2)}(z) + f^{(2)}(z) - \frac{\partial \phi^{(2)}}{\partial z}Az + g^{(1)}(z)u \; - \frac{\partial \phi^{(2)}}{\partial z}Bu + O(z,u)^3 \tag{9}$$

Now we introduce some notation. The Lie bracket of two vector fields is another vector field defined by:

$$[f,g] = \frac{\partial g}{\partial x}f - \frac{\partial f}{\partial x}g$$

So (9) can be written as:

$$\dot{z} = Az + Bu + f^{(2)}(z) - [Az,\phi^{(2)}(z)] + g^{(1)}(z)u - [Bu,\phi^{(2)}(z)] + O(z,u)^3 \tag{10}$$

With the following choice for $\phi^{(2)}$ all the second order terms of (10) will vanish and the approximation will be accurate to second order:

$$f^{(2)}(z) = [Az,\phi^{(2)}(z)] \tag{11a}$$

$$g^{(1)}(z)u = [Bu,\phi^{(2)}(z)] \tag{11b}$$

which must hold for all constant u. Eqn. (11a) is called the *Homological Equation* [6]. A solution to (11) has to be found by using the freedom in the choice of u; we use a feedback of the following form [2]:

$$u = \alpha^{(2)}(x) + (I + \beta^{(1)}(x))v \tag{12}$$

where $\alpha^{(2)}(x)$ is an m×1 vector of order 2 polynomials, I is an m×m identity matrix, and $\beta^{(1)}(x)$ is an m×m matrix of first order terms. A new input in the linearized coordinates is designated as v. Note that v agrees with u to first order. With the introduction of this feedback, f and g of Eqn. (1) are redefined:

$$\tilde{f}(x) = f(x) + g(x)\alpha^{(2)}(x) \tag{13a}$$

$$\tilde{g}(x) = g(x) + g(x)\beta^{(1)}(x) \tag{13b}$$

The Taylor expansion of (13) yields

$$\tilde{f}(x) = Ax + B\alpha^{(2)}(x) + f^{(2)}(x) + O^3(x) \tag{14a}$$

$$\tilde{g}(x) = B + g^{(1)}(x) + B\beta^{(1)}(x) + O^2(x) \tag{14b}$$

and

$$\tilde{f}^{(2)}(x) = B\alpha^{(2)}(x) + f^{(2)}(x) \tag{15a}$$

$$\tilde{g}^{(1)}(x) = \; g^{(1)}(x) + B\beta^{(1)}(x) \tag{15b}$$

Reiterating the steps of Eqns. (3) through (11) we find:

$$\tilde{f}^{(2)}(z) = [Az,\phi^{(2)}(z)] \tag{16a}$$

$$\tilde{g}^{(1)}(z) = \; [Bu,\phi^{(2)}(z)] \tag{16b}$$

The distinction between Eqns. (11) and (16) is seen when (16) is rewritten as:

$$f^{(2)}(z) = -\,B\alpha^{(2)}(z) + [Az,\phi^{(2)}(z)] \tag{17a}$$

$$g^{(1)}(z)v = -\,B\beta^{(1)}(z)v + [Bv,\phi^{(2)}(z)] \tag{17b}$$

In the generalized homological equation of (17), the second order terms $f^{(2)}(z)$ and $g^{(1)}(z)v$ can be cancelled out under certain solvability conditions by proper choice of $\phi^{(2)}(z)$, $\alpha^{(2)}(z)$, and $\beta^{(1)}(z)$. For this second order linearization we have a system of $n^2(n + 1)/2 + n^2m$ linear algebraic equations in $n^2(n + 1)/2 + mn(n + 1)/2 + m^2n$ unknowns. When a solution can be found, the resulting system becomes:

$$\dot{z} = Az + Bv + O(z,v)^3 \tag{18}$$

In order to find an approximation of the next higher order, we rewrite (18) by reverting to the variables x and u:

$$\dot{x} = Ax + Bu + O(x,u)^3 \tag*{(18)'}$$

Now we are asuming that in the given nonlinear system the second order terms have been already been canceled as outlined above. Then we seek a new transformation of the form:

$$x = z + \phi^{(3)}(z) \tag{19}$$

Note that transformation (19) will not introduce any terms of degree less than 3. Then the same procedure outlined above is repeated, with the feedback:

$$u = \alpha^{(3)}(x) + (I + \beta^{(2)}(x))v \tag{20}$$

which results in:

$$f^{(3)}(z) = -B\alpha^{(3)}(z) + [Az,\phi^{(3)}(z)] \tag{21a}$$

$$g^{(2)}(z)v = -B\beta^{(2)}(z)v + [Bv,\phi^{(3)}(z)] \tag{21b}$$

These results can be generalized as follows. Given a system which is accurate to only order p-1, i. e.

$$\dot{x} = Ax + Bu + O(x,u)^p \tag{22}$$

a coordinate change is sought as:

$$x = z + \phi^{(p)}(z) \tag{23}$$

along with feedback:

$$u = \alpha^{(p)}(x) + (I + \beta^{(p-1)}(x))v \tag{24}$$

which yields the homological equation to be solved:

$$f^{(p)}(z) = -B\alpha^{(p)}(z) + [Az,\phi^{(p)}(z)] \tag{25a}$$

$$g^{(p-1)}(z)v = -B\beta^{(p-1)}(z)v + [Bv,\phi^{(p)}(z)] \tag{25b}$$

In (25), $\phi^{(p)}$, $f^{(p)}$, $\alpha^{(p)}$, $g^{(p-1)}$ and $\beta^{(p-1)}$ are, respectively, homogeneous vector fields of orders corresponding to their superscripts. The resulting system is accurate up to order p:

$$\dot{z} = Az + Bv + O(z,v)^{p+1} \tag{26}$$

3. Linearizing Transformations for Systems with Small Parameters

In this section, we consider a control system of the form:

$$\dot{x} = f(x,\varepsilon) + g(x,\varepsilon)u \tag{27a}$$

$$x(0) = x^*. \tag{27b}$$

where ε is a small parameter that characterizes the way parasitic effects or disturbances enter into the system. We will develop a method of linearizing transformation for this type of system, similar to that of Section 2. First, (27) is expanded as follows:

$$\dot{x} = Ax + Bu + \varepsilon(f^{(1)}(x) + g^{(1)}(x)u) + O(\varepsilon)^2 \tag{28}$$

In (28), the nonlinear function is expanded and grouped in powers of ε. Thus, the superscripts of f and g now correspond to the powers of ε these functions multiply, in contrast with the notation of Section 2. A coordinate change is assumed of the following form:

$$x = z + \varepsilon\phi^{(1)}(z) \tag{29}$$

where the form and the polynomial order of the function $\phi^{(1)}(z)$ is not determined yet. Repeating the calculations similar to the steps of Eqns. (4) through (9) of Section 2 yields:

$$\dot{z} = Az + Bu + \varepsilon(f^{(1)}(z) - [Az,\phi^{(1)}(z)] + g^{(1)}(z)u - [Bu,\phi^{(1)}(z)]) + O(\varepsilon)^2 \tag{30}$$

An input for the control system of Eqn. (28) is chosen as:

$$u = v + \varepsilon(\alpha^{(1)}(x) + \beta^{(1)}(x)v) \tag{31}$$

After a sequence of calculations similar to Eqns. (13) through (17), the homological equations are found:

$$f^{(1)}(z) = -B\alpha^{(1)}(z) + [Az,\phi^{(1)}(z)] \tag{32a}$$

$$g^{(1)}(z)v = -B\beta^{(1)}(z)v + [Bv,\phi^{(1)}(z)] \tag{32b}$$

This result can be generalized for an arbitrary power of ε in the same fashion: A solution to

$$f^{(p)}(z) = -B\alpha^{(p)}(z) + [Az,\phi^{(p)}(z)] \tag{33a}$$

$$g^{(p)}(z)v = -B\beta^{(p)}(z)v + [Bv,\phi^{(p)}(z)] \tag{33b}$$

will yield

$$\dot{z} = Az + Bv + O(\varepsilon)^{p+1} \tag{34}$$

Even though Eqns. (33) and (25) look very similar, there are some fundamental differences. All the variables in Eqn. (33) have different definitions than those of Eqn. (25), as mentioned at the beginning of this section. Moreover, the solvability conditions of (33) are not the same as the conditions of Eqn. (25). Actually, both (32) and (33) may represent an infinite family of equations as opposed to the finite dimensional set of expressions that arise from (25).

Any nonlinear system expressed in the form of in Eqn. (1) can always be transformed into the form of (27) as follows: First, consider the expansion of (1) as

$$\dot{x} = Ax + f^{(2)}(x) + Bu + g^{(1)}(x)u + O(x,u)^3 \tag{35}$$

Scale the coordinates and the input with:

$$\xi = \varepsilon^{-1}x$$

$$\mu = \varepsilon^{-1}u$$

introducing the above into (35) yields

$$\dot{\xi} = A\xi + B\mu + \varepsilon(\,\bar{f}^{(2)}(\xi) + \bar{g}^{(1)}(\xi)\mu\,) + O(\varepsilon)^2 \tag{36}$$

This equation is of the form of Eqn. (28), except for the difference in the way expansions of f and g are defined. We use the overbar notation to emphasize this point. The input

$$u = \alpha^{(2)}(x) + (I + \beta^{(1)}(x))v \tag{12}$$

is also transformed with an additional scaling $\eta = \varepsilon^{-1}v$:

$$\mu = \eta + \varepsilon(\bar{\alpha}^{(2)}(\xi) + \bar{\beta}^{(1)}(\xi)\eta) \tag{37}$$

With this scaling of coordinates, a linearization problem given as in Section 2 can be alternatively solved with the procedure outlined in this section.

4. Form of the Nonlinear Compensation

After a higher order linearization is obtained, the next step is to choose a feedback law to achieve closed-loop pole assignment. Consider the approximation of Section 2 where

$$\dot{x} = Ax + Bu + O(x,u)^p \tag{22}$$

has been transformed by the coordinate change

$$x = z + \phi^{(p)}(z) \tag{23}$$

and feedback

$$u = \alpha^{(p)}(x) + (I + \beta^{(p-1)}(x))v \tag{24}$$

into

$$\dot{z} = Az + Bv + O(z,v)^{p+1} \tag{26}$$

A closed-loop pole assignment can be made with state feedback of the form

$$v = Fz + r \tag{38}$$

where r is an open loop control. The approximation of (26) then becomes

$$\dot{z} = (A + BF)z + Br. \tag{39}$$

Notice that (23) agrees with the identity transformation up to order p–1, so it is easily inverted at least up to order p:

$$z = x - \phi^{(p)}(x) \tag{40}$$

The input u of Eqn. (24) in the original coordinate becomes, with the aid of (22), (38), and (40):

$$\begin{aligned} u &= \alpha^{(p)}(x) + (I + \beta^{(p-1)}(x))(F(x - \phi^{(p)}(x)) + r) \\ &= Fx + r + \{\alpha^{(p)}(x) + \beta^{(p-1)}(x)(Fx - F\phi^{(p)}(x) + r\,) \\ &\quad - F\phi^{(p)}(x)\}. \end{aligned} \tag{41}$$

Thus the control function has the form of a pole assignment for the linear part of (22) plus some correction terms of higher order (grouped in the bracket of Eqn. (41)). This result clearly shows the *purpose and nature* of the nonlinear feedback.

6. Conclusion

In this paper we have presented an alternative approach to the analysis and design of nonlinear control systems. The procedure consists of finding a coordinate change by an appropriate feedback to achieve higher order linear approximations to nonlinear systems. Because of space limitations, we have not presented the details of the solvability conditions. The method of solving for the linearizing transformations is based on the normal forms approach of Poincaré, which is a widely used technique in the analysis of bifurcations in nonlinear vector fields. This suggests the applicability of these powerful bifurcation methods in nonlinear control systems analysis. Aeyels [10] and Abed and Fu [11] have studied the local stabilization problem for nonlinear systems with this approach. In other words, the method is an appropriate tool for the analysis of nonlinear systems in which plant parameter variations cause fundamental changes in the structure of the system. Another important issue is the following: When a solution exists, the functions $\alpha^{(p)}(x)$, $\beta^{(p-1)}(x)$, $\phi^{(p)}(x)$ are not necessarily unique. The question of what is the best choice, or even what is a reasonable choice among the possible solutions needs more investigation.

The equations that need to be solved for finding the transformations are a set of linear algebraic equations. However, the number of equations grow rapidly with increasing orders of linearization and with higher dimensional systems.For example, for a second order linearization and with n states and m inputs we have a system of $n^2(n + 1)/2 + n^2 m$ linear algebraic equations in $n^2(n + 1)/2 + mn(n + 1)/2 + m^2 n$ unknowns. With the use of symbolic algebraic manipulation packages and with the availibility of increasingly powerful computers, this is not considered as a serious setback. A symbolic algebra program that automatically solves these transformations on the computer is in preparation.

References

[1] Krener, A. J., "On the Equivalence of Control Systems and the Linearization of Nonlinear Systems", *SIAM J. Control* **11** (1973) 670–676.

[2] Krener, A. J., "Approximate Linearization by State Feedback and Coordinate Change", *Systems and Control Letters* 5 (1984) 181–185.

[3] Hunt, L. R. and R. Su, "Linear Equivalents of Nonlinear Time Varying Systems", *Proceedings MTNS Symposium*, Santa Monica (1981) 119–123.

[4] Jakubczyk, B. and W. Respondek, "On the Linearization of Control Systems", *Bull. Acad. Polon. Sci. Ser. Math. Astron. Physics* 28 (1980) 517–522.

[5] Krener, A. J., "New Approaches to the Design of Nonlinear Compensators", *Proceedings of the Berkeley Ames Conf. on Nonlinear Problems, Aerodynamics and Flight Control* (Math. Sci. Press, 1984).

[6] Arnold, V. I., *Geometric Methods in the Theory of Ordinary Differential Equations* (Springer Verlag, New York, 1983).

[7] Guckenheimer, J. and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer Verlag, New York, 1983).

[8] Brockett, R. W., "Feedback Invariants for Nonlinear Systems" *Proceedings, IFAC Congress*, Helsinki (1978).

[9] Sommer, R., "Control Design for Multivariable Non-linear Time-varying Systems", *Int. J. Control* 31 (1980) 883–891.

[10] Aeyels, D., "Stabilization of a Class of Nonlinear Systems by a Smooth Feedback Control", *Systems and Control Letters* 5 (1985) 289-294.

[11] Abed, E. H., and J. H. Fu, "Local Feedback Stabilization and Bifurcation Control, I. Hopf Bifurcation", *Systems and Control Letters* 7 (1986) 11–17.

AFOSR·TR· 89 - 1689

Determining Torque and Velocity Limits on Joint Actuators for Robot Manipulators with Coupled Joint Motion*

Sinan Karahan
Department of Mechanical Engineering
University of California, Davis, CA 95616

Abstract: *In robots with remotely driven links, the relationships between some of the actuator and joint motions are coupled. In this paper robot dynamic equations for these types of manipulators are analyzed. The equations are expressed in actuator coordinates by means of the coupling matrix between joint and actuator coordinates. The motivation is that, in the actuator coordinates, the actuator velocity and torque limits can be calculated. The procedure is demonstrated by an example.*

Keywords: Robot manipulator, torque limit, coupled motion, robot dynamics.

1. Introduction

In the formulation of manipulator dynamic equations, two well-known methods, Newton-Euler [4] and Lagrangian [1,5] have been the most popular. While the Newton-Euler method is computationally more efficient, it is the Lagrangian method that allows better insight into the the dynamics, as well as into the analysis and design of control systems for manipulators. In the Lagrangian formulation, the Denavit-Hartenberg convention [2] is often used to assign link parameters. Within this framework, the joint coordinates are measured with respect to the next lower link in the open kinematic chain, regardless of the mechanical characteristics of the robot. While for the more recent direct drive robot designs there is a one-to-one relationship between joint and actuator coordinates except for the gear ratio conversions (thus the name "direct drive"), conventional type manipulators generally have remotely driven links due to limitations on cost, weight, and other design criteria. In such robots, the relationships between some of the actuator and joint motions are coupled, with an actuator rotation or displacement resulting in the motion of more than one link. Avoidance of the upper bounds on actuator torques or forces is a concern for high speed motions and optimal control schemes, because the end effector of the robot will not be able to follow the commanded trajectory if the actuator limits are exceeded. Formulating the equations for remotely driven manipulators in joint coordinates will result in complicated calculations for the upper bounds for some of the joints, thus making it difficult to check the torque/force limits at the actuator level with given motion commands at the joint level.

The goal of this paper is to reformulate the Lagrangian dynamic equations for robots in actuator coordinates, and to show that the problem of checking for the torque/force upper bounds can be simplified.

2. Manipulator Dynamic Equations and their transformations

The equations of motion for an n-link open chain manipulator arm are given in vector-matrix form as:

$$D(\theta)\ddot{\theta} + \begin{bmatrix} \dot{\theta}^t C_1 \dot{\theta} \\ \dot{\theta}^t C_2 \dot{\theta} \\ \cdot \\ \cdot \\ \dot{\theta}^t C_n \dot{\theta} \end{bmatrix} + G(\theta) = \tau \qquad (1)$$

where θ, $\dot{\theta}$, $\ddot{\theta}$ are, respectively, (nx1) joint displacement, velocity, and acceleration vectors. $D(\theta)$ represents the (nxn) configuration dependent inertia matrix. $C_1, C_2, \ldots, C_n$ are the (nxn) Coriolis- centrifugal force coefficient matrices, $G(\theta)$ is the (nx1) vector of gravitational forces acting on each joint, and τ is the vector of forces applied by the actuators. We note here that, for a manipulator with joint motion coupling, τ is composed of the forces applied to the joints, not the actuator output forces. For the purposes of the derivations to be performed later, it is preferable to rewrite (1) in the following form:

* Research supported by the Air Force Office of Scientific Research under grant number 85-0267.

$$D(\theta)\ddot{\theta} + [\,I_n \otimes \dot{\theta}^t] \begin{bmatrix} C_1 \\ C_2 \\ . \\ . \\ C_n \end{bmatrix} \dot{\theta} + G(\theta) = \tau \qquad (2)$$

In (2), the symbol $\otimes$ represents the Kronecker product of matrices, and I_n is the (nxn) identity matrix. Obviously, the matrix $[\,I_n \otimes \dot{\theta}^t]$ has dimensions (nxn^2) and the Coriolis-centrifugal coefficient matrix of C_i's is (n^2xn). We refer the reader to [3] as an excellent reference on Kronecker algebra of matrices.

Let the coupling between joint and the actuator displacements be given by:

$$\theta = H\varphi \qquad (3)$$

where φ is the (nx1) vector of actuator displacements, and H is the (nxn) coupling matrix. H is constant and non-singular. For a direct-drive robot arm, H would reduce to a diagonal matrix with the corresponding gear reduction ratios as the entries. By differentiating (3) with respect to time, we obtain:

$$\dot{\theta} = H\dot{\varphi} \qquad (4)$$

and

$$\ddot{\theta} = H\ddot{\varphi} \qquad (5)$$

On the other hand, the relationship between the actuator forces and the joint forces is:

$$\tau_a = H^t\tau \qquad (6)$$

where τ_a is the force input vector in actuator coordinates.

Introducing the relations (4) and (5) into the equations of motion (2) yields:

$$DH\ddot{\varphi} + [\,I_n \otimes \dot{\varphi}^t H^t] \begin{bmatrix} C_1 \\ C_2 \\ . \\ . \\ C_n \end{bmatrix} H\dot{\varphi} + G = \tau \qquad (7)$$

Multiplying each side of Eqn. (7) by H^t we get:

$$H^tDH\ddot{\varphi} + H^t[\,I_n \otimes \dot{\varphi}^t H^t] \begin{bmatrix} C_1 \\ C_2 \\ . \\ . \\ C_n \end{bmatrix} H\dot{\varphi} + H^tG = H^t\tau \quad (8)$$

In the above, we first note that the right hand side is the same as Eqn. (6), i.e. τ_a. By using the identity $(A\otimes B)(E\otimes F) = (AE\otimes BF)$ (this is possible whenever A, E and B, F are conformable pairs; see [3]), we rewrite the term involving the Kronecker product as:

$$H^t[\,I_n \otimes \dot{\varphi}^t H^t\,] = [\,I_n \otimes \dot{\varphi}^t\,][H^t \otimes H^t\,] \qquad (9)$$

The term $[H^t \otimes H^t]$ can also be written as $(H^t)^{[2]}$, the superscript in brackets being the notation for Kronecker square of matrices [3]. With this result, Eqn. (8) becomes:

$$D_a\ddot{\varphi} + [\,I_n \otimes \dot{\varphi}^t] \begin{bmatrix} C_{a1} \\ C_{a2} \\ . \\ . \\ C_{an} \end{bmatrix} \dot{\varphi} + G_a = \tau_a \qquad (10)$$

where the parameters corresponding to the actuator coordinates have been assigned the subscript "a". Comparing with Eqns. (8) and (9), these are:

$D_a = H^tDH$ Inertia matrix

$$\begin{bmatrix} C_{a1} \\ C_{a2} \\ . \\ . \\ C_{an} \end{bmatrix} = (H^t)^{[2]} \begin{bmatrix} C_1 \\ C_2 \\ . \\ . \\ C_n \end{bmatrix} H$$ Coriolis-centrifugal terms

$G_a = H^tG$ Gravitational forces

and the torque term is given by Eqn. (6).

3. Application to velocity and torque limit calculations

For a practical implementation, consider the problem of calculating torque commands to the actuators for a given trajectory to be followed. Let the maximum force/torque he actuators can deliver be τ_{amax}. Additionally, there is usualy a maximum velocity at which the actuators can be driven, which we denote by $\dot{\varphi}_{max}$. For a typical DC motor, the maximum torque is limited by the maximum current rating of the motor, while the maximum velocity is proportional to the voltage applied. The actuator dynamic effects are ignored in this context. We assume that the sensors placed at the actuator or joint level measure the displacement and the velocity and that the acceleration can be approximated by:

$$\ddot{\varphi}^i = \frac{\dot{\varphi}^i - \dot{\varphi}^{i-1}}{\Delta t}$$

where superscripts represent the ith and i-1st measurements, and Δt is the sampling period. We need to calculate the left-hand side of Eqn. (10) with these velocities and accelerations. This will yield the total forces/torques *required* at each joint to achieve the motion commanded.

Call this quantity A. Comparing the vector A obtained from this calculation with the actuator limits τ_{amax} term by term, which are the maximum forces/torques that are *available*, it is possible to determine whether any of these limits are exceeded. Whenever a torque or velocity limit is exceeded, one needs to renormalize the actuator commands in the following manner:

$$N_1 = \mathrm{Max}\left\{\frac{A_i}{\tau_{amax\,i}}\right\} \qquad i = 1,\dots,n$$

$$N_2 = \mathrm{Max}\left\{\frac{\dot{\varphi}_i}{\dot{\varphi}_{max\,i}}\right\} \qquad i=1,\dots,n$$

$$N = \mathrm{Max}\{N_1, N_2\}$$

and if N > 1, in which case either a torque or velocity limit is exceeded, then set

$$\dot{\varphi}_{i_{new}} = \frac{\dot{\varphi}_i}{N} \qquad i=1,\dots,n$$

This will renormalize all the joint velocity commands with the same ratio, thus ensuring that the manipulator end effector stays on the commanded path, even though at a slower velocity. Using such an algorithm, if a limit is reached, usually only one of the actuators will be at its maximum.

4. An Example

In this section, the calculations of the joint-actuator coordinate transformations will be demonstrated with a simple example. Consider the 3 d-o-f revolute manipulator of Fig.1:

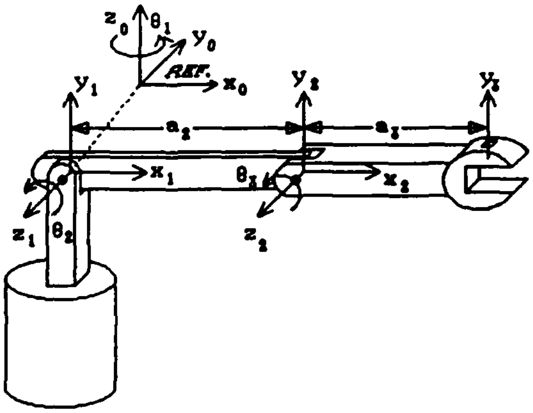


Figure 1. A 3 d-o-f revolute manipulator.

The coordinate system on the robot and link parameters are assigned in accordance with the Denavit-Hartenberg notation. It is assumed that joint 3 is driven remotely from the base, and that the motion is transmitted through a mechanical connection. For this robot, Eqn. (3) becomes:

$$\begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\end{bmatrix} = \begin{bmatrix}\frac{1}{k_1} & 0 & 0\\ 0 & \frac{1}{k_2} & 0\\ 0 & -\frac{1}{k_2} & \frac{1}{k_3}\end{bmatrix}\begin{bmatrix}\varphi_1\\ \varphi_2\\ \varphi_3\end{bmatrix} \qquad (11)$$

In the above, k_1, k_2, k_3, are the gear transmission ratios between the actuators and the joints. By taking the time derivative of (11) the counterparts of Eqns. (4) and (5) are also obtained. For the purposes of the foregoing calculations the explicit forms of the inertia matrix elements and the other terms in Eqn. (2) will not be necessary. Therefore, the inertia matrix is expressed in its symbolic form:

$$D = \begin{bmatrix}D_{11} & D_{12} & D_{13}\\ D_{12} & D_{22} & D_{23}\\ D_{13} & D_{23} & D_{33}\end{bmatrix}$$

Calculation of the inertia matrix in actuator coordinates yields:

$$D_a = H^tDH = \begin{bmatrix}\frac{D_{11}}{k_1^2} & \frac{D_{12}-D_{13}}{k_1k_2} & \frac{D_{13}}{k_1k_3}\\ \frac{D_{12}-D_{13}}{k_1k_2} & \frac{D_{22}-2D_{23}+D_{33}}{k_2^2} & \frac{D_{23}-D_{33}}{k_2k_3}\\ \frac{D_{13}}{k_1k_3} & \frac{D_{23}-D_{33}}{k_2k_3} & \frac{D_{33}}{k_3^2}\end{bmatrix}$$

Similarly, the Coriolis-centrifugal terms are calculated as:

$$\begin{bmatrix}C_{a1}\\ C_{a2}\\ C_{a3}\end{bmatrix} = [H^t \otimes H^t]\begin{bmatrix}C_1\\ C_2\\ C_3\end{bmatrix}H = \begin{bmatrix}\frac{H^tC_1H}{k_1}\\ \frac{H^t(C_2-C_3)H}{k_2}\\ \frac{H^tC_3H}{k_3}\end{bmatrix}$$

Note that in the above C_i are (n x n) matrices of Coriolis-centrifugal force coefficients. The gravitational terms become:

$$\begin{bmatrix}G_{a1}\\ G_{a2}\\ G_{a3}\end{bmatrix} = H^t\begin{bmatrix}G_1\\ G_2\\ G_3\end{bmatrix} = \begin{bmatrix}\frac{G_1}{k_1}\\ \frac{G_2-G_3}{k_2}\\ \frac{G_3}{k_3}\end{bmatrix}$$

And, finally, the torque terms are transformed as:

$$\begin{bmatrix} \tau_{a1} \\ \tau_{a2} \\ \tau_{a3} \end{bmatrix} = H^{t} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} = \begin{bmatrix} \frac{\tau_1}{k_1} \\ \frac{\tau_2 - \tau_3}{k_2} \\ \frac{\tau_3}{k_3} \end{bmatrix}$$

5. Conclusion

In this paper, manipulator dynamics equations formulated in joint coordinates were transformed into actuator coordinates by use of the actuator-joint coordinate coupling matrix. This resulted in the expressions for the transformation of inertia and Coriolis-centrifugal terms, as presented in Section 2. Then, the final form of the equations allows an easier implementation of an algorithm that checks the velocity and torque limits, and renormalizes the actuator commands so as to prevent tracking errors.

Acknowledgements

This work was partially supported by the Air Force Office of Scientific Research under grant number 85-0267. Some results were developed during the author's stay at ATHTEC Systems, Inc. of Sacramento, California. The algorithm described in Section 3 has been suggested by the author's colleague Scott Maxwell, now with the Mechanical Engineering Department of M.I.T.

References

[1] Hollerbach, J., "Dynamics", *Robot Motion: Planning and Control*, eds. M. Brady, J. Hollerbach, T. Lozano-Perez, and M. T. Mason. MIT Press, Cambridge, Mass. 1982, pp. # 51-73.

[2] Denavit, J. and S. Hartenberg, "A Kinematic Notation for Lower-pair Mechanisms Based on Matrices", J. Applied Mechanics, 77(2), 1955, pp. # 215-22.

[3] Brewer, J., "Kronecker Products and Matrix Calculus in System Theory", IEEE Transactions on Circuits and Systems, 25(9), 1978, pp. # 772-781.

[4] Luh, J. Y. S., M. W. Walker, and R. P. C. Paul, "On-line Computational Scheme for Mechanical Manipulators", ASME Journal of Dynamic Systems, Measurement, and Control, 102(2), June 1980, pp. # 69-76.

[5] Paul, R. P. *Robot Manipulators: Mathematics, Programming, and Control.* MIT Press, Cambridge, Mass. 1981.

Application of An Efficient Nonlinear Filter*

Ruggero FREZZA , Sinan KARAHAN, Arthur J. KRENER, Mont HUBBARD

Abstract- *In this paper we present an application of a new filtering technique based on geometric linearization and asymptotic analysis. The technique is compared to the conventional extended Kalman filter to demonstrate its computational efficiency.*

Introduction

While the theory of linear filtering has been well developed and understood, practical nonlinear filtering has typically relied on heuristic techniques. One of these is the extended Kalman filter which is based on a linear approximation of the system equations around a trajectory. The asymptotic geometric nonlinear filtering technique was developed by Krener in [5]. It is based on the so-called observer normal form which linearizes the dynamics by a change of state coordinates and output injection. The development of the observer normal form is due to Krener and Isidori [1], Krener and Respondek [2], Bestle and Zeitz [8], Zeitz [11,12], Fritz and Keller [9], Keller [6,7], and Li and Tao [10].

The approach is geometric and consists of finding a change of coordinates, in general nonlinear, such that the equations of the system transform to their most linear form. Once the system is in "nearly linear" form it will be possible to apply asymptotically the theory of linear filtering. This has many advantages. In fact it is possible to formally define optimality and asymptotic stability. Moreover, the gains of the filter may be computed off-line because the Riccati differential equation is independent of the states. This reduces the on-line computational burden of the filter.

The only drawback of the technique is that the change of coordinates generally requires a very heavy algebraic computational effort. One solution to this is to use already existing software packages for symbolic computations. Naturally the technique is not applicable to all nonlinear systems; otherwise we would have discovered that everything in nature is linear in some appropriate coordinate set. The class of "nearly linearizable" systems is substantial, and gives to

Research supported by the Air Force Office of Scientific Research under 85-0267.

generally requires a very heavy algebraic computational effort. One solution to this is to use already existing software packages for symbolic computations. Naturally the technique is not applicable to all nonlinear systems; otherwise we would have discovered that everything in nature is linear in some appropriate coordinate set. The class of "nearly linearizable" systems is substantial, and gives to the method a certain flavor of generality.

In this paper we will illustrate the technique for a specific example. We will estimate the height, the velocity and the ballistic coefficient of a falling object in an atmosphere with variable density. We will also implement the extended Kalman filter for the same problem and compare the two filtering techniques in terms of performance and computation time. Throughout the paper we will refer to the geometric asymptotic nonlinear filter, the new technique, as the GANF and to the extended Kalman filter as the EKF.

1- Brief description of the EKF and the GANF techniques

We say that a given nonlinear system without input:

$$\begin{aligned} \dot{\zeta} &= f(\zeta) \, , \\ y &= h(\zeta) \, . \end{aligned} \qquad (1)$$

where $\zeta \, \varepsilon \, \mathbf{R}^n$ and $y \, \varepsilon \, \mathbf{R}^1$, is observable if it can be transformed into observable normal form. This corresponds, in some sense, to the property of observability for a linear system. Normal forms have the advantage of making transparent the effect of an input on the dynamics of the system. There are four such normal forms: Observer, observable, controller and controllable. In the present application, and in the case of nonlinear observers in general, we will have occasion to use only the first two:

Observable form:

$$\dot{x} = Ax - B\alpha(x)$$
$$y = Cx \qquad (2)$$

Observer form:

$$\dot{x} = Ax - \alpha(Cx)$$
$$y = \gamma(Cx) \qquad (3)$$

In some particular cases (in the application treated in this paper, for example) we will require a modified observer form.

Modified observer form:

$$\dot{x} = Ax - \alpha(Cx, CAx, ..., CA^{n-1}x)$$
$$y = \gamma(Cx) \qquad (4)$$

where A, B, C are the standard matrices of the Brunovsky canonical form. For the case of a three dimensional system with a single output:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \qquad (5)$$

The system (1) can be transformed into observable form if y and its first n-1 time derivatives are local coordinates on the state space.

A physical system will never satisfy exactly a set of differential equations like (1). In reality each of the states is affected by random process noise, the parameters are not known exactly, and the measurement of the output will be affected by observation noise due to the physical and technical limitations on the measurement procedure. The system (1) can then be expressed in its stochastic equivalent:

$$\begin{aligned} d\zeta &= f(\zeta)dt + Bdw \\ dy &= h(\zeta)dt + Ddv \\ \zeta(0) &= N(\zeta_0, P_0) \end{aligned} \tag{6}$$

Where w and v are standard Wiener processes. The covariances of the driving and measurement noise are, respectively, $Q = BB^T$ and $R = D^2$. For our work we will assume the measurement noise to be small, i.e. $D = \varepsilon$ where ε is a small parameter. Then the problem is to estimate the states of (6) at time t given the measurement of y at time less than or equal to t. It is clear that in the absence of noise and if (1) is observable the problem is easily solved because the states could be computed exactly as nonlinear function of derivatives up to order n - 1 of the output.

We compute the estimates by introducing the filter:

$$\begin{aligned} d\hat{\zeta} &= f(\hat{\zeta})dt + g(\hat{\zeta})(dy - h(\hat{\zeta})dt) \\ d\hat{y} &= h(\hat{\zeta})dt \end{aligned} \tag{7}$$

The conventional extended Kalman filter technique computes the estimates from:

$$\begin{aligned} d\hat{\zeta} &= f(\hat{\zeta})dt + P(t)H^T(\hat{\zeta}, t)R^{-1}(dy - h(\hat{\zeta})dt) \\ d\hat{y} &= h(\hat{\zeta})dt \end{aligned} \tag{8}$$

where

$$\dot{P} = F(\hat{\zeta}, t)P + PF^T(\hat{\zeta}, t) + Q(t) - PH^T(\hat{\zeta}, t)R^{-1}H(\hat{\zeta}, t)P$$
$$P(t_0) = P_0 \tag{9}$$

and

$$H(\hat{\zeta}, t) = \left\{ \frac{\partial h_i(\zeta, t)}{\partial \zeta_j} \bigg|_{\zeta = \hat{\zeta}} \right\} \tag{10}$$

$$F(\hat{\zeta}, t) = \left\{ \frac{\partial f_i(\zeta, t)}{\partial \zeta_j} \bigg|_{\zeta = \hat{\zeta}} \right\} \tag{11}$$

and P, Q and R are, respectively, the covariance matrices of the estimate error, the process noise and the observation noise.

The GANF computes the filter in observer form coordinates. In these coordinates the system (6) can be written as:

$$dx = (Ax + \alpha(y))dt + Gdw$$
$$dy = Cxdt + Ddv \tag{12}$$
$$x(0) = N(x_0, \bar{P}_0)$$

where, if J is the Jacobian of the change of coordinates, G = J B. If the measurement noise is small, and $\alpha(y)$ is smooth enough, then $\alpha(y)$ is approximately equal to $\alpha(C\hat{x})$ and we can use the filter equations

$$d\hat{x} = A\hat{x}dt + \alpha(C\hat{x})dt + PC^TR^{-1}(d\bar{y} - C\hat{x}\,dt\,) \tag{13}$$

with $\bar{P}$ being obtained from the solution to the Riccati differential equation

$$\dot{\bar{P}} = A\bar{P} + \bar{P}A^T + \bar{Q} - \bar{P}C^TR^{-1}C\bar{P}$$

$$\bar{P}(0) = \bar{P}_0 \tag{14}$$

Where $\bar{Q} = GG^T$. The dynamics of the error are given by:

$$d\tilde{x} = (A + PC^TR^{-1}C)\,\tilde{x}dt + (\alpha(y) - \alpha(C\hat{x}))dt + Gdw + PC^TR^{-1}C\varepsilon dv \tag{15}$$

The Riccati equation (14) is state independent and hence can be integrated off-line. The dynamics of the error are nearly linear up to output injection, and the covariance of $\tilde{x}$ asymptotically equals $\bar{P}(t)$. We can assume, without loss of generality, that the output injection term can be expanded in a Taylor series starting from the second order term:

$$\alpha(y) - \alpha(C\hat{x}) = \frac{d^2\alpha(C\hat{x})}{dy^2}(C\tilde{x} + \varepsilon dv)^2 + O(3) \tag{16}$$

In fact, we can combine any linear term of the above in the $(A + PC^TR^{-1}C)\tilde{x}$ term of (15). Then, if the second derivative of α is small, the output injection term can be neglected. Practically, we are requiring α to have a small "curvature" in some sense.

One should note that these two filters are not being compared on exactly the same grounds because the noise covariances of the EKF and the GANF are state independent in both ζ and x coordinates. However the comparison is justified if the Jacobian of the change of coordinates J is nearly constant along the trajectory since $\bar{Q} = JQJ^T$ and R is invariant under the change of coordinates. The main advantage of the GANF over the EKF is that the

Riccati equation (14) can be solved off-line. This allows the computation of $g(\zeta)$ off-line, while when using the EKF it is necessary to integrate (9) and compute (10) and (11) on-line. In particular if the problem can be transformed into observer form with $\alpha(y) = 0$ then the GANF is the optimal filter while the extended Kalman filter is generally not.

2- An Application of GANF

We consider a falling object in an atmosphere of varying density. The problem is to estimate the position, the velocity and the ballistic coefficient of the object. This problem has been discussed previously by Gelb [3] and Wishner et al [4].

The dynamic model consists of:

$$\begin{aligned} y &= \zeta_1 \\ \dot{\zeta}_1 &= \zeta_2 \\ \dot{\zeta}_2 &= -g + \rho(\zeta_1)\zeta_2^2\zeta_3 \\ \dot{\zeta}_3 &= 0. \end{aligned} \tag{17}$$

$$\zeta_1(0) = \zeta_1^0$$

$$\zeta_2(0) = \zeta_2^0$$

$$\zeta_3(0) = \zeta_3^0$$

where ζ_1 is the vertical position, ζ_2 the velocity, ζ_3 the inverse of the ballistic coefficient and

$$\rho(\zeta_1) = \rho_0 \exp(-\frac{\zeta_1}{k_\rho}) \tag{18}$$

represents the variation of the atmospheric density with the height ζ_1. It can be verified that this system is observable according to the observability condition defined in section 1. In fact

$$dh = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

$$dL_f h = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

$$dL_f^2 h = \begin{bmatrix} -\rho(\zeta_1)\frac{\zeta_2^2}{k_\rho}\zeta_3 & \rho(\zeta_1)2\zeta_2\zeta_3 & \rho(\zeta_1)\zeta_2^2 \end{bmatrix} \tag{19}$$

span $\mathbf{R}^3$, except at the singular point $\zeta_2 = 0$ where, the object being stationary, it is impossible to observe any of its dynamics. Hence it is possible to write (17) in Observable form:

$$y = \xi_1$$

$$\dot{\xi}_1 = \xi_2$$

$$\dot{\xi}_2 = \xi_3$$

$$\dot{\xi}_3 = f_3(\xi) = -g\frac{\xi_2}{k_\rho} + (2\frac{g}{\xi_2} - \frac{\xi_2}{k_\rho})\xi_3 + \frac{2}{\xi_2}\xi_3^2 \tag{20}$$

The Jacobian of the transformation between ζ and ξ is

$$\frac{\partial \xi}{\partial \zeta}(\zeta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\rho\frac{\zeta_2^2\zeta_3}{k_\rho} & 2\rho\zeta_2\zeta_3 & \rho\zeta_2^2 \end{bmatrix} \tag{21}$$

Unfortunately (20) does not satisfy the conditions for the transformation to observer form, but it can be transformed into the modified observer form.

If we let

$$\alpha_1(\xi_2) = -\frac{(\xi_2^0)^2}{2\xi_2^0 - \bar{\xi}_2} \qquad \bar{\xi}_2 = \xi_2^0\left(2 - \frac{\xi_2^0}{\xi_2}\right)$$

$$\alpha_2(\xi_2) = (\xi_2^0)^2\left(k_\rho \ln\left(\frac{\xi_2}{\xi_2^0}\right) + \frac{g}{(\xi_2)^2} - \frac{g}{(\xi_2^0)^2}\right)$$

$$\alpha_3(\xi_2) = \frac{g(\xi_2^0)^2}{k_\rho \xi_2} \tag{22}$$

and

$$x_1 = \xi_1$$

$$x_2 = \xi_2$$

$$x_3 = \frac{\xi_2^0}{\xi_2}\xi_3 + \alpha_2(\xi_2) \tag{23}$$

then the system can be written in modified observer form:

$$\begin{aligned} y &= x_1 \\ \dot{x}_1 &= x_2 - \alpha_1(\xi_2) \\ \dot{x}_2 &= x_3 - \alpha_2(\xi_2) \\ \dot{x}_3 &= -\alpha_3(\xi_2) \end{aligned} \tag{24}$$

In some sense this change of coordinates linearizes the system "as much as possible".

Assuming the model to be affected by random process and measurement noises, the corresponding stochastic differential equations become:

$$
\begin{aligned}
dy &= x_1 dt + Ddv \\
dx_1 &= (x_2 - \alpha_1(\xi_2))dt + B_1 dw_1 \\
dx_2 &= (x_3 - \alpha_2(\xi_2))dt + B_2 dw_2 \\
dx_3 &= - \alpha_3(\xi_2)dt + B_3 dw_3
\end{aligned}
\qquad (25)
$$

with

$$
\begin{aligned}
x_1(0) &= N(x_1^0, P(1,1)) \\
x_2(0) &= N(x_2^0, P(2,2)) \\
x_3(0) &= N(x_3^0, P(3,3))
\end{aligned}
$$

$$Q = BB^T \quad R = DD^T$$

where R is the measurement noise covariance and Q(1,1), Q(2,2), Q(3,3) are the process noise variances. P(1,1), P(2,2), P(3,3) are the variances of the errors in the initial conditions.

The filter dynamic equations are:

$$
\begin{aligned}
d\hat{y} &= \hat{x}_1 dt \\
d\hat{x}_1 &= (\hat{x}_2 - \alpha_1(\hat{\xi}_2))dt + K_1(dy - d\hat{y}) \\
d\hat{x}_2 &= (\hat{x}_3 - \alpha_2(\hat{\xi}_2))dt + K_2(dy - d\hat{y}) \\
d\hat{x}_3 &= - \alpha_3(\hat{\xi}_2)dt + K_3(dy - d\hat{y})
\end{aligned}
\qquad (26)
$$

Rewriting the system in the original coordinates, the filter equations become:

$$
\begin{aligned}
d\hat{\zeta}_1 &= \hat{\zeta}_2 dt + K_1(dy - \hat{\zeta}_1 dt) \\
d\hat{\zeta}_2 &= (\, -g + \rho(\hat{\zeta}_1)\hat{\zeta}_2^2\hat{\zeta}_3)dt + (\frac{\hat{\zeta}_2}{\hat{\zeta}_2^0})^2 K_2(dy - \hat{\zeta}_1 dt) \\
d\hat{\zeta}_3 &= (\frac{\hat{\zeta}_3}{k_\rho} K_1 - \frac{\hat{\zeta}_2}{k_\rho \rho(\hat{\zeta}_1)(\hat{\zeta}_2^0)^2} K_2 + \frac{1}{\rho(\hat{\zeta}_1)(\hat{\zeta}_2^0)^2} K_3)(dy - \hat{\zeta}_1 dt)
\end{aligned}
\qquad (27)
$$

where K_1, K_2, K_3, are the gains computed from (13) and (14) with $K=PC^T R^{-1}$.

The Jacobian of the transformation from physical coordinates to observer form is:

$$
J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & (\frac{\zeta_2}{\zeta_2^0})^2 & 0 \\ \frac{\zeta_3}{k_\rho} & \frac{-\zeta_2}{k_\rho \rho(\zeta_2^0)^2} & \frac{1}{\rho(\zeta_2^0)^2} \end{bmatrix} \qquad \rho = \rho_0 e^{-\zeta_1/k_\rho} \qquad (28)
$$

The Jacobian turns out to be close to a unitary operator. This is clear for the 2x2 upper left minor. In fact, simulations were computed both assuming the Jacobian constant evaluated at ζ^0, and computing its actual value on line. The results were not affected in an appreciable manner.

3 - Simulations and Results

In this section we present the results of some simulations, run for different noise regimes and with the following constants:

initial conditions:

- initial height = 30,000 m
- initial velocity = - 2,000 m/sec
- initial inverse ballistic coefficient = $1.025*10^{-4}$ m^2/kg.

physical parameters:

- atmospheric decay constant k_ρ = 10,000 m
- atmospheric density at sea level ρ_0 = 1.230 kg/m^3
- gravity acceleration g = 9.81 m / sec^2

Additionally, the results correspond to the following noise regime:

measurement noise

- R = 100 m^2sec

process noises

- Q(1,1) = 100 m^2
- Q(2,2) = 100 m^2/sec^2
- Q(3,3) = $1.0*10^{-10}$ m^4 / kg^2

uncertainties on the initial conditions

- P(1,1) = 5000 m^2
- P(2,2) = 2000 m^2/ sec^2
- P(3,3) = $1.0*10^{-8}$ m^4/ kg^2

The object falls from an initial height of 30,000 m with an initial downward speed of 2,000 m/sec. It has an initial ballistic coefficient of 9,756 kg/m^2, which is equivalent to the object weighing approximately ten metric tons per square meter of surface perpendicular to the direction of the fall.

The errors in the initial conditions are, probably, not unrealistic for a radar tracking problem. The process noise covariances are also close to reality if we consider the possible random effects of changes in the conditions of the atmosphere and wind encountered on the way to the ground along the trajectory. Finally, the process noise on the ballistic coefficient could be interpreted as variations of the shape or orientation of the object during the fall.

The behavior of the real system is presented in Fig. 1, which portrays the trajectories of the three states affected by the noises. We remind the reader that the first state is the height of the object, the second the velocity and the third the inverse ballistic coefficient.

In Fig. 2(a) are shown the errors between the estimates of the height of the two filters and the height of the real system (Fig. 1(a)). The EKF and the GANF, compared in terms of performance in the estimate of the height, are nearly equivalent. The EKF performed slightly better, but the time history of the error is nearly identical. Similarly, in Figs. 2(b) and 2(c) are shown the estimates of the velocity and the inverse ballistic coefficient, respectively. Although the two filters showed a very similar behavior, the EKF performed slightly better, the difference probably being mainly due to the approximation introduced in assuming that the noise covariances were constant in observer coordinates for the GANF. In Fig. 2(d) is shown the behavior of the real inverse ballistic coefficient and of its estimates to give the reader an idea of the initial error in the estimates and of the effect of the process noise. Without process noise the real coefficient would be constant.

Shown in Fig. 3(a) is the logarithm of the average covariance of the error in the estimates of the height for 25 Monte Carlo runs. As can be seen from the Figs. 3(a) and 2(a) the recovery from errors in the initial guess is very fast after which the covariance settles down to values close to process noise covariance for the height. There is no appreciable difference in the behavior of the two filters. In Figs. 3(b) and 3(c) are shown the logarithms of the average covariance, for 25 Monte Carlo runs, of the error in the estimates of the velocity and the inverse ballistic coefficient, respectively. Again the two filters performed very similarly. In Fig. 3(c) the two filters behaved so similarly that the two curves are almost indistinguishable. All these results were checked by insuring that the covariances of the errors were near the values of the covariances theoretically predicted by the solution of the Riccati differential equation.

Conclusion:

The simulation results demonstrate that the GANF filter performed practically as well as the extended Kalman filter from the point of view the accuracy of the estimates.

In terms of the algebraic efforts in developing the filter equations, the EKF is obviously more straightforward, since one need only evaluate a first order approximation to the nonlinear equations along the trajectory. The GANF, on the other hand, relies on differential geometric concepts, and require considerably more difficult algebraic computations off-line. However, the development of the algebra may eventually become a simple exercise in computer programming if currently popular symbolic manipulation programs like MACSYMA or SMP are used.

The real advantage of the GANF filter over the extended Kalman filter is in its computational efficiency. In fact, as mentioned previously, we can compute the gains of the GANF filter off-line, whereas the gains of the extended Kalman filter must be calculated on-line. For the particular simulation presented in this paper, written in FORTRAN language and executed on a VAX 785 computer running under the VMS operating system, the integration of the GANF filter along the entire trajectory required 1.06 seconds of CPU time, while the integration of the extended Kalman filter took 3.2 seconds. Thus the GANF filter has performed three times faster. With higher order systems the computational advantage will be further emphasized since the on-line computational burden of the extended Kalman filter grows as $(n^2 + 3n)/2$ while that of the asymptotic nonlinear filter grows only as n. In fact, solving the Riccati differential equation on-line requires the integration of $(n^2 + n)/2$ scalar differential equations.

List of figures:

ballistic coefficient together with the actual inverse ballistic coefficient.

Fig. 3 Variances of the errors of the extended Kalman filter and of the geometric asymptotic nonlinear filter in the estimates of the states of the falling object after 25 Monte Carlo runs. Plotted using logarithmic scale. (a): height, (b): velocity, (c): inverse ballistic coefficient.

References:

[1] Krener, A. J., and A. Isidori, "Linearization by Output Injection and Nonlinear observers", Systems and Control Letters 3 (1983), pp. 47-52.

[2] Krener, A. J., and W. Respondek, "Nonlinear Observers with Linearizable Error Dynamics", SIAM J. Control and Optimization 23(2) (1985), pp. 197-216.

[3] Gelb, A. (ed.), *Applied Optimal Estimation* (MIT Press, Cambridge, 1974).

[4] Wishner, R. P., J. A. Tabaczynski, and M. Athans, "A Comparison of Three Non-Linear Filters", Automatica 5 (4) (1969), pp. 487-496.

[5] Krener, A. J., "The Asymptotic Approximation of Nonlinear Filters by Linear Filters", in C. I. Byrnes and A. Lundquist, editors, *Theory and Application of Nonlinear Control Systems*, Elsevier Science Publishers B. V. (North Holland) (1986), pp. 359-378.

[6] H. Keller, "Nonlinear Observer Design by Transformation into a Generalized Observer Canonical Form", University of Karlsruhe, preprint.

[7] H. Keller, "Nonlinear Observer Design via Two Canonical Forms", University of Karlsruhe, preprint.

[8] Bestle, D. and M. Zeitz, "Canonical Form Observer Design for Non-linear Time-variable Systems", International Journal of Control, 38(2) (1983), pp. 419-431

[9] Fritz, H. and H. Keller, "Design of Nonlinear Observers by a Two-step Transformation", in M. Fliess and M. Hazewinkel, editors, *Algebraic and Geometric Methods in Nonlilnear Control Theory*, D. Riedel, Dordrecht, 1986.

[10] Li, C. W. and L. W. Tao, "Observing Non-linear Time-variable Systems Through a Canonical Form Observer", International Journal of Control, 44(6) (1986), pp. 1703-1713.

[11] Zeitz, M., "Canonical Forms for Nonlinear Systems", in *Proceedings of the Conference on the Geometric Theory of Nonlinear Control Systems*, Wroclaw Technical University Press, Wroclaw, 1985.

[12] Zeitz, M., "The Extended Luenberger Observer for Nonlinear Systems", preprint.

HEIGHT OF THE OBJECT

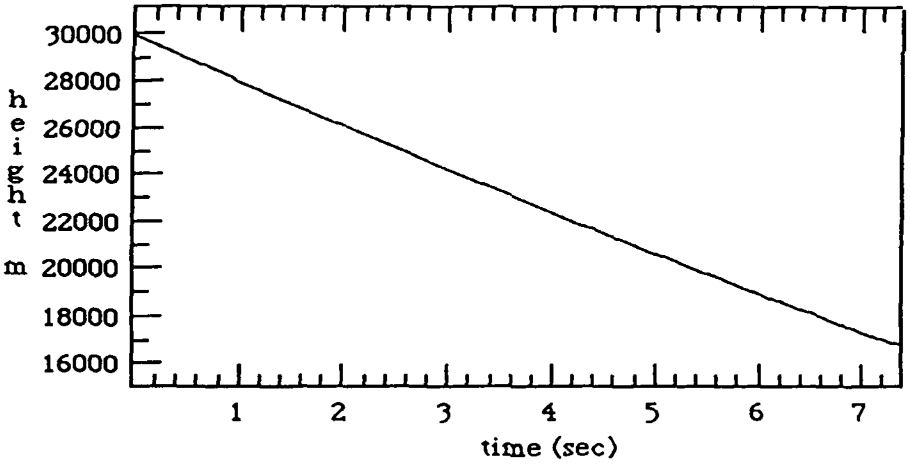


Fig. 1.a.

VELOCITY OF THE OBJECT

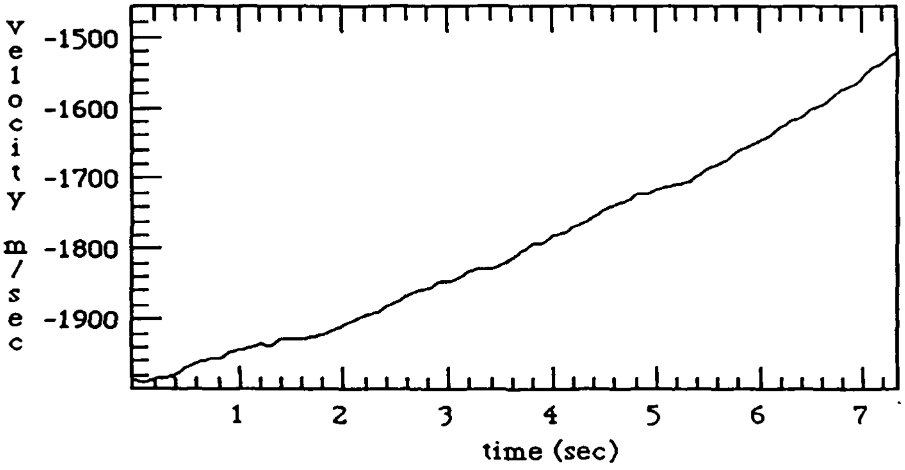


Fig. 1.b.

INV. BALLISTIC COEFFICIENT

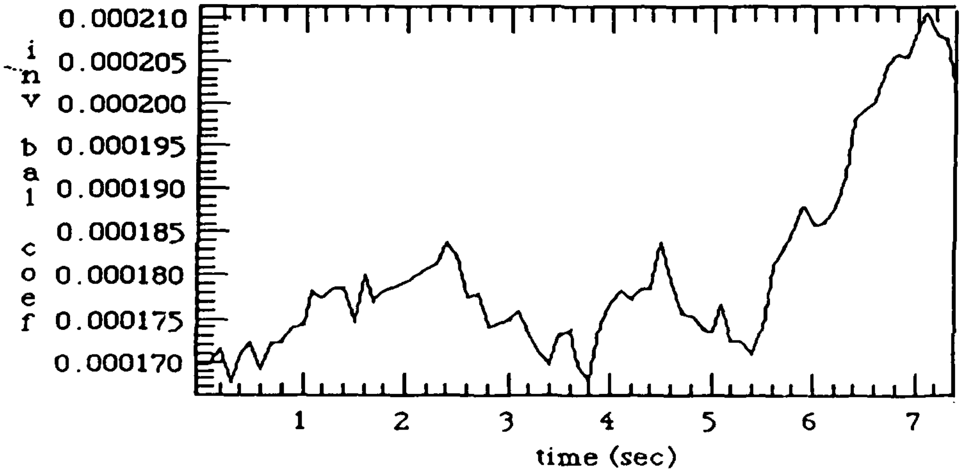
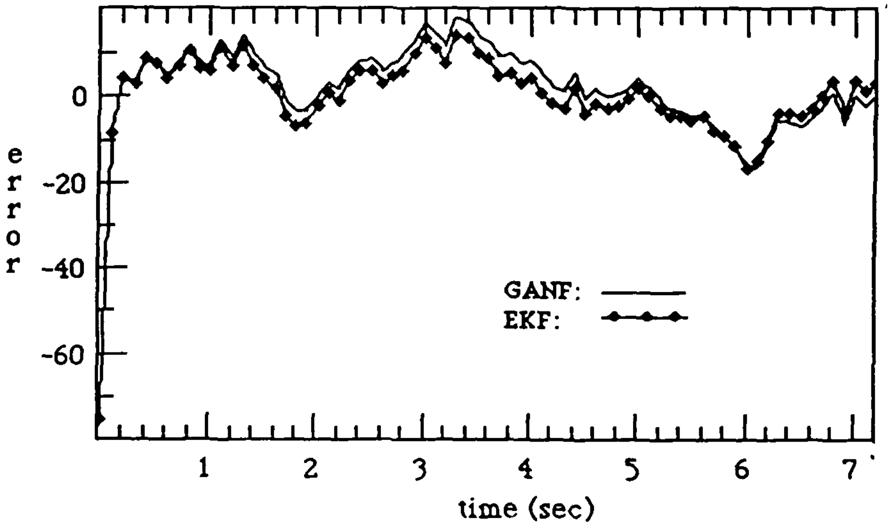
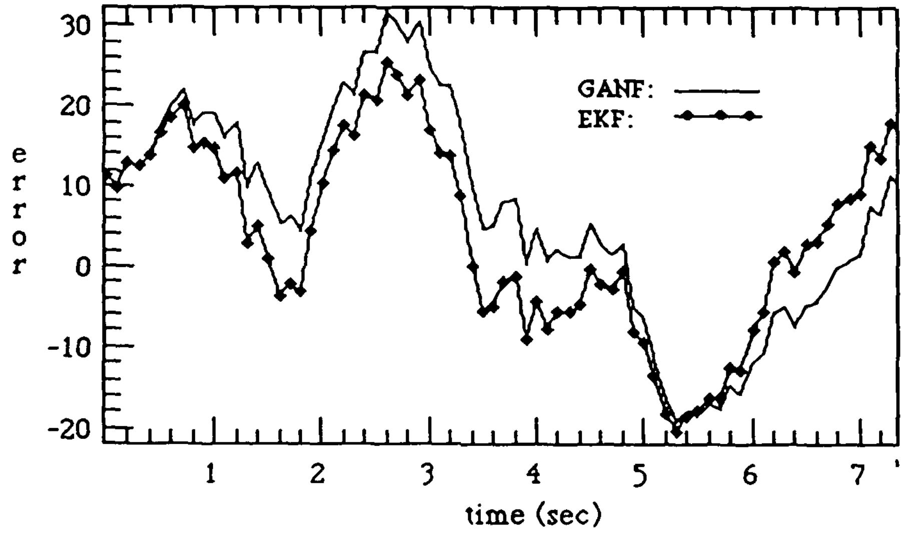


Fig. 1.c.



ERROR IN THE HEIGHT

Fig. 2.a.



ERROR IN THE VELOCITY

Fig. 2.b.

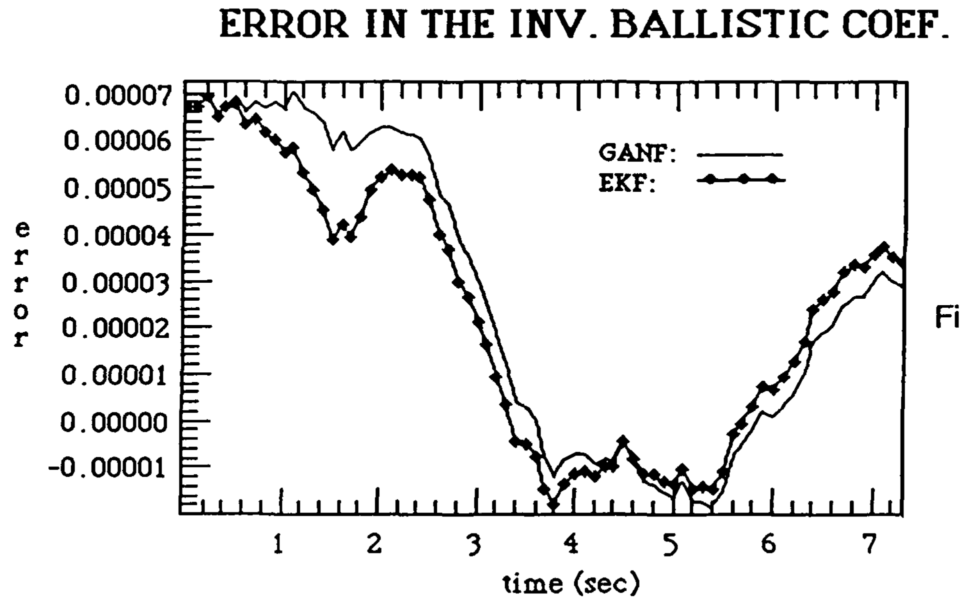

ERROR IN THE INV. BALLISTIC COEF.
error
0.00007
0.00006
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
-0.00001
1
2
3
4
5
6
7
time (sec)
GANF:
EKF:

Fig. 2.c.

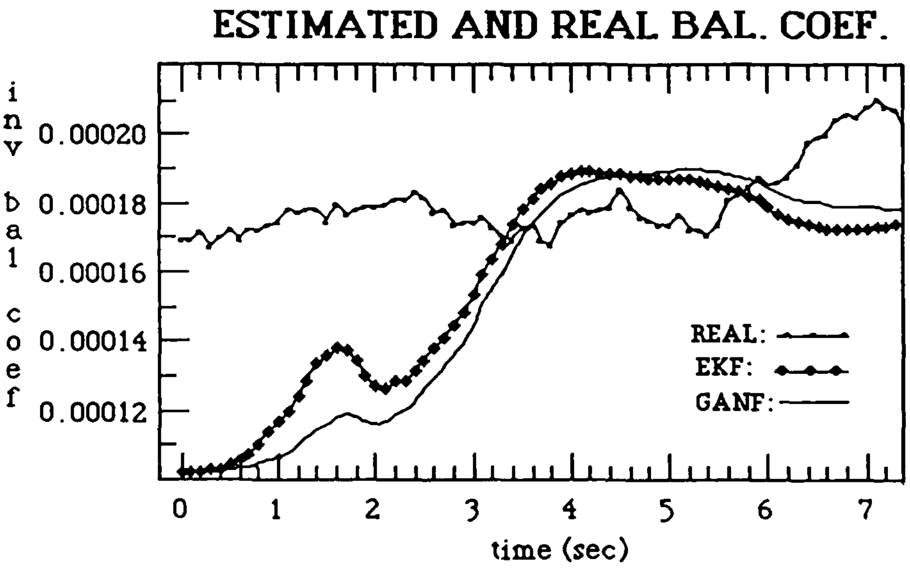

ESTIMATED AND REAL BAL. COEF.
inv bal coef
0.00020
0.00018
0.00016
0.00014
0.00012
0
1
2
3
4
5
6
7
time (sec)
REAL:
EKF:
GANF:

Fig. 2.d.

COVARIANCE OF THE ERROR IN THE HEIGHT

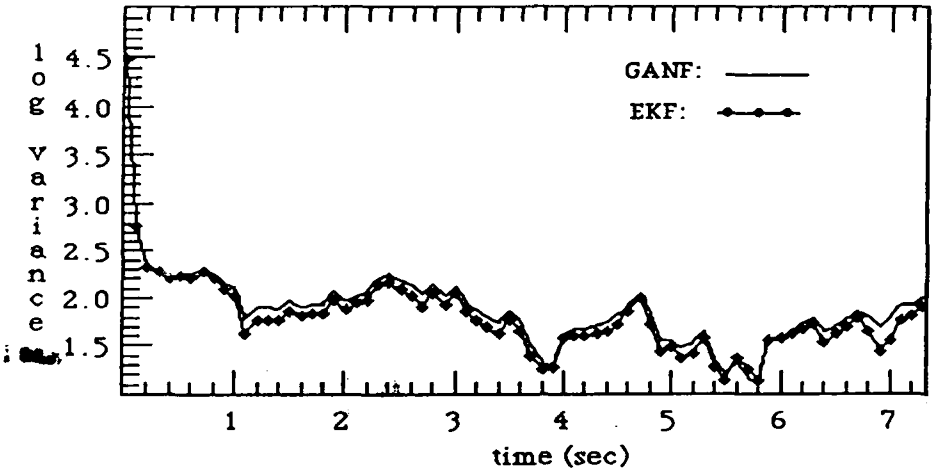


Fig. 3.a.

COVARIANCE OF THE ERROR IN THE VELOCITY

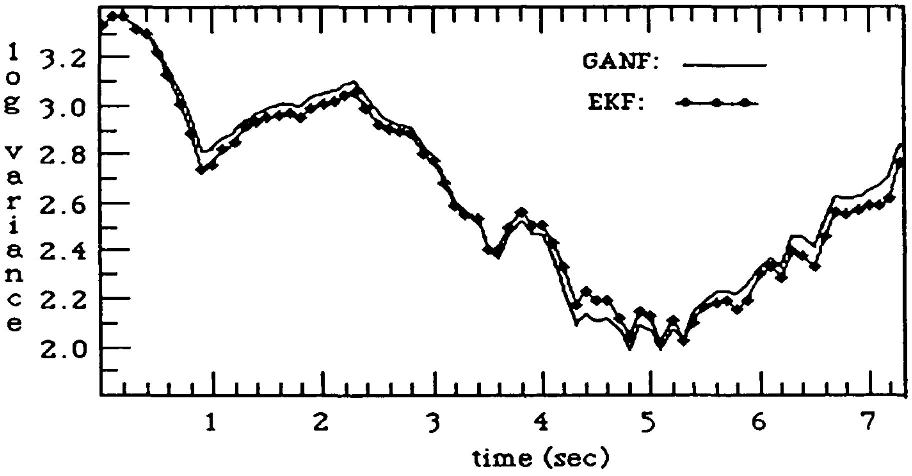


Fig. 3.b.

COV. OF THE ERROR IN THE INV. BAL. COEF.

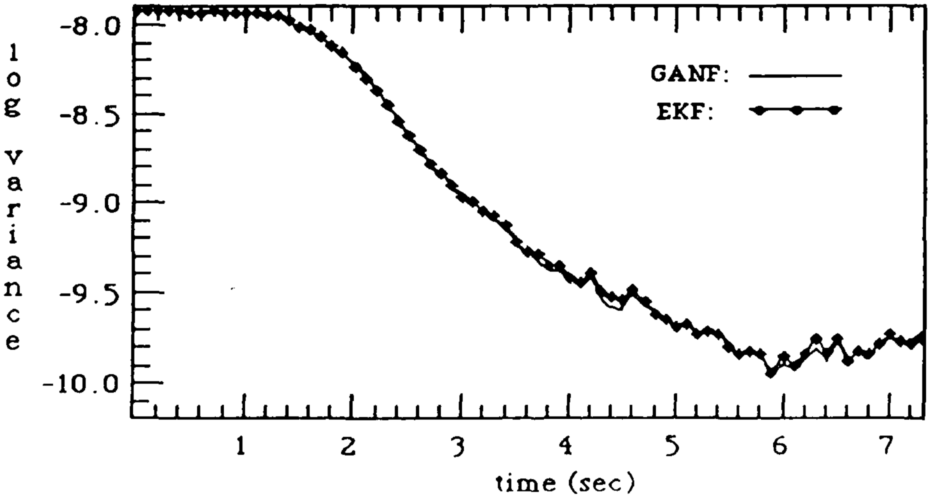


Fig. 3.c.

Computation of Observer Normal Form Using *Macsyma*[*]

by Andrew R. Phelps[†] and Arthur J. Krener[‡]

1. Introduction

The computation of linear observers has become relatively routine, and computer packages exist which make these computations straightforward and accessible. When it comes to nonlinear observers, however, the picture has not been so bright. Algorithms for this sort of calculation have been published [1], [2], [4], [5], [6]. In general, these are limited by one or more steps involving difficult computations.

The Ph.D. thesis of Phelps [8] provides a breakthrough in the nonlinear observer algorithm. In particular, Lie bracket calculations are no longer required to perform changes of state coordinates, and the computation becomes straightforward. *Macsyma* was instrumental in developing the new approach. A prototype of this new algorithm has been implemented in *Macsyma*.

We consider an uncontrolled dynamical system with partial state observation:

$$\begin{aligned} \dot{\xi} &= f(\xi)\,, \\ y &= h(\xi)\,. \end{aligned} \tag{1}$$

The state space is in $\mathbf{R}^n$ and the output space is in $\mathbf{R}^p$. Generally, this may be put in observa*ble* normal form:

$$\begin{aligned} \dot{\xi} &= A\,\xi - B\,\alpha(\xi)\,, \\ y &= C\,\xi\,. \end{aligned} \tag{2}$$

The problem is to see if, in fact, it supports observer normal form:

$$\begin{aligned} \dot{x} &= A\,x - \alpha(C\,x)\,, \\ \bar{y} &= C\,x\,. \end{aligned} \tag{3}$$

Here the A, B and C are matrices given in Brunovský canonical form.

[*] Research supported in part under AFOSR 85-0267

[†] University of California, Berkeley. Current address:
Dept. of Math. and Computer Sci.
San Jose State University
San Jose, CA 95192.

[‡] Address:
Dept. of Mathematics
University of California
Davis, CA 95616.

The algorithm in question is determined by the conditions required for conversion of a system (1) to observer form (3). This paper is based on the approach in [4] and [5], as modified in [8]. These conditions are:

Observable form Must be able to convert system to observable form (2);
Output coordinate change Must satisfy d.e. for $y = y(\bar{y})$;
Polynomial degree Observable form polynomials $f_j(\xi)$, for $1 \leq j \leq p$, (the entries of $B\,\alpha(\xi)$ in (2)) must have degree $\leq \ell_j$;
Coefficient compatibility Observable form polynomials must evaluate to certain integrals of differential expressions in injection terms (the entries of $\alpha(C\,x)$ in (3)).

Note. An earlier version of these conditions replaces the coefficient compatibility condition with the condition that the certain brackets vanish. Let q_j be the unit vector in the $\xi_{j:\ell_j}$ direction. All brackets of elements in $\{ad_{-f}^{i-1} q_j \ : \ 1 \leq i \leq \ell_j\}$ must vanish.

The approach in [1], [2] and [6], which is not used here, has been developed only for the case when $p = 1$ and there is no change of output coordinates required. It calls for the existence of observable form (2) (but not its computation) and replaces the last two conditions with a requirement that $d(ad_{-f}^{\ell} q) \in$ span $\{dh\}$, plus a slight technical adjustment.

We will elaborate the theorems and computations associated with the coefficient compatibility condition, which will obviate the need for extensive bracket computations.

2. Coefficient Compatibility in Standard Coordinates

For simpucity's sake, we first describe the results in the case that we have the "right" output $\bar{y}$. In fact, we will see that this could be considered sufficient for an improved algorithm, since the relevant change of state coordinates will be entirely determined by the tranformation $y = y(\bar{y})$ of the corresponding outputs.

In section 3, however, we will indicate a result expressed directly in observable form coordinates (ξ, y).

We may compute observ*able* form $((\bar{\xi}, \bar{y})$ coordinates), relative to the output $\bar{y}$ which is the solution to the output d.e.'s:

$$\dot{\bar{\xi}} = A\,\bar{\xi} - B\,\alpha(\bar{\xi})\,, \qquad \bar{y} = C\,\bar{\xi}\,. \tag{4}$$

We call the coordinates (4) *standard coordinates.*

This computation does not constitute a major burden on our algorithm (given *Macsyma*), since we only require an iterated set of Lie differentiations of functions and back-substitutions to get the transformation $\xi = \xi(\bar{\xi})$.

To annotate our coordinate systems, we adopt certain conventions. We describe the state variables by $\xi_{i:j}$, indicating that it is the $(j-1)$-th time derivative of the output variable y_i, the same going for $(x, \bar{y})$ and $(\bar{\xi}, \bar{y})$ coordinates. The injection functions $\alpha_{i:j}$ are written similarly. If $p = 1$, we omit the "1 :" to simplify the notation.

Furthermore, the coefficient

$$\bar{a}_{m:\cdots}(\bar{y}) := \bar{a}_{m:\cdots \underbrace{| \cdots \underbrace{i_{j:k}+1 \cdots i_{j:k}+1}_{e_{j:k} \text{ times}} \cdots |}_{(j)} \cdots}(\bar{y})$$

of the monomial

$$\xi^{\sharp} := \prod_{j=1}^{p} \prod_{k=1}^{r_j} \bar{\xi}_{j:i_{j:k}+1}^{e_{j:k}} \tag{5}$$

is characterized by having *degree* $i_{j:k}$ and *exponent* $e_{j:k}$ with respect to its factor $\xi_{j:i_{j:k}+1}^{e_{j:k}}$, for $1 \leq k \leq r_j$ and $1 \leq j \leq p$. The vertical bars '—' separate the subscript into parts according to the the output j involved. We also have cumulative indices

$$e_j := \sum_{k=1}^{r_j} e_{j:k}\,, \quad e := \sum_{j=1}^{p} e_j \quad \text{and} \quad w := \sum_{j=1}^{p} \sum_{k=1}^{r_j} i_{j:k}\, e_{j:k}\,.$$

To simplify the notation, we also occasionally represent functions as their own "0-th" derivatives and we write "$\alpha_{m:i}$," for $i \leq 0$, as a null symbol indicating a contribution which vanishes.

The natural way to describe the change from observable form coordinates ξ to observer form coordinates x is to compute the d.e.'s which determine it. These get quite complicated, since they involve the change of coordinates matrix $J = \dfrac{\partial y}{\partial \bar{y}}$ and its iterated derivatives. The choice of standard coordinates, however, causes all these terms to vanish and makes these d.e. 's accessible.

Using, for convenience, the simplifying assumption that each observability index ℓ_j is equal to some ℓ, we compute the equations $x = x(\bar{\xi})$ governing the change of state coordinates:

$$\begin{aligned}
\bar{\xi}_{j:1} &= x_{j:1}\,, \\
\bar{\xi}_{j:2} &= x_{j:2} - \alpha_{j:1}(\bar{y})\,, \\
\bar{\xi}_{j:3} &= x_{j:3} - L_{\bar{f}}\,\alpha_{j:1}(\bar{y}) - \alpha_{j:2}(\bar{y})\,, \\
&\vdots \\
\bar{\xi}_{j:\ell_j} &= x_{j:\ell_j} - \sum_{i=1}^{\ell_j - 1} L_{\bar{f}}^{\ell_j - i - 1}\,\alpha_{j:i}(\bar{y})\,, \\
\bar{f}_j(\bar{\xi}) &= -\sum_{i=1}^{\ell_j} L_{\bar{f}}^{\ell_j - i}\,\alpha_{j:i}(\bar{y})\,,
\end{aligned} \tag{6}$$

for $1 \leq j \leq p$, where $\bar{f}$ is f in the $\bar{\xi}$ coordinates.

From the expansion (6) we can, in effect, read off the $\bar{a}$ polynomial coefficients in terms of the α injection functions.

Example 2.1. *Coefficient solutions, for the case $p = 1$ and $\ell = 3$.*

The expansion (6) gives us:

$$\begin{aligned} f_1(\xi) = {} & a_{23}(y)\,\xi_2\,\xi_3 + a_3(y)\,\xi_3 \\ & + a_{222}(y)\,\xi_2^3 + a_{22}(y)\,\xi_2^2 + a_2(y)\,\xi_2 \\ & + a_1(y)\,. \end{aligned}$$

This leads to the coefficient solutions:

$$\begin{aligned} \bar{a}_1 &= -\alpha_3\,, \\ \bar{a}_2 &= -\frac{d\alpha_2}{d\bar{y}}\,, \\ \bar{a}_{22} &= -\frac{d^2\alpha_1}{d\bar{y}^2}\,, \\ \bar{a}_{222} &= 0\,, \\ \bar{a}_{23} &= 0\,, \\ \bar{a}_3 &= -\frac{d\alpha_1}{d\bar{y}}\,. \end{aligned}$$

The α's may be computed from the $\bar{a}$'s by a simple integration. △

The pattern described in the above example is easily extended to the case where we have p *equal* indices.

Theorem 2.2. *Suppose that all the indices are equal,* i.e., *$\ell_j = \ell$ for $1 \le j \le p$. Then the polynomial coefficient $\bar{a}_{m:\cdots}(\bar{y})$ is given by*

$$-\left(\frac{w!}{\prod_{j=1}^{p}\prod_{k=1}^{r_j} e_{j:k}!\, i_{j:k}!^{e_{j:k}}}\right)\frac{\partial^{e}\alpha_{m:\ell-w}(\bar{y})}{\partial\bar{y}_1^{e_1}\cdots\partial\bar{y}_p^{e_p}}\,. \tag{7}$$

The existence of an injection vector $\alpha(\bar{y})$ compatible with all *the coefficients of $f_1(\xi)$, as given in (7), together with the observable form, output coordinate change and polynomial degree conditions, constitute necessary and sufficient conditions for the existence of observer normal form (supposing all indices equal).*

Proof
Outline

We use induction on ℓ. We make a counting argument with combinatorics to trace the typical coefficient of a monomial term.

Body of Proof

Let m, $1 \le m \le p$, be fixed.

First of all, if $\ell = 1$, then (4) simply becomes $\dot{\xi}_{m:1} = \bar{f}_m(\xi) = \bar{f}_m(\bar{y})$. Thus $\bar{a}_{m:1} = \bar{f}_m(\bar{y}) = -\alpha_{m:1}(\bar{y})$, where we take $\alpha_{m:1} := -\bar{f}_m$. But this matches formula (7).

Note that the derivation in example 2.1 gives a result in accordance with formula (7). That illustrates the pattern we use for our induction.

As induction hypothesis, we assume that formula (7) holds for $\ell = \mu$. We target the coefficients in the expansion for $\bar{\xi}_{m:\mu} - x_{m:\mu}$ in the case $\ell = \mu + 1$ which are the *source* via Lie differentiation of the coefficients of $\bar{f}_m$ we seek to evaluate.

But, when we examine the p.d.e. expansion (6), we find the same expression,

$$-\sum_{i=1}^{\mu} L_{\bar{f}}^{\mu-i}\, \alpha_{m:i}\,,$$

for $\bar{\xi}_{m:\mu} - x_{m:\mu}$, in the expansion with $\ell = \mu + 1$, as we find for $\bar{f}_m$ in the expansion with $\ell = \mu$. This means that the induction assumption will enable us to *know* those "target" coefficients, which, when Lie differentiated, give contributions to the expansion for $\bar{f}_m$ in the case when the multi-index ℓ is $\mu + 1$. These coefficients evaluate to nothing but the coefficients of the terms of $\bar{f}_m$ in the case (given by induction assumption) that the multi-index ℓ is μ.

Let $\xi^\sharp$, given as in (5), be a monomial in $\bar{f}_m(\bar{\xi})$. We wish to determine its coefficient $\bar{a}_{m:\cdots}$.

An expression which under Lie differentiation by $\bar{f}$ can give a term like $\xi^\sharp$ may have two forms.

Case 1. It may have no increment in the exponent $e_{j:1}$ of $\bar{\xi}_{j:2}$, for $1 \le j \le p$.

In this instance, it will come from Lie differentiation of a term of the form

$$\xi^\sharp\, \bar{\xi}_{\eta:\iota_{\eta:k}}\, \bar{\xi}^{-1}_{\eta:\iota_{\eta:k}+1}\,. \tag{8}$$

By induction assumption, term (8) has a numerical coefficient which calculates back from our projected coefficient for $\xi^\sharp$ as

$$-\frac{\iota_{\eta:k}!\, e_{\eta:k}}{(\iota_{\eta:k}-1)!\, e^\sharp}\left(\frac{1}{w}\right)\left(\frac{w!}{\prod_{j=1}^{p}\prod_{k=1}^{r_j} e_{j:k}!\, i_{j:k}!^{\,e_{j:k}}}\right).$$

Here we take

$$e^\sharp := \begin{cases} e_{\eta:k-1} + 1\,, & \text{if } \iota_{\eta:k-1} = \iota_{\eta:k} - 1\,; \\ 1\,, & \text{otherwise}\,. \end{cases}$$

The differentiation process contributes an extra factor of $e^\sharp$. Thus, the partial numerical contribution from this term is

$$\iota_{\eta:k}\, e_{\eta:k}\left(\frac{1}{w}\right). \tag{9}$$

In this case, the "α" part of the coefficient is carried through unchanged. Moreover, it was already of the required form, since e_η has not been altered by adding and subtracting 1 in (8).

Case 2. It may have an increment in the coefficient of $\bar{\xi}_{j:2}$ for $j = \eta$.

The source of the terms of this type is partial differentiation of the "α" coefficient by $\bar{y}_\eta$. This will give an additional factor of $\bar{\xi}_{\eta:2}$, while not affecting the numerical coefficient. The prior exponent of $\bar{\xi}_{\eta:2}$ will have been $e_{\eta:1} - 1$.

Therefore, the monomial term prior to Lie differentiation was $\xi^\sharp \, \bar{\xi}_{\eta:2}^{-1}$. By induction assumption, its numerical contribution was

$$- \iota_{\eta:1}! \, e_{\eta:1} \left(\frac{1}{w}\right) \left(\frac{w!}{\prod_{j=1}^{p} \prod_{k=1}^{r_j} e_{j:k}! \, i_{j:k}!^{\, e_{j:k}}}\right) .$$

Its partial numerical contribution is therefore

$$\iota_{\eta:1} \, e_{\eta:1} \left(\frac{1}{w}\right) , \tag{10}$$

since $\iota_{j:1}! = 1 = \iota_{j:1}$ when $j = \eta$.

Note that in this case the e_η increments by 1, so that the "α" term adjusts as prescribed by formula (7).

These two cases exhaust the possibilities. The "α" coefficients are as required. And, combining (9) and (10), we get a total contribution to the numerical coefficient of a factor of

$$\left(\sum_{j=1}^{p} \sum_{k=1}^{r_j} \iota_{j:k} \, e_{j:k}\right) \left(\frac{1}{w}\right) = 1 ,$$

which is also as required. △

In the general case, where the observability indices are given arbitrarily, we adopt a recursive method for calculating the $\bar{a}$ coefficients in terms of the α injection functions.

We may take a system with arbitrary indices $\ell_1, \ldots, \ell_p$, and prolong it to a system of dimension $p \cdot \ell_p$, to which theorem 2.2 applies. Retracing the prolongation step-by-step, we can track the (increasingly complex) form of the formulas for the $\bar{a}$ coefficients. We rely on the following prolongation lemma (for a proof, see [8]):

Lemma 2.3 (Krener-Respondek-Phelps). *Suppose an uncontrolled system, given in observable form, has 2 distinct multi-indices ℓ_1, ℓ_2 of multiplicities p_1, p_2 and, further, that it may be transformed by change of output coordinates $y = y(\bar{y})$ to observer form. Then it may be prolonged to a system in observable form, having multi-indices $\lambda_1 := \ell_1 + 1$ and $\lambda_2 := \ell_2$, of the above multiplicities. Furthermore, the transformation $y = y(\bar{y})$ and the injection function $\alpha(\cdot)$ both prolong trivially to functions which will take the prolonged system over into observer form.*

We formulate our recursive coefficient calculation as follows:

Algorithm 2.4 (Coefficient Prolongation Algorithm). *Suppose we have solved for the coefficients belonging to the indices $\ell_1, \ldots, \ell_k$ and moreover $\ell_k < \ell_k + 1 = \ldots = \ell_s$, where $s \leq p$. We may construct a "quasi-solution" using the method of theorem 1 applied to the prolonged system with s indices all equal to ℓ_s. By back-substitution we may then express the prolonged version of $\bar{f}_j(\bar{\xi})$ with α's as coefficients, for $k+1 \leq j \leq s$. We may then substitute $L^i_{\bar{f}}\bar{\xi}_{h:\ell_h}$ (expressed in terms of the α's, and using the solutions previously derived for $1 \leq j \leq k$) for the "quasi-variables" $\bar{\xi}_{h:\ell_h+i}$ where $1 \leq i \leq \ell_s - \ell_h$. Finally, we may "read off" the coefficients of the monomials thus derived.*

We may now combine theorem 2.2 and algorithm 2.4 to get the coefficient compatibility theorem for $\bar{\xi}$ coordinates:

Theorem 2.5. *Suppose we have arbitrary indices $\ell_1, \ldots, \ell_p$. Then we may derive the coefficients $\bar{a}_{m:\ldots}(\bar{y})$ in terms of the $\alpha(\bar{y})$ injection functions by application of the Coefficient Prolongation Algorithm. The existence of an injection vector $\alpha(\bar{y})$ compatible with all the coefficients of $f_1(\xi)$, together with the observable form, output coordinate change and polynomial degree conditions, constitute necessary and sufficient conditions for the existence of observer normal form.*

Proof

This has already been done in theorem 2.2 for the case where all indices are the same. Algorithm 2.4 enables us to extend this result inductively whenever $\ell_h < \ell_{h+1}$. △

It can also be shown that we can back-solve for the α injection functions in terms of the $\bar{a}$ coefficients by iterated integrations (see [8]).

For the generic case, where there are two distinct indices, differing by 1, we state the formula:

Corollary 2.6. *For the generic case of two different multi-indices of size λ_1 and $\lambda_2 := \lambda_1 + 1$ and multiplicities p_1 and p_2, the coefficient $\bar{a}_{m:\ldots}(\bar{y})$ is given by*

$$\left(\frac{w!}{\prod_{j=1}^{p} \prod_{k=1}^{r_j} e_{j:k}!\, i_k!^{e_{j:k}}} \right) \left[\sum_{i=1}^{p_1} \left(\frac{\partial^e \alpha_{i:\lambda_1 - w}(\bar{y})}{\partial \bar{y}_1^{e_1} \cdots \partial \bar{y}_p^{e_p}} \frac{\partial \alpha_{m:\ell_m - \lambda_1}(\bar{y})}{\partial \bar{y}_i} \right) - \frac{\partial^e \alpha_{m:\lambda_2 - w}(\bar{y})}{\partial \bar{y}_1^{e_1} \cdots \partial \bar{y}_p^{e_p}} \right].$$

In particular, $\bar{a}_{m:\ldots}(\bar{y})$ is given by theorem 1 for $1 \leq m \leq p_1$ and for degree $\geq \lambda_1$ *when $p_1 + 1 \leq m \leq p$.*

Proof

This is directly calculated using theorem 2.2 and one application of algorithm 2.4. △

To conclude this section, we give an example of coefficients for the (simplest) non-generic case. Note that it is trivial to back-solve for the derivatives of the α 's and integrate.

Example 2.7. *Coefficient solutions, for $p = 2$, $\ell_1 = 1$, $\ell_2 = 3$.*

$$\begin{aligned}
\bar{a}_{1:1} &= -\,\alpha_{1:1}\,, \\
\bar{a}_{2:1} &= -\,\alpha_{1:1}^{2}\,\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_1^2} - \alpha_{1:}\,\frac{\partial\alpha_{1:1}}{\partial\bar{y}_1}\,\frac{\partial\alpha_{2:1}}{\partial\bar{y}_1} + \alpha_{1:1}\,\frac{\partial\alpha_{2:2}}{\partial\bar{y}_1} - \alpha_{2:3}\,, \\
\bar{a}_{2:|2} &= 2\,\alpha_{1:1}\,\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_1\,\partial\bar{y}_2} + \frac{\partial\alpha_{1:1}}{\partial\bar{y}_2}\,\frac{\partial\alpha_{2:1}}{\partial\bar{y}_1} - \frac{\partial\alpha_{2:2}}{\partial\bar{y}_2}\,, \\
\bar{a}_{2:|22} &= -\,\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_2^2}\,, \\
\bar{a}_{2:|3} &= -\,\frac{\partial\alpha_{2:1}}{\partial\bar{y}_2}\,.
\end{aligned}$$

△

3. Coefficient Compatibility—General Case

We have seen the expansion (6) in the simplified situation of standard coordinates. This can be converted to general observable form coordinates (2) by a not-too-inconvenient calculation. However, if we have *Macsyma* or some other facility that enables us to do suggestive examples expeditiously, we can find patterns in the results that suggest direct solutions for the general version of (6). For instance, consider:

Example 3.1. *Coefficient solutions, in general observable form coordinates, for* $p = 1$, $\ell = 4$.

We have the following pattern, which mimics the result in theorem 2.2:

$$\begin{aligned}
a_1 &= -\frac{dy}{d\bar{y}}\,\alpha_4\,, \\
a_2 &= -\frac{dy}{d\bar{y}}\,\frac{d\alpha_3}{dy}\,, \\
a_{23} &= -3\,\frac{dy}{d\bar{y}}\,\frac{d^2\alpha_1}{dy^2}\,, \\
a_{22} &= -\frac{dy}{d\bar{y}}\,\frac{d^2\alpha_2}{dy^2}\,, \\
a_3 &= -\frac{dy}{d\bar{y}}\,\frac{d\alpha_2}{dy}\,, \\
a_{222} &= -\frac{dy}{d\bar{y}}\,\frac{d^3\alpha_1}{dy^3}\,, \\
a_4 &\quad \frac{dy}{d\bar{y}}\,\frac{d\alpha_1}{dy}\,.
\end{aligned}$$

We have another pattern, which relates to the degree ℓ terms that vanish in standard coordinates:

$$
\begin{aligned}
a_{2222} &= \frac{1}{4}\frac{d^2a_{24}}{dy^2} - \frac{3}{16}a_{24}\frac{da_{24}}{dy} + \frac{1}{64}a_{24}^3\,, \\
a_{223} &= \frac{3}{2}\frac{da_{24}}{dy} - \frac{3}{8}a_{24}^2\,, \\
a_{33} &= \frac{3}{4}a_{24}\,.
\end{aligned}
$$

△

To describe the intricacies of the degree ℓ terms in example 3.1, we introduce the following notational scheme. Let $P(m)$ be the *partitions* of m. Write a partition π of $e-1$ by

$$e - 1 = \sum_{j=1}^{s} c_j n_j\,.$$

Define $c := \sum_{j=1}^{s} c_j$ as the *number of pieces* of the partition.

With these annotations, we formulate the pattern of example 3.1 for the following theorem on coefficient compatibility:

Theorem 3.2. *A single-output system has an observer normal form iff the observable form, output coordinate change and polynomial degree conditions hold, and moreover in observable form* (ξ) *coordinates the coefficient* $a_{\dots}(y)$ *equals*

$$-\left(\frac{w!}{\prod_{k=1}^{r} e_k!\, i_k!^{e_k}}\right)\frac{d^e\alpha_{\ell-w}(y)}{dy^e}\,\frac{dy}{d\bar{y}}\,,$$

for terms of degree less than ℓ, *and*

$$-\sum_{\pi\in P(e-1)}\left[\left(\frac{w!\,(e-1)!}{\prod_{k=1}^{r} e_k!\, i_k!^{e_k}\,\prod_{j=1}^{s} c_j!\, n_j!^{c_j}}\right)\left(\frac{-1}{\ell}\right)^c \prod_{j=1}^{s}\left(\frac{d^{n_j-1}a_{2\ell}(y)}{dy^{n_j-1}}\right)^{c_j}\right],$$

for terms of degree equal to ℓ.

A proof of this theorem will appear in a forthcoming paper of Phelps [7].

Using Lemma 2.3 and the above theorem (adjusted to the case of p equal indices), we may in principle compute the general transformation $x = x(\xi)$, relating observer form (3) to observable form (2).

4. Conclusion

Two points need to be made here.

First, the "coefficient compatibility" approach to nonlinear observer calculations simplifies in principle the theory and makes unwieldy bracket calculations unnecessary.

Second, the use of *Macsyma* made it possible to do the rather extended calculations of examples that made the patterns in the data stand out. Every aspect of the algorithms for nonlinear observer calculation is readily accessible to *Macsyma* programming, and converting the algorithm from its abstract form of "algorithm-in-principle" to a concrete "algorithm-in-fact" is naturally done in this milieu.

References

[1] D. Bestle and M. Zeitz. Canonical form observer design for non-linear time-variable systems. *International Journal of Control*, 38(2):419–431, 1983.

[2] H. Fritz and H. Keller. Design of nonlinear observers by a two-step transformation. In M. Fliess and M. Hazewinkel, editors, *Algebraic and Geometric Methods in Nonlinear Control Theory*, D. Riedel, Dordrecht, 1986.

[3] A. J. Krener. Normal forms for linear and nonlinear systems. In M. Luksik, C. Martin, and W. Shadwick, editors, *Differential Geometry, the Interface between Pure and Applied Mathematics*, American Mathematical Society, Providence, RI. In press.

[4] A. J. Krener and A. Isidori. Linearization by output injection and nonlinear observers. *Systems & Control Letters*, 3:47–52, 1983.

[5] A. J. Krener and W. Respondek. Nonlinear observers with linearizable error dynamics. *SIAM Journal on Control and Optimization*, 23(2):197–216, 1985.

[6] C. W. Li and L. W. Tao. Observing non-linear time-variable systems through a canonical form observer. *International Journal of Control*, 44(6):1703–1713, 1986.

[7] A. R. Phelps. Polynomial coefficients of nonlinear observable form compatible with observer normal form. In preparation.

[8] A. R. Phelps. *A Simplification of Nonlinear Observer Theory.* PhD thesis, University of California, Berkeley, 1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 120-147, January 1989

THE STRUCTURE OF SMALL-TIME REACHABLE SETS IN LOW DIMENSIONS*

ARTHUR J. KRENER† AND HEINZ SCHÄTTLER‡

Abstract. This paper outlines a general method to determine the geometric structure of small-time reachable sets for a single-input control system with a bounded linear control. The authors' analysis relies on free nilpotent systems as a guide, and hence their techniques only apply to nondegenerate situations. The paper illustrates the effectiveness of the method in low dimensions. Among other results is given a precise description of the small-time reachable set for a system $\dot{x} = f(x) + g(x)u$, $|u| \leqq 1$ in dimension four, under the generic assumption that the constant controls $u \equiv +1$ and $u \equiv -1$ are not singular. As a corollary, a local synthesis is obtained in dimension three for the time-optimal control problem under the analogous generic condition.

Key words. nonlinear systems, nilpotent approximation, reachable sets, bang-bang trajectories, singular arcs

AMS(MOS) subject classifications. 49B10, 93B10

1. Introduction. In this paper we study the qualitative structure of small-time reachable sets in low dimensions for a single-input system with a bounded linear control. More precisely, we consider a system of the form

$$(1) \qquad \Sigma: \dot{x} = f(x) + g(x)u, \quad |u| \leqq 1, \quad x \in \mathbb{R}^n$$

where f and g are smooth (C^∞) or analytic vector fields and admissible controls are measurable functions with values in $[-1, 1]$ almost everywhere. A trajectory of the system corresponding to a control $u(\cdot)$ is an absolutely continuous curve $x(\cdot)$ such that $\dot{x}(t) = f(x(t)) + g(x(t))u(t)$ almost everywhere. We say a point q is reachable from a point p within time T if and only if there exists a trajectory $x(\cdot)$ defined on an interval $[0, t]$, $t \leqq T$, such that $x(0) = p$ and $x(t) = q$. The set of all such points q is denoted by Reach $(p, \leqq T)$; Reach (p, T) denotes the set of points that are reachable exactly at time T. The reachable set from p, Reach (p), is the set of all points that are reachable from p within some time T.

Reachable sets play an important role in control theory. If a system can be stabilized to a given point by a feedback control law, then that point must be in the reachable set of every other point. In optimal control problems, if the cost is added as another coordinate, then the optimal trajectories must lie in the boundary of the set of reachable points. For this reason the Pontryagin Maximum Principle plays an important role in studying the boundaries of reachable sets.

The problem of describing a reachable set and the extremal trajectories that generate its boundary is closely related to the problem of regular synthesis in the sense of Boltyansky [1] and others [5], [18]. While the problem has been studied extensively for many years, only a few examples of regular syntheses have been described, for instance, [24]. Even in low dimensions, the reachable set of a general control system can be extremely complicated.

* Received by the editors June 22, 1987; accepted for publication (in revised form) April 11, 1988. This research was partly supported by National Science Foundation grant DMS-8601635 and Air Force Office of Scientific Research grant 85-0267.

† University of California at Davis, Department of Mathematics, Davis, California 95616.

‡ Washington University, Department of Systems Science and Mathematics, St. Louis, Missouri 63130.

We shall attempt to avoid this difficulty by considering only "nondegenerate" systems. By a nondegenerate system we mean one where (i) f, g, and the low-order Lie brackets of f and g span as many dimensions as is possible given the dimensions of the state space; and where (ii) no nontrivial equality relations hold between those vector fields (for instance, if n is the space dimension, then any relation saying that n vector fields are dependent at a point is considered a nontrivial equality relation, whereas a relation that simply expresses the fact that a vector field can be written in terms of a basis is considered trivial).

This is in the spirit of Lobry [14], who described the small-time reachable set of (1) in dimension three under the assumption that f, g, and $[f, g]$ are linearly independent. The method described below is an attempt to extend Lobry's result to higher dimensions. As will be seen, it is successful in the four-dimensional case, but in higher-dimensional cases obstacles still have to be overcome. These obstacles, however, are not due to our general approach, but they lie in the fact that, at the moment, too little is known about the structure of extremal trajectories. We shall return to this question at the end of the paper. In the paper we shall give a precise description of the small-time reachable set in dimension four assuming that the constant controls $u=+1$ and $u=-1$ are not singular on the boundary of the reachable set. It can easily be seen (cf. § 4) that this is equivalent to an independence assumption on the vector fields f, g, $[f, g]$, and $[f+g,[f, g]]$, respectively, $[f-g,[f, g]]$. As a corollary we are able to improve on recent results of Bressan [4], Schättler [17], and Sussmann [21] on time-optimal control in dimension three.

Throughout this paper we will use nilpotent systems as a guide to the general situation. A system is nilpotent of order k if all brackets of orders greater than k vanish and if k is the smallest integer with this property. In a certain sense these systems play the same role as the polynomials do within the class of smooth functions. Nilpotent systems are the low-order part of the coordinate free Taylor series expansion of a general system.

To be more precise, we must define the Lie jet of system (1). At a point p the Lie jet consists of a list of the values at p of the Lie brackets of f and g written down in some prescribed order. Of course, because of the skew-symmetry and Jacobi relation

$$[f, g]+[g, f]=0, \qquad [f,[g, h]]+[g,[h, f]]+[h,[f, g]]=0,$$

we need only consider a list of distinct brackets. These brackets can be partially ordered by the total number of vector fields involved; for example, f is a bracket of order one and $[f, g]$ is of order two. The Lie jet of order k is a list of values at p of the distinct brackets of f and g of order less than or equal to k. The Lie jets of orders one through four are given below:

Order one: $\{f(p), g(p)\}$,
Order two: $\{f(p), g(p),[f, g](p)\}$,
Order three: $\{f(p), g(p),[f, g](p),[f,[f, g]](p),[g,[f, g]](p)\}$,
Order four: $\{f(p), g(p),[f, g](p),[f,[f, g]](p),[g,[f, g]](p),$
$[f,[f,[f, g]]](p),[f,[g,[f, p]]](p),[g,[g,[f, g]]](p)\}$.

If $N(k)$ is the number of distinct brackets of f and g of order k or less, then the kth-order Lie jet of (1) at p is a point in the vector bundle consisting of the Whitney sum of $N(k)$ copies of the tangent bundle.

A basic result of Krener [12], later proved in other contexts by Rothschild and Stein [15], Hermes [10], Crouch [8], Bressan [3], and Sussmann [20], [21] is that for analytic systems of the form (1), the kth-order Lie jet at p determines the trajectories emanating from p up to order $O(t^{k+1})$ and up to diffeomorphisms of the state space.

Sussmann [22], [23], Bressan [4], and Schättler [16], [17] have shown that the local structure of time-optimal controls in dimension two or three is determined in nondegenerate situations by the second, respectively, third-order Lie jet at a reference point. In degenerate situations higher-order jets need to be considered [16], [17], [23].

On the basis of these results we might conjecture that in nondegenerate situations the kth-order Lie jet at p determines the structure of the set of small-time reachable points where the Hörmander or controllability condition is satisfied, i.e., the rank of the kth-order Lie jet at p equals the dimension of the state space. And maybe the qualitative structure of the reachable set can be obtained by looking at a kth-order nilpotent approximation. Unfortunately, as we mention in the last section, these conjectures are not completely true, but they do motivate much of our work.

The paper is organized as follows. The next section reviews the Pontryagin Maximum Principle as applied to the system (1). This also gives us a chance to introduce some notation and terminology. In § 3, we will describe the main ideas and outline the general structure of our techniques by looking at the trivial two-dimensional case. We will also give a brief proof of Lobry's three-dimensional result. The main part of the paper is § 4, where we determine the geometric structure of the small-time reachable set for the nondegenerate four-dimensional system (assuming that both quadruples $(f, g, [f, g], [f+g, [f, g]])$ and $(f, g, [f, g], [f-g, [f, g]])$ consist of independent vectors at p). We also draw the obvious corollaries about time-optimal control in dimension three. Section 5 concludes with a brief discussion of the free nilpotent five-dimensional system and explains why the general nondegenerate five-dimensional case is different from this one.

2. The maximum principle. The Maximum Principle [13] gives necessary conditions for a point to lie on the boundary of the reachable set. Let $u(\cdot)$ be an admissible control defined on an interval $[0, T]$ and let $x(\cdot)$ be the corresponding trajectory starting at p. If $x(T) \in \partial$ Reach (p), then $x(t) \in \partial$ Reach (p) for all $t \in [0, T]$ and there exists an absolutely continuous curve $\lambda : [0, T] \to \mathbb{R}^n$, which does not vanish anywhere such that

$$\dot{\lambda}(t)^T = -\lambda(t)^T(Df(x(t)) + Dg(x(t)) \cdot u(t)), \tag{2}$$

$$\langle \lambda(t), g(x(t))\rangle u(t) = \operatorname*{Min}_{|v| \leq 1} \langle \lambda(t), g(x(t))\rangle v, \tag{3}$$

$$H = \langle \lambda(t), f(x(t)) + g(x(t))u(t)\rangle \equiv 0 \tag{4}$$

almost everywhere on $[0, T]$. (We write vectors as columns, $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product on $\mathbb{R}^n$, and Df and Dg denote the Jacobian matrices of f and g, respectively.) Any trajectory for which an adjoint variable $\lambda(\cdot)$ exists such that (2)-(4) are satisfied is called an extremal trajectory. The optimality condition (3) determines the control $u(t)$ whenever $\phi(t) := \langle \lambda(t), g(x(t))\rangle \neq 0$; ϕ is called the switching function and $u \equiv -1$ $(u \equiv +1)$ on intervals where ϕ is positive (negative). Trajectories corresponding to these constant controls are called bang arcs and are denoted by X $(=f-g)$ and Y $(=f+g)$, respectively. A concatenation of bang arcs is a *bang-bang trajectory*. Observe that $\langle \lambda(t), f(x(t))\rangle = 0$ at switching times t, i.e., where $\langle \lambda(t), g(x(t))\rangle = 0$. At these times (3) gives no information about the optimal control. If, however, ϕ vanishes on an open interval I, then all the derivatives of ϕ also vanish on I and this may determine the control u. We have

$$\dot{\phi}(t) = \langle \lambda(t), [f, g](x(t))\rangle,$$

$$\ddot{\phi}(t) = \langle \lambda(t), [f + gu, [f, g]](x(t))\rangle,$$

and if $\langle\lambda(t), [g, [f, g]](x(t))\rangle$ does not vanish on I, we can solve for u in $\ddot{\phi}=0$ as follows:

$$u(t) = -\frac{\langle\lambda(t), [f, [f, g]](x(t))\rangle}{\langle\lambda(t), [g, [f, g]](x(t))\rangle}.$$

A control of this type is called singular and the corresponding trajectory is a *singular arc.*

This suggests that concatenations of bang and singular arcs are the natural candidates for trajectories in the boundary of the reachable set (but of course no such regularity statement can be drawn from the Maximum Principle alone). We denote concatenations of bang and singular arcs by the corresponding letter sequence; for instance, we simply write XSY for a concatenation of an X-arc, followed by a singular arc and a Y-trajectory, etc.

3. The main ideas of the technique: the nondegenerate two- and three-dimensional cases. In this section we analyze the (well-known) structure of small-time reachable sets in a nondegenerate situation in dimensions two and three. These cases are easy and give us an opportunity to outline the general ideas of our technique without getting preoccupied with technical details.

Suppose Σ is a system of the form (1) in dimension two and assume that f and g are independent at a reference point p (see Fig. 1). It is clear how the small-time reachable set from p will look. If we let Γ^+ (respectively, Γ^-) be the integral curves of the vector fields $f+g$ (respectively, $f-g$) for positive times, then for sufficiently small T, Reach $(p, \leqq T)$ is the union of Γ^+, Γ^-, and the open sector R between Γ^+ and Γ^- into which $f(p)$ points. It is easy to see that any point in R is reachable from p; for instance, if $q \in R$, just run a trajectory of Σ corresponding to the control $u \equiv +1$ backward in time until it hits Γ^-. The important point is that this is all of the small-time reachable set. This follows immediately from the Maximum Principle since only trajectories corresponding to the constant controls $u \equiv +1$ or $u \equiv -1$ can lie in the boundary of the reachable set. (There cannot be a junction, since then both $\langle\lambda(t), f(x(t))\rangle$ and $\langle\lambda(t), g(x(t))\rangle$ vanish, contradicting the nontriviality of λ.)

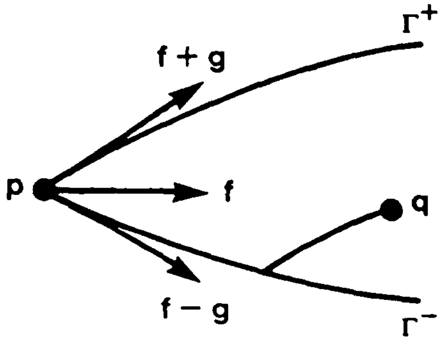


FIG. 1

Generalized to higher dimensions, the quintessence of this argument is to have two hypersurfaces Γ^* and Γ_* which are generated by extremal trajectories, have a common relative boundary and "enclose" a region R. Then, to prove that R is actually the reachable set Reach $(p, \leqq T)$, we must show (i) trajectories cannot leave R through Γ^* or Γ_*, and (ii) all points in the sector are reachable. The latter is immediate if we have a drift vector field f with $f(p) \neq 0$. This is exactly the same argument as in the two-dimensional case. Take any point q inside R and run a trajectory of Σ corresponding to the control $u \equiv 0$ (or for that matter corresponding to any control) backward in time. Since $f(p) \neq 0$, this trajectory will hit Γ^* or Γ_*. So basically (i) must be checked; this is mostly a matter of computing tangent spaces, as will be shown below. This is the general strategy of our technique.

All technical issues left aside for a moment, the key question is how to come up with the surfaces Γ^* and Γ_*. We propose an inductive procedure. Let us explain it at the next step, which is the case of a three-dimensional system Σ, where we assume that f, g, and $[f, g]$ are independent at a reference point p. (This is the example considered by Lobry [14].)

Choose coordinates $x = (x_1, x_2, x_3)$ such that $\langle dx, (f(p), g(p), [f, g](p)) \rangle = \mathrm{Id}$, the identity matrix. The projection of Σ into the (x_1, x_2)-plane is then the two-dimensional system considered above and we know the structure of its small-time reachable set. Our aim is to find two hypersurfaces Γ^* and Γ_* consisting of extremal trajectories that project onto the reachable set $\tilde{R}$ of the two-dimensional system in dimension three. If Γ^* and Γ_* have a common relative boundary that projects onto $\partial \tilde{R}$ and if Γ^* and Γ_* do not intersect in their relative interior, then it is clear that these surfaces "enclose" a region R. Then we must check whether trajectories can leave R. If this is impossible, R is the small-time reachable set.

The Maximum Principle gives preliminary information about Γ^* and Γ_* because it describes necessary conditions for trajectories to lie in the boundary of the reachable set. In this three-dimensional case it actually determines Γ^* and Γ_* precisely, but in higher dimensions this is no longer true. It is then that we will use nilpotent systems as our guide to find candidates for Γ^* and Γ_*. More on that appears in § 4.

Now that we have outlined the general approach, let us also illustrate the basic technical arguments by reproving Lobry's result. It follows from the Maximum Principle that all trajectories that lie on the boundary of the reachable set are bang-bang. For, if the switching function vanishes at some t, i.e., if $\langle \lambda(t), g(x(t)) \rangle = 0$, then also $\langle \lambda(t), f(x(t)) \rangle = 0$, and hence $\dot{\phi}_\Gamma(t) = \langle \lambda(t), [f, g](x(t)) \rangle$ cannot vanish by the independence of f, g, and $[f, g]$ and the nontriviality of λ. For dimensionality reasons it is therefore reasonable to consider the following two surfaces as candidates for Γ^* and Γ_*:

$$\Gamma^* = \{p \exp(s_1(f-g)) \exp(s_2(f+g)) \colon s_i \geqq 0, s_1 + s_2 \text{ small}\},$$

$$\Gamma_* = \{p \exp(t_1(f+g)) \exp(t_2(f-g)) \colon t_i \geqq 0, t_1 + t_2 \text{ small}\}.$$

We write flows of vector fields as exponentials and we let the diffeomorphisms act on the right, i.e., $p \exp(tf)$ denotes the point obtained by following the integral curve of f that passes through p at time zero for t units of time.

It is clear that Γ^* and Γ_* are two-dimensional surfaces with boundary. In both cases the boundary consists of the two curves corresponding to the trajectories of $f+g$ and $f-g$ and the point p. Furthermore, by the Campbell-Hausdorff formula [11]

$$\begin{aligned} &p \exp(s_1(f-g)) \exp(s_2(f+g)) \\ &\quad = p \exp((s_1+s_2)f + (s_2-s_1)g + s_1 s_2 [f, g] + s_1 s_2 \cdot O(T)), \end{aligned}$$

$$p \exp((t_1(f+g)) \exp(t_2(f-g)) = p \exp(t_1+t_2)f + (t_1-t_2)g - t_1 t_2 [f, g] + t_1 t_2 \cdot O(T))$$

where $O(T)$ stands for terms that are linear in the total time T. This shows that Γ^* and Γ_* do not intersect in their relative interior. So Γ^* and Γ_* enclose a region R.

To prove that the enclosed sector R is the small-time reachable set we must show that there cannot be any other points in the reachable set. As in the two-dimensional case we have two options: either we show that we have exhausted all trajectories that

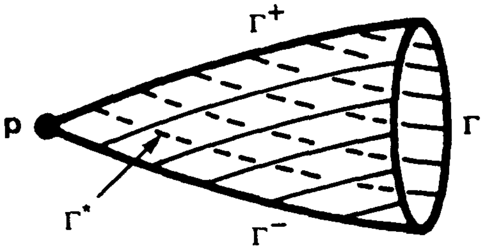


FIG. 2

possibly can lie on the boundary of the reachable set, or we show that trajectories starting at points on Γ^*, Γ_*, Γ^+, or Γ_- cannot leave $R \cup \Gamma^* \cup \Gamma_* \cup \Gamma^+ \cup \Gamma_-$. As it turns out, this is the same argument, only viewed differently.

Let us first show that we have exhausted all possible trajectories that can lie in the boundary of the small-time reachable set, i.e., that such a trajectory is bang-bang with at most one switching. Let γ be a bang-bang trajectory with two switches, say of the form XYX, with junctions p_0 and p_1 at times $t_0 < t_1$. If $\bar\lambda = \lambda(t_1)$, then we have $\langle \bar\lambda, g(p_1)\rangle = 0$ and $\langle \bar\lambda, f(p_1)\rangle = 0$. Also $\langle \lambda(t_0), g(p_0)\rangle = 0$ or, equivalently, if we move g ahead along the flow of the vector field Y we get $\langle \bar\lambda, \exp(-(t_1-t_0)\,\mathrm{ad}\,Y)X(p_0)\rangle = 0$. But $\bar\lambda \neq 0$ and so these three vectors are dependent: p_0 and p_1 are *conjugate points* (Sussmann [22]). Therefore

$$X(p_1) \wedge Y(p_1) \wedge \exp(-\Delta t\,\mathrm{ad}\,Y)X(p_0) = 0$$

i.e., $X(p_1) \wedge Y(p_1) \wedge [X, Y](p_1) + O(\Delta t) = 0$, where $\Delta t = t_1 - t_0$. But such a relation cannot hold in small time by the independence of X, Y, and $[X, Y]$. Similarly it follows that YXY-concatenations cannot satisfy the Maximum Principle.

This computation can also be viewed in the following way. Define a map $F: (t_1, t_2, t_3) \mapsto p \exp(t_1 X) \exp(t_2 Y) \exp(t_3 X)$ for t_i small. Then this map has full rank if $t_i > 0$. For, if we compute the tangent space to the image, but pull back to $p \exp(t_1 X) \exp(t_2 Y)$, we get exactly the vectors $\exp(-t_2\,\mathrm{ad}\,Y)\,X$, Y, and X. Therefore $F(t_1, t_2, t_3)$ is an interior point of the reachable set. Finally, if we pull back the tangent space one step further to $p \exp(t_1 X)$ we have the vectors X, Y, and $\exp(t_2\,\mathrm{ad}\,Y)X = X - t_2[X, Y] + O(t_2^2)$. The minus sign at $[X, Y]$ implies that X-trajectories point inside R at points on Γ^*. Similarly, it follows that Y-trajectories steer the system into R from Γ^*. And this proves that trajectories of the system cannot leave R through Γ^*, Γ_*, Γ^+, or Γ_-. (Because of the Maximum Principle we can restrict ourselves to just looking at these regular controls instead of having to consider arbitrary measurable functions. For, if any trajectory would leave R, then there will also have to be additional trajectories lying on the boundary of the reachable set and these must be bang-bang.)

The structure of the small-time reachable set as a stratified set can easily be described using the following notation. For $n \in \mathbb{N}$ let

$$\begin{aligned} S_{n,-} := \{&p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \\ &\cdots \exp(s_n B) \colon s_i > 0,\ B = X \text{ if } n \text{ is odd},\ B = Y \text{ if } n \text{ is even}\}, \\ S_{n,+} := \{&p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \\ &\cdots \exp(t_n B) \colon t_i > 0,\ B = X \text{ of } n \text{ is even } B = Y \text{ if } n \text{ is odd}\}. \end{aligned}$$

In a nondegenerate situation each of the $S_{n,\pm}$ is a n-dimensional smooth manifold. (Certainly this will be true in all the cases we consider here.) In the three-dimensional case the boundary of the small-time reachable set consists of the two two-dimensional strata $S_{2,\pm}$ which have in their boundary the two one-dimensional strata $S_{1,\pm}$ and the zero-dimensional stratum $S_0=\{p\}$. S_0 also lies in the boundary of $S_{1,\pm}$. If we restrict the total time to be $\leqq T$ we must make the obvious adjustments. In particular, we must add the strata $\hat{S}_{n,\pm}:=S_{n,\pm}\cap \text{Reach}\,(p,T)$ for $n=1,2$.

4. The nondegenerate four-dimensional systems. In this section we determine the geometric structure of the small-time reachable sets from a point p for a system Σ of the form (1) in dimension four, where we assume that the constant controls $u\equiv+1$ and $u\equiv-1$ are not singular. These conditions can easily be expressed in terms of independence assumptions on f, g, and lower-order brackets of f and g. For, a constant control $u\equiv u^0$ is singular on an interval I if and only if there exists an adjoint multiplier λ such that $\langle\lambda,f\rangle$, $\langle\lambda,g\rangle$, $\langle\lambda,[f,g]\rangle$, and $\langle\lambda[f+gu^0,[f,g]]\rangle$ vanish identically on I. By the nontriviality of λ this is impossible if f, g, $[f,g]$, and $[f+gu^0,[f,g]]$ are independent. Therefore in terms of the vector fields X and Y our conditions are equivalent to

(A) X, Y, $[X,Y]$ and $[X,[X,Y]]$ are independent near p;

(B) X, Y, $[X,Y]$ and $[Y,[X,Y]]$ are independent near p.

If we write $[X,[X,Y]]$ as a linear combination of X, Y, $[X,Y]$ and $[Y,[X,Y]]$ as

$$[X,[X,Y]]=\alpha X+\beta Y+\gamma[X,Y]+\delta[Y,[X,Y]],$$

then (A) is equivalent to $\delta\neq 0$.

The cases $\delta>0$ and $\delta<0$ are significantly different: if $\delta>0$ only bang-bang trajectories can lie in the boundary of the reachable set, if $\delta<0$ singular arcs are possible. Intuitively this is clear. If u is singular on an interval I, then (omitting the arguments t and $x(t)$)

$$\begin{aligned}\ddot{\phi}&=\langle\lambda,[f+gu,[f,g]]\rangle\\&=\tfrac{1}{4}\langle\lambda,(1-u)[X,[X,Y]]+(1+u)[Y,[X,Y]]\rangle\\&=\tfrac{1}{4}((1-u)\delta+(1+u))\cdot\langle\lambda,[Y[X,Y]]\rangle\neq 0\end{aligned}$$

and so $u=(\delta+1)/(\delta-1)$. This is an admissible control only if $\delta\leqq 0$. Note that the singular vector field is given in feedback form as

$$S=f+\frac{\delta+1}{\delta-1}g=\frac{1}{1-\delta}X+\frac{-\delta}{1-\delta}Y,\quad \delta<0.$$

4.1. The totally bang-bang case: $\delta>0$. This is the generalization of Lobry's example to dimension four. We treat only the general case here, but we remark that the structure of the small-time reachable set is the same as for a nilpotent system where f, g, $[f,g]$, and $[f,[f,g]]$ form a basis and all other brackets vanish. In appropriate coordinates the latter system is linear.

The key observation again is that the Maximum Principle precisely determines the possible trajectories that can lie in the boundary of the small-time reachable set.

LEMMA 1. *If γ is a trajectory that lies in the boundary of the small-time reachable set, then γ is bang-bang with at most two switches.*

Proof. We first exclude bang-bang trajectories with more switches. Let γ be a $YXYX$-trajectory with switching points p_1, p_2, and p_3 and let s_1, s_2, s_3, s_4 be the length of the times along the respective X-arcs or Y-arcs. At every junction we have $\langle\lambda,X(p_i)\rangle=0$ and $\langle\lambda,Y(p_i)\rangle=0$. This gives rise to four conditions on λ.

If $\tilde{\lambda}$ is the value of the adjoint vector at the switching time at p_2, we have

$$\langle\tilde{\lambda}, X(p_2)\rangle = \langle\tilde{\lambda}, Y(p_2)\rangle = 0,$$

$$\langle\tilde{\lambda}, \exp(-s_2 \operatorname{ad} Y)X(p_1)\rangle = 0,$$

and

$$\langle\tilde{\lambda}, \exp(s_3 \operatorname{ad} X)Y(p_3)\rangle = 0.$$

Again, the nontriviality of $\tilde{\lambda}$ implies that these four vectors are dependent ("conjugate points"). So we get (dividing out s_2 and s_3)

$$\begin{aligned} 0 &= X \wedge Y \wedge \left(\frac{\exp(s_3 \operatorname{ad} X) - 1}{s_3}\right) Y \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y) - 1}{-s_2}\right) X \\ (5) \qquad &= X \wedge Y \wedge [X, Y] + \frac{1}{2}s_3[X, [X, Y]] + O(s_3^2) \wedge -[X, Y] + \frac{1}{2}s_2[Y, [X, Y]] + O(T^2) \\ &= \frac{1}{2}\sigma(s_2, s_3)(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])_{|p_2} \end{aligned}$$

where T is the total time along γ and $O(T^2)$ stands for terms that are quadratic in T; σ is a smooth function of s_2 and s_3. If we express $[X, [X, Y]]$ in terms of X, Y, $[X, Y]$, and $[Y, [X, Y]]$, we see that

$$(6) \qquad \sigma(s_2, s_3) = s_2 + s_3\delta + O(T^2)$$

where δ is evaluated at p_2. In a sufficiently small neighborhood of p, δ is bounded away from zero and so the linear terms dominate quadratic remainders in small time. Hence $\sigma(s_2, s_3)$ is positive for s_i small; in particular, it cannot vanish, a contradiction.

Analogously, if $\tilde{\gamma}$ is a $XYXY$-concatenation with switching points q_1, q_2, and q_3 and if t_1, t_2, t_3, t_4 are the times along the respective trajectories, then we get

$$(7) \qquad \begin{aligned} 0 &= X \wedge Y \wedge \left(\frac{\exp(-t_2 \operatorname{ad} X) - 1}{-t_2}\right) Y \wedge \left(\frac{\exp(t_3 \operatorname{ad} Y) - 1}{t_3}\right) X \\ &= \frac{1}{2}\tau(t_2, t_3)(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])_{|q_2} \end{aligned}$$

where

$$(8) \qquad \tau(t_2, t_3) = -t_3 - t_2\delta + O(T^2)$$

is a smooth function of t_2 and t_3 near the origin. Again, since δ is bounded away from zero near p this function is negative for small times, a contradiction.

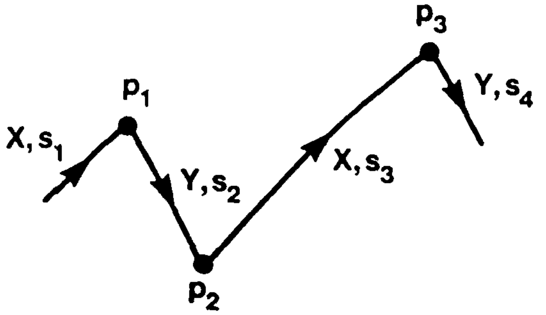


FIG. 3

It now follows that, in fact, any trajectory that lies in the boundary of the small-time reachable set is bang-bang. This is an easy but slightly technical argument. We will do it here rigorously since we will need the computations later on anyway. The point is that we do not have a priori knowledge about regularity properties of the controls, e.g., that they are piecewise constant. This is the case if and only if the zero set $Z(\phi)$ of the switching function ϕ is finite. If it were infinite, then the set N_ϕ of limit points of $Z(\phi)$ would be nonempty. In fact, it is a closed, nowhere dense, perfect set. (If $t_1 < t_2$ are points in $N(\phi)$ then, since ϕ cannot vanish identically, ϕ is different from zero somewhere in (t_1, t_2) and by continuity it is different from zero on a whole interval. It is perfect, i.e., every point $t \in N(\phi)$ is a limit point of points $t_n \in N(\phi)$, $t_n \neq t$, since $N(\phi)$ cannot have isolated points. We can see that this is so, since we know already that bang-bang trajectories with more than three switchings do not lie in the boundary of the small-time reachable set!) Suppose $t_1 < t_2$ are times in $N(\phi)$. There exists a $\tilde{t} \in (t_1, t_2)$ such that $\phi(\tilde{t}) \neq 0$. Let $\tilde{t}_1 := \sup([t_1, \tilde{t}] \cap N(\phi))$ and let $\tilde{t}_2 := \inf([\tilde{t}_1, t_2] \cap N(\phi))$. Then $\tilde{t}_1 < \tilde{t}_2$, $\tilde{t}_i \in N(\phi)$, and $Z(\phi) \cap [\tilde{t}_1, \tilde{t}_2]$ is finite. This implies that γ contains subarcs of the form $*B\cdot$ and $\cdot B*$, where B denotes a bang arc (X or Y), $\cdot$ stands for any switching, and $*$ stands for a junction in $N(\phi)$. Observe that $\dot{\phi}(t) = 0$ if $t \in N(\phi)$. We will now show that none of these concatenations can lie in the boundary of the reachable set and this will prove the lemma.

Without loss of generality we consider a concatenation of the form $*X\cdot$ with switching points p_0 and p_1 and let t be the time along X. Then, if $\tilde{\lambda}$ is the value of the adjoint vector at the switching time corresponding to p_0, we have

$$\langle \tilde{\lambda}, X(p_0) \rangle = \langle \tilde{\lambda}, Y(p_0) \rangle = \langle \tilde{\lambda}, [X, Y](p_0) \rangle = 0.$$

Also $\langle \tilde{\lambda}, \exp(-t \operatorname{ad} X) Y(p_1) \rangle = 0$ and so by nontriviality of $\tilde{\lambda}$ we again get

$$\begin{aligned} 0 &= X \wedge Y \wedge [X, Y] \wedge Y - t[X, Y] + \tfrac{1}{2}t^2[X, [X, Y]] + O(t^3) \\ &= \tfrac{1}{2}t^2(1 + O(t))(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])_{|p_0}. \end{aligned} \tag{9}$$

This cannot hold in small time. Analogously it follows that no $*B\cdot$ or $\cdot B*$ concatenation can lie in the boundary of the small-time reachable set if $\delta \neq 0$. This proves the lemma (and note that the argument is valid in general under assumptions (A) and (B)). □

It is now clear that the surfaces Γ^* and Γ_* must be as follows:

$$\Gamma^* = \{p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \colon s_i \geqq 0, \text{small}\},$$

$$\Gamma_* = \{p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \colon t_i \geqq 0, \text{small}\}.$$

Γ^* and Γ_* are three-dimensional surfaces with common boundary C that has precisely the structure of the boundary of the small-time reachable set in dimension three. It is the union of two two-dimensional surfaces made out of XY- and YX-trajectories respectively, glued together along the X- and Y-trajectories.

We will now show that Γ^* and Γ_* do not intersect away from C, in particular that they enclose an open region that will be the interior of the small-time reachable set.

Definition. We say a point q is an entry point (respectively, an exit point) of a (closed) set S for a vector field Z if for some $\varepsilon > 0$, $S \cap \{q \exp(tZ) \colon -\varepsilon \leqq t \leqq 0\} = \{q\}$ (respectively, if $S \cap \{q \exp(tZ) \colon 0 \leqq t \leqq \varepsilon\} = \{q\}$).

Lemma 2. *For sufficiently small T the points in Γ^* are entry points for the small-time reachable set from p for $[Y, [X, Y]]$. The points in Γ_* are exit points.*

Proof. If q is an exit (entry) point for $\operatorname{Reach}(p, \leqq T)$ that does not lie in $\operatorname{Reach}(p, T)$, i.e., exit or entry is not due to the time restriction, then the corresponding trajectory is extremal and the adjoint multiplier satisfies the transversality condition $\langle\lambda,[Y,[X,Y]](q)\rangle\leqq 0$ $(\langle\lambda,[Y,[X,Y]](q)\rangle\geqq 0)$. We claim that necessarily

$$q\in\Gamma_* \ (q\in\Gamma^*).$$

Recall that the second derivative of the switching function is given by

$$\begin{aligned}\ddot{\phi}(t)&=\langle\lambda(t),[f+gu,[f,g]](x(t))\rangle\\ &=\tfrac{1}{4}(1-u(t))\langle\lambda,[X,[X,Y]](x(t))\rangle\\ &\quad+\tfrac{1}{4}(1+u(t))\langle\lambda,[Y,[X,Y]](x(t))\rangle.\end{aligned}\tag{10}$$

Expressing $[X,[X,Y]]$ in terms of X, Y, $[X,Y]$, and $[Y,[X,Y]]$, we get a linear combination of terms $\langle\lambda,X\rangle$, $\langle\lambda,Y\rangle$, $\langle\lambda,[X,Y]\rangle$, and $\langle\lambda,[Y,[X,Y]]\rangle$, where the coefficient at $\langle\lambda,[Y,[X,Y]]\rangle$ is

$$\tfrac{1}{2}(1-u)\delta+\tfrac{1}{2}(1+u)\geqq \mathrm{Min}\,(1,\delta)>0.$$

Suppose γ is a bang-bang trajectory with two junctions. Then the two junctions determine a multiplier λ up to a positive constant multiple. Normalize such that $\|\lambda(0)\|_2=1$. Because γ has two junctions $\langle\lambda,X\rangle$, $\langle\lambda,Y\rangle$, and $\langle\lambda,[X,Y]\rangle$ vanish somewhere on $[0,T]$, $T=t_1+t_2+t_3$. For sufficiently small T these functions will be bounded in absolute value on $[0,T]$ by any $\varepsilon>0$. Because of (B) $|\langle\lambda(t),[Y,[X,Y]](x(t))\rangle|$ can be bounded away from zero on $[0,T]$. By choosing ε, i.e., T small enough, $\langle\lambda,[Y,[X,Y]]\rangle$ dominates all other terms in (10), that is, we have in small time: $\ddot{\phi}$ has constant sign equal to sign $(\langle\lambda,[Y,[X,Y]]\rangle)$. But $\langle\lambda,[Y,[X,Y]]\rangle>0$ allows only for XYX-trajectories and $\langle\lambda,[Y,[X,Y]]\rangle<0$ permits only YXY-concatenations. This proves our claim.

We still need to show that points in Γ^* and Γ_* in fact have these optimization properties. Suppose γ is a XYX trajectory. Then the tangent space at the endpoint is spanned by X, $\exp(-t_3\,\mathrm{ad}\,X)Y$ and $\exp(-t_3\,\mathrm{ad}\,X)\exp(-t_2\,\mathrm{ad}\,Y)X$. Note that $[Y,[X,Y]]$ always points to one side of the tangent space since

$$\begin{aligned}&X\wedge\exp(-t_3\,\mathrm{ad}\,X)Y\wedge\exp(-t_3\,\mathrm{ad}\,X)\exp(-t_2\,\mathrm{ad}\,Y)X\wedge[Y,[X,Y]]\\ &\quad=-t_2\Bigg(X\wedge\exp(-t_3\,\mathrm{ad}\,X)Y\wedge\exp(-t_3\,\mathrm{ad}\,X)\left(\frac{\exp(-t_2\,\mathrm{ad}\,Y)-1}{-t_2}\right)X\\ &\qquad\qquad\wedge[Y,[X,Y]]\Bigg)\\ &\quad=t_2(X\wedge Y-t_3[X,Y]+O(t_3^2)\wedge[X,Y]+O(T)\wedge[Y,[X,Y]])\\ &\quad=t_2(1+O(T))(X\wedge Y\wedge[X,Y]\wedge[Y,[X,Y]]).\end{aligned}\tag{11}$$

If we write the defining equations for Γ^* and Γ_* in terms of canonical coordinates of the second kind, that is, as products of the flows of the vector fields X, Y, $[X,Y]$, $[Y,[X,Y]]$ in the form

$$p\exp(x_1X)\exp(x_2Y)\exp(x_3[X,Y])\exp(x_4[Y,[X,Y]]),\tag{12}$$

then this implies that we can think of Γ^* as the graph of a function $x_4=\psi(x_1,x_2,x_3)$. It also follows from (12) that the integral curve of $[Y,[X,Y]]$ through p and the compact set Reach (p,T) are disjoint for small positive T. Therefore, given T, there exists a $\tilde{T}\leqq T$ with the following property. Any integral curve of $[Y,[X,Y]]$ that passes through a point on $\Gamma^*(\tilde{T})$, the set of all trajectories in Γ^* of total time $\leqq\tilde{T}$, does not meet Reach (p,T). This implies that the points on $\Gamma^*(\tilde{T})$ are entry points for the small-time reachable set. For, if $q\in\Gamma^*(\tilde{T})$ is not an entry point, then by compactness there exists an entry point of Reach $(p, \leqq T)$ of the form $q \exp r[Y,[X, Y]]$. Since this flow does not meet Reach (p, T) this point must lie on Γ^* and this contradicts the graph property. Analogously the result follows for Γ_*. □

An easy computation shows that, if Γ^* and Γ_* would intersect away from C, then it would have to happen transversally. This would contradict Lemma 2.

The geometric structure of the small-time reachable set is now clear. It is the exact analogue of Figs. 1 and 2 in four dimensions. Its boundary consists of the surfaces Γ_* and Γ^* that match up along C, the set of points reachable by a bang-bang trajectory with at most one switch. The open region enclosed by Γ^* and Γ_* is the interior of the reachable set. A stratification of its boundary is given by S_0 and $S_{n,\pm}$ for $n=1, 2, 3$ (see § 3).

Remark. This qualitative structure of the small-time reachable set for a totally bang-bang system generalizes to arbitrary dimensions under the conditions of Krener's and Sussmann's nonlinear bang-bang theorem [19]. Suppose that the vector fields f and $\mathrm{ad}^i f(g)$, $i=0, \cdots, n-1$ are independent at p and that for $i=0, \cdots, n-1$ there exist smooth functions α_{ij} and β_i with $|\beta_i(p)|<1$ such that

$$[g, \mathrm{ad}^i f(g)]=\sum_{j=0}^{i} \alpha_{ij}\, \mathrm{ad}^j f(g)+\beta_i\, \mathrm{ad}^{i+1} f(g).$$

Then it follows that for sufficiently small-time T all trajectories that lie in the boundary of the reachable set from p are bang-bang with at most n switchings. A stratification of the boundary is given by the strata $S_0=\{p\}$ and $S_{k,\pm}$, $k=1, \cdots, n$. In particular, points in $S_{n,+}$ are exit points of the reachable set for $(-1)^{n-1}\mathrm{ad}^{n-1} f(g)$, points in $S_{n,-}$ are entry points. Given the results on the structure of trajectories in the boundary, this is a straightforward generalization of the argument above. All the difficult work has been carried out by Sussmann in [19], specifically in the proof of Lemma 3 there.

4.2. The bang-bang singular case: $\delta<0$. This case is a nontrivial extension of Lobry's result. Here not all the extremal trajectories actually lie in the boundary of the small-time reachable set. It is therefore not clear how we should choose Γ^* and Γ_*. We now use the structure of the small-time reachable set for the corresponding free nilpotent system as a guide. The only reasonable nilpotent approximation to choose is one where all brackets of orders greater than or equal to 4 vanish. Note that f, g, $[f, g]$, and $[g,[f, g]]$ are always independent in this case. Since we want to work with a system as simple as possible, we also assume $[f,[f, g]] \equiv 0$. This is an equality relation in the third-order Lie jet, but in a slightly more general setup (weighted Lie algebra) this would be a free nilpotent system. Therefore we refer to this system as the "free" nilpotent case. We will first analyze a model of this "free" nilpotent case, and then we will show that the general case has the same qualitative behavior.

4.2.1. The reachable set in the "free" nilpotent case. To simplify some computations we restrict ourselves to the following model $\tilde{\Sigma}$:

$$\dot{x}_0=1, \quad \dot{x}_1=u, \quad \dot{x}_2=x_1, \quad \dot{x}_3=\tfrac{1}{2}x_1^2. \tag{13}$$

Note that $[g, f](x)=(\partial/\partial x_2)+x_1(\partial/\partial x_3)$, $[g,[g, f]] \equiv \partial/\partial x_3$ and all other brackets vanish identically. It is clear that the qualitative structure of the reachable set from the origin at any time is the same as for the small-time reachable set: one is a rescaling of the other. (If u is a control defined on $[0, T]$ and x is the corresponding trajectory, then the time 1 reachable set can be obtained from the time T reachable set by letting $\bar{u}(t):=u(t/T)$ and $\bar{x}_i(t):=T^i x_i(t/T)$ for $i=1, 2, 3$.) To determine the reachable set it therefore suffices to look at time slices $T=$ constant, and without loss of generality we can assume $T=1$.

If $\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)^T$ is an adjoint vector for an extremal trajectory $x(\cdot)$, then λ_1 is the switching function and

$$\dot{\lambda}_1 = -\lambda_2 - \lambda_3 x_1, \quad \dot{\lambda}_2 = 0, \quad \dot{\lambda}_3 = 0,$$

and in particular $\ddot{\lambda}_1 = \lambda_3 u$, i.e., $u \equiv 0$ is the only singular control. Note that, if $\lambda_3 = 0$, then λ_1 is a linear function and the extremal trajectory is uniquely determined. By a theorem of Bressan [2] this implies that the reachable set is convex in direction of $(0, 0, 0, 1)^T$ or equivalently in the direction of $[g, [g, f]] = \frac{1}{2}[X, [X, Y]]$; that is, if (p_0, p_1, p_2, a) and (p_0, p_1, p_2, b) lie in the reachable set, then the whole segment $\{(p_0, p_1, p_2, c): a \leqq c \leqq b\}$ lies in the reachable set. It is therefore clear what the surfaces Γ^* and Γ_* have to be: Γ^* consists of trajectories which are exit points for $[X, [X, Y]]$ and Γ_* of those which are entry points. Equivalently, we can speak of trajectories that maximize/minimize the coordinate x_3.

For extremal trajectories that give rise to entry/exit points for $[X, [X, Y]]$, an additional transversality condition was to hold. One of the directions $\pm[X, [X, Y]]$ can be separated from an approximating cone to the reachable set at this point. In our case these conditions simply say that $\lambda_3 \geqq 0$ for trajectories that minimize x_3 and $\lambda_3 \leqq 0$ for those that maximize x_3. In particular $\lambda_3 = 0$ for those that do both and these trajectories are bang-bang with at most one switching. So again the common boundary of Γ^* and Γ_* will be a set C that has the structure of the boundary of the small-time reachable set in dimension three.

We now determine Γ_*. We can assume $\lambda_3 > 0$ and without loss of generality normalize λ_3 to 1. Thus, $\ddot{\lambda}_1 = -u$ and so λ_1 is strictly convex and positive along X, strictly concave and negative along Y. Singular controls satisfy the generalized Legendre-Clebsch condition [13]: $\langle \lambda, [g, [f, g]] \rangle = -\lambda_3 < 0$. It follows that the only extremal trajectories are concatenations of a bang arc, followed by a singular arc and another bang arc. We now restrict to the time slice $T = 1$. Define

$$\Gamma_{-0-} := \{0 \exp(s_1 X) \exp(s_2 f) \exp(s_3 X): s_i \geqq 0,\ s_1 + s_2 + s_3 = 1\},$$

$$\Gamma_{-0+} := \{0 \exp(s_1 X) \exp(s_2 f) \exp(s_3 Y): s_i \geqq 0,\ s_1 + s_2 + s_3 = 1\},$$

$$\Gamma_{+0-} := \{0 \exp(t_1 Y) \exp(t_2 f) \exp(t_3 X): t_i \geqq 0,\ t_1 + t_2 + t_3 = 1\},$$

$$\Gamma_{+0+} := \{0 \exp(t_1 Y) \exp(t_2 f) \exp(t_3 Y): t_i \geqq 0,\ t_1 + t_2 + t_3 = 1\}.$$

We will show that these are two-dimensional surfaces with boundary which match up and together form Γ_* with

$$\begin{aligned}\partial\Gamma_* = {} & \{0 \exp(s_1 X) \exp(s_2 Y): s_i \geqq 0,\ s_1 + s_2 = 1\} \\ & \cup \{0 \exp(t_1 Y) \exp(t_2 X): t_i \geqq 0,\ t_1 + t_2 = 1\}.\end{aligned}$$

LEMMA 3. *Each of the sets $\Gamma_{\pm0\pm}$ is a two-dimensional surface with boundary. For any two of them the images of the open simplices are disjoint. Furthermore,*

$$\Gamma_{-0-} \cap \Gamma_{-0+} = \Gamma_{-0} = \{0 \exp(s_1 X) \exp(s_2 f): s_i \geqq 0,\ s_1 + s_2 = 1\},$$

$$\Gamma_{-0-} \cap \Gamma_{+0-} = \Gamma_{0-} = \{0 \exp(s_1 f) \exp(s_2 X): s_i \geqq 0,\ s_1 + s_2 = 1\},$$

$$\Gamma_{-0-} \cap \Gamma_{+0+} = \Gamma_0 = \{0 \exp(sf): 0 \leqq s \leqq 1\} = \Gamma_{-0+} \cap \Gamma_{+0-},$$

$$\Gamma_{-0+} \cap \Gamma_{+0+} = \Gamma_{0+} = \{0 \exp(s_1 f) \exp(s_2 Y): s_i \geqq 0,\ s_1 + s_2 = 1\},$$

$$\Gamma_{+0-} \cap \Gamma_{+0+} = \Gamma_{+0} = \{0 \exp(s_1 Y) \exp(s_2 f): s_i \geqq 0,\ s_1 + s_2 = 1\}.$$

Graphically, these relations can be illustrated as shown in Fig. 4.

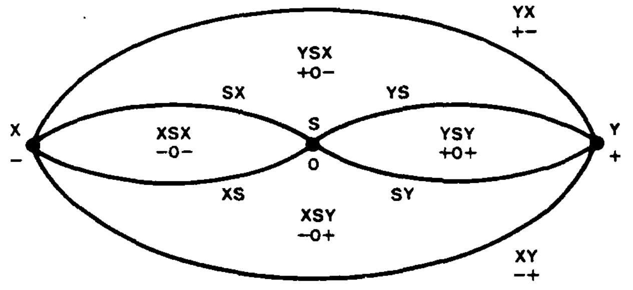


FIG. 4

The proof of the lemma consists of straightforward computations that we shall only illustrate in one case. It is easy to see that all the maps are regular with rank 2 in the interior, and it is clear how the maps behave on the boundary. So the $\Gamma_{\pm 0 \pm}$ are two-dimensional surfaces with boundary. To prove that the images of the open simplex under different maps are disjoint, we choose a way that does not use the specific form of the equations, but works with a basis provided by the vector fields f, g, $[f, g]$, and $[g, [f, g]]$. This also gives an idea how the analogous argument in the general case runs. We rewrite the defining equations in terms of canonical coordinates of the second kind as products of the flows of the vector fields f, g, $[f, g]$, and $[g, [f, g]]$. Since in this case

$$\exp(f+g) = \exp([g,[f,g]]/3)\exp([f,g]/2)\exp(g)\exp(f), \tag{14}$$

we get, for instance, for Γ_{+0+}:

$$\begin{aligned} &0 \exp(t_1(f+g)) \exp(t_2 f) \exp(t_3(f+g)) \\ &\quad = 0 \exp((\tfrac{1}{3}t_1^3[g,[f,g]]) \exp(\tfrac{1}{2}t_1^2[f,g]) \exp(t_1 g) \exp((t_1+t_2)f)) \\ &\qquad \times \exp(\tfrac{1}{3}t_3^3[g,[f,g]]) \exp(\tfrac{1}{2}t_3^2[f,g]) \exp(t_3 g) \exp(t_3 f) \\ &\quad = 0 \exp((\tfrac{1}{6}(t_1+t_3)^3 + t_2 t_3(t_1+\tfrac{1}{2}t_3))[g,[f,g]]) \exp((\tfrac{1}{2}(t_1+t_3)^2 + t_2 t_3)[f,g]) \\ &\qquad \times \exp((t_1+t_3)g) \exp(f). \end{aligned}$$

Analogously we have for Γ_{-0+}:

$$\begin{aligned} &0 \exp(s_1(f-g)) \exp(s_2 f) \exp(s_3(f+g)) \\ &\quad = 0 \exp((\tfrac{1}{3}s_1^3 - s_1^2 s_3 + \tfrac{1}{3}s_3^3 + \tfrac{1}{2}s_2 s_3^2 - s_1 s_2 s_3)[g,[f,g]]) \\ &\qquad \times \exp((-\tfrac{1}{2}s_1^2 + \tfrac{1}{2}s_3^2 + (s_1+s_2)s_3)[f,g]) \exp((s_3 - s_1)g) \exp(f). \end{aligned}$$

A simple computation shows that the equations we obtain by equating the coordinates have no positive solution. Similarly this is shown for all pairs of surfaces. The statements about the intersections are then clear. □

This shows that Γ_* is a two-dimensional stratified set with its one-dimensional relative boundary $\partial\Gamma_*$ made out of bang-bang trajectories with at most one switching. Figure 4 gives a precise description of the stratification. We now show that the points on Γ_* are, in fact, the points that have the smallest x_3 coordinate among all points of Reach (0, 1) with a fixed (x_0, x_1, x_2).

Let us first compute the tangent spaces to the surfaces $\Gamma_{\pm 0 \pm}$. Note that in each case the pullback of the tangent space to the endpoint of the singular arc simply consists of the space spanned by the vectors g and $[f, g]$ evaluated there (remember that we are working in the time slice $T = 1$). This implies that $[X, [X, Y]] = 2[g, [g, f]]$ always points to one side of the tangent space. In fact,

$$\exp(-t \operatorname{ad}(f \pm g))g \wedge \exp(-t \operatorname{ad}(f \pm g))[f, g] \wedge [g, [f, g]] \equiv 1(g \wedge [f, g] \wedge [g, [f, g]]).$$

In the limit this also holds for the one-dimensional strata. Therefore $[g, [g, f]]$ always points to one side of the stratified surface Γ_*. It is easy to see that, in fact, we can think of Γ_* as the graph of a piecewise defined function $x_3 = \psi(x_1, x_2)$. (The projections of the images onto (x_1, x_2) intersect only along the projections of the intersections of the surfaces $\Gamma_{\pm 0 \pm}$.) Since we have exhausted all possible extremal trajectories that can minimize the coordinate x_3 with Γ_*, it is now clear that given $(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \Gamma_*$ any other point $(x_1, x_2, x_3) \in \text{Reach}(0, 1)$ with $x_1 = \bar{x}_1$ and $x_2 = \bar{x}_2$ must satisfy $x_3 > \bar{x}_3$. This concludes the analysis of Γ_*.

Next we will determine Γ^*. Here we can assume $\lambda_3 = -1$ and so $\ddot{\lambda}_1 = u$, i.e., the switching function ϕ is convex when ϕ is negative and concave when ϕ is positive. This clearly suggests bang-bang extremals. However, now the situation is significantly different from all previous cases: it will turn out that the times along bang arcs are no longer free, which in turn will mean that we cannot a priori exclude bang-bang trajectories with a large number of switchings. In general, it is a very difficult problem to eliminate extremal trajectories with a large number of switchings (cf. [4] or [16]). It turns out that in our approach we do not even have to address this issue.

Let us start by showing that the times along bang arcs can no longer vary freely. Suppose we have a concatenation of a Y-trajectory followed by an X-arc with switchings at the beginning and the end $(\cdot XY \cdot)$. Call the switching points p_0, p_1, and p_2 and let s and t be the times along X and Y, respectively. Then p_0, p_1, and p_2 are conjugate points and therefore

$$\begin{aligned} 0 &= \exp(-s \operatorname{ad} X) Y \wedge X \wedge Y \wedge \exp(t \operatorname{ad} Y) X \\ &= \left(\frac{\exp(-s \operatorname{ad} X) - 1}{-s}\right) Y \wedge X \wedge Y \wedge \left(\frac{\exp(t \operatorname{ad} Y) - 1}{t}\right) X \\ &= X \wedge Y \wedge [X, Y] + s[g, [f, g]] \wedge [Y, X] - t[g, [f, g]] \\ &= (s - t)(X \wedge Y \wedge [X, Y] \wedge [g, [f, g]]). \end{aligned} \tag{15}$$

Hence $s = t$ and the same is true for a $\cdot YX \cdot$-concatenatiou. Therefore, so as not to violate the Maximum Principle, and since we do not expect any degeneracies in the structure of the reachable set, we restrict ourselves to the following two surfaces:

$$\tilde{\Gamma}^- = \{0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \colon s_i \geqq 0,\ s_1 + s_2 + s_3 = 1,\ s_1 \leqq s_2,\ s_3 \leqq s_2\},$$

$$\tilde{\Gamma}^+ = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \colon t_i \geqq 0,\ t_1 + t_2 + t_3 = 1,\ t_1 \leqq t_2,\ t_3 \leqq t_2\}.$$

Our aim is to build Γ^* out of trajectories from $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$. However, as they are at the moment, we still have too many extremal trajectories. The surfaces $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ have a nontrivial intersection $\tilde{\gamma}$. To see this let us rewrite the defining maps in terms of canonical coordinates as follows:

$$\begin{aligned} 0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) = {} & 0 \exp(s_1 s_2 (s_2 - s_1)[g, [f, g]]) \exp(s_1 s_2 [X, Y]) \\ & \times \exp(s_2 Y) \exp((s_1 + s_3) X), \end{aligned}$$

$$0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) = 0 \exp(t_2 t_3(2t_1 - t_2 + t_3)[g,[f,g]]) \exp(t_2 t_3[X,Y]) \times \exp((t_1 + t_3)Y) \exp(t_2 X).$$

If we equate the coordinates, it follows easily that $s_1 = t_3$, $s_2 = t_2$, and $s_3 = t_1$. It follows that $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ also intersect along the one-dimensional curve

$$\tilde{\gamma} = \{0 \exp(sX) \exp(Y/2) \exp((\tfrac{1}{2} - s)X) : 0 \leqq s \leqq \tfrac{1}{2}\}.$$

We need to analyze the intersection more closely. Let

$$q = 0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \in \tilde{\gamma}.$$

Then the tangent space to $\tilde{\Gamma}^-$ at q is spanned by (recall that $s_3 = 1 - s_1 - s_2$)

$$\exp(-s_3 \operatorname{ad} X) \exp(-s_2 \operatorname{ad} Y)X - X = s_2([X,Y] + (2s_3 - s_2)[g,[f,g]]),$$
$$\exp(-s_3 \operatorname{ad} X)Y - X = 2g - s_3[X,Y] - s_3^2[g,[f,g]].$$

The point q also lies on $\tilde{\Gamma}^+$ and a tangent vector to $\tilde{\Gamma}^+$ at q is

$$\ell = \exp(-t_3 \operatorname{ad} Y)X - Y = -2g + t_3[X,Y] - t_3^2[g,[f,g]].$$

In the intersection $t_3 = s_1 =: s$, $s_2 = \frac{1}{2}$ and $s_3 = \frac{1}{2} - s$. Thus

$$T_q\tilde{\Gamma}^- \wedge \ell = A(2g \wedge [X,Y] \wedge [g,[f,g]])$$

where

$$A = \begin{vmatrix} 1 & s - \frac{1}{2} & -(s - \frac{1}{2})^2 \\ .0 & 1 & \frac{1}{2} - 2s \\ -1 & s & -s^2 \end{vmatrix} = 2s(s - \tfrac{1}{2}) \leqq 0.$$

Hence $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect transversally except at the endpoints of $\tilde{\gamma}$ ($s = 0$, $s = \frac{1}{2}$). Observe that the endpoints are characterized by the condition that the conjugate point relation $s = t$ ($= \frac{1}{2}$) holds. We need to know which surface has a larger x_3-coordinate. It follows from

$$T_q\tilde{\Gamma}^- \wedge [g,[g,f]] \equiv -2g \wedge [X,Y] \wedge [g,[f,g]]$$

that ℓ and $[g,[g,f]]$ point to the same side of $\tilde{\Gamma}^-$ at q. Observe that $x_1 = 0$ for points on $\tilde{\gamma}$. Since the coefficient of ℓ at g is negative, the points of $\tilde{\Gamma}^+$ for which $x_1 < 0$ have a larger x_3-coordinate than those points on $\tilde{\Gamma}^-$. Conversely for $x_1 > 0$ the x_3-coordinate of points on $\tilde{\Gamma}^-$ is larger. Therefore we define

$$\Gamma^- := \{0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) : s_i \geqq 0, s_1 + s_2 + s_3 = 1, s_2 \geqq \tfrac{1}{2}\},$$
$$\Gamma^+ := \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) : t_i \geqq 0, t_1 + t_2 + t_3 = 1, t_2 \geqq \tfrac{1}{2}\}.$$

Observe that Γ^- has the Y-trajectory in its boundary and that the X-trajectory lies in the boundary of Γ^+. Define $\Gamma^* := \Gamma^- \cup \Gamma^+$. It follows from above that $[X,[X,Y]] = 2[g,[g,f]]$ always points to one side of $\tilde{\Gamma}^-$, and similarly this holds for $\tilde{\Gamma}^+$. Since $x_1 \geqq 0$ for points in Γ^-, $x_1 \leqq 0$ for points in Γ^+ and $x_1 = 0$ exactly on the intersection, it follows that Γ^* is a piecewise defined function $x_3 = \psi(x_1, x_2)$.

It is obvious that $\partial\Gamma^*$ consists of all trajectories that are bang-bang with at most one switching, i.e., $\partial\Gamma^* = \partial\Gamma_*$. Graphically, the structure is illustrated in Fig. 5.

By directional convexity it is clear that the whole set R between Γ_* and Γ^* lies in Reach (0, 1). We need to show that it lies nowhere else. The points of $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ that we deleted lie in the interior of R. (We delcted those points on $\tilde{\Gamma}^+$, respectively, $\tilde{\Gamma}^-$ that lie below $\tilde{\Gamma}^-$, respectively, $\tilde{\Gamma}^+$ in the direction of $[X,[X,Y]]$.) But this implies that the endpoints of bang-bang trajectories with more than two switchings lie in the interior of the reachable set. Suppose we have an extremal $XYXY$-trajectory with

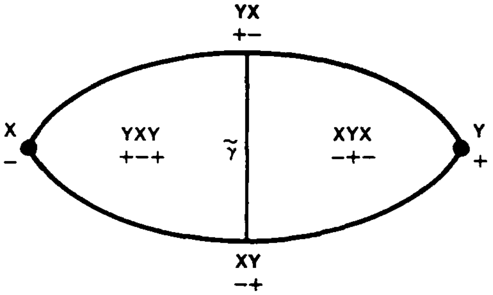


FIG. 5

times s_1, s_2, s_3, and s_4 along the trajectories. Then $s_2 = s_3$ by the conjugate point relation, and thus $s_2 < s_1 + s_3$. By the invariance of the structure of the reachable set it follows that $0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \in \operatorname{int} \operatorname{Reach}(0, s_1 + s_2 + s_3)$. (This is a point of the type we deleted!) Hence the trajectories that define Γ^+ and Γ^- are the only extremal trajectories that can lie on the boundary of the reachable set. This proves $R = \operatorname{Reach}(0, 1)$.

Summary. For every time t the time $-t-$ reachable set is a stratified set that is topologically a sphere. Its boundary consists of two hemispheres $\Gamma^*(t)$ and $\Gamma_*(t)$ whose common relative boundary $\partial\Gamma^*(t)$ consists of all points reachable in time t by a bang-bang trajectory with at most one switch. $\Gamma^*(t)$ consists of all bang-bang trajectories with at most two switchings for which the time along the intermediate arc is greater than or equal to the sum of the times of the adjacent arcs. $\Gamma_*(t)$ consists of all trajectories that are concatenations of a bang arc, followed by a singular arc and another bang arc, where the times along these trajectories are free subject to $0 \leqq \text{time} \leqq t$. The stratification of its boundary is given in Figs. 4 and 5.

4.2.2. The general case. We now show that the qualitative structure of the small-time reachable set does not change in the general case. Clearly, some of the arguments will have to be adjusted; for instance, the correct generalization of the arguments using directional convexity now use the integral curves of $[X, [X, Y]]$. However, finding a general version for the explicit computations in the analysis of the bang-bang extremal trajectories is crucial.

We first define Γ_*. Recall that the singular control is given in feedback form as $u = (\delta + 1)/(\delta - 1)$ and since $\delta < 0$ we have no problems with u hitting the control constraint $|u| = 1$ in small time. Let $\rho = 1/(1 - \delta)$, $\rho \in (0, 1)$, and let $S := f + (\delta + 1)/(\delta - 1)g = \rho X + (1 - \rho)Y$, be the singular vector field. Define

$$\Gamma_{-s-} := \{p \exp(s_1 X) \exp(s_2 S) \exp(s_3 X) \colon s_i \geqq 0, \text{small}\},$$

$$\Gamma_{-s+} := \{p \exp(s_1 X) \exp(s_2 S) \exp(s_3 Y) \colon s_i \geqq 0, \text{small}\},$$

$$\Gamma_{+s-} := \{p \exp(t_1 Y) \exp(t_2 S) \exp(t_3 X) \colon t_i \geqq 0, \text{small}\},$$

$$\Gamma_{+s+} := \{p \exp(t_1 Y) \exp(t_2 S) \exp(t_3 Y) \colon t_i \geqq 0, \text{small}\},$$

$$\Gamma_* := \Gamma_{-s-} \cup \Gamma_{-s+} \cup \Gamma_{+s-} \cup \Gamma_{+s+}.$$

If we replace f by S in Lemma 3, then the statement stays true verbatim for $\Gamma_{\pm s\pm}$ instead of $\Gamma_{\pm 0\pm}$. (The computations are a straightforward though somewhat messy extension of the computation in the "free" nilpotent case and we omit them.) So again Γ_* is a stratified two-dimensional surface; its one-dimensional relative boundary $\partial\Gamma_*$ is made out of the bang-bang trajectories with at most one switching.

LEMMA 4. *For sufficiently small T the points on Γ_* are entry points of* Reach $(p, \leqq T)$ *for* $[X,[X, Y]]$.

Proof. The strategy is the same as in the proof of Lemma 2. We first show that the extremals on Γ_* satisfy the necessary transversality condition for entry points (which are not due to the time constraint). Then we show that Γ_* actually is a graph with the coefficient of the flow of $[X,[X, Y]]$ as dependent variable. As in Lemma 2 this suffices to prove our result.

If γ is any trajectory containing a singular arc then, for sufficiently small time, $\langle\lambda,[X,[X, Y]]\rangle$ will dominate $\langle\lambda, X\rangle$, $\langle\lambda, Y\rangle$, and $\langle\lambda,[X, Y]\rangle$, in particular, it has constant sign. Along the singular arc $\langle\lambda,[X,[X, Y]]\rangle = 2\delta/(1-\delta)\,\langle\lambda,[g,[X, Y]]\rangle$ and the generalized Legendre-Clebsch condition implies that $\langle\lambda,[X,[X, Y]]\rangle$ is positive. This shows that points in Γ_* satisfy the necessary transversality condition. An argument analogous to the one made in the proof of Lemma 1 shows that, in fact, any extremal trajectory for which $\langle\lambda,[X,[X, Y]]\rangle$ is positive has to be of the form BSB, that is, we have exhausted all possible candidates. To prove that indeed each point on Γ_* has the entry property, we show again that we can think of Γ_* as the graph of a piecewise defined function $x_3 = \psi(x_0, x_1, x_2)$, where (x_0, x_1, x_2, x_3) are canonical coordinates of the second kind, and x_3 is the coefficient at the flow of $[X,[X, Y]]$. Let us consider, for instance, $\Gamma_{+,-}$. It is easier to compute the pullback of the tangent space to the endpoint of the singular arc. It is spanned by X, S, and $\exp(-t_2 \operatorname{ad} S)X$. Note that $S = \rho X + (1+\rho)Y$ and it follows by induction that $\operatorname{ad}^n S(X) = \alpha_n X + \beta_n Y + \gamma_n[X, Y]$ with smooth functions $\alpha_n, \beta_n, \gamma_n$:

$$\begin{aligned}[S, \operatorname{ad}^{n-1} S(X)] &= [\rho X + (1-\rho)Y, \alpha_{n-1}X + \beta_{n-1}Y + \gamma_{n-1}[X, Y]] \\ &= \gamma_{n-1}\underbrace{(\rho[X,[X, Y]] + (1-\rho)[Y,[X, Y]])}_{=\rho(\alpha X + \beta Y + \gamma[X, Y]).} + f, g \text{ or } [f, g] \text{ terms}\end{aligned}$$

Also $[S, X] = [\rho X + (1-\rho)Y, X] = 2L_x(\rho)g + (\rho - 1)[X, Y]$. Therefore

$$X \wedge S \wedge \exp(-t_2 \operatorname{ad} S)X = (1-\rho)^2 t_2(1 + O(t_2)) \cdot (f \wedge g \wedge [X, Y]).$$

Now if we take the wedge-product with $[X,[X, Y]]$ pulled back along X, t_3 this yields

$$\begin{aligned}&X \wedge S \wedge \exp(-t_2 \operatorname{ad} S)X \wedge \exp(t_3 \operatorname{ad} X)([X,[X, Y]]) \\ &\quad = (1-\rho)^2 t_2(1 + O(T)) \cdot (f \wedge g \wedge [f, g] \wedge [X,[X, Y]])\end{aligned}$$

and there are no problems with dominance since t_2 factors. Hence $[X,[X, Y]]$ always points to one side of $\Gamma_{+,-}$ in the interior. Analogously it follows for the other surfaces. By continuity this also follows for the one-dimensional strata. Straightforward but slightly more tedious computations show also that the projections of the relative interiors of the sets $\Gamma_{\pm,\pm}$ onto (x_0, x_1, x_2)-space are pairwise disjoint. Therefore Γ_* is a graph in canonical coordinates. This proves the lemma. □

The analysis of the bang-bang extremals is more difficult. We start by computing the conjugate point relations. Suppose γ is a $\cdot XYX\cdot$-concatenation starting at p with junctions at p, p_1, p_2, p_3 and times s_1, s_2, s_3 along the respective trajectories. Then we have (the vector fields are evaluated at p_1):

$$\begin{aligned}0 &= X \wedge Y \wedge \left(\frac{\exp(-s_1 \operatorname{ad} X) - 1}{-s_1}\right) Y \wedge \left(\frac{\exp(s_2 \operatorname{ad} Y) - 1}{s_2}\right) X \\ &= X \wedge Y \wedge [X, Y] - \frac{1}{2}s_1[X,[X, Y]] + O(s_1^2) \wedge -[X, Y] - \frac{1}{2}s_2[Y,[X, Y]] + O(s_2^2) \qquad (16) \\ &= \frac{1}{2}\sigma(s_1, s_2)(X \wedge Y \wedge [X, Y] \wedge [Y,[X, Y]])|_{p_1}\end{aligned}$$

where $\sigma(s_1, s_2) = -s_1\delta - s_2 + O(2)$.

The equation $\tilde{\sigma}(s_1, s_2)=0$ has a unique solution $\overline{s_1}(s_2)$ and in general XYX-trajectories only satisfy the necessary conditions of the Maximum Principle if $s_1 \leqq \overline{s_1}(s_2)$. Note that $\underset{\sim}{\sigma}(0, s_2)<0$ and so this is equivalent to $\underset{\sim}{\sigma}(s_1, s_2) \leqq 0$. (Using an argument analogous to (9) it can be shown that extremal trajectories do indeed have switchings at $s_1=\overline{s_1}$, but we will not need this.) Furthermore,

$$\begin{aligned}
0 &= X(p_2) \wedge Y(p_2) \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y)-1}{-s_2}\right) X(p_1) \\
&\quad \wedge \left(\frac{\exp(s_3 \operatorname{ad} X)-1}{s_3}\right) Y(p_3) \\
&= X \wedge Y \wedge -[X, Y]+\frac{1}{2}s_2[Y,[X, Y]]+\cdots \wedge [X, Y]+\frac{1}{2}s_3[X,[X, Y]]+\cdots \\
&= \frac{1}{2}\tilde{\sigma}(s_2, s_3)(X \wedge Y \wedge [X, Y] \wedge [Y,[X, Y]])|_{p_2}
\end{aligned}$$

where $\tilde{\sigma}(s_2, s_3)=-s_2-s_3\delta+O(T^2)$.

Again the equation $\tilde{\sigma}(s_2, s_3)=0$ can be solved by $\overline{s_3}(s_2)$, and YXY-concatenations only satisfy the Maximum Principle if $s_3 \leqq \overline{s_3}(s_2)$. Since $\tilde{\sigma}(s_2, 0)<0$ this is equivalent to $\tilde{\sigma}(s_2, s_3) \leqq 0$.

Therefore we define

$$\tilde{\Gamma}^- = \{p \exp(s_1X) \exp(s_2Y) \exp(s_3X)\colon s_i \geqq 0, \text{ small, } s_2 \text{ is free,}$$
$$\underset{\sim}{\sigma}(s_1, s_2) \leqq 0, \tilde{\sigma}(s_2, s_3) \leqq 0\}.$$

Analogously we must compute the conjugate point relations along a $\cdot YXY\cdot$-concatenation which yields

$$\tilde{\Gamma}^+ := \{p \exp(t_1 Y) \exp(t_2X) \exp(t_3 Y)\colon t_1 \geqq 0, \text{ small } t_2 \text{ is free,}$$
$$\underset{\sim}{\tau}(t_1, t_2) \geqq 0 \Leftrightarrow t_1 \leqq \bar{t}_1(t_2) \tilde{\tau}(t_2, t_3) \geqq 0 \Leftrightarrow t_3 \leqq \bar{t}_3(t_2)\}$$

where

$$\underset{\sim}{\tau}(t_1, t_2)=-t_1-t_2\delta+O(T^2), \qquad \tilde{\tau}(t_2, t_3)=-t_2\delta-t_3+O(T^2)$$

and $\bar{t}_1$ and $\bar{t}_3$ are the solutions of $\underset{\sim}{\tau}=0$ and $\tilde{\tau}=0$, respectively. $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ are three-dimensional surfaces with relative boundary made up entirely of bang-bang trajectories with at most one switch.

LEMMA 5. *The surfaces $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect along a two-dimensional surface $\hat{\Gamma}$.*

The intersection of $\hat{\Gamma}$ with the relative boundaries $\partial\tilde{\Gamma}^-$ and $\partial\tilde{\Gamma}^+$ are the following one-dimensional curves:

$$\tilde{\gamma} = \{p \exp(s_1X) \exp(s_2Y)\colon s_2 \geqq 0, \textit{ small, } s_1=\bar{s}_1(s_2)\},$$
$$\underset{\sim}{\gamma} = \{p \exp(t_1 Y) \exp(t_2X)\colon t_2 \geqq 0, \textit{ small, } t_1=\bar{t}_1(t_2)\}$$

(i.e., the trajectories corresponding to the conjugate points). Away from $\underset{\sim}{\gamma}$ and $\tilde{\gamma}$ the surface entirely lies in the relative interior of $\tilde{\Gamma}^-$, respectively, $\tilde{\Gamma}^+$ and there the intersection is transversal.

Proof. We want to solve the equation

$$(17) \qquad p \exp(s_1X) \exp(s_2Y) \exp(s_3X)=p \exp(t_1 Y) \exp(t_2X) \exp(t_3 Y).$$

Suppose a point q in the relative interior of $\tilde{\Gamma}^+$ or $\tilde{\Gamma}^-$ lies on $\hat{\Gamma}$. We claim that (16) can be solved in terms of t_1 and t_2 near q. This follows from the Implicit Function Theorem if the Jacobian with respect to (s_1, s_2, s_3, t_3) is nonsingular at q. If we compute these derivatives and pull the vectors back along X we get

$$\begin{aligned}
&\exp(-s_2 \operatorname{ad} Y)X \wedge Y \wedge X \wedge \exp(s_3 \operatorname{ad} X)Y \\
&\quad = s_2 s_3\left(X \wedge Y \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y)-1}{-s_2}\right) X \wedge \left(\frac{\exp(s_3 \operatorname{ad} X)-1}{s_3}\right) Y\right) \\
&\quad = \frac{1}{2} s_2 s_3 \cdot \tilde{\sigma}(s_2, s_3)(X \wedge Y[X, Y] \wedge [Y,[X, Y]])\big|_{p_2 = p \exp(s_1 X) \exp(s_2 Y)}.
\end{aligned}$$

But in int $(\tilde{\Gamma}^-)$ s_2 and s_3 are positive and also $\tilde{\sigma}(s_2, s_3) < 0$ since the conjugate point relation does not hold. So we can solve in terms of t_1 and t_2. This computation shows also that $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ intersect transversally in int $(\tilde{\Gamma}^+)$ or int $(\tilde{\Gamma}^-)$.

Next we show that points q of this type exist. For that we rewrite both sides of (17) in terms of canonical coordinates of the second kind. A short computation (cf., for instance, [16]) shows that

$$\begin{aligned}
p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) &= p \exp\left(\tfrac{1}{2} s_1 s_2 (s_1 \delta + s_2 + O(S^2))[Y,[X, Y]]\right) \\
&\quad \cdot \exp(s_1 s_2(1 + O(S))[X, Y]) \\
&\quad \cdot \exp((s_2 + O(S^3))Y) \exp((s_1 + s_3 + O(S^3))X), \\
p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) &= p \exp\left(\tfrac{1}{2} t_2 t_3 (2t_1 + t_3 + t_2 \delta + O(T^2))[Y,[X, Y]]\right) \\
&\quad \cdot \exp(t_2 t_3(1 + O(T))[X, Y]) \\
&\quad \cdot \exp((t_1 + t_3 + O(T^3))Y) \exp((t_2 + O(T^3))X)
\end{aligned}$$

where $O(S^k)$ or $O(T^k)$ stand for terms of order greater than or equal to k in the total time, $S = s_1 + s_2 + s_3$, $T = t_1 + t_2 + t_3$, and δ is evaluated at p. Equating coefficients we get

$$(18) \quad \begin{aligned}
&\text{(i)} \quad s_1 + s_3 + O(S^3) = t_2 + O(T^3), \\
&\text{(ii)} \quad s_2 + O(S^3) = t_1 + t_3 + O(T^3), \\
&\text{(iii)} \quad s_1 s_2(1 + O(S)) = t_2 t_3(1 + O(T)), \\
&\text{(iv)} \quad s_1 s_2(s_1 \delta + s_2 + O(S^2)) = t_2 t_3(2t_1 + t_3 + t_2 \delta + O(T^2)).
\end{aligned}$$

If we assume that all switching times are comparable, i.e., of order T, then (18(i), (ii)), and

$$\text{(iv}') \quad s_1 \delta + s_2 + O(S^2) = 2t_1 + t_3 + t_2 \delta + O(T^2)$$

can easily be solved for s in terms of t modulo higher-order terms:

$$(19) \quad \begin{aligned}
s_1 &= t_2 + \frac{1}{\delta} t_1 + O(T^2), \\
s_2 &= t_1 + t_3 + O(T^3), \\
s_3 &= -\frac{1}{\delta} t_1 + O(T^2).
\end{aligned}$$

With these times the conjugate point relations cannot hold since

$$(20) \quad \tilde{\sigma}(s_2, s_3) = -s_2 - s_3 \delta + O(T^2) = -t_3 + O(T^2)$$

is negative. So the corresponding point q lies in fact in the relative interior and therefore it is possible to solve for t_3 in terms of t_1 and t_2:

$$t_3 = -t_1 - \delta t_2 + O(T^2). \tag{21}$$

This gives a solution to (18). Note that

$$t_2 = \rho T + O(T^2) = \frac{T}{1-\delta} + O(T^2). \tag{22}$$

As long as (t_1, t_2, t_3) are bounded away from the boundary of the simplex $t_1 + t_2 + t_3 \leqq T$, the times are comparable, these computations are justified, and we get a two-dimensional intersection that we can parametrize by t_1 and t_2. The problem is whether it extends all the way to the boundary. But the equations (19) and (21) are well defined for $t_1 \to 0$ (in a time-slice $t_1 + t_2 + t_3 = T$ it follows that $t_3 \to -\delta t_2 + O(t_2^2)$, i.e., to a limit of order T. By (20) this implies that the two-dimensional surface defined by these functions of (t_1, t_2) stays away from the conjugate point condition $\tilde{\sigma}(s_2, s_3) = 0$. Hence the implicit function theorem is still applicable.) Therefore $\hat{\Gamma}$ extends all the way out to $t_1 = 0$, i.e., to the XY boundary surface.

A precise characterization of $\tilde{\Gamma}^+ \cap \tilde{\Gamma}^- \cap \{p \exp(s_1 X) \exp(s_2 Y) : s_i \geqq 0, \text{small}\}$ is possible. Clearly these are points such that $t_1 = 0$, $t_2 = s_1$, $t_3 = s_2$, and $0 = s_3$. Since $(s_1, s_2, 0) \in \operatorname{dom} \tilde{\Gamma}^-$ we have $\underset{\sim}{\sigma}(s_1, s_2) \leqq 0$, and since $(0, s_1, s_2) \in \operatorname{dom} \tilde{\Gamma}^+$ we have $\tilde{\tau}(s_1, s_2) \geqq 0$. But in this case $\underset{\sim}{\sigma}(s_1, s_2) = \tilde{\tau}(s_1, s_2)$ (cf. (16) and the analogous formula for $\tilde{\tau}$). Therefore $\underset{\sim}{\sigma}(s_1, s_2) = 0$, i.e., $s_1 = \overline{s_1}(s_2)$, the conjugate point relation.

This proves that $\tilde{\Gamma}^- \cap \tilde{\Gamma}^+$ extends all the way out to the XY-boundary surface and that the intersection with the XY-surface is the one-dimensional curve $\tilde{\gamma}$ consisting of the conjugate points.

Analogously we can show that (17) can also be solved in terms of s_1 and s_2 in $\operatorname{int}(\tilde{\Gamma}^-)$. Using these formulas we can show that $\tilde{\Gamma}^- \cap \tilde{\Gamma}^+$ extends all the way up to the YX-boundary surface and that the intersection of $\tilde{\Gamma}^- \cap \tilde{\Gamma}^+$ with the YX-surface consists of the curve γ. □

Note that in a time-slice $t_1 + t_2 + t_3 = T$ the qualitative geometric structure of $\tilde{\Gamma}^- \cup \tilde{\Gamma}^+$ is exactly as in the free nilpotent case. Only the condition $t_2 = T/2$ is replaced by $t_2 \doteq (1/(1-\delta))T$ (modulo higher terms) which shifts $\hat{\Gamma}$ away from the center. This is illustrated in Fig. 6.

The surface $\hat{\Gamma}$ bisects $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ and only one of the two components has the Y-, respectively, X-trajectory in its boundary. We define Γ^- and Γ^+ to be these components

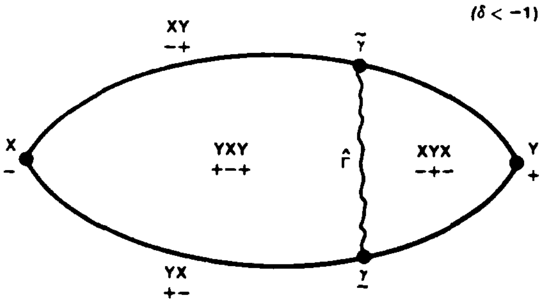


FIG. 6

and let $\Gamma^* = \Gamma^+ \cup \Gamma^-$. It is then clear that Γ^* is a three-dimensional stratified surface whose relative boundary consists of all bang-bang trajectories with at most one switching, i.e., $\partial\Gamma^* = \partial\Gamma_*$.

LEMMA 6. *The points in Γ^* are exit points of the small-time reachable set for* $[X, [X, Y]]$.

Proof. It is easy to see (cf. (10)) that, for sufficiently small time, all extremals on $\tilde{\Gamma}^-$ or $\tilde{\Gamma}^+$ satisfy the necessary transversality condition $\langle \lambda, [X, [X, Y]]\rangle \leqq 0$.

We show first that the points that we deleted from $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ are not exit points (see Fig. 7). Let

$$q = p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) = p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y)$$

be a point in the relative interior of $\hat{\Gamma}$. $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect transversally. It follows as in the proof of Lemma 2 (cf. (11)) that the XYX- and YXY-surfaces are graphs $x_4 = \psi(x_1, x_2, x_3)$ in canonical coordinates of the second kind with x_4 the coefficient at the flow of $[X, [X, Y]]$. This inherits on $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$. To prove that the parts of $\tilde{\Gamma}^-$ (respectively, $\tilde{\Gamma}^+$) that we delete are not exit points, it suffices to show that these parts lie below $\tilde{\Gamma}^+$ (respectively, $\tilde{\Gamma}^-$) in direction of $[X, [X, Y]]$.

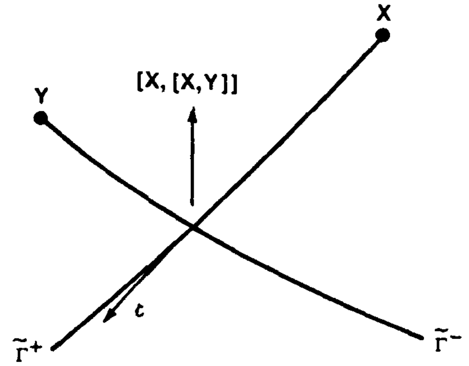


FIG. 7

The tangent space to $\tilde{\Gamma}^-$ at q is spanned by X, $\exp(-s_3 \operatorname{ad} X)Y$ and $\exp(-s_3 \operatorname{ad} X)(\exp((-s_2 \operatorname{ad} Y) - 1)/-s_2)X$. To show that the part of $\tilde{\Gamma}^+$ that we deleted lies below $\tilde{\Gamma}^-$ near q it suffices to show that $[X, [X, Y]]$ and a tangent vector t to $\tilde{\Gamma}^+$ that is oriented toward the sector of $\tilde{\Gamma}^+$ that we deleted point to opposite sides of $T_q\tilde{\Gamma}^-$. We get such a vector t if we lengthe: the time along the last Y leg. (We delete the piece that contains in its boundary the trajectories corresponding to the conjugate point relation $t_3 = \bar{t}_3(t_2)$.)

Instead of computing at q we pull back all vectors along X, s_3 and get

$$\begin{aligned}
&\exp(+s_3 \operatorname{ad} X)(T_q\tilde{\Gamma}^-) \wedge \exp(+s_3 \operatorname{ad} X)[X, [X, Y]] \\
&\quad = \left(X \wedge Y \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y) - 1}{-s_2} \right) X \wedge \exp(s_3 \operatorname{ad} X)[X, [X, Y]] \right) \\
&\quad = -(\delta + O(T))(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])|_{p_2 = p \exp(s_1 X) \exp(s_2 Y)},
\end{aligned}$$

$$\begin{aligned}&\exp(s_3 \operatorname{ad} X)(T_q\tilde{\Gamma}^-)\wedge \exp(s_3 \operatorname{ad} X)Y\\ &\quad = s_3\left(X\wedge Y\wedge\left(\frac{\exp(-s_2 \operatorname{ad} Y)-1}{-s_2}\right)X\wedge\left(\frac{\exp(s_3 \operatorname{ad} X)-1}{s_3}\right)Y\right)\\ &\quad =\frac{1}{2}s_3\tilde{\sigma}(s_2,s_3)(X\wedge Y\wedge[X,Y]\wedge[Y,[X,Y]])|_{p_2}.\end{aligned}$$

But q is a point in $\hat{\Gamma}$, and $\hat{\Gamma}$ lies entirely in the relative interior of $\tilde{\Gamma}^-$ except for the obvious boundary curves $\tilde{\gamma}$ and γ. In particular (cf. also the proof of Lemma 5) the conjugate point relation $s_3=\bar{s}_3(\tilde{s}_2)$ does not hold, or equivalently, $\tilde{\sigma}(s_2,s_3)<0$. So these wedge-products have opposite signs, which proves our claim. This also implies that the portion of $\tilde{\Gamma}^-$ that we delete lies below $\tilde{\Gamma}^+$, and since there is no other intersection this holds for all the points we deleted.

The stratified sets Γ^* and Γ_* enclose a region R that lies in the small-time reachable set. In particular, the portions of $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ that we deleted therefore lie in the interior of the reachable set. Since these pieces contain the trajectories corresponding to the conjugate points $t_3=\bar{t}_3(t_2)$ and $s_3=\overline{s_3}(s_2)$, it follows that no bang-bang trajectory with more than two switchings lies in the boundary of the small-time reachable set. Hence the points in Γ^* are the only possible exit points of the small-time reachable set for $[X,[X,Y]]$. It follows from the construction of Γ^- and Γ^+ that Γ^* is also a graph. Again, the projections onto (x_1,x_2,x_3)-space are disjoint. Therefore it follows as in Lemma 2 that the points on Γ^* have the exit property for sufficiently small time. □

Finally, Γ^* and Γ_* do not intersect in their relative interiors. It is now clear how the *small-time reachable* set looks: It is the set of points enclosed by the two three-dimensional stratified surfaces Γ^* and Γ_*. Γ^* consists of bang-bang trajectories with at most two switchings such that modulo higher-order terms

$$t_1+\delta t_2+t_3\leqq 0 \tag{23}$$

if t_1, t_2, and t_3 are the consecutive times along a YXY arc and

$$s_1\delta+s_2+s_3\delta\geqq 0 \tag{24}$$

if s_1, s_2, s_3 are consecutive times along XYX. Γ_* consists of all concatenations of a bang arc, followed by a singular arc and another bang arc where the time along the trajectories is free. Γ^* and Γ_* have a common relative boundary C consisting of all trajectories that are bang-bang with at most one switching. For sufficiently small-time T a time-slice of the reachable set has exactly the same qualitative geometric structure as for the free nilpotent system (13). Furthermore, if $\delta(\cdot)$ is an integral curve of $[X,[X,Y]]$ such that $\delta(t_1)$ and $\delta(t_2)$, $t_1<t_2$, lie in the small-time reachable set, then so does the whole curve $\delta(t)$, $t_1\leqq t\leqq t_2$. The points on Γ_* are entry points for $[X,[X,Y]]$; the points on Γ^* are exit points.

Remark. We emphasize that the result is not what might be expected intuitively. From dimensionality we could conjecture the occurrence of bang-bang trajectories with two switchings, respectively, *BSB* trajectories in the boundary of the small-time reachable set. Also, this is essentially what was partially known from earlier results. However, we see no simple reasoning that could explain why, in fact, *some* of these *bang-bang trajectories with two switchings are not a part of the boundary.* This is only revealed by our analysis.

4.3. Time-optimal control in dimension three. Our results have immediate implications on time-optimal control in dimension three. Suppose the triples $(g,[f,g],[f+g,[f,g]])$ and $(g,[f,g],[f-g,[f,g]])$ consist of independent vectors at a point p in $\mathbf{R}^3$.

Equivalently, suppose that the constant controls $u \equiv +1$ and $u \equiv -1$ are not singular. If we augment the three-dimensional system Σ to a four-dimensional system $\hat{\Sigma}$ by introducing time as a coordinate, $\dot{x}_0 = 1$, $x_0(0) = 0$, i.e.,

$$\hat{f} = \begin{pmatrix} 1 \\ f \end{pmatrix}, \qquad \hat{g} = \begin{pmatrix} 0 \\ g \end{pmatrix},$$

then if a Σ-trajectory $x(\cdot):[0, T] \to \mathbb{R}^3$ steering p to q is time-optimal, the augmented trajectory $\hat{x}$ lies in the boundary of the reachable set from p. The augmented system $\hat{\Sigma}$ satisfies our assumptions $(\underset{\sim}{A})$ and $(\underset{\sim}{B})$, and therefore time-optimal trajectories are bang-bang with at most two switchings or concatenations of a bang-arc, followed by a singular arc and one more bang arc. Under additional assumptions this result was obtained earlier by Bressan [4], who studied only trajectories emanating from an equilibrium point of f and by Sussmann [22] and Schättler [17] who both assumed in addition also that f, g and $[f, g]$ were independent. Our analysis shows that the vector field f is irrelevant and we do not have to make any assumptions about it. Our results are also more precise in the sense that we can exclude the optimality of those bang-bang trajectories with two switchings that violate (23) (respectively, (24)) in the bang-bang singular case. We summarize in the following corollary.

COROLLARY. *Suppose the vector fields g, $[f, g]$ and $[f+g, [f, g]]$ are independent near a reference point $p \in \mathbb{R}^3$. Write*

$$[f-g, [f, g]] = ag + b[f, g] + c[f+g, [f, g]]$$

and assume that c does not vanish. Then we have in small time:

(i) *If $c > 0$, then time-optimal trajectories are bang-bang with at most 2 switches.*

(ii) *If $c < 0$, then time-optimal trajectories are bang-bang with at most two switchings or are concatenations of a bang arc, a singular arc, and another bang arc. Time-optimal XYX (respectively, YXY) concatenations satisfy modulo higher-order terms*

$$c(s_1 + s_3) + s_2 \geqq 0 \qquad (\textit{resp.}, \ t_1 + t_3 + ct_2 \leqq 0)$$

where s_1, s_2, s_3 (respectively, t_1, t_2, t_3) are the consecutive times along the bang arcs.

5. A brief outlook to higher dimensions. We have outlined a general method to determine the structure of the small-time reachable sets and proved its effectiveness in nondegenerate cases in small dimensions. One of the difficulties that will become more and more prominent in higher dimensions is that the necessary conditions of the Maximum Principle will not restrict the class of extremal trajectories sufficiently enough to give the candidates for Γ^* and Γ_*.

Under assumptions $(\underset{\sim}{A})$ and $(\underset{\sim}{B})$ in dimension four, we could overcome this problem by taking a corresponding "free" nilpotent system of the same dimension as a guide. We do not expect this to happen in general. In fact, for the five-dimensional system Σ, where we assume that f, g, $[f, g]$, $[f, [f, g]]$, and $[g, [f, g]]$ are independent, the small-time reachable set has extremal trajectories in its boundary that do not appear in the analogous five-dimensional free nilpotent system. The reason for this lies in a qualitatively different behavior of the singular controls, specifically, in the fact that singular controls can now hit the control constraint $|u| = 1$ and may have to be terminated. Nevertheless, the free nilpotent system contains most of the information about the small-time reachable set, though it does not characterize it completely. To be more specific, we will briefly describe (without proofs) the structure of the reachable set for the free nilpotent system in dimension five and how the general case differs from it.

We take as our model:

$$\dot{x}_0 = 1, \quad \dot{x}_1 = u, \quad \dot{x}_2 = x_1, \quad \dot{x}_3 = x_2, \quad \dot{x}_4 = \tfrac{1}{2}x_1^2.$$

It is no problem whatsoever to carry out the analysis within our technique as in the construction in § 4.2.1. Now the reachable set is convex in direction of $(0, 0, 0, 0, 1)^T = [g, [g, f]]$ and Γ^*, respectively, Γ_* will consist of those trajectories that are exit, respectively, entry points.

It follows from the generalized Legendre-Clebsch condition that Γ_* contains concatenations with singular arcs, whereas Γ^* will consist of bang-bang trajectories only. Singular controls are constant, but now they can take on any value in $[-1, 1]$.

Let $\Gamma_* = \Gamma_{-u-} \cup \Gamma_{-u+} \cup \Gamma_{+u-} \cup \Gamma_{+u+}$, where

$$\Gamma_{-u-} := \{0 \exp(s_1 X) \exp(s_2(f + ug)) \exp(s_3 X) \colon s_i \geqq 0,\ s_1 + s_2 + s_3 = 1,\ u \in [-1, 1]\},$$

etc. (By the invariance property of the reachable set we can restrict to the time-slice $T = 1$.) The points on Γ_* are precisely the ones that minimize the coordinate x_4.

For a fixed value u_0 of the singular control, $-1 < u_0 < +1$, the qualitative structure of $\Gamma_{*,u_0} = \Gamma_*$ restricted to values $u = u_0$ is precisely as in 4.2.2, Fig. 4 (see Fig. 8).

For $u_0 = +1$, $\Gamma_{-u-} \restriction u = 1$ reduces to Γ_{-+-} and all other strata become trivial whereas for $u_0 = -1$, $\Gamma_{+u+} \restriction u = -1 = \Gamma_{+-+}$ and the remaining strata are trivial. For each of these two-dimensional surfaces (u_0 fixed) the relative boundary consists of all bang-bang trajectories with at most one switching. The surfaces Γ_{*,u_0} themselves interpolate between Γ_{+-+} for $u_0 = -1$ and Γ_{-+-} for $u_0 = 1$. Topologically Γ_* is a stratified sphere

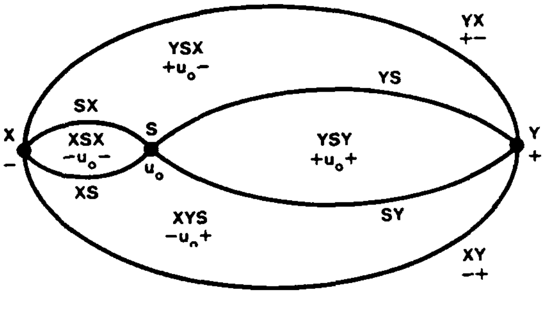


FIG. 8

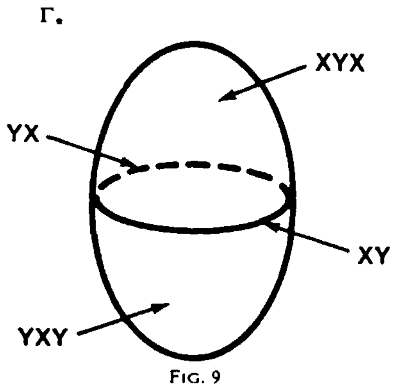


FIG. 9

with $\partial\Gamma_* = \Gamma_{-+-} \cup \Gamma_{+-+}$, i.e., all bang-bang trajectories with at most two switchings (see Fig. 9).

The surface Γ^* consists of bang-bang trajectories analogous to the bang-bang singular case in dimension four. Now

$$\tilde{\Gamma}^- = \{0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \exp(s_4 Y)\colon s_i \geqq 0,\ s_1+s_2+s_3+s_4 = 1,\\ s_1 \leqq s_3,\ s_4 \leqq s_2\text{-conjugate point relations}\},$$

$$\tilde{\Gamma}^+ = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \exp(t_4 X)\colon t_i \geqq 0,\ t_1+t_2+t_3+t_4 = 1,\\ t_1 \leqq t_3,\ t_4 \leqq t_2\text{-conjugate point relations}\}.$$

$\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect in a two-dimensional surface $\hat{\Gamma}$, which consists of those trajectories for which

$$(s_1+s_3)^2 - (s_1+s_3) + 2s_2s_3 = 0,$$

respectively,

$$(t_1+t_3)^2 - (t_1+t_3) + 2t_2t_3 = 0.$$

The intersection is transversal except at those points that lie on the relative boundary of $\tilde{\Gamma}^-$ or $\tilde{\Gamma}^+$. These points are again characterized by the conjugate point relation

$$\hat{\Gamma} \cap \Gamma_{-+-} = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \exp(t_4 X)\colon t_1 = 0,\ t_4 = t_2\},$$

$$\hat{\Gamma} \cap \Gamma_{+-+} = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \exp(t_4 X)\colon t_1 = t_3,\ t_4 = 0\}.$$

We define Γ^- (respectively, Γ^+) as the component of $\tilde{\Gamma}^-$ ($\tilde{\Gamma}^+$) containing the YX-curve $= \{0 \exp(s_2 Y) \exp(s_3 X)\colon s_i \geqq 0,\ s_2+s_3 = 1\}$ (respectively, the XY-curve) in its boundary. Then $\Gamma^* := \Gamma^- \cup \Gamma^+$ consists precisely of those points that maximize x_4 on the reachable set. Note that topologically Γ^* also is a stratified sphere with $\partial\Gamma^* = \Gamma_{-+-} \cup \Gamma_{+-+}$, the set of all bang-bang trajectories with at most two switchings (see Fig. 10).

The key fact here is that it is still obvious that $\partial\Gamma^*$ and $\partial\Gamma_*$ match up. They are identical. It is therefore clear that Reach (0, 1) is the set of all points that lie between Γ^* and Γ_*.

It is precisely this simple reasoning that breaks down in the general case. The cause for this lies in the structure of the singular controls. The analysis of the bang-bang trajectories carries over to the general case with only one minor change in the structure. Whereas in the free nilpotent system the two curves $\hat{\Gamma} \cap \Gamma_{+-+}$ and $\hat{\Gamma} \cap \Gamma_{-+-}$ both have points corresponding to the X- and Y-trajectories as endpoints, this need no longer be true: $\hat{\Gamma} \cap \Gamma_{+-+}$ is a curve starting at $0 \exp(1 \cdot Y)$ but which in general no longer ends in $0 \exp(1 \cdot X)$ but rather on a point in the XY-curve (respectively, YX-curve). This distortion is due to the presence of fourth-order brackets. One possible case is depicted in Fig. 11.

Still the relative boundary of Γ^* consists of all bang-bang trajectories with at most two switchings. The structure breaks down in the analysis of the singular surface Γ_* for u near ±1. The reason is that in the presence of fourth-order brackets the singular controls are no longer constant, and thus the analogue of Γ_{*,u_0} for $u_0 = -1$ does not reduce to Γ_{+-+}, i.e., to bang-bang trajectories with two switchings. For instance, it may not be at all possible to start a singular control with $u_0 = -1$. This is the case if $\dot{u} < 0$ at $u_0 = -1$, which happens under generic assumptions on fourth-order brackets.

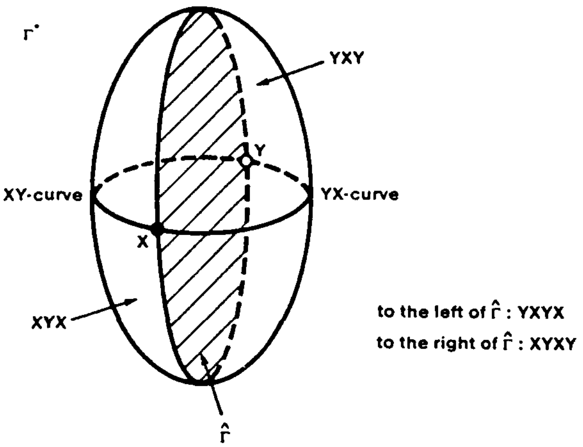

Γ^*
YXY
Y
XY-curve
YX-curve
X
XYX
to the left of $\hat{\Gamma}$: YXYX
to the right of $\hat{\Gamma}$: XYXY
$\hat{\Gamma}$

FIG. 10

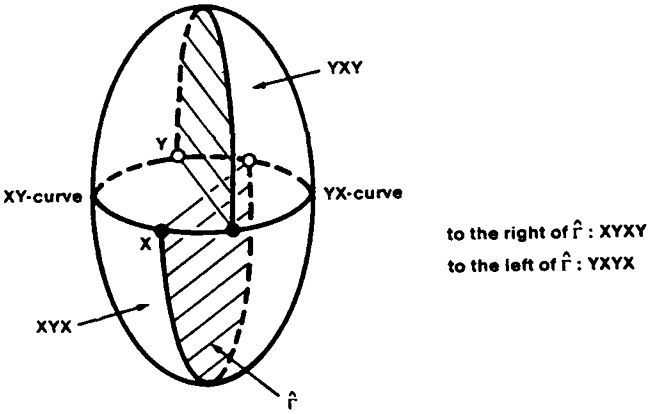

YXY
Y
XY-curve
YX-curve
X
XYX
to the right of $\hat{\Gamma}$: XYXY
to the left of $\hat{\Gamma}$: YXYX
$\hat{\Gamma}$

FIG. 11

For the same reason, singular controls with u_0 close to ±1 may have to be terminated when they become one in absolute value. If the singular control becomes saturated (i.e., hits the constraint and cannot be continued) then this determines the subsequent structure of the trajectory and it is easy to see that concatenations such as *BSBB* or *BBSB*, which are not present in the free nilpotent system, come into play. Therefore Γ_* has trajectories in its relative boundary that contain singular arcs. The main challenge in applying our technique to higher dimensions seems to be finding a way to decide whether structurally different trajectories, such as a bang-bang trajectory, and a concatenation that contains a singular arc steer a system to the same point. Once $\partial\Gamma^*$ and $\partial\Gamma_*$ can be identified, it is clear that the set they enclose is the small-time reachable set.

Note, however, that this structural instability only happens near $\Gamma_{*,-1}$ and $\Gamma_{*,+1}$. The structure of most of the trajectories in the boundary is still the same as in the free nilpotent systems. And it is intuitively clear that the structure of the exceptional trajectories will come up in a higher-dimensional nilpotent system. Therefore, in our view, the study of the structure of the reachable sets for nilpotent systems will be the key to the general problem.

6. Summary. We have described an approach to determining the qualitative structure of the small-time reachable set in a nondegenerate situation. It is a nontrivial extension of a construction done by Lobry in dimension three. In dimension four we succeed completely in determining the small-time reachable set. For higher dimensions obstacles still have to be overcome. However, they do not lie in the general structure of our approach, but in the fact that too little is known about the structure of extremal trajectories in higher dimensions. For instance, in the five-dimensional case, what is the precise structure of extremal trajectories that contain a saturated singular arc? For dimensions six and beyond, the crucial new ingredient appears to be the incorporation of chattering arcs, another structure of extremal trajectories about which little is still known.

REFERENCES

[1] V. G. BOLTYANSKY, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326-361.

[2] A. BRESSAN, *Directional convexity and finite optimality conditions*, J. Math. Anal. Appl., 125 (1987), pp. 234-246.

[3] ———, *Local asymptotic approximation of nonlinear control systems*, Tech. Report, University of Wisconsin, Madison, WI, 1984.

[4] ———, *The generic local time-optimal stabilizing controls in dimension* 3, SIAM J. Control Optim., 24 (1986), pp. 177-190.

[5] P. BRUNOVSKY, *Existence of regular synthesis for general problems*, J. Differential Equations, 38 (1980), pp. 317-343.

[6] C. BYRNES AND P. CROUCH, *Local accessibility, local reachability, and representations of compact groups*, Math. Systems Theory, 19 (1986), pp. 43-65.

[7] P. E. CROUCH AND P. C. COLLINGWOOD, *The observation space and realizations of finite Volterra series*, SIAM J. Control Optim., 25 (1987), pp. 316-333.

[8] P. E. CROUCH, *Solvable approximations to control systems*, SIAM J. Control Optim., 22 (1984), pp. 40-45.

[9] W. H. FLEMING AND R. W. RISHEL, *Deterministic and stochastic optimal control*, Applications of Mathematics, Vol. 1, Springer-Verlag, New York, 1975.

[10] H. HERMES, *Lie algebras of vector fields and local approximation of attainable sets*, SIAM J. Control Optim., 16 (1978), pp. 715-727.

[11] N. JACOBSON, *Lie Algebras*, Dover, New York, 1979.

[12] A. KRENER, *Local approximation of control systems*, J. Differential Equations, 19 (1975), pp. 125-133.

[13] ———, *The higher order maximum principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256-293.

[14] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573-605.

[15] L. P. ROTHSCHILD AND E. M. STEIN, *Hypoelliptic differential operators and nilpotent groups*, Acta Math., 137 (1977), pp. 247-320.

[16] H. SCHÄTTLER, *On the local structure of time-optimal bang-bang trajectories in* $\mathbf{R}^3$, SIAM J. Control Optim., 26 (1988), pp. 186-204.

[17] ———, *The local structure of time-optimal trajectories in dimension three under generic conditions*, SIAM J. Control Optim., 26 (1988), pp. 899-918

[18] H. SUSSMANN, *Analytic stratifications and control theory*, in Proc. International Congress of Mathematics, Helsinki, 1978, pp. 865-871.

[19] ———, *A bang-bang theorem with bounds on the number of switchings*, SIAM J. Control Optim., 17 (1979), pp. 629-651.

[20] ———, *Lie-Volterra expansion for nonlinear systems*, in Mathematical Theory of Networks and Systems, P. Fuhrman, ed., Lecture Notes in Control and Information Science, 58, Springer-Verlag, Berlin, 1984, pp. 822-828.

[21] ———, *Lie brackets and real analyticity in control theory*, in Mathematical Control Theory, Banach Center Publications, Vol. 14, Warsaw, 1984, pp. 515-542.

[22] H. SUSSMANN, *Envelopes, conjugate points and optimal bang-bang extremals*, in Proc. 1985 Paris Conference on Nonlinear Systems, M. Fliess, M. Harewinkel, eds., D. Reidel, Dordrecht, the Netherlands, 1986.

[23] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 856-905.

[24] ———, *Regular synthesis for time-optimal control of single-input analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145-1162.

Analysis and Control of Nonlinear Systems
C.I. Byrnes, C.F. Martin and R.E. Saeks (editors)
© Elsevier Science Publishers B.V. (North-Holland), 1988

Reciprocal Processes, Second Order Stochastic Differential Equations and PDE's of Conservation and Balance

Arthur J. Krener
Department of Mathematics
and
Institute of Theoretical Dynamics
University of California
Davis, CA 95616

1. INTRODUCTION

The close connection between Markov processes, diffusions and parabolic partial differential equations is of course well–known. In this paper we shall describe the beginning of a new theory which links reciprocal processes, second order diffusions and the partial differential equations of fluid mechanics, i.e., the continuity, Euler and energy balance equations.

2. RECIPROCAL PROCESSES

In the early thirties E. Schrödinger [2,3] introduced a new class of stochastic processes in attempt to formalize the stochastic aspects of Quantum Mechanics. This concept was formalized by S. Bernstein [1] in an address to the International Congress of Mathematicians in Zurich in 1932. Bernstein defined a reciprocal process $x(t)$ as one where conditioned on the values $x(t_0)$ and $x(t_1)$ of the process at two times $t_0 \leq t_1$, the process exterior to $[t_0,t_1]$ is independent of the process interior to $[t_0,t_1]$. This is readily seen to be a generalization of the Markov property, i.e., conditioned on single time t_0 the process before t_0 is independent of the process after t_0. Hence every Markov process is reciprocal but the converse is not true.

The reciprocal property is the specialization to one dimension of P. Levy's definition of a Markov random field [20]. There are two other ways of viewing the reciprocal property. Suppose $x(t)$ is a random process taking values in $\mathbb{R}^n$ and defined for $t \in [0,T]$. We define another process $X(t_0,t_1) = (x(t_0),x(t_1))$ taking values in $\mathbb{R}^{2n}$. We view this process as parametrized by pairs (t_0,t_1) where $t_0 \leq t_1$ or equivalently by subintervals (t_0,t_1). Subintervals are partially ordered by inclusion. It is easy to see

*Research supported in part by NSF under DMS–8601635 and AFOSR under NM85–0267.

that the original process x(t) is reciprocal iff the two time process $X(t_0,t_1)$ is Markov relative to this partial ordering.

Alternatively we can view a reciprocal process as being conditionally Markov in the following sense. Given any $t_0 \in [0,T]$ and $x^0 \in \mathbb{R}^n$, we define a conditioned process $\tilde{x}(t\,|\,t_0,x^0)$ consisting of all sample paths of x(t) satisfying $x(t_0) = x^0$ with the conditional probabilty measure. The process x(t) is said to be conditionally Markov if every $\tilde{x}(t\,|\,t_0,x^0)$ is Markov for $t \in [0,t_0]$ and is also Markov for $t \in [t_0,T]$. (It need not be Markov on [0,T].) It is straightforward to note that a process x(t) is reciprocal iff it is conditionally Markov.

To essentially specify a stochastic process one must describe all finite dimensional distributions of the process, e.g., give the probability distribution of $x(t_0),x(t_1), \ldots ,x(t_n)$ where $0 \le t_1 \le \ldots \le t_n \le T$. One reason that Markov processes are so well–studied is that they are completely determined by only two functions. The first is $\rho_0(x^0)$, the probability density of x(0). (Throughout we assume that probability densities exist although the discussion can be easily extended using probability distributions.) The second p(s,x;t,y) is the Markov transition density of x(t) = y given that x(s) = x. By Bayes' formula the probability density of $x(t_1) = x^1, \ldots , x(t_n) = x^n$ where $0 \le t_1 \le t_2 \le \ldots \le t_n \le T$ is given by

$$\rho(t_1,x^1, \ldots , t_n,x^n) = \int \rho_0(x^0)\; p(0,x^0;t_1,x^1) \ldots p(t_{n-1},x^{n-1};t_n,x^n)\; dx^0.$$

A function p(s,x;t,y) is a Markov transition density iff it satisfies the well–known Chapman–Kolmogorov relations, i.e.,

$$\int p(s,x;t,y)\; dy = 1$$

and

$$p(s,x;u,z) = \int_{\mathbb{R}^n} p(s,x;t,y)\; p(t,y;u,z)\; dy$$

where $0 \le s \le t \le u \le T$.

There is a similar development for reciprocal process due to Schrödinger [2] and Jamison [5]. A reciprocal process x(t) is completely determined by the joint density $\rho_{0,T}(x^0,x^T)$ of the end points x(0) and x(T) and a reciprocal transition density q(s,x;t,y;u,z). The latter is the probability density of x(t) = y given that x(s) = x and x(u) = z where $0 \le s \le t \le u \le T$. The finite dimensional densities of x(t) are then given by

$$\rho(t_1,x^1,\ldots,t_n,x^n) = \int \rho_{0,T}(x^0,x^T)\, q(t_0,x^0;t_1,x^1;T,x^T)$$

$$q(t_1,x^1;t_2,x^2;T,x^T)\ldots q(t_{n-1},x^{n-1};t_n,x^n;T,x^T)\, dx^0 dx^T$$

To be a reciprocal transition density, q(s,x;t,y;u,z) must satisfy the Schrödinger–Jamison relations

$$\int q(s,x;t,y;u,z)\, dy = 1$$

and

$$q(r,w;s,x;u,z)\, q(s,x;t,y;u,z) = q(r,w;t,y;u,z)\, q(r,w;s,x;t,y)$$

where $0 \leq r \leq s \leq t \leq u \leq T$ and $w,x,y,z \in \mathbb{R}^n$.

Suppose x(t) is a reciprocal process and $X(t_0,t_1)$ is the associated two time process which is Markov relative to the inclusion partial ordering. One can show that the Chapman–Kolmogorov relations for the Markov transition density of $X(t_0,t_1)$ are equivalent to the Schrödinger–Jamison relations for the reciprocal transition density of x(t).

Schrödinger realized that there is Bayesian way of constructing a reciprocal transition density q from a Markov transition density p,

$$q(s,x;t,y;u,z) = \frac{p(s,x;t,y)\, p(t,y;u,z)}{p(s,x;t,y)}$$

Of course the conditionally Markov property allows one to reverse the process and define a Markov transition density p from a reciprocal transition density q,

$$p(s,x;t,y) = q(s,x;t,y;T,x^T).$$

If we start with a reciprocal transition density q, which we use to define a Markov transition density p which we use to define another reciprocal density $\bar{q}$ then by the second Schrödinger–Jamison relation, $\bar{q} = q$. If we start with a Markov transition density p which we use to define a reciprocal transition density q which we use to define another Markov transition density $\bar{p}$, it does not follow that $\bar{p} = p$.

Schrödinger used a Markov transition density p to construct reciprocal transition density q. With this and an end point density $\rho_{0,T}$ he was able to construct reciprocal processes. Jamison [6] showed that the resulting reciprocal process is actually Markov iff the end point density satisfies

$$\rho_{0,T}(x^0,x^T) = \pi_0(x^0)\ \pi_T(x^T)\ p(0,x^0;T,x^T)$$

for some nonnegative functions $\pi_0(x^0)$ and $\pi_T(x^T)$.

Jamison [6] also studied one dimensional stationary Gaussian reciprocal processes. He showed that covariance r(t) of such a process must satisfy a second order linear differential equation

$$\frac{d^2}{dt^2} r = a\, r$$

where a is a constant. He then used this in an attempt to classify all such processes and this program was successfully completed by Chay [8] and Carmichael–Masse–Theodorescu [11].

The author became interested in reciprocal process through his study of acausal linear systems [14] driven by white noise and satisfying independent random boundary conditions of the form

$$dx = A(t)\, x\, dt + B(t)\, dw$$

$$v = V^0\, x(0) + V^1\, x(t).$$

Here x(t) is an n dimensional Gaussian process, w(t) is a standard n dimensional Wiener process and v is an n dimensional random vector independent of w(t). We assume that the above boundary value problem is well–posed so that the Green's matrix $\Gamma(t,s)$ exists. We can express the solution of the stochastic differential equation as

$$x(t) = \Phi(t,0)\, v + \int_0^T \Gamma(t,s)\, B(s)\, dw(s)$$

where the integral is a Wiener integral and $\Phi(t,s)$ is the fundamental matrix solution of $\dot{x} = Ax$. We have normalized so that $V^0 + V^1\, \Phi(T,0) = I$.

We have proved [14] that the solution of such a stochastic boundary value problem is a reciprocal process and we speculated that every Gaussian reciprocal process is the solution of such a stochastic boundary value problem. This conjecture was motivated by the fact that every Gaussian Markov process is the solution of a stochastic initial value problem, i.e., $V^0 = I$ and $V^1 = 0$. This conjecture is not true and this led us to discover a theory of reciprocal diffusions and stochastic differential equations of second order.

3. Diffusions

We recall the Feller postulates for a Markov diffusion x(t). First some notation, let

x^* denote the transpose of a n dimensional column vector x and x^{*2} the outer product of x with itself $x^{*2} = xx^*$, this is n x n matrix. The forward difference operator d^+ is defined by

$$d^+x(t;dt) = x(t+dt) - x(t)$$

where $dt > 0$ is a small positive quantity. Typically we suppress arguments as in d^+x. Conditional expectation given that $x(t) = x$ is denoted by

$$E_{x(t)}(\cdot) = E(\cdot \mid x(t) = x)$$

The symbol $O(dt)^k$ denotes a function of x,t and dt for which there exist $\epsilon,\delta>0$ such that if $dt < \delta$ then $|O(dt)| < \epsilon\, dt^k$ for all $x \in \mathbb{R}^n$ and $t \in (0,T)$. The symbol $o(dt)^k$ denotes function of x,t and dt which for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $dt < \delta$ then $|o(dt)^k| < \epsilon\, dt^k$.

A Markov process x(t) is a Markov diffusion if there exists n x 1 and n x m valued functions f(x,t) and g(x,t) such that

(MD1) $\text{Prob}\,\{|\, x(t+dt) - x| > \epsilon \mid x(t) = x\} = O(dt)$

(MD2) $E_{x(t)}(d^+x) = f(x,t)\, dt + o(dt)$

(MD3) $E_{x(t)}(d^+x)^{*2} = (g(x,t))^{*2}\, dt + o(dt)$

(MD4) Third and higher centered conditional moments of dx vanish like o(dt).

The interpertation of these postulates is that conditioned on $x(t) = x$, the forward increment d^+x of the process has a mean value approximately equal to f dt and variance approximately equal to g^{*2} dt. In other words x(t) is mean differential but the individual sample paths are not for they have an extremely large standard deviation $O(dt)^{1/2}$.

From these postulates one can deduce that the density $\rho(x,t)$ of x(t) satisfies the Fokker–Plank equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_i}(\rho f_i) - \frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}(\rho\, g_{ik} g_{jk}) = 0.$$

Moreover using the Ito stochastic integral we can realize x(t) as the solution of the stochastic differential equation

$$d^+x = f(x,t)\, dt + g(x,t)\, d^+w$$

$$x(0) = x^0.$$

We now sketch out the foundations of a parallel theory of reciprocal diffusions that we have recently developed. More details can be found in [21]. We need some more notation. We define the centered average, centered first difference and centered second difference as

$$\bar{x}(t;dt) = \frac{x(t+dt) + x(t-dt)}{2}$$

$$dx(t;dt) = \frac{x(t+dt) - x(t-dt)}{2}$$

$$d^2x(t;dt) = x(t+dt) - 2\,x(t) + x(t-dt)$$

Frequently we suppress argument as in $\bar{x}(t)$ or dx. We also introduce another conditional expectation operation

$$E_{\bar{x}(t)}\,(\cdot) = E\,(\,\cdot\mid\bar{x}(t;dt) = x)$$

A reciprocal process x(t) is a reciprocal diffusion if there exists n x 1 valued functions f(x,t) and u(x,t), n x n valued functions g(x,t) and $\pi(x,t)$ and n x m valued function h(x,t) such that

(RD1) $\text{Prob}\,\{\,|x(t)-x| > \epsilon \mid \bar{x}(t;dt) = x\,\} = O(dt)$

(RD2) $E_{\bar{x}(t)}\,(dx) = u(x,t)\,dt + o(dt)$

(RD3) $E_{\bar{x}(t)}\,(d^2x) = (f(x,t) + g(x,t)\,u(x,t))\,dt^2 + o(dt)^2$

(RD4) $E_{\bar{x}(t)}\,(dx)^{*2} = \frac{1}{2}\,(h(x,t))^{*2}\,dt + \pi(x,t)\,dt^2 + o(dt)^2$

(RD5) $E_{\bar{x}(t)}\,(d^2x)^{*2} = 2\,(h(x,t))^{*2}\,dt + o(dt)^2$

(RD6) $E_{\bar{x}(t)}\,(d^2x\,dx^*) = \frac{1}{2}\,g(x,t)(h(x,t))^{*2}\,dt^2 + o(dt)^2$

(RD7) Third and higher joint centered conditional moments of dx and d^2x vanish like $o(dt)^2$

Basically these postulates assert that the first and second joint conditional moments of dx and d^2x exist and have the indicated expansions in power series in dt. They define the coefficients f,g,h,u and π of the power series and they imply certain relation between these coefficients. These definitions and relations are as follows:

(i) The dt part of RD2 defines u.

(ii) The dt^2 part of RD3 defines f + g u.

(iii) The dt and dt^2 parts of RD4 defines h^{*2} and π.

(iv) The dt^2 part of RD6 defines g h^{*2}.

(v) The dt part of RD3 vanishes.

(vi) The dt^2 part of RD5 vanishes and the dt part is four times the dt part of RD4.

Any process satisfying (RD1–6) is called a second order diffusion.

We refer to u as the mean velocity, ρ u as the mean momentum, f + g u as the mean acceleration, h as the noise coefficient and $\rho\ \pi$ as the mean momentum flux of the process x(t). A related quantity $\rho\ \sigma = \rho(u\ u^* - \pi)$ is called the stress tensor. The reason for the terminology will become apparent in a moment.

A reciprocal diffusion satisfying RD1–7 is said to be a solution of the second order stochastic differential equation.

$$d^2x = f(x,t)\ dt^2 + g(x,t)\ dx\ dt + h(x,t)\ d^2w$$

where w(t) is a standard m dimensional Wiener process. This is a (partial) mnemonic for the above portulates. In particular applying $E_{\bar{x}(t)}\ (\cdot)$ we obtain RD3 from RD2 under the assumption that d^2w is independent of $\bar{x}(t)$. Applying $E_{\bar{x}(t)}\ (\cdot)$ to $(d^2x)^{*2}$ yields RD5. Finally RD6 follows from applying $E_{\bar{x}(t)}\ (\cdot)$ to $d^2x\ dx^*$ using RD4.

To get a feeling for these axioms it is convenient to introduce another conditional expectation

$$E_{x(t\pm dt)}\ (\cdot) = E\ (\ \mid x(t\pm dt) = x\pm v\ dt\ )$$

Suppose x(t) is a reciprocal diffusion which also satisfies the stronger conditions.

(RD3*) $E_{x(t\pm dt)}\ (d^2x) = (f(x,t) + g(x,t)\ v)\ dt^2 + o(dt)^2$

(RD4*) $E_{x(t\pm dt)}\ (d^2x)^{*2} = 2(h(x,t))^{*2}\ dt + o(dt)^2$

(RD6) $E_{\bar{x}(t)}\ (d^2x\ dx^*) = \frac{1}{2}\ g(x,t)(h(x,t))^{*2}\ dt^2$
$+ f(x,t)\ u(x,t)^* + g(x,t)\ \pi(x,t)\ dt^3 + o(dt)^3$

Then x(t) is called a strongly reciprocal diffusion.

Conditioned on x(t±dt) = x±v dt the mean sample path of the process over the time interval [t–dt, t+dt] traces out a parabola in (t,x) space passing through (t±dt, x ± v dt) and with second derivative equal to f(x,t) + g(x,t) v. Hence the mean path deviates from the straight line between (t±dt, x±v dt) by $O(dt)^2$. Compared to this, the standard deviation of sample paths from the mean path is very large, $O(dt)^{1/2}$.

Conditioning on $\bar{x}(t;dt) = x$ rather than x(t) = x is crucial to the above development. Even for very nice processes, such as an Ornstein Uhlenbeck process, the quantity

$E_{x(t)}\,(d^2x)$ is $O(dt)$ rather than $O(dt)^2$. In the stochastic mechanics of Nelson [18] the dt part of this quantity is twice the osmotic velocity. Nelson's current velocity, the dt part of $E_{x(t)}\,(dx)$ is generally equal to our mean velocity $u(x,t)$ from (RD2).

The first question that comes to mind is "Are there any reciprocal diffusions?". In [21] we showed that answer is decidely yes. In particular we showed that any reciprocal Gaussian process with smooth covariance $R(t,s)$ satisfying certain technical conditions is a strongly reciprocal diffusion. This includes such Markov processes. For a Gaussian reciprocal process the second order stochastic differential is linear of the form

$$d^2x = F(t)\,x\,dt^2 + G(t)\,dx\,dt + H(t)\,d^2w$$

where

$$f(x,t) = F(t)\,x = \left(\frac{\partial^2 R}{\partial t^2}(t,t) - G(t)\,\frac{\partial R}{\partial t}(t,t)\right) x$$

$$g(x,t) = G(t) = \left(\frac{\partial^2 R}{\partial t^2}(t,t) - \frac{\partial^2 R^*}{\partial s^2}(t,t)\right)\left(\frac{\partial R}{\partial t}(t,t) - \frac{\partial R^*}{\partial s}(t,t)\right)^{-1}$$

$$(h(x,t))^{*2} = (H(t))^{*2} = -\left(\frac{\partial R}{\partial t}(t,t) - \frac{\partial R^*}{\partial s}(t,t)\right)$$

The principle technical conditions are that $R(t,t) = I$ and $H(t)^{*2}$ is invertible. The other quantities $u(x,t)$ and $\pi(x,t)$ are given by

$$u(x,t) = U(t)\,x = \frac{1}{2}\left(\frac{\partial R}{\partial t}(t,t) + \frac{\partial R^*}{\partial s}(t,t)\right) x$$

$$\pi(x,t) = u(x,t)\,u^*(x,t) - \sigma(x,t)$$

$$\sigma(x,t) = -\frac{1}{2}\left(\frac{\partial^2 R}{\partial t\,\partial s}(t,t) + \frac{\partial^2 R^*}{\partial t\,\partial s}(t,t)\right) + U(t)\,U^*(t).$$

All of the above evaluations are at $s = t^-$

Suppose $x(t)$ is Gaussian process and a solution of the first order stochastic boundary value problem

$$dx = A(t)\,x\,dt + B(t)\,dw$$

$$v = V^0\,x(0) + V^1\,x(t)$$

in the sense defined above using the Green's matrix. Assume $R(t,t) = I$ and $B(t)$ is invertible. Then $x(t)$ is a reciprocal diffusion satisfying the second order linear stochastic differential equation above with

$$\begin{aligned}
&H(t) = B(t)\\
&G(t) = -(A^2(t) - A^{*2}(t) + \dot{A}(t) - \dot{A}^*(t))\,(B(t)\,B^*(t))^{-1}\\
&F(t) = A^2(t) + \dot{A}(t) - G(t)\,A(t).
\end{aligned}$$

Because of the complexity of these relations, it is possible for a process to satisfy a relatively simple first order equation and a relatively complicated second order equation or vice versa. The latter is the case for the Brownian Bridge or pinned Wiener process $x(t)$ which satisfies the first order equation

$$d^+x = \frac{-1}{(1-t)}\,x\,dt + d^+w$$

$$x(0) = 0$$

and the second order equation

$$d^2x = d^2w$$

$$x(0) = x(1) = 0.$$

The density ρ of a Markov diffusion satisfies the Fokker–Plank equation. For a strongly reciprocal diffusion the density ρ, mean momentum $\rho\,u$ and mean momentum flux $\rho\,\pi$ satisfy at least in a weak sense a system of hyperbolic conservation laws similar to the continuity, Euler and kinetic energy balance equations of fluid mechanics. They are

$$\frac{\partial}{\partial t}\rho = -\frac{\partial}{\partial x_k}(\rho\,u_k)$$

$$\frac{\partial}{\partial t}(\rho\,u_i) = \rho\,(f + g\,u)_i - \frac{\partial}{\partial x_k}(\rho\,\pi_{ik})$$

and

$$\frac{\partial}{\partial t}(\rho\,\pi_{ij}) = \rho\,(f\,u^* + uf^* + g\pi + \pi g^*)_{ij}$$

$$-\frac{\partial}{\partial x_k}(\rho\,(u_i\,u_j\,u_k - \sigma_{ij}\,u_k - \sigma_{ik}\,u_j - \sigma_{jk}\,u_i))$$

with summation on repeated indices understood. A second order or reciprocal diffusion need only satisfy the first two of these equations.

Suppose we consider a volume with boundary in x–space. If we integrate ρ over this volume we obtain the probability measure of the volume. The first equation states that the time rate of change of the probability of the volume is equal to the flux of particles through the boundary due to the mean velocity.

If we integrate ρ u over the volume, we obtain the total momentum in the volume. The second equation states that the time rate of change of momentum in the volume is equal to the forces acting on the particles in the volume plus the net flux of momentum through the boundary.

If we integrate $\rho\ \pi$ over the volume we obtain the total momentum flux in the volume. Physically this is somewhat hard to comprehend but for smooth processes the contraction $\frac{1}{2}\rho\ \pi_{ii}$ is the kinetic energy. Hence we view $\frac{1}{2}\rho\ \pi_{ij}$ as a tensor form of kinetic energy. More precisely, if λ_i is a constant n vector then the scalar valued process $z(t) = \lambda_i x_i(t)$ has kinetic energy equal to $\frac{1}{2}\rho\ \pi_{ij}\ \lambda_i\lambda_j$. With this interpertation the third equation states that time rate of change of tensor kinetic energy in the volume is equal to the mean work done on the particles in the volume by the force d^2x/dt^2 acting through the distance dx plus the flux of tensor kinetic energy through the surface of the volume. This flux is due to mean tensor kinetic energy $\frac{1}{2}\rho\ u_i\ u_j$ (called internal energy) transported by mean velocity u_k, random tensor kinetic energy $-\frac{1}{2}\rho\ \sigma_{ij}$ transported by mean velocity u_k and mixed random/mean kinetic energy transported by random velocity. The latter represented by the last two terms of the flux are usually described as viscosity or stress terms in fluid dynamics. They represent the transport of energy due to random jumps of particles between regions of differing mean velocity.

The third equation expresses kinetic energy balance at the standard time scale, i.e., the dt^2 part of $E_{\bar{x}(t)}\ (dx)^{*2}$. There is also a form of energy at a fast time scale, i.e., the dt part $E_{\bar{x}(t)}\ (dx)^{*2}$. We call this hyperkinetic energy and its balance is described by another conservation law

$$\frac{\partial}{\partial t}(\rho\, h\, h^*)_{ij} = \frac{\rho}{2}(g\, h\, h^* + h\, h^*\, g^*)_{ij}$$

$$-\frac{\partial}{\partial x_k}(\rho\,(h\ \ h^*)_{ij}\, u_k)$$

which is also satisfied by second order diffusions.

Notice that the first three equations can be viewed independently of this last. We chose the name "hyperkinetic" to suggest a hyperkinetic child sitting at his school desk whose endless fidgeting is to no net effect (except possibly on his teacher).

In [21] we formally derived the four conservation laws from the postulates of a strongly reciprocal diffusion. Although they can be thought of in physical terms, they are not consequences of physical principles or assumptions. We verified that these conservation laws are satisfied by the reciprocal Gaussian processes discussed above.

REFERENCES

[1] Bernstein, S., Sur les laisons entre les grandeurs aleatoires, Proc. of Int. Cong. of Math., Zurich (1932) 288–309.

[2] Schrödinger, E., Uber die Umkehrung der Naturgesetze, Sitz. Ber der Preuss. Akad. Wissen., Berlin Phys. Math. 144 (1931).

[3] Schrödinger, E., Theorie relativiste de l'electron et l'interpretation de la mechanique quantique, Ann. Inst. H. Poincare 2, 269–310 (1932).

[4] Slepian, D., First passage time for a particular Gaussian process, Ann. Math. Statist. 32, 610–612 (1961).

[5] Jamison, B., Reciprocal Processes, Z. Wahrsch. Gebiete 30 (1974) 65–86.

[6] Jamison, B., The Markov Processes of Schrödinger, Z. Wahrsch. Gebiete 32 (1975) 323–331.

[7] Jamison, D., Reciprocal processes: the stationary Gaussian case, Ann. Math. Stat. 41 (1970) 1624–1630.

[8] Chay, S. C., On quasi–Markov random fields, J. Multivar. Anal. 2 (1972) 14–76.

[9] Abrahams, J. and Thomas, J. B., "Some comments on conditionally Markov and reciprocal Gaussian processes", IEEE Trans. Information Theory, vol. IT–27, 523–525, 1981.

[10] Adler, R. J., The geometry of random fields. New York: Wiley, 1981.

[11] Carmichael, J. P., Masse, J. C., and Theodorescu, R., "Processus gaussiens stationnaires reciproques sur un intervalle", C. R. Acad. Sci. Paris, Ser. I, vol. 295, 291–293, 1982.

[12] Carmichael, J. P., Masse, J. C., and Theodorescu, R., "Multivariate reciprocal stationary Gaussian processes", Preprint, Laval Univ., Dept. Math., Quebec, 1984.

[13] Carmichael, J. P., Masse, J. C., and Theodorescu, R., Representations for multivariate reciprocal Gaussian processes, Preprint, Laval Univ., Dept. Math., Quebec, 1986.

[14] Krener, A. J., Reciprocal Processes and the Stochastic Realization Problem for Acausal Systems, in Modeling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North–Holland, Amsterdam, 1986, 197–211.

[15] Krener, A. J., Realizations of Reciprocal Processes, Proceedings IIASA Conf. on Modeling and Adaptive Control, Sopron, 1986.

[16] Minoshin, R. N., (1979). Second–order Markov and reciprocal stationary Gaussian processes. Theor. Prob. Appl. 24, 845–852.

[17] Abrahams, J. (1984). On Miroshin's second–order reciprocal processes. SIAM J. Appl. Math. 44, 190–192.

[18] Nelson, E., Quantum Fluctuations, Princeton Univ. Press, Princeton, NJ, 1985.

[19] Guerra, F. and Morato, L. M., Quantization of dynamical systems and stochastic control theory, Phys. Rev. D. 27 (1983) 1774–1786.

[20] Lévy, P., A special problem of Brownian motion and a general theory of Gaussian random functions, Proc. Third Berkeley Symp. Math Stat. and Prob. 2 (1956) 133–175.

[21] Krener, A. J., Reciprocal diffusions and stochastic differential equations of second order. Preprint, 1987. (Appendixed)

[22] Kushner, H. J., Necessary conditions for continuous parameter stochastic optimization problems, SIAM J. Control and Opt., 10 (1972) 550–565.

[23] Hausmann, U. G., On the stochastic maximum principle, SIAM J. Control and Opt., 19 (1978) 252–269.

[24] Bismut, J. M., Théorie probabiliste du contrôe des diffusions, Mem. AMS 4 (1976) No. 167.

[25] Imre Fényes, Eine wahrscheinlichkeitstheoretische Begründung und

Interpretation der Quantenmechanik, Zeitschrift für Physik 132 (1952) 81–106.

[26] Kunio Yasue, Stochastic calculus of variations, J. of Functional Analysis 41 (1981) 327–340.

[27] Krener, A. J., The asymptotic approximation of nonlinear filters by linear filters. In Theory and Application of Nonlinear Control System, C. I. Byrnes and A. Lindquist, eds. North–Holland, Amsterdam (1986) 359–378.

[28] Frezza, R., S. Karahan, A. J. Krener and M. Hubbard. Application of an efficient nonlinear filter, Preprint, submitted for publication.

[29] Phelps, A. and A. J. Krener, Computation of observer normal form using MACSYMA, Preprint, submitted for publication.

[30] Krener, A. J., S. Karahan, M. Hubbard and R. Frezza, Higher order linear approximations to nonlinear control systems. Proceedings, IEEE Conf. on Decision and Control, Los Angeles (1987).

[31] Krener, A. J., Acausal realization theory, Part. 1; Linear deterministic systems, SIAM J. Control and Optimization, 25 (1987) 499–525.

[32] Krener, A. J., Reciprocal processes and the stochastic realization problem for acausal systems. In Modeling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds. North–Holland, Amsterdam (1986) 197–210.

[33] Krener, A. J., Realizations of reciprocal processes, Preprint, submitted for publication.

[34] Krener, A. J., Reciprocal diffusions and stochastic differential of second order, Preprint, submitted for publication.

Stochastics, Vol. 24, pp. 393–422
Reprints available directly from the publisher
Photocopying permitted by license only

Printed in Great Britain

Reciprocal Diffusions and Stochastic Differential Equations of Second Order*

ARTHUR J. KRENER

Department of Mathematics and Institute of Theoretical Dynamics, University of California, Davis, CA 95616, USA

(Received 27 October 1987; in final form 15 March 1988)

We develop a theory of second order diffusion processes and associated stochastic differential equations of second order. We show that equations of evolution of the density, mean velocity and momentum flux are a family of first order conservation laws similar to those of continuum mechanics. We verify that the theory is satisfied for a large class of reciprocal Gaussian processes.

KEY WORDS: Reciprocal process, second order diffusion, second order stochastic differential equation, conservation law, statistical mechanics, continuum mechanics, stochastic mechanics.

AMS(MOS) (1970): Primary 60H10, Secondary 60J60, 73B99, 82A05, 35L65.

1. RECIPROCAL DIFFUSIONS

One of the most beautiful parts of modern mathematics is the rich and wonderful interplay between Markov diffusion processes, linear parabolic partial differential equations and stochastic differential equations of first order. We shall describe the foundations of a

*Research supported in part by NSF under DMS-8601635 and AFOSR under 85-0267.

parallel theory involving reciprocal diffusion processes, nonlinear conservation laws and stochastic differential equations of second order.

Let $(\Omega, \mathscr{F}, \Pr)$ be a probability triple consisting of a sample space Ω, σ-algebra of events $\mathscr{F}$ and probability measure Pr. Let E denote expectation with respect to Pr. Throughout we let $x(t)$ denote a stochastic process over this triple defined for $t \in [0, T]$ and taking values in $\mathbb{R}^{n \times 1}$. We assume that

$$E\int_0^T |x(t)|^2 \, dt < \infty.$$

Given $0 \leqq t_0 \leqq t_1 \leqq T$, let $\mathscr{T}(t_0, t_1)$, $\mathscr{E}(t_0, t_1)$ and $\mathscr{B}(t_0, t_1)$ be the σ subalgebras of $\mathscr{F}$ generated by $x(t)$ interior to, exterior to and on the boundary of the interval defined by t_0 and t_1. In other words

$$\mathscr{T}(t_0, t_1) = \sigma\{x(t): t \in [t_0, t_1]\}$$

$$\mathscr{E}(t_0, t_1) = \sigma\{x(t): t \in [0, t_0] \cup [t_1, T]\}$$

$$\mathscr{B}(t_0, t_1) = \sigma\{x(t_0), x(t_1)\}.$$

We denote by $\bar{\mathscr{T}}(t_0, t_1)$, $\bar{\mathscr{E}}(t_0, t_1)$ and $\bar{\mathscr{B}}(t_0, t_1)$ the space of square integrable random variables which are measurable with respect to $\mathscr{T}(t_0, t_1)$, $\mathscr{E}(t_0, t_1)$ and $\mathscr{B}(t_0, t_1)$ respectively.

The concept of a reciprocal process was introduced by Bernstein [1] following ideas of Schrödinger [2,3]. A process $x(t)$ is *reciprocal* if on every subinterval of $[0, T]$, the interior and exterior are conditional independent given the boundary. More precisely, if $\phi \in \bar{\mathscr{T}}(t_0, t_1)$ and $\psi \in \bar{\mathscr{E}}(t_0, t_1)$ then

$$E(\phi\psi \mid \mathscr{B}(t_0, t_1)) = E(\phi \mid \mathscr{B}(t_0, t_1)) E(\psi \mid \mathscr{B}(t_0, t_1)).$$

We refer the reader to [4–17] for more detailed discussions of reciprocal processes. The have also been called quasi Markov or Bernstein processes. They are closely related to conditionally Markov processes. Following Schrödinger's original motivation and Nelson's stochastic mechanics, Zambrini [20–22] has related reciprocal processes to quantum mechanics.

It follows immediately from the definition that Markov processes are reciprocal but not vice versa [5].

To define reciprocal diffusions we must introduce some notation. Given a process $x(t)$ and a small time increment $dt > 0$, define the centered average evaluation $\bar{x}$, centered first difference dx and centered second difference d^2x as

$$\bar{x}(t; dt) = \frac{x(t+dt) + x(t-dt)}{2}$$

$$dx(t; dt) = \frac{x(t+dt) - x(t-dt)}{2}$$

$$d^2x(t; dt) = x(t+dt) - 2x(t) + x(t-dt).$$

Frequently when the context is clear we suppress the argument dt as in $\bar{x}(t)$, or both t and dt as in dx. We also have the forward d^+x and backward d^-x first differences

$$d^+x(t; dt) = x(t+dt) - x(t)$$

$$d^-x(t; dt) = x(t) - x(t-dt).$$

In contrast to the standard conditional expectation of Markov theory,

$$E_x(\cdot) = E_{x(t)}(\cdot) = E(\cdot \mid x(t) = x)$$

we shall utilize

$$E_{\bar{x}}(\cdot) = E_{\bar{x}(t;dt)}(\cdot) = E(\cdot \mid \bar{x}(t; dt) = x)$$

and occasionally the stronger conditioning

$$E_{\bar{x},dx}(\cdot) = E_{x(t \pm dt)}(\cdot) = E(\cdot \mid x(t \pm dt) = x \pm v\,dt)$$

$$= E(\cdot \mid \bar{x}(t; dt) = x, dx(t; dt) = v\,dt).$$

We now give the second order analogs of the Feller postulates for a first order diffusion. A stochastic process $x(t)$ is a *second order diffusion* if there exist functions $f_i(x,t)$, $g_{ij}(x,t)$, $h_{ik}(x,t)$, $u_i(x,t)$ for $i, j=1,\ldots,m$ such that

$$E_{\bar{x}}(dx_i)=u_i(x,t)\,dt+o(dt) \tag{1.1a}$$

$$E_{\bar{x}}(dx_i\,dx_j)=\tfrac{1}{2}h_{ik}(x,t)h_{jk}(x,t)\,dt+\pi_{ij}(x,t)\,dt^2+o(dt)^2 \tag{1.1b}$$

$$E_{\bar{x}}(d^2x_i)=(f_i(x,t)+g_{ij}(x,t)u_j(x,t))\,dt^2+o(dt)^2 \tag{1.1c}$$

$$E_{\bar{x}}(d^2x_i\,d^2x_j)=2h_{ik}(x,t)h_{jk}(x,t)\,dt+o(dt)^2 \tag{1.1d}$$

$$E_{\bar{x}}(d^2x_i\,dx_j)=\tfrac{1}{2}g_{ik}(x,t)h_{kl}(x,t)h_{jl}(x,t)\,dt^2+o(dt)^2 \tag{1.1e}$$

The higher conditional moments of dx and d^2x agree to the lowest nonzero powers of dt with those of Gaussians with the above first and second moments. (1.1f)

In the above we have utilized the summation convention. Conditioned on $\bar{x}(t;dt)=x$ for fixed x,t, the expression $o(dt)^k$ is a deterministic quantity $y(dt;x,t)$ which vanishes faster than dt^k as $dt\to 0$ uniformly in x and t. In other words for every $\varepsilon>0$ there exist $\delta>0$ such that $|y(dt;x,t)|<\varepsilon\,dt^k$ for all $dt<\delta, x\in\mathbb{R}^n$ and $t\in(0,T)$. We denote by $0(dt)^k$ a quantity $y(dt;x,t)$ for which there exist $\varepsilon,\delta>0$ such that $|y(dt;x,t)|<\varepsilon\,dt^k$ if $dt<\delta$.

If $x(t)$ is both a reciprocal process and a second order diffusion then we say it is a *reciprocal diffusion.*

A stochastic process $x(t)$ satisfies the *second order stochastic differential equation*

$$d^2x=f(x,t)\,dt^2+g(x,t)\,dx\,dt+h(x,t)\,d^2w \tag{1.2}$$

where $w(t)$ is a standard m dimensional Wiener process if $x(t)$ is a reciprocal diffusion satisfying (1.1a–f). Actually (1.2) is a mnemonic description of (1.1) in the same way that the first order stochastic differential equation

$$d^+x=f(x,t)\,dt+h(x,t)\,d^+w \tag{1.3}$$

is a mnemonic for the axioms of a first order diffusion.

$$E_x(d^+x_i) = f_i(x,t)\,dt + o(dt) \tag{1.4a}$$

$$E_x(d^+x_i\,d^+x_j) = h_{ik}(x,t)h_{jk}(x,t)\,dt + o(dt) \tag{1.4b}$$

all higher centered conditional moments of dx vanish like $o(dt)$. (1.4c)

In particular (1.1c) asserts that conditional mean acceleration equals $f+gu$ where u is the conditional mean velocity given by (1.1a). Note the difference between the conditional expectations in (1.2) and (1.4). Conditioning on $\bar{x}(t;dt)=x$ is an essential part of the second order stochastic calculus. If we were to condition on $x(t)=x$, we would find that generally $E_x(d^2x)$ is order dt rather than dt^2. In fact it is precisely this conditioning which distinguishes our work from that of Nelson [18] and Zambrini [20–22]. In Nelson's stochastic mechanics the order dt part of $E_x(dx)$ is a vector field $v(x,t)$ called the current velocity while the order dt part of $1/2E_x(d^2x)$ is another vector field $u(x,t)$ called the osmotic velocity. In our work the order dt part of $E_{\bar{x}}(dx)$ is a vector field $u(x,t)$ called the conditional mean velocity. For a Gaussian process with a smooth covariance, Nelson's current velocity equals our conditional mean velocity and we suspect that this is true whenever both exist. On the other hand in our theory $E_{\bar{x}}(d^2x)$ is postulated to be of order dt^2. Hence the coefficient of dt^2 can be viewed as an acceleration. For this reason it differs from Nelson's osmotic velocity. Zambrini's work also uses the Nelson framework.

We call the $n\times 1$ vector field $u(x,t)$ the *mean velocity*. The density of $x(t)$ is denoted by $\rho(x,t)$. The $n\times n$ tensor field $\rho(x,t)\ \pi(x,t)$ is called the *momentum flux tensor*. A related $n\times n$ tensor field, $\rho(x,t)$ $\sigma_{ij}(x,t)=\rho(x,t)(u_i(x,t)\ u_j(x,t)-\pi_{ij}(x,t))$, is called the *stress tensor*. The reason for these names will become apparent in Section Four.

Formulas (1.1,2) suggest that the random parts of dx and d^2x conditioned on $\bar{x}(t;dt)=x$ are given by

$$\widetilde{dx} = dx - E_{\bar{x}}(dx) = dx - u(x,t)\,dt$$

$$\widetilde{d^2x} = d^2x - E_{\bar{x}}(d^2x) = g(x,t)\widetilde{dx} + h(x,t)\,d^2w.$$

We use x^* denote the transpose of x and x^{*2} to denote the symmetric square or outer product, $x^{*2}=xx^*$.

Note that $u(x,t)$ and $\pi(x,t)$ from (1.1) do not appear explicitly in (1.2). As we shall see in Section 4, this is because these quantities and the density $\rho(x,t)$ satisfy a system of nonlinear conservation laws determined by f, g and h. This system of four first order partial differential equations is very similar to the equations of fluid and continuum mechanics. They express conservation of probability, balance of momentum and balance of a tensor form of work in two time scales. They replace the familiar Fokker–Planck equation for a first order diffusion (1.3, 4).

Equations (1.1a–e) assert that the conditional moments of dx and d^2x can be expanded in powers of dt as shown. These formulas give names to the coefficients. The only constraints on the coefficients are found by comparing (1.1b, d and e). Equation (1.1a) asserts that the conditional mean velocity exists and gives it a name $u(x,t)$. Equation (1.1b) describes the variance of dx. The order dt part arises from the fluctuation of the second difference $d^2w\ (t;dt)$ of the Wiener process that appears in (1.2). The factors of 1/2 and 2 in the dt part of (1.1b, d) are explained by

$$E_{\bar{x}}(dw)^{*2}=\tfrac{1}{2}I\,dt$$

$$E_{\bar{x}}(d^2w)^{*2}=2I\,dt.$$

The second order part $\pi\,dt^2$ of (1.1b) has a deterministic contribution $u^{*2}\,dt^2$ from (1.1a) and a stochastic contribution $-\sigma\,dt^2=(\pi-u^{*2})\,dt^2$ from the noise throughout $[0,T]$.

The immediate question that arises is "Are there any second order or reciprocal diffusions?". We answer this in the affirmative in the next section by showing that under mild technical assumptions Gaussian processes with smooth covariances are second order diffusions, and Gaussian reciprocal processes with smooth covariances are reciprocal diffusions. Of course the latter includes Gauss–Markov processes. We derive explicit formulas for the quantities appearing in (1.1) in terms of the covariance matrix of the process.

In Section Three we explore how second order diffusions transform under change of variables and in Section Four we derive the conservation and balance laws which are described above. In Section 5 we verify that Gaussian reciprocal processes of Section 2 satisfy these laws.

2. RECIPROCAL AND GAUSSIAN PROCESSES

Let $x(t)$ be a Gaussian process defined on $[0, T]$ and taking values in $\mathbb{R}^{n \times 1}$. For convenience we assume $x(t)$ is zero mean and we denote by $R(t,s)$ its covariance matrix,

$$E(x(t)) = 0$$

$$E(x(t)x^*(s)) = R(t,s).$$

We shall assume that $R(t,s)$ is a smooth (C^∞) function of t, s in the triangle $0 \leqq s \leqq t \leqq T$ and the limits of R and its partial derivatives exist and are continuous on the closed triangle. Because $R(t,s) = R^*(s,t)$ we need only consider R in this triangle.

We shall also assume that

$$R(t,t) = I \tag{2.1a}$$

$$\begin{bmatrix} I & R^*(t+\tau, t-\tau) \\ R(t+\tau, t-\tau) & I \end{bmatrix} \text{is nonsingular for small } \tau > 0 \tag{2.1b}$$

$$\frac{\partial R}{\partial t}(t,t) + \frac{\partial R^*}{\partial s}(t,t) < 0. \tag{2.1c}$$

All evaluations of R and its partials at $t = s$ are limits of values in the interior of the triangle $0 \leqq s \leqq t \leqq T$. Notice that (2.1a) is merely a normalization assuming $R(t,t)$ is invertible. Moreover (2.1c) essentially implies (2.1a, b) holds for almost all t.

In [15] we showed that any stationary reciprocal Gaussian process satisfying (2.1) has a C^∞ covariance in the above sense and moreover the covariance $R(t-s) = R(t,s)$ satisfies a pair of second order matrix differential equations

$$\ddot{R} = G\dot{R} + FR \tag{2.2a}$$

$$\ddot{R}^* = -G\dot{R}^* + FR^*. \tag{2.2b}$$

We now extend this to the nonstationary case. We refer the reader to [14, 15] for full details.

Let $0 \leqq s < t < T$ then by the Gaussian and reciprocal properties the covariance $R(t,s)$ satisfies for $\tau > 0$ sufficiently small

$$R(t,s)=[R(t,t-\tau)R(t,t+\tau)]\begin{bmatrix} I & R^*(t+\tau,t-\tau) \\ R(t+\tau,t-\tau) & I \end{bmatrix}^{-1}$$

$$\times\begin{bmatrix} R(t-\tau,s) \\ R(t+t,s) \end{bmatrix}. \tag{2.3}$$

We define $K_i(t,\tau)$ in the obvious fashion so that this becomes

$$R(t,s)=[K_1(t,\tau)K_2(t,\tau)]\begin{bmatrix} R(t-\tau,s) \\ R(t+\tau,s) \end{bmatrix}$$

We differentiate this twice with respect to τ to obtain

$$0=[K_1(t,\tau)K_2(t,\tau)]\begin{bmatrix} \dfrac{\partial^2}{\partial t^2}R(t-\tau,s) \\ \dfrac{\partial^2}{\partial t^2}R(t+\tau,s) \end{bmatrix}$$

$$+2\left[\frac{\partial K_1}{\partial \tau}(t,\tau)\frac{\partial K_2}{\partial \tau}(t,\tau)\right]\begin{bmatrix} \dfrac{\partial}{\partial t}R(t-\tau,s) \\ \dfrac{\partial}{\partial t}R(t+\tau,s) \end{bmatrix}$$

$$+\left[\frac{\partial^2 K_1}{\partial \tau^2}(t,\tau)\frac{\partial^2 K_2}{\partial \tau^2}(t,\tau)\right]\begin{bmatrix} R(t-\tau,s) \\ R(t+\tau,s) \end{bmatrix}.$$

By (2.1) and arguments similar to those of [15] we verify that the limits of $K(t,\tau)$, $\partial/\partial\tau K(t,\tau)$ and $(\partial/\partial\tau)^2K(t,\tau)$ exist as $\tau \to 0$. In particular $K_i(t,0)=\frac{1}{2}I$ and so we obtain for all $0 \leqq s < t < T$

$$\frac{\partial^2}{\partial t^2}R(t,s)=G(t)\frac{\partial R}{\partial t}(t,s)+F(t)R(t,s) \tag{2.4}$$

where

$$G(t)=-2\left(\frac{\partial K_1}{\partial \tau}(t,0)-\frac{\partial K_2}{\partial \tau}(t,0)\right) \tag{2.5a}$$

$$F(t) = -\left(\frac{\partial^2 K_1}{\partial \tau^2}(t,0) + \frac{\partial^2 K_2}{\partial \tau^2}(t,0)\right). \tag{2.5b}$$

But notice in our derivation of (2.4) we did not use the fact that $t > s$, we only used the fact that $t \in (t-\tau, t+\tau)$ and $s \notin (t-\tau, t+\tau)$. Hence (2.4) must also hold for $0 < t < s \leqq T$. Since $R(t,s) = R^*(s,t)$ we conclude that for $0 < t < s \leqq T$,

$$\frac{\partial^2}{\partial t^2} R^*(s,t) = G(t)\frac{\partial R^*}{\partial t}(s,t) + F(t)R^*(s,t).$$

By interchanging the symbols t and s, we see that for $0 < s < t \leqq T$ we have

$$\frac{\partial^2}{\partial s^2} R^*(t,s) = G(s)\frac{\partial R^*}{\partial s}(t,s) + F(s)R^*(t,s) \tag{2.6}$$

By continuity (2.4) and (2.6) must hold for $0 \leqq s \leqq t \leqq T$. We also obtain alternative formulas for $G(t)$ and $F(t)$, namely

$$G(t) = \left(\frac{\partial^2 R}{\partial t^2}(t,t) - \frac{\partial^2 R^*}{\partial s^2}(t,t)\right)\left(\frac{\partial R}{\partial t}(t,t) - \frac{\partial R^*}{\partial s}(t,t)\right)^{-1} \tag{2.7a}$$

$$F(t) = \frac{\partial^2 R}{\partial t^2}(t,t) - G(t)\frac{\partial R}{\partial t}(t,t) = \frac{\partial^2 R^*}{\partial s^2}(t,t) - G(t)\frac{\partial R^*}{\partial s}(t,t) \tag{2.7b}$$

We define $H(t)H^*(t)$ by

$$H(t)H^*(t) = -\left(\frac{\partial R}{\partial t}(t,t) - \frac{\partial R^*}{\partial s}(t,t)\right). \tag{2.7c}$$

We have proved

THEOREM 2.1 *Let $x(t)$ be a zero mean Gaussian reciprocal process with smooth covariance $R(t,s)$ satisfying* (2.1). *Then $R(t,s)$ satisfies the matrix differential equations* (2.4) *and* (2.6) *on $0 \leqq s \leqq t \leqq T$ where $F(t)$ and $G(t)$ are given by* (2.7*a*, *b*).

Every Gauss–Markov process $x(t)$ with a smooth covariance is an Ornstein–Uhlenbeck process, i.e., a solution of a first order linear stochastic differential equation of the form

$$d^+x = F(t)x\,dt + H(t)\,d^+w.$$

The next theorem shows that every Gaussian reciprocal process $x(t)$ with smooth covariance satisfying (2.1) is a solution of a second order linear stochastic differential equation

$$d^2x = F(t)x\,dt^2 + G(t)\,dx\,dt + H(t)\,d^2w.$$

THEOREM 2.2 *Let $x(t)$ be a Gaussian process with smooth covariance $R(t,s)$ satisfying* (2.1). *Then $x(t)$ is a second order diffusion. If $x(t)$ is also reciprocal then it is a reciprocal diffusion. In either case $f(x,t)=F(t)x$, $g(x,t)=G(t)$ and $h(x,t)=H(t)$ of* (2.7) *and $u(x,t)$, $\pi(x,t)$ are given by*

$$u(x,t)=U(t)x \tag{2.8a}$$

$$U(t)=\frac{1}{2}\left(\frac{\partial R}{\partial t}(t,t)+\frac{\partial R^*}{\partial s}(t,t)\right) \tag{2.8b}$$

$$\pi(x,t)=u(x,t)u^*(x,t)-\sigma(x,t) \tag{2.8c}$$

$$\sigma(x,t)=\sigma(t)=-\frac{1}{2}\left(\frac{\partial^2 R}{\partial t\,\partial s}(t,t)+\frac{\partial^2 R^*}{\partial t\,\partial s}(t,t)\right)+U(t)U^*(t). \tag{2.8d}$$

Proof The proof of this theorem is a straightforward exercise in computing conditional expectations of Gaussian random variables. We shall sketch the details.

By assumption (2.1)

$$E(x(t))^{*2}=R(t,t)=I \tag{2.9a}$$

so

$$0=\frac{\partial R}{\partial t}(t,t)+\frac{\partial R}{\partial s}(t,t) \tag{2.9b}$$

and

$$0=\frac{\partial^2 R}{\partial t}(t,t)+2\frac{\partial^2 R}{\partial t\,\partial s}(t,t)+\frac{\partial^2 R}{\partial s^2}(t,t). \tag{2.9c}$$

In particular (2.9b) insures that HH^* (2.7a) is symmetric and $U(t)$ (2.8b) is skew symmetric.

Next

$$\begin{aligned}E(dx(t))^{*2}=\tfrac{1}{4}(R(t+dt,t+dt)-R(t+dt,t-dt)\\ -R^*(t+dt,t-dt)+R(t-dt,t-dt))\end{aligned}$$

By Taylor series expansion and (2.9) we obtain

$$\begin{aligned}E(dx(t))^{*2}=&-\frac{1}{2}\left(\frac{\partial R}{\partial t}(t,t)-\frac{\partial R^*}{\partial s}(t,t)\right)dt\\ &+\frac{1}{2}\left(\frac{\partial^2 R}{\partial t\,\partial s}(t,t)+\frac{\partial^2 R^*}{\partial t\,\partial s}(t,t)\right)dt^2+o(dt)^2.\end{aligned} \tag{2.10a}$$

In a similar fashion we obtain

$$E(\bar{x}(t))^{*2}=I+\frac{1}{2}\left(\frac{\partial R}{\partial t}(t,t)+\frac{\partial R^*}{\partial s}(t,t)\right)dt+o(dt) \tag{2.10b}$$

$$E(d^2x(t))^{*2}=-2\left(\frac{\partial R}{\partial t}(t,t)-\frac{\partial R^*}{\partial s}(t,t)\right)dt+o(dt)^2 \tag{2.10c}$$

$$E(dx(t)\bar{x}(t)^*)=\frac{1}{2}\left(\frac{\partial R}{\partial t}(t,t)+\frac{\partial R^*}{\partial s}(t,t)\right)dt+o(dt) \tag{2.10d}$$

$$E(d^2x(t)\bar{x}(t)^*)=\frac{1}{2}\left(\frac{\partial^2 R}{\partial t^2}(t,t)+\frac{\partial^2 R^*}{\partial s^2}(t' t)\right)dt^2+o(dt)^2 \tag{2.10e}$$

$$E(d^2x(t)\,dx(t)^*)=-\frac{1}{2}\left(\frac{\partial^2 R}{\partial t^2}(t,t)-\frac{\partial^2 R^*}{\partial s^2}(t,t)\right)dt^2$$

$$+\frac{1}{2}\left(\frac{\partial^3 R}{\partial s\,\partial t^2}(t,t)+\frac{\partial^3 R^*}{\partial t\,\partial s^2}\right)dt^3+o(dt)^3. \quad (2.10f)$$

Therefore we obtain

$$\begin{aligned}
E_{\bar{x}}(dx) &= E(dx\bar{x}(t)^*)(E(\bar{x}(t))^{*2})^{-1}x\\
&= \frac{1}{2}\left(\frac{\partial R}{\partial t}(t,t)+\frac{\partial R^*}{\partial s}(t,t)\right)x\,dt+o(dt)\\
&= U(t)x\,dt+o(dt) \qquad (2.11a)
\end{aligned}$$

and in a similar fashion

$$\begin{aligned}
E_{\bar{x}}(d^2x) &= \frac{1}{2}\left(\frac{\partial^2 R}{\partial t^2}(t,t)+\frac{\partial^2 R^*}{\partial s^2}(t,t)\right)x\,dt^2+o(dt)^2\\
&= (F(t)+G(t)U(t))x\,dt^2+o(dt)^2. \qquad (2.11b)
\end{aligned}$$

One can also show that

$$E_{\bar{x},dx}(d^2x)=(F(t)x+G(t)v)\,dt^2+o(dt)^2. \qquad (2.11c)$$

To obtain the conditional second moments of dx and d^2x, we utilize the particular property of Gaussian random variables that the conditional variance is independent of the conditioning and so

$$E_{\bar{x}}(dx-E_{\bar{x}}(dx))^{*2}=E(dx-E_{\bar{x}}(dx))^{*2}.$$

Hence

$$\begin{aligned}
E_{\bar{x}}(dx)^{*2} &= E(dx)^{*2}+(E_{\bar{x}}(dx))^{*2}-E(E_{\bar{x}}(dx))^{*2}\\
&= -\frac{1}{2}\left(\frac{\partial R}{\partial t}(t,t)-\frac{\partial R^*}{\partial s}(t,t)\right)dt+\left(\frac{1}{2}\left(\frac{\partial^2 R}{\partial t\,\partial s}(t,t)+\frac{\partial^2 R^*}{\partial t\,\partial s}(t,t)\right)\right.\\
&\quad \left.+\,U(t)(xx^*-I)U^*(t)\right)dt^2+o(dt)^2
\end{aligned}$$

$$=\tfrac{1}{2}H(t)H^*(t)\,dt+\pi(x,t)\,dt^2+o(dt)^2. \tag{2.12a}$$

In a similar fashion we conclude that

$$E_{\bar{x}}(d^2x)^{*2}=E(d^2x)^{*2}+(E_{\bar{x}}(d^2x))^{*2}-E(E_{\bar{x}}(d^2x))^{*2}$$

$$=2H(t)^{*2}\,dt+o(dt)^2 \tag{2.12b}$$

$$E_{\bar{x},dx}(d^2x)^{*2}=2H(t)^{*2}\,dt+o(dt)^2 \tag{2.12c}$$

and

$$E_{\bar{x}}(d^2x\,dx^*)=E(d^2x\,dx^*)+E_{\bar{x}}(d^2x)(E_{\bar{x}}(dx))^*$$

$$-E(E_{\bar{x}}(d^2x)(E_{\bar{x}}(dx))^*)$$

$$=-\frac{1}{2}\left(\frac{\partial^2 R}{\partial t^2}(t,t)-\frac{\partial^2 R^*}{\partial s^2}(t,t)\right)dt^2$$

$$+\frac{1}{2}\left(\frac{\partial^3 R}{\partial s\,\partial t^2}(t,t)+\frac{\partial^3 R^*}{\partial t\,\partial s^2}(t,t)\right.$$

$$\left.+(F(t)+G(t)U(t))(xx^*-I)U^*(t)\right)dt^3+o(dt)^3$$

$$=\tfrac{1}{2}G(t)H(t)H(t)^*\,dt^2+o(dt)^2. \tag{2.12d}$$

If $x(t)$ reciprocal then by utilizing the sum of partials of (2.4) and (2.6) with respect to s and t respectively we obtain

$$E_{\bar{x}}(d^2x\,dx^*)=\tfrac{1}{2}G(t)H(t)H^*(t)\,dt^2$$

$$+((F(t)+G(t)U(t))xx^*U^*(t)$$

$$-G(t)\sigma(t))\,dt^3+o(dt)^3. \tag{2.12e}$$

Of course (1.1f) follows from the Gaussian assumption. Q.E.D.

It is enlightening to apply the above formulas to particular classes of reciprocal processes. For example suppose $x(t)$ is a stationary

Gauss–Markov process satisfying (2.1) with covariance $R(t,s)=R(t-s)$. It is well-known that $R(t)=\exp(At)$ where the spectrum of A lies in the open left half of the complex plane and that $x(t)$ satisfies on $[0.\infty]$ the first order stochastic differential equation

$$d^+x=Ax\,dt+Bd^+w \tag{2.13a}$$

$$x(0)\sim N(0,I) \tag{2.13b}$$

where $w(t)$ is an n dimensional standard Wiener process independent of $x(0)$ and the fluctuation-dissipation relation is satisfied,

$$A+A^*+BB^*=0. \tag{2.13c}$$

By the above discussion this process also satisfies the second order stochastic differential equation

$$d^2x=Fx\,dt^2+G\,dx\,dt+H\,d^2w \tag{2.14a}$$

where

$$F=A^2-GA \tag{2.14b}$$

$$G=-(A^2-A^{*2})(HH^*)^{-1} \tag{2.14c}$$

$$HH^*=BB^*. \tag{2.14d}$$

Moreover

$$U=\tfrac{1}{2}(A-A^*) \tag{2.15a}$$

$$\sigma=\tfrac{1}{2}(A^2+A^{*2})+UU^*. \tag{2.15b}$$

The stationary Gaussian reciprocal one dimensional processes have been completely classified [7,8,11]. See also [15]. The covariance $R(t)$ must satisfy (2.2) which in the case of scalars reduces to

$$\ddot{R}=FR.$$

There are three cases $F>0$, $F=0$ and $F<0$. If $F>0$ there are after various normalizations only 3 possibilities.

1.a) Ornstein–Uhlenbeck Process: $R(t)=e^{-t}$, $t>0$, which satisfies the second order equation.

$$d^2x = x\,dt^2 + \sqrt{2}\,d^2w \qquad x(0) \sim N(0,1)$$

and $U=0$, $\sigma=1$. Of course this is also Markov and satisfies the first order equation

$$d^+x = -x\,dt + \sqrt{2}\,d^+w.$$

1.b) Cosh Process: $R(t)=\cosh(\frac{1}{2}-t)/\cosh\frac{1}{2}$ for $0 \leqq t \leqq 1$ which satisfies

$$d^2x = x\,dt^2 + \sqrt{2\tanh\tfrac{1}{2}}\,d^2w \qquad x(0)=x(1) \sim N(0,1)$$

and $U=0$, $\sigma=1$. This process is not Markov but it does have a realization by a stochastic differential equation with an independent boundary condition [15].

1.c) Sinh Process: $R(t)=\sinh(\frac{1}{2}-t)/\sinh\frac{1}{2}$ for $0 \leqq t \leqq 1$ which satisfies

$$d^2x = x\,dt^2 + \sqrt{2\coth\tfrac{1}{2}}\,d^2w \qquad x(0) = -x(1) \sim N(0,1)$$

and $U=0$, $\sigma=1$. Again this is not Markov but can be realized by a stochastic differential equation with an independent boundary condition [15].

2) Slepian Process [4]: $R(t)=1-2t$ for $0 \leqq t \leqq 1$ which satisfies

$$d^2x = 2\,d^2w$$

$$x(0) = -x(1) \sim N(0,1)$$

and $U=0$, $\sigma=0$. This is not Markov and again has a stochastic boundary realization [15].

3.a) Cosine Process: $R(t)=\cos t$ for $-\infty<t<\infty$ which satisfies

$$d^2x = -x\,dt^2 \qquad x(0) = -x(\pi) \sim N(0,1)$$

and $U=0$, $\sigma=-1$. This is not Markov but the sample paths are completely determined by $x(t_1)$, $x(t_2)$ where t_1-t_2 is not a multiple of π.

3.b) Shifted Cosine Process: $R(t)=\cos(t+\tau)/\cos\tau$ for $0\leqq t\leqq\pi-2\tau$. To be a covariance, τ must satisfy $0\leqq\tau<\pi/2$. The process $x(t)$ satisfies

$$d^2x=-x\,dt^2+\sqrt{2\tan\tau}\,d^2w \qquad x(0)=-x(\pi-2\tau)\sim N(0,1)$$

and $U=0$, $\sigma=1$. This process is not Markov and cannot be realized by a scalar first order stochastic differential equation with initial or boundary condition.

We close this section with another interesting example. The Brownian bridge or pinned Wiener process $x(t)$ is obtained from a standard Wiener process by conditioning that $x(0)=x(1)=0$. Another representation is $x(t)=w(t)-tw(1)$ where $w(t)$ is a standard Wiener process. This is a zero mean Gauss–Markov process with covariance $R(t,s)=s(1-t)$ for $0\leqq s\leqq t\leqq 1$. It satisfies the first order differential equation

$$d^+x=\frac{-x}{(1-t)}dt+d^+w \qquad x(0)=0$$

and also satisfies the second order differential equation

$$d^2x=d^2w \qquad x(0)=x(1)=0.$$

Note that this is essentially the same differential equation as that of the Slepian Process.

3. CHANGE OF VARIABLES

In this section we develop some formulas that we shall need in the next. Suppose $x(t)$ is a second order diffusion satisfying (1.1) and (1.2). Let $\phi(x,t),\psi(x,t)$ be C^∞ scalar valued functions and define $\phi(t)=\phi(x(t),t)$ $\psi(t)=\psi(x(t),t)$. We compute the centered mean, first difference and second difference of $\phi(t)$ using the identities

$$dx(t) = \pm(x(t \pm dt) - \bar{x}(t)) \tag{3.1a}$$

$$d^2x(t) = 2(\bar{x}(t) - x(t)). \tag{3.1b}$$

Now

$$\bar{\phi}(t) = \frac{\phi(t+dt) + \phi(t-dt)}{2}$$

$$= \phi(\bar{x}(t), t) + \tfrac{1}{2}(\phi(t+dt) - \phi(\bar{x}(t), t) + \phi(t-dt) - \phi(\bar{x}(t), t))$$

$$\bar{\phi}(t) = \phi(\bar{x}(t), t) + \frac{1}{2}\frac{\partial^2 \phi}{\partial x_i \partial x_j}(\bar{x}(t), t)\, dx_i\, dx_j + O(dx)^4 + O(dt)^2. \tag{3.2a}$$

The symbols $O(dx)^4$ and $O(dt)^2$ denote quantities that go to zero as fast as $|dx|^4$ and dt^2 respectively. Next

$$d\phi = \frac{\phi(t+dt) - \phi(t-dt)}{2} + \frac{\phi(\bar{x}(t), t) - \phi(\bar{x}(t), t)}{2}.$$

By a similar Taylor expansion we obtain

$$\begin{aligned} d\phi = {} & \frac{\partial \phi}{\partial x_i}(\bar{x}(t), t)\, dx_i + \frac{\partial \phi}{\partial t}(\bar{x}(t), t)\, dt \\ & + \frac{1}{6}\frac{\partial^3 \phi(\bar{x}(t), t)}{\partial x_i \partial x_j \partial x_k} dx_i\, dx_j\, dx_k \\ & + \frac{1}{2}\frac{\partial^3 \phi(\bar{x}(t), t)}{\partial x_i \partial x_j \partial t} dx_i\, dx_j\, dt \\ & + \frac{1}{2}\frac{\partial^3 \phi(\bar{x}(t), t)}{\partial x_i \partial t^2} dx_i\, dt^2 \\ & + \frac{1}{120}\frac{\partial^5 \phi(\bar{x}(t), t)}{\partial x_i \partial x_j \partial x_k \partial x_l \partial x_m} dx_i\, dx_j\, dx_k\, dx_l\, dx_m \\ & + O(dx)^6 + O(dx)^4 O(dt) + O(dx)^2 O(dt)^2 + O(dt)^3 \end{aligned} \tag{3.2b}$$

and

$$d^2\phi=\frac{\partial\phi}{\partial x_i}(\bar{x}(t),t)\,d^2x_i+\frac{\partial^2\phi}{\partial x_i\,\partial x_j}(\bar{x}(t),t)\left(dx_i\,dx_j-\frac{1}{4}d^2x_i\,d^2x_j\right)$$

$$+2\frac{\partial^2\phi}{\partial t\,\partial x_i}(\bar{x}(t),t)\,dx_i\,dt+\frac{\partial^2\phi}{\partial t^2}(\bar{x}(t),t)\,dt^2$$

$$+\frac{1}{24}\frac{\partial^3\phi(\bar{x}(t),t)}{\partial x_i\,\partial x_j\,\partial x_k}d^2x_i\,d^2x_j\,d^2x_k$$

$$+\frac{1}{12}\frac{\partial^4\phi(\bar{x}(t),t)}{\partial x_i\,\partial x_j\,\partial x_k\,\partial x_l}\left(dx_i\,dx_j\,dx_k\,dx_l-\frac{1}{16}d^2x_i\,d^2x_j\,d^2x_k\,d^2x_l\right)$$

$$+\frac{1}{3}\frac{\partial^4\phi(\bar{x}(t),t)}{\partial x_i\,\partial x_j\,\partial x_k\,\partial t}dx_i\,dx_j\,dx_k\,dt$$

$$+\frac{1}{1920}\frac{\partial^5\phi(\bar{x}(t),t)}{\partial x_i\,\partial x_j\,\partial x_k\,\partial x_l\,\partial x_m}d^2x_i\,d^2x_j d^2x_k\,d^2x_l\,d^2x_m$$

$$+O(d^2x)^6+O(dx)^6+O(dx)^4O(dt)+O(dx)^2O(dt)^2+O(dt)^3 \quad (3.2c)$$

Hence it follows from (1.1) that

$$E_{\bar{x}}(\bar{\phi}(t))=\phi+\frac{1}{4}\frac{\partial^2\phi}{\partial x_i\,\partial x_j}h_{ik}h_{jk}dt+o(dt) \quad (3.3a)$$

$$E_{\bar{x}}(d\phi)=\frac{\partial\phi}{\partial x_i}u_i\,dt+\frac{\partial\phi}{\partial t}dt+o(dt) \quad (3.3b)$$

$$E_{\bar{x}}(d^2\phi)=\frac{\partial\phi}{\partial x_i}(f_i+g_{ij}u_j)\,dt^2$$

$$+\frac{\partial^2\phi}{\partial x_i\,\partial x_k}\pi_{ik}\,dt^2+2\frac{\partial^2\phi}{\partial t\,\partial x_i}u_i\,dt^2+\frac{\partial^2\phi}{\partial t^2}dt^2+o(dt)^2 \quad (3.3c)$$

$$E_{\bar{x}}(d\phi\,d\psi)=\frac{\partial\phi}{\partial x_i}\frac{\partial\psi}{\partial x_j}(\tfrac{1}{2}h_{ik}h_{jk}\,dt+\pi_{ij}\,dt^2)+\frac{\partial\phi}{\partial t}\frac{\partial\psi}{\partial t}dt^2+o(dt)^2 \quad (3.3d)$$

$$E_{\bar{x}}(d^2\phi\, d^2\psi) = 2\frac{\partial\phi}{\partial x_i}\frac{\partial\psi}{\partial x_j}h_{ir}h_{jr}\,dt + o(dt)^2 \tag{3.3e}$$

$$E_{\bar{x}}(d^2\phi\, d\psi) = \frac{1}{2}\frac{\partial\phi}{\partial x_i}\frac{\partial\psi}{\partial x_j}g_{ik}h_{kr}h_{jr}\,dt^2 + \frac{\partial\psi}{\partial x_j}u_k h_{ir}h_{jr}\,dt^2$$

$$+\frac{\partial^2\phi}{\partial t\,\partial x_i}\frac{\partial\psi}{\partial x_j}h_{ir}h_{jr}\,dt^2 + o(dt)^2. \tag{3.3f}$$

The right sides of the above are evaluated at $(\bar{x}(t), t) = (x, t)$. Note that these are not in the form of (1.1) in that the mean differences of $\phi(t)$ are conditioned by $\bar{x}(t)$ rather than $\bar{\phi}(t)$.

4. CONSERVATION LAWS

Suppose $x(t)$ is a Markov diffusion satisfying the first order stochastic differential equation.

$$d^+x = f\,dt + h\,d^+w. \tag{4.1}$$

The probability density $\rho(x,t)$ of $x(t)$ satisfies the Fokker–Planck equation,

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x_i}(\rho f_i) - \frac{1}{2}\frac{\partial}{\partial x_i \partial x_j}(\rho h_{ik}h_{jk}) = 0. \tag{4.2}$$

This is a second order parabolic partial differential equation.

Suppose $x(t)$ is a strongly reciprocal diffusion satisfying the second order stochastic differential equation.

$$d^2x = f\,dt^2 + g\,dx\,dt + h\,d^2w \tag{4.3}$$

then we shall demonstrate that the density $\rho(x,t)$, mean velocity $u(x,t)$ and momentum flux tensor $\rho\pi(x,t)$ satisfy, at least in the weak sense, a system of conservation laws, very similar to those of continuum mechanics,

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x_k}(\rho u_k) = 0 \tag{4.4a}$$

$$\frac{\partial}{\partial t}(\rho u_i)=\rho(f_i+g_{ik}u_k)-\frac{\partial}{\partial x_k}(\rho\pi_{ik}) \tag{4.4b}$$

$$\frac{\partial}{\partial t}(\rho h_{ir}h_{jr})=\frac{\rho}{2}(g_{ik}h_{kl}h_{jl}+g_{jk}h_{kl}h_{il})-\frac{\partial}{\partial x_k}(\rho u_k h_{ir}h_{jr}). \tag{4.4c}$$

In addition we shall show in Section 5 that at least for the reciprocal Gaussian processes of Section 2 an additional conservation condition must hold

$$\frac{\partial}{\partial_t}(\rho\pi_{ij})=\rho(f_iu_j+u_jf_i+g_{ik}\pi_{kj}+\pi_{ik}g_{jk})$$

$$-\frac{\partial}{\partial x_k}(\rho(u_iu_ju_k-\sigma_{ij}u_k-\sigma_{ik}u_j-\sigma_{jk}u_i)). \tag{4.4d}$$

Since $\sigma_{ij}=u_iu_j-\pi_{ij}$, (4.4a, b, d) appear to be a complete set of equations for the unknowns ρ, u, π in terms of f and g in the Gaussian case. But, we doubt that (4.4d) holds for all reciprocal diffusions.

Before we derive these equations, let's take a look at their meaning. Equation (4.4a) express the conservation of probability under the mean flow described by u. This corresponds to conservation of mass in continuum mechanics. A similar equation relates the density and current velocity of Nelson [18].

Equation [4.4b) expresses the balance of momentum ρu. If we integrate the left side over a volume in x-space we obtain the time rate of change of momentum within the volume. The integral of the right side of (4.4b) has contributions from two sources. The first integral involving $\rho(f_i+g_{ik}u_k)$ is the change of momentum due to the mean acceleration of the particles inside the volume, the random acceleration produces no net change of momentum. The integral of the second term over the volume can be converted to a surface integral over its boundary by Stokes' Theorem. The integral over the surface bounding the volume of $\rho\pi_{ik}$ contracted with outward unit normal is the total flux through the boundary. Recall the definition (1.1b) of π^{ij} as the dt^2 part of $E_{\bar{x}}(dx_idx_j)$. This tensor has a deterministic and a random component, $\rho\pi_{ij}=\rho u_iu_j-\rho\sigma_{ij}$ and each contributes to the momentum flux. Notice that the order dt part of $E_{\bar{x}}(dx_i\,dx_j)$ in (1.1b) does not contribute anything to the momentum flux. Intuitively this is because these changes are so fast and so random, they cannot transport momentum.

In continuum mechanics the contraction $\frac{1}{2}\rho\pi_{ii}$ describes the density of kinetic energy. This has two parts, the first $\frac{1}{2}\rho u_i u_i$ due to deterministic part of the velocity and the other $-(\rho/2)\sigma_{ii}$ due to the random part. The latter is frequently called the internal energy density.

The tensor $\frac{1}{2}\rho\pi_{ij}$ describes the density of kinetic energy in every component of the x process. If λ_i is a constant n vector then the scalar valued process $z(t)=\lambda_i x_i(x)$ has kinetic energy density given by $\frac{1}{2}\rho\pi_{ij}\lambda_i\lambda_j$. For this reason we call $\frac{1}{2}\rho\pi_{ij}$ the tensor kinetic energy.

There is an alternative definition of π_{ij} as $\frac{1}{2}E_{\bar{x}}(d^+x_i\,d^-x_j+d^+x_j d^-x_i)=\pi_{ij}d^2+o(dt)^2$ which reduces to the standard one for smooth process. Based on this and Eq. (4.4c) which we discuss in a moment, we define the kinetic energy density to be $(\rho/2)\pi_{ii}$. This definition of kinetic energy is similar in spirit but different from that of Guerra-Morato [19]. It is interesting to note that some of the examples of Section 2 have negative or zero kinetic energy. For the Ornstein–Uhlenbeck, Cosh and Sinh process, $\pi=-1$, and for the Slepian process $\pi=0$. The Brownian bridge has both negative and positive kinetic energy depending on x and t.

Equation (4.4d), which may hold only for Gaussian processes, is a tensor form of the balance of kinetic energy and work. In other words Eq. (4.4d) expresses the balance of kinetic energy and work for every scalar process $z(t)=\lambda_i x_i(t)$. The momentum flux or *tensor kinetic energy* is $(\rho/2)\pi_{ij}$, the dt^2 part of $(\rho/2)E_{\bar{x}}(dx_i\,dx_j)$. The tensor part of the rate of work done or power is the dt^2 part of $(\rho/2)E_{\bar{x}}(d^2x_i\,dx_j+d^2x_i\,dx_j)$ which explains the first term on the right side of (4.4c). The second, flux term represents the flow of tensor kinetic energy across the boundary of the volume under consideration. This flux has contributions both from the deterministic and random parts of the motion. The first term $u_i u_j u_k$ represents the flux due to strictly deterministic motion, the others due to a mix of deterministic and stochastic motion. In continuum mechanics, $(\rho/2)\sigma_{ii}u_k$ is the flux of internal energy and $\rho\sigma_{ik}u_i$ is the flux of energy due to viscosity or stress. In our stochastic model, $(\rho/2)\sigma_{ii}u_k$ is the flux of random kinetic energy transported by the mean velocity and

$\rho\sigma_{ik}u_i$ is the flux of energy due to the random motion of particles between regions of differing mean velocity.

But (4.4d) only expresses a balance between tensor kinetic energy and tensor work terms of size dt^2. The quantities involved also have terms of size dt and the balance of these is expressed by (4.4d). We view $\rho h_{ir}h_{jr}/4$ as the tensor form of the *hyperkinetic energy* due to *hypervelocity* part of dx, namely, $\widetilde{dx} = 0(dt)^{1/2}$. The tensor $(\rho/2)(g_{ik}h_{kr}h_{jr} + g_{jk}h_{kr}h_{ir})$ is the *hyperpower* and $(\rho/2)h_{ir}h_{jr}u_k$ is the flux tensor of hyperkinetic energy. Of course this extra equation leads to an overdetermined system of equations for ρ, u, and π but if we consider $h_{ir}h_{jr}$ as an unknown also, this problem disappears. An interesting question which we don't address is that of boundary and/or initial conditions for (4.4).

In particular, (4.4c) implies that we cannot find processes satisfying the second order stochastic differential equation (1.2) for arbitrary choices of f, g and h. Notice that if h is constant in x and t, (4.4c) and (4.4a) imply that the tensor field $g(x,t)$ is skew-symmetric relative to the symmetric tensor field $h^{*2}(x,t)$, i.e.,

$$ghh^* + hh^*g^* = 0. \tag{4.5}$$

We shall derive Eqs. (4.4a, b, c) using no *a priori* assumptions of conserved quantities. Rather they shall follow from basic mathematical facts. We warn the reader that our derivation is somewhat formal, we shall interchange limiting operations, neglect small quantities, etc. In the next section we shall verify that the reciprocal Gaussian diffusions treated in Section 2 satisfy (4.4a, b, c) and also (4.4d).

Before we start we list some basic formlulas about centered differences that will be useful. Let $x(t)$, $y(t)$ be n-dimensional processes defined on $[0, T]$. Suppose $0 < \tau_0 < \tau_1 < T$ and $t_r = \tau_0 + (r - \frac{1}{2})\,dt$, $\tau_1 = t_N + \frac{1}{2}\,dt$. Then

$$\sum_{r=1}^{N} dx(t_r; dt) = \bar{x}(\tau_1; dt/2) - \bar{x}(\tau_0; dt/2) \tag{4.6a}$$

$$\sum_{r=1}^{N} d^2x(t_r; dt) = 2(dx(\tau_1; dt/2) - dx(\tau_0; dt/2)) \tag{4.6b}$$

$$d(xy^*)(t;dt)=\bar{x}(t;dt)\,dy^*(t;dt)+dx(t;dt)\,\bar{y}^*(t;dt) \tag{4.6c}$$

$$d\bar{x}(t;dt)=\overline{dx}(t;dt)=\tfrac{1}{2}\,dx(t;2\,dt) \tag{4.6d}$$

$$d(dx)(t,dt)=\tfrac{1}{4}\,d^2x(t;2\,dt) \tag{4.6e}$$

$$\begin{aligned} d(dx\,dy^*)(t;dt)=\tfrac{1}{8}(dx(t;2\,dt)\,d^2y^*(t;2\,dt)\\ +d^2x(t;2\,dt)\,dy^*(t;2\,dt)).\end{aligned} \tag{4.6f}$$

Let $x(t)$ be a second order diffusion satisfying (1.1a–f) with density $\rho(x,t)$, mean velocity $u(x,t)$, momentum flux $\rho(x,t)\ \pi(x,t)$ and stress $\rho(x,t)\ \sigma(x,t)$. We assume that as $|x|\to\infty$, ρ goes to zero faster than every rational function of $|x|$ uniformly for all $t\in[0,T]$. We also assume that $|u|$, $|\pi|$ and $|\sigma|$ are bounded above by some polynomial in $|x|$ for all $t\in[0,T]$. Let $\phi(x,t)$ be a smooth scalar valued function also bounded in norm by a polynomial in $|x|$ and suppose $\phi(t,x)$ vanishes off some closed subinterval of (τ_0,τ_1). Finally we assume that density $\bar{\rho}(x,t,dt)$ of $\bar{x}(t;dt)$ converges to the density $\rho(x,t)$ of $x(t)$ as $dt\to 0$.

Using (4.6a) we have

$$0=E\sum_{r=1}^{N} d\phi(t_r;dt)=E\sum_{1}^{N} E_{\bar{x}}(d\phi(t_r;dt))$$

We employ (3.3b) and let $dt\to 0$ to obtain

$$0=\int\!\!\int\left(\frac{\partial\phi}{\partial x_k}u_k+\frac{\partial\phi}{\partial t}\right)\rho\,dt\,dx$$

Integration by parts yields a weak form of (4.4b),

$$0=\int\!\!\int\phi\left(\frac{\partial\rho}{\partial t}+\frac{\partial}{\partial x_k}\rho u_k\right)dt\,dx.$$

In similar fashion (4.6b) yields

$$0=\frac{1}{dt}E\sum_{r=1}^{N} d^2\phi(t_r;dt)=\frac{1}{dt}E\sum E_{\bar{x}}(d^2\phi(t_r;dt))$$

We employ (3.3c) and let $dt \to 0$ to obtain

$$0 = \int\int\left(\frac{\partial\phi}{\partial x_i}(f_i + g_{ij}u_j) + \frac{\partial^2\phi}{\partial x_i x_k}\pi_{ik} + 2\frac{\partial^2\phi}{\partial t\,\partial x_i}u_i + \frac{\partial^2\phi}{\partial t^2}\right)\rho\,dt\,dx.$$

Integrating by parts yields

$$0 = \int\int\frac{\partial\phi}{\partial x_i}\left(\rho(f_i + g_{ij}u_j) - \frac{\partial}{\partial x_k}(\rho\pi_{ik}) - \frac{\partial}{\partial t}(\rho u)\right) - \frac{\partial\phi}{\partial t}\left(\frac{\partial}{\partial x_i}\rho u_u + \frac{\partial\rho}{\partial t}\right)dt\,dx.$$

By (4.4a) this reduces to a weak form of (4.4b),

$$0 = \int\int\frac{\partial\phi}{\partial x_i}\left(\rho(f_i + g_{ij}u_j) - \frac{\partial}{\partial x_k}(\rho\pi_{ik}) - \frac{\partial}{\partial t}(\rho u)\right)dt\,dx.$$

Finally we start with (4.6f) applied to ϕ, which we sum and divide by dt to obtain

$$0 = \frac{1}{8\,dt}E\sum_{r=1}^{N} d^2\phi(\tau_r; 2\,dt)\,d\phi(\tau_r; 2\,dt)$$

$$= \frac{1}{8\,dt}E\sum_{1}^{N} E_x(d^2\phi(\tau_r; 2\,dt)\,d\phi(\tau_r; 2\,dt)).$$

So by (3.3f)

$$0 = E\sum_{1}^{N}\frac{1}{2}\frac{\partial\phi}{\partial x_i}\frac{\partial\phi}{\partial x_j}g_{ik}h_{kl}h_{jl}dt + \frac{\partial^2\phi}{\partial x_i\,\partial x_k}\frac{\partial\phi}{\partial x_j}u_k h_{ir}h_{jr}\,dt$$

$$+ \frac{\partial^2\phi}{\partial t\,\partial x_i}\frac{\partial\phi}{\partial x_j}h_{ir}h_{jr}\,dt + o(dt). \tag{4.7}$$

As $dt \to 0$ we obtain

$$0 = \int\int\left(\frac{1}{2}\frac{\partial\phi}{\partial x_i}\frac{\partial\phi}{\partial x_j}g_{ik}h_{kr}h_{jr} + \frac{\partial^2\phi}{\partial x_i\,\partial x_k}\frac{\partial\phi}{\partial x_j}u_k h_{ir}h_{jr} + \frac{\partial^2\phi}{\partial t\,\partial x_i}\frac{\partial\phi}{\partial x_j}h_{ir}h_{jr}\right)\rho\,dt\,dx.$$

We symmetrize this with respect to i and j,

$$0=\frac{1}{2}\int\int\bigg(\frac{1}{2}\frac{\partial\phi}{\partial x_i}\frac{\partial\phi}{\partial x_j}(g_{ik}h_{kr}h_{jr}+g_{jk}h_{kr}h_{ir})$$

$$+\frac{\partial}{\partial x_k}\left(\frac{\partial\phi}{\partial x_i}\frac{\partial\phi}{\partial x_j}\right)u_k h_{ir}h_{jr}+\frac{\partial}{\partial t}\left(\frac{\partial\phi}{\partial x_i}\frac{\partial\phi}{\partial x_j}\right)h_{ir}h_{jr}\bigg)\rho\,dt\,dx$$

and integrate by parts

$$0=\frac{1}{2}\int\int\frac{\partial\phi}{\partial x_i}\frac{\partial\phi}{\partial x_j}\bigg(\frac{\rho}{2}(g_{ik}h_{kr}h_{jr}+g_{jk}h_{kr}h_{ir})$$

$$-\frac{\partial}{\partial x_k}(\rho u_k h_{ir}h_{jr})-\frac{\partial}{\partial t}(\rho h_{ir}h_{jr})\bigg)dt\,dx$$

which we recognize as a weak form of (4.4c).

5. RECIPROCAL AND GAUSSIAN PROCESSES REVISITED

In this section we verify that the Gaussian (reciprocal Gaussian) processes discussed in Section 2 satisfy the three (four) conservation laws of Section 4. Let $x(t)$ be a Gaussian process with smooth covariance $R(t,s)$ satisfying (2.1), (2.4) and (2.6) where $F(t)$, $G(t)$ and $H(t)$ are defined by (2.7) and $u(x,t)$, $\pi(x,t)$ and $\sigma(x,t)$ by (2.8). Since $R(t,t)=I$ (2.1a),

$$\rho(x,t)=(2\pi)^{-n/2}\exp-\tfrac{1}{2}|x|^2$$

which satisfies

$$\frac{\partial\rho}{\partial t}=0 \tag{5.1a}$$

$$\frac{\partial\rho}{\partial x_k}=-\rho x_k \tag{5.1b}$$

consider the conservation of probability (4.4a)

$$\frac{\partial \rho}{\partial t}+\frac{\partial}{\partial x_k}(\rho U_{kj}x_j)=0. \tag{5.2}$$

By (5.1) and (2.8) this reduces to

$$\rho(x_k U_{kj} x_j + U_{kk})=0$$

which holds since U is skew symmetric.

Next we return to the balance of momentum

$$\frac{\partial}{\partial t}\rho U_{ij}x_j=\rho(F_i+G_{ir}U_{rj})x_j-\frac{\partial}{\partial x_k}(\rho\pi_{ik}). \tag{5.3}$$

The left hand side is

$$\rho\frac{d}{dt}(U_{ij})x_j=\frac{\rho}{2}\left(\frac{\partial^2 R_{ij}}{\partial t^2}+\frac{\partial^2 R_{ij}}{\partial t\,\partial s}+\frac{\partial^2 R_{ji}}{\partial t\,\partial s}+\frac{\partial^2 R_{ji}}{\partial s^2}\right)x_j.$$

All evaluations of R and its partials are always at $t=s$. It is straightforward to verify that

$$\frac{d}{dt}U_{ij}=[F+GU]_{ij}-\sigma_{ij}+[UU^*]_{ij} \tag{5.4}$$

where $[\cdot]_{ij}$ denotes the *i-j* entry of the enclosed matrix. Hence the left side of (5.3) equals

$$\rho[F+GU-\sigma+UU^*]_{ij}x_j.$$

The right side equals

$$\rho[F+GU]_{ij}x_j+\rho x_k([Uxx^*U^*-\sigma]_{ik})-\rho\frac{\partial}{\partial x_k}[Uxx^*U^*]_{ik}$$

$$=\rho[F+GU]_{ij}x_j-\rho G_{ik}x_k$$

$$+\rho[Ux]_i(x^*U^*x)-\rho[Ux]_iU_{kk}-\rho[UU]_{ij}x_j$$

which equals the left side by the skew symmetry of U.

Next we verify (4.4d) for reciprocal Gaussian processes. In this case (4.4d) becomes

$$\begin{aligned}\frac{\partial}{\partial t}\rho\pi_{ij}&=\rho[Fxx^*U^*+Uxx^*F^*+G\pi+\pi G^*]_{ij}\\&\quad-\frac{\partial}{\partial x_k}\rho(U_{ir}U_{jm}U_{kl}x_rx_mx_l-\sigma_{ij}U_{kr}x_r\\&\quad-\sigma_{ik}U_{jr}x_r-\sigma_{jk}U_{ir}x_r).\end{aligned}\tag{5.5}$$

It is convenient to break up each side of this equation into terms that are time varying multiples of ρ and terms that are time varying multiples of ρxx^*, there are no others. We refer to these as constant and quadratic terms.

On the left side, the constant terms are

$$\frac{\rho}{2}\left[\frac{\partial^3R}{\partial s\,\partial t^2}+\frac{\partial^3R}{\partial t\,\partial s^2}+\frac{\partial^3R^*}{\partial s\,\partial t^2}+\frac{\partial^3R^*}{\partial t\,\partial s^2}\right]_{ij}-\rho\left[\frac{d}{dt}UU^*\right]_{ij}.$$

By (2.4) (2.6) this equals

$$\begin{aligned}&\frac{\rho}{2}\left[G\left(\frac{\partial^2R}{\partial t\,\partial s}+\frac{\partial^2R^*}{\partial t\,\partial s}\right)+\left(\frac{\partial^2R}{\partial t\,\partial s}+\frac{\partial^2R^*}{\partial t\,\partial s}\right)G^*\right.\\&\qquad\qquad\left.+F\left(\frac{\partial R^*}{\partial t}+\frac{\partial R}{\partial s}\right)+\left(\frac{\partial R}{\partial t}+\frac{\partial R^*}{\partial s}\right)F^*\right]_{ij}\\&-\rho\left[\frac{d}{dt}UU^*\right]_{ij}\end{aligned}$$

By (2.8b, d) and (5.4) this equals

$$\begin{aligned}&\rho[G[UU^*-\sigma)+(UU^*-\sigma)G^*+FU^*+UF^*]_{ij}\\&\quad-\rho[(F+GU+UU^*-\sigma)U^*+U(F+GU+UU^*-\sigma)^*]_{ij}\\&=\rho\,[(U-G)\sigma+\sigma(U-G)^*]_{ij}\end{aligned}$$

which equals the constant terms on the right side of (5.5).

The quadratic terms on the left side of (5.5) are

$$\rho\frac{d}{dt}(Uxx^*U^*)_{ij}=\rho[(F+GU-\sigma+UU^*)xx^*U^*$$
$$+Uxx^*(F+GU-\sigma+UU^*)]_{ij}$$

On the right side (5.5) the quadratic terms are

$$\rho[(F+GU)xx^*U^*+Uxx^*(F+GU)^*]_{ij}$$
$$+\rho x_kU_{kk}x_r[Uxx^*U^*]_{ij}-\rho x_k(\sigma_{ik}U_{jr}x_r+\sigma_{jk}U_{ir}x_r)$$
$$-\rho[UUxx^*U^*+Uxx^*U^*U^*]_{ij}$$
$$-\rho U_{kk}[Uxx^*U^*]_{ij}$$
$$=\rho[(F+GU-\sigma+UU^*)xx^*U^*$$
$$+Uxx^*(F+GU-\sigma+UU^*)^*]_{ij}$$

as desired to prove (5.5).

Finally we verify (4.4c) for Gaussian processes which reduces to

$$\frac{\partial}{\partial t}(\rho HH^*)=\frac{\rho}{2}(GHH^*+HH^*G^*)-\frac{\partial}{\partial x_k}(\rho U_{kr}x_r)HH^*. \tag{5.6}$$

By (5.2) this becomes

$$\frac{d}{dt}HH^*=\tfrac{1}{2}(GHH^*+HH^*G^*).$$

By (2.7c) the left side equals

$$-\left(\frac{\partial^2R}{\partial t^2}+\frac{\partial^2R}{\partial t\,\partial s}-\frac{\partial^2R^*}{\partial t\,\partial s}-\frac{\partial^2R^*}{\partial s^2}\right).$$

Since $R(t,t)=I, d^2/dt^2R=0$ and so

$$\frac{1}{2}\frac{\partial^2 R}{\partial t^2}+\frac{\partial^2 R}{\partial t\,\partial s}=-\frac{1}{2}\frac{\partial^2 R}{\partial s^2}.$$

This reduces the left side of the above to

$$-\frac{1}{2}\left(\frac{\partial^2 R}{\partial t^2}-\frac{\partial^2 R}{\partial s^2}-\frac{\partial^2 R^*}{\partial s^2}+\frac{\partial^2 R^*}{\partial t^2}\right)$$

which equals the right by (2.7, a, c).

6. CONCLUSION

We have described a theory of stochastic differential equations of second order and have demonstrated that the theory is not vacuous, it includes the reciprocal Gaussian processes which satisfy some mild assumptions. We have also demonstrated that the density, mean velocity and momentum flux obey a system of nonlinear conservation laws similar to those of continuum mechanics.

Obviously considerable work remains to be done including the following.

1) A theory of stochastic integration for second order stochastic differential equations.
2) Further study of nonlinear second order stochastic differential equations and non-Gaussian reciprocal processes.
3) Possible applications in statistical mechanics, continuum and fluid mechanics and quantum stochastic mechanics.

Acknowledgements

The work has greatly benefited from the comments, criticisms and enthusiasm of my colleagues, Joel Keizer, Blake Temple, Craig Tracy, Roger Wets and particularly Marc Mangel. I greatly appreciate their help.

References

[1] S. Bernstein, Sur les laisons entre les grandeurs aleatoires, *Proc. of Int. Cong. of Math., Zurich* (1932), 288–309.

[2] E. Schrödinger, Uber die Umkehrung der Naturgesetze, Sitz. Ber. der Preuss. Akad. Wissen., *Berlin Phys. Math. 144* (1931).

[3] E. Schrödinger, Theorie relativiste de l'electron et l'interpretation de la mechanique quantique, *Ann. Inst. H. Poincare* **2** (1932), 269–310.

[4] D. Slepian, First passage time for a particular Gaussian process, *Ann. Math. Statist.* **32** (1961), 610–612.

[5] B. Jamison, Reciprocal processes, *Z. Wahrsch. Gebiete* **30** (1974), 65–86.

[6] B. Jamison, The Markov processes of Schrodinger, *Z. Wahrsch. Gebiete* **32** (1975), 323–331.

[7] D. Jamison, Reciprocal processes: The stationary Gaussian case, *Ann. Math. Stat.* **41** (1970), 1624–1630.

[8] Chay, S. C., On quasi-Markov random fields, *J. Multivar. Anal.* **2** (1972), 14–76.

[9] J. Abrahams and J. B. Thomas, Some comments on conditionally Markov and reciprocal Gaussian processes, *IEEE Trans. Information Theory*, **IT-27**, 523–525, 1981.

[10] R. J. Adler, *The Geometry of Random Fields*, New York, Wiley, 1981.

[11] J. P. Carmichael, J. C. Masse and R. Theodorescu, Processus Gaussiens stationnaires reciproques sur un intervalle, *C.R. Acad. Sci. Paris, Ser, I*, **295**, 291-293, 1982.

[12] J. P. Carmichael, J. C. Masse and R. Theodorescu, Multivariate reciprocal stationary Gaussian processes, Preprint, Laval Univ., Dept. Math., Quebec, 1984.

[13] J. P. Carmichael, J. C. Masse and R. Theodorescu, Representations for multivariate reciprocal Gaussian processes, Preprint, Laval Univ., Dept. Math., Quebec, 1986.

[14] A. J. Krener, Reciprocal processes and the stochastic realization problem for acausal systems, in: *Modelling, Identification and Robust Control*, C.I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, 197–211.

[15] A. J. Krener, Realizations of reciprocal processes, *Proceedings IIASA Conf. on Modeling and Adaptive Control, Sopron, 1986*.,

[16] *R. N. Miroshin, Second-order Markov and reciprocal stationary Gaussian processes. Theor. Prob. Appl.* **24** (1979), 845–852.

[17] J. Abrahams, On Miroshin's second-order reciprocal processes. *SIAM J. Appl. Math.* **44** (1984), 190–192.

[18] E. Nelson, *Quantum Fluctuations*, Princeton Univ. Press, Princeton, NJ, 1985.

[19] F. Guerra and L. M. Morato, Quantization of dynamical systems and stochastic control theory, *Phys. Rev. D.* **27** (1983), 1774–1786.

[20] J. C. Zambrini, Stochastic mechanics according to E. Schrödinger, *Phys. Rev.* **A33** (1986), 1532-1548.

[21] J. C. Zambrini, Variational processes and stochastic versions of mechanics, *J. Math. Phys.* **27** (1986), 2307-2330.

[22] J. C. Zambrini, Euclidean quantum mechanics, *Phys. Rev.* **A35** (1987), 3631-3649.

APPROXIMATE NORMAL FORMS OF NONLINEAR SYSTEMS*

Arthur J. Krener, Sinan Karahan, Mont Hubbard

Institute of Theoretical Dynamics
University of California
Davis, CA 95616

Abstract A first degree approximation by a linear system is the standard approach for treating most nonlinear systems. Exact transformation of certain nonlinear systems into linear systems is possible under nonlinear state feedback and coordinate change, as shown by Jakubzcyk and Respondek [13] and Hunt and Su [10]. The approximation of nonlinear systems to higher degrees by linear systems has been treated in [16] and recently in [17]. In this paper, we develop a method of solution to find such higher degree approximations by reducing the linearization problem into the solution of a set of linear equations. We suggest a solution that, in some sense, minimizes the error in the approximation.

1. Introduction

In the analysis of scientific and engineering systems, one often encounters situations which do not lend themselves to exact solutions by conventional methods. The assumption of linearity in most control system models, for example, is an oversimplification at best, and it reflects the difficulties one would rather avoid in dealing with an otherwise nonlinear model. Seldom a technique can be found to solve a given nonlinear problem exactly. Since the control system designer is equipped with powerful methods and tools for attacking linear control systems, the motivation for "linearizing" a nonlinear problem is clearly very strong.

Therefore, whenever possible, a nonlinear control problem must be suitably transformed to bring it into an appropriate form that enables the implementation of linear control design techniques. However, the systematics of such modifications by transformation are usuallly not self-evident. The simplest of these modifications is a first degree linear approximation by calculating a series expansion at a nominal operating point. The validity of this approximation depends on the relative size of the second degree terms. In systems where nonlinearities are strong, the higher degree terms cannot be neglected, and the approximation fails.

The earliest example on the question of whether a nonlinear system can be equivalent to a linear system under some group of transformations such as change of coordinates was solved by Poincaré [19]. Various researchers in [6,8,9,15,20,21] discussed the question of when a nonlinear control system can be transformed into a linear system by a change of state coordinates. Jakubczyk and Respondek [13], and Hunt and Su [10,11,12] independently considered the full state feedback and coordinate change problem. Related work also appeared in [2,22,23,24]. In [16], Krener investigated an approximate linearization considering the second and higher degree terms in the truncated series expansion of a the vector field, and proved a weakened version of the Hunt–Su linearization condition. In [17], further results in an attempt to solve for the resulting transformations were presented. An application for nonlinear observers also appeared in [4].

The objectives of this paper are to: 1) Present a solution to the approximate linearization problem, 2) Suggest a method to solve the Homological equations to minimize the error in the approximation in some sense. For further work see [14].

2. Higher Degree Approximations to Autonomous Systems: Normal Form Theorem

In this section, the normal form theorem of Poincaré will be introduced. The approximate linearization problem for a control system will be formulated later in a similar spirit. As a reminder of the connection between the two, we continue to use the term "Homological Equations" (after Arnold [1]).

Let us consider an autonomous system:

$$\dot{x} = f(x) \tag{1a}$$

$$x(0) = x^*. \tag{1b}$$

where $x \in \Re^n$ and the system is assumed to be at rest at the origin, i.e. $f(0) = 0$. Without loss of generality we will assume $x^* = 0$. First, consider the linearization of (1) at x^*:

$$\dot{x} = Fx \tag{2a}$$

$$F = \frac{\partial f}{\partial x}(0) \tag{2b}$$

We will seek a coordinate change for (1) of the form identity plus higher degree terms, such that the resulting system will agree with (1) up to an error of degree $O(x)^{\rho+1}$ where ρ is the degree of approximation. The following treatment is for $\rho = 2$. The results can be easily generalized to any arbitrary degree ρ by induction.

We assume a transformation of the form:

$$z = x - \phi^{(2)}(x) \tag{3}$$

where z denotes a new set of coordinates. $\phi^{(2)}(x)$ is a polynomial of degree 2. The function $f(x)$ in (1a) is expanded in a series:

$$\begin{aligned} f(x) &= f^{(1)}(x) + f^{(2)}(x) + O(x)^3 \\ &= Fx + f^{(2)}(x) + O(x)^3 \end{aligned} \tag{4}$$

The goal of the transformation (3) is to choose $\phi^{(2)}(x)$ such that in z coordinates the dynamics of the system is represented by

* Supported in part by grant AFOSR-85-0267

$$\dot{z} = Fz + O(z)^3 \qquad (5)$$

namely the second degree terms in the series expansion (4) vanish under the coordinate change. We take the time derivative of (3) and using (1a), (4) and (5) evaluate each side by ignoring $O(x)^3$ and higher terms:

$$F(x - \phi^{(2)}(x)) = Fx + f^{(2)}(x) - \frac{\partial \phi^{(2)}(x)}{\partial x} Fx \qquad (6)$$

Now we introduce some notation. The Lie bracket of two vector fields f, g is another vector field defined by:

$$[f,g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g \qquad (7)$$

Rearranging and cancelling terms in (6), and using (7) we obtain

$$f^{(2)}(x) = [Fx, \phi^{(2)}(x)] \qquad (8)$$

Equation (8) is called the *Homological Equation* [1]. In [5], a similar derivation is also presented. The Lie bracket operation in the above defines a mapping

$$[Fx, \cdot\,] : \phi^{(2)}(x) \rightarrow [Fx, \phi^{(2)}(x)] \qquad (9)$$

Obviously, (9) represents a linear mapping from $n^2(n+1)/2$ dimensional parameter space of the coefficients of $\phi^{(2)}(x)$ to an $n^2(n+1)/2$ dimensional parameter space. The question is whether $f^{(2)}(x)$ in the range of this mapping, i.e. can we always find a $\phi^{(2)}(x)$ that will satisfy (8)? This problem was first solved by Poincaré [19]. In the following, we present a slightly modified proof that closely follows [1,5]:

Suppose F has a full set of linearly independent eigenvectors. Then we can take the right eigenvectors of F as a set of basis vectors, and the left eigenvectors as a set of coordinates, which are defined by

$$Fv^k = \lambda_k v^k \qquad (10a)$$

$$w_i F = \lambda_i w_i \qquad (10b)$$

where $v^k \in \mathbb{C}^{n \times 1}$, $w_i \in \mathbb{C}^{1 \times n}$ and $\lambda_i, \lambda_k \in \mathbb{C}$. We define a basis for n-dimensional vector valued polynomials of degree 2 as follows:

$$\phi^k_{ij}(x) = v^k (w_i x)(w_j x) \qquad \text{for } j, k = 1, \ldots, n\,;\; i = 1, \ldots j. \qquad (11)$$

Using this basis for the polynomials in Eqn. (8), we evaluate the Lie bracket:

$$[Fx, \phi^k_{ij}(x)] = (\lambda_i + \lambda_j - \lambda_k)\phi^k_{ij}(x) \qquad (12)$$

The mapping (9) is onto if $(\lambda_i + \lambda_j - \lambda_k) \neq 0$ for all $j, k = 1, \ldots, n$; $i = 1, \ldots j$. In the literature, this is called the *resonance condition*. We note that this is only a sufficient condition. A general proof for the case when F does not have a full set of independent eigenvectors may be found in [1].

The above can easily be extended to an arbitrary degree of linearization ρ. We present the final form:

$$[Fx, \phi^k_{i_1,\ldots,i_\rho}(x)] = (\lambda_{i_1} + \cdots + \lambda_{i_\rho} - \lambda_k)\phi^k_{i_1,\ldots,i_\rho}(x) \qquad (13)$$

with $(\lambda_{i_1} + \cdots + \lambda_{i_\rho} - \lambda_k) \neq 0$ the condition of no resonance.

3. Higher Degree Approximations to Control Systems

In this section we will seek a solution to the problem of linearization for control. Full state observability is implicitly assumed. Consider a nonlinear system affine in control:

$$\dot{x} = f(x) + g(x)u \qquad (14a)$$

$$x(0) = x^*. \qquad (14b)$$

where $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. The system is assumed to be at rest at the nominal operating point (x^*; $u^* = 0$). Again, we will assume $x^* = 0$. First, consider the linearization of (14) at x^*:

$$\dot{x} = Fx + Gu \qquad (15a)$$

$$F = \frac{\partial f}{\partial x}(0), \quad G = g(0). \qquad (15b)$$

We want a coordinate change for (14) of the form identity plus higher degree terms, such that the resulting linear plant will agree with (14) up to an error of degree $O(x,u)^{\rho+1}$ (i.e. terms of $O(x)^{\rho+1}$ and $O(x,u)^{\rho}$) where ρ is the degree of approximation. When ρ = 1, the first degree approximation (15) is obtained. Similar to the previous section, the case for ρ = 2 will be derived first, and the results will be generalized to an arbitrary degree ρ by induction. First, the functions f and g are expanded in a series:

$$\begin{aligned} f(x) &= f^{(1)}(x) + f^{(2)}(x) + O(x)^3 \\ &= Fx + f^{(2)}(x) + O(x)^3 \end{aligned} \qquad (16)$$

$$\begin{aligned} g(x) &= g^{(0)}(x) + g^{(1)}(x) + O(x)^2 \\ &= G + g^{(1)}(x) + O(x)^2 \end{aligned} \qquad (17)$$

and the nonlinear system (14a) is rewritten as

$$\dot{x} = Fx + f^{(2)}(x) + \left(G + g^{(1)}(x)\right)u + O(x,u)^3 \qquad (18)$$

We assume a transformation similar to the one proposed in Sec. 2:

$$z = x - \phi^{(2)}(x) \qquad (19)$$

In addition, a new input v is chosen as

$$v = \alpha^{(2)}(x) + (\mathrm{I} + \beta^{(1)}(x))u \qquad (20)$$

where $\alpha^{(2)}(x)$ is an $n \times 1$ vector of second degree polynomials, and $\mathrm{I} + \beta^{(1)}(x)$ is an $m \times m$ identity matrix plus first degree terms with nonsingular $\beta(x)$. Now we want the system to become, in z coordinates,

$$\dot{z} = Fz + Gv + O(z,v)^3 \qquad (21)$$

We take the time derivative of (19), and introducing (18), (19), (20) and (21) we obtain:

$$f^{(2)}(x) = [Fx, \phi^{(2)}(x)] + G\alpha^{(2)}(x) \qquad (22a)$$

$$g^{(1)}(x)u = [Gu, \phi^{(2)}(x)] + G\beta^{(1)}(x)u \qquad \forall \text{ constant } u. \qquad (22b)$$

Because of its similarity to the homological equations derived in the previous section, we call (22) the *Generalized Homological Equations*. For a detailed derivation of (22), see [17]. In the same reference, the above approximation is extended to an arbitrary degree as

$$f^{(\rho)}(x) = [Fx, \phi^{(\rho)}(x)] + G\alpha^{(\rho)}(x) \qquad (23a)$$

$$g^{(\rho-1)}(x)u = [Gu, \phi^{(\rho)}(x)] + G\beta^{(\rho-1)}(x)u \qquad \forall \text{ constant } u. \qquad (23b)$$

The resulting system is accurate up to degree ρ:

$$\dot{z} = Fz + Gv + O(z,v)^{\rho+1} \qquad (24)$$

Once a higher degree linear approximation is obtained, one of the important issues is the stability of the closed loop system. Thus one may choose, for instance, a linear state feedback for the approximate model

$$\dot{z} = Fz + Gv \tag{25}$$

by setting $v = Kz$. The gain matrix K is chosen such that in the closed loop the system gives the desired performance. If we assume that the model has been linearized up to second degree, using the feedback $v = Kz$ and Eqn. (19) we evaluate (20):

$$Kx - K\phi^{(2)}(x) = \alpha^{(2)}(x) + \left(1 + \beta^{(1)}(x)\right)u \tag{26}$$

and calculate the feedback u as:

$$\begin{aligned} u &= \left(1 + \beta^{(1)}(x)\right)^{-1}\left(Kx - K\phi^{(2)}(x) - \alpha^{(2)}(x)\right) \\ &= Kx - \left\{\beta^{(1)}(x)Kx + K\phi^{(2)}(x) + \alpha^{(2)}(x)\right\} + O(x,u)^3. \end{aligned} \tag{27}$$

In the above, purpose of the feedback u becomes immediately clear. In addition to the linear feedback, there are *second degree correction terms* (inside curly brackets in (27)). While the purpose of the feedback $u = Kx$ is to achieve stability, pole placement, etc. for the first degree approximation (15a) to get

$$\dot{x} = (F + GK)x, \tag{28}$$

the feedback (27) cancels certain second degree terms to achieve a second degree approximation (accurate to second degree compared with a linear model) toward the same feedback design goals:

$$\begin{aligned} \dot{x} = (F + GK)x &+ f^{(2)}(x) + g^{(1)}(x)Kx - G\left\{\beta^{(1)}(x)Kx \right. \\ &\left. + K\phi^{(2)}(x) + \alpha^{(2)}(x)\right\} + O(x,u)^3. \end{aligned} \tag{29}$$

An important feature of the feedback (27) and the resulting closed loop system (29) is that one need not transform the state variables into the new coordinates z that was introduced for the sake of calculations. Feedback design can be performed in the natural coordinates in which the system is originally presented. If some of the states are not observable, one can estimate the unavailable state variables by means of an observer, and apply the same procedure. For further work on this problem, see [18].

4. Analysis of the Linear Mapping

In the homological equations (22) of Sec. 3, the second degree terms $f^{(2)}(x)$ and $g^{(1)}(x)u$ can be cancelled out under certain solvability conditions by proper choice of $\phi^{(2)}(x)$, $\alpha^{(2)}(x)$, and $\beta^{(1)}(x)$. When the *coefficients* of the like terms in (22) are set equal, a linear mapping is obtained as

$$\begin{Bmatrix} \phi^{(2)}(x) \\ \alpha^{(2)}(x) \\ \beta^{(1)}(x) \end{Bmatrix} \longrightarrow \begin{Bmatrix} f^{(2)}(x) \\ g^{(1)}(x) \end{Bmatrix} \tag{30}$$

A simple count yields the dimensions of the domain and the range:

$$\frac{n^2(n+1)}{2} + \frac{mn(n+1)}{2} + m^2 n \longrightarrow \frac{n^2(n+1)}{2} + n^2 m \tag{31}$$

To analyze the mapping, we make a table for the dimensions:

For $m = 1$: For $m = 2$:

State Space	Domain	Range	State Space	Domain	Range
$n = 2$	11	10	$n = 2$	20	14
$n = 3$	27	27	$n = 3$	42	36
$n = 4$	54	56	$n = 4$	76	72
.	.	.	$n = 5$	125	125
.	.	.	.	.	.

Dimensions of the domain and the range become equal whenever $n = 2m + 1$. However, this does not imply that the mapping is of full rank. For example, when $m = 1$, $n = 3$ the rank is 26, not 27. In general, for a single input system, the rank of the mapping is always one less than the dimension of the domain for $n \geq 3$.

We will restrict our analysis to the second degree *linearization problem with a single input* u, i.e. $m = 1$. We will start with the analysis of the linear mapping.

A necessary condition for finding a coordinate change-feedback pair for a nonlinear control system is the local controllability condition at the nominal point. For the system (18) with a scalar input, this implies

$$\text{rank}\,\{G \;\; FG \;\; \ldots \;\; F^{n-1}G\} = n. \tag{32}$$

On the other hand, we define a $1 \times n$ matrix K such that

$$KF^{i-1}G^{j} = \begin{cases} 0 & 1 \leq i < n \\ 1 & i = n \end{cases} \tag{33}$$

Then, the following collection of one forms is of full rank:

$$\text{rank}\,\{K \;\; KF \ldots KF^{n-1}\} = n. \tag{34}$$

(32) and (34) together imply that we can define a basis for the second and first degree monomials as follows. Define as a basis

$$v^k = F^{k-1}G \tag{35a}$$

and a co–basis

$$w_i = KF^{i-1} \tag{35b}$$

Now we define a basis for second degree monomials as

$$\phi^k_{ij}(x) = v^k(w_i x)(w_j x) \quad \text{for } j, k = 1, \ldots, n\,;\; i = 1, \ldots, j. \tag{36}$$

and a basis for first degree monomials as

$$\phi^k_i(x) = v^k(w_i x) \qquad \text{for } k = 1, \ldots, n\,;\; i = 1, \ldots, n. \tag{37}$$

Using the definitions (36) and (37) is a great convenience for calculating the Lie brackets that appear in the generalized homological equations (22). Calculation of (22a) gives

$$[Fx, \phi^k_{ij}(x)] = \begin{cases} \phi^k_{i+1\,j} + \phi^k_{i\,j+1} - \phi^{k+1}_{ij} & 1 \leq i \leq j < n;\; 1 \leq k < n \\ \phi^k_{i+1\,j} - \phi^{k+1}_{ij} & 1 \leq i < j = n;\; 1 \leq k < n \\ -\phi^{k+1}_{ij} & i = j = n\,;\; 1 \leq k < n \end{cases} \tag{38}$$

In the evaluation of (38), when $k = n$, the expressions become slightly more complicated. However, transforming the control system into a Brunovsky canonical form [3] prior to the linearization helps simplify the expressions [14].

Next, we calculate (22b)

$$[G, \phi^k_{ij}(x)] = \begin{cases} 0 & i, j < n \\ \phi^n_i & i < j = n \\ 2\phi^n_n & i = j = n \end{cases} \tag{39}$$

These two formulas are used to compute the kernel and co-kernel of the mapping

$$\begin{Bmatrix} \phi^{(2)}(x) \\ \alpha^{(2)}(x) \\ \beta^{(1)}(x) \end{Bmatrix} \longrightarrow \begin{Bmatrix} f^{(2)}(x) \\ g^{(1)}(x) \end{Bmatrix} \tag{30}$$

and we now obtain a set of linear equations expressed in matrix form:

$$L\begin{bmatrix}\phi^{(2)}\\ \alpha^{(2)}\\ \beta^{(1)}\end{bmatrix}=\begin{bmatrix}f^{(2)}\\ g^{(1)}\end{bmatrix} \qquad (40)$$

In (40), L is a constant coefficient matrix of $n^2(n+1)/2 + n^2$ rows by $n^2(n+1)/2 + n(n+1)/2 + n$ columns that is found from the above evaluation of the Lie brackets of the mapping. In $\begin{bmatrix}\phi^{(2)}\\ \alpha^{(2)}\\ \beta^{(1)}\end{bmatrix}$ and $\begin{bmatrix}f^{(2)}\\ g^{(1)}\end{bmatrix}$ the constant coefficients of their corresponding second degree terms are stacked in a consistent lexicographic ordering. For the single input linearization problem, the column rank of L is $(n^2(n+1)/2+n(n+1)/2+n-1)$.

A solution to the linearization problem is developed as follows. First, we note that since the mapping (40) is deficient in rank for $n > 2$, a control system with nonlinear terms $f^{(2)}(x)$ and $g^{(1)}(x)u$ will not, in general, have an exact solution to yield a second degree linearization. In fact, the Hunt-Su result [9] (or Krener's extension of the same to the approximate case in [16]) is a test for precisely this condition. Consequently, Eqn. (40) will not usually have an exact solution. For a system with $n=3$, $m=1$:

$$2f_3^{33} - g_3^2 + g_2^3 = 0 \qquad (41)$$

is the condition for exact linearizability up to second degree. In (41), f_k^{ij} represents the coefficient of an element of $f^{(2)}$ in the basis $\phi_{ij}^k(x)$ *for second degree monomials obtained* when the system is in Brunovsky canonical form. Similarly, g_k^i is the coefficient of the corresponding element of $g^{(1)}$ in the basis ϕ_i^k for first degree monomials. Eqn. (41) is called the *co-kernel equation.*

When an exact solution does not exist, it is reasonable to seek an approximate solution which will minimize the error in the linearization with respect to some norm. In order to give a precise meaning to this problem, first assume that we have adequate knowledge about the operating regime of the control system and the desired accuracy as determined by

$\rho(x,u)$: A probability density function; typically uniform over some compact set, or Gaussian.

Q: A sensitivity matrix, positive definite.

And define the "error"

$$\left|\left|\begin{pmatrix}f^{(2)}\\ g^{(1)}\end{pmatrix}-\begin{pmatrix}\tilde{f}^{(2)}\\ \tilde{g}^{(1)}\end{pmatrix}\right|\right|^2$$

$$\overset{\text{def}}{=}\iint\left|f^{(2)}-\tilde{f}^{(2)}+(g^{(1)}-\tilde{g}^{(1)})u\right|_Q^2\rho(x,u)\mathrm{d}x\mathrm{d}u \qquad (42)$$

We want to choose $\begin{pmatrix}\tilde{f}^{(2)}\\ \tilde{g}^{(1)}\end{pmatrix}$ such that the above error is minimized. Note that this term is in the range of the mapping, i.e. it satisfies the generalized homological equations

$$\tilde{f}^{(2)}(x) = [Fx,\phi^{(2)}(x)] + G\alpha^{(2)}(x) \qquad (43a)$$

$$\tilde{g}^{(1)}(x)u = [Gu,\phi^{(2)}(x)] + G\beta^{(1)}(x)u \qquad \forall \text{ constant } u. \qquad (43b)$$

Furthermore, we wish to choose the smallest $\begin{pmatrix}\phi^{(2)}\\ \alpha^{(2)}\\ \beta^{(1)}\end{pmatrix}$ that will achieve the above. Again, we choose positive definite matrices S, R and minimize

$$\iint\left|\phi^{(2)}\right|_S^2 + \left|\alpha^{(2)}(x) + \beta^{(1)}(x)u\right|_R^2\rho(x,u)\mathrm{d}x\mathrm{d}u \qquad (44)$$

or one can take a weighted combination of the above. In fact, S can be taken to be equal to Q of (42), but the choice is not limited to this case. We illustrate the minimization in Figs. (1) and (2). Fig. 1 represents the $\frac{n^2(n+1)}{2}+n^2$ dimensional parameter space for the range of the mapping. The coefficients of the second degree terms in the control system define a point in this space, denoted by $\begin{pmatrix}f^{(2)}\\ g^{(1)}\end{pmatrix}$. The range of L is represented by a straight line going through the origin. Those points in the range space of L that exactly satisfy (40) will lie on this line. Among these infinitely many points we want to find the one (shown as $\begin{pmatrix}\tilde{f}^{(2)}\\ \tilde{g}^{(1)}\end{pmatrix}$ on the figure) which will minimize, with respect to a norm as defined earlier, the *error between the actual system that is being approximately linearized and a model which is exactly linearizable* (up to degree 2) by the coordinate change and feedback. Fig. 2 shows the $n^2(n+1)/2 + n(n+1)/2 + n$ dimensional domain space of the mapping, and the minimization done in the domain space.

The numerical solution to (40) is found by linear algebraic methods. For illustration purposes consider a mapping

$$A : \mathbb{R}^N \longrightarrow \mathbb{R}^M \qquad (45)$$

and solve

$$Ax = b. \qquad (46)$$

If the mapping is not of full rank, it can be expanded as follows

$$\begin{pmatrix}I\\ A\end{pmatrix} : x \longrightarrow \begin{pmatrix}x\\ Ax\end{pmatrix} \in \mathbb{R}^{N+M} \qquad (47)$$

where I is the identity matrix of appropriate dimension. The mapping (47) is always of full column rank. Now we solve for

$$\begin{bmatrix}I\\ A\end{bmatrix}x = \begin{bmatrix}0\\ b\end{bmatrix} \qquad (48)$$

Then one can choose a metric G on $\mathbb{R}^{N+M}$

$$G = \begin{bmatrix}G_{11} & 0\\ 0 & G_{22}\end{bmatrix} \qquad (49)$$

and find a solution to

$$\min_{x\in\mathbb{R}^N}\left|\left|\begin{matrix}x\\ Ax-b\end{matrix}\right|\right|_G^2 . \qquad (50)$$

The well-known solution of (50) is

$$x = \left([I \;\; A^T]G\begin{bmatrix}I\\ A\end{bmatrix}\right)^{-1}[I \;\; A^T]G\begin{bmatrix}0\\ b\end{bmatrix} \qquad (51)$$

Finally, we note the following correspondence between the dimensions and variables in (51) and in the linearization problem:

$$M : \frac{n^2(n+1)}{2} + n^2 ; \qquad N : n^2(n+1)/2 + n(n+1)/2 + n$$

$$x : \begin{bmatrix} \phi^{(2)} \\ \alpha^{(2)} \\ \beta^{(1)} \end{bmatrix} ; \qquad b : \begin{bmatrix} f^{(2)} \\ g^{(1)} \end{bmatrix}$$

$$A : L ; \qquad G : \begin{bmatrix} S & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & Q \end{bmatrix}$$

5. An Example

In this section we linearize the following nonlinear plant using the method outlined in Sec. 3.:

$$\dot{x}_1 = x_2 + 0.5x_1^2 + x_1 u \tag{52a}$$

$$\dot{x}_2 = x_3 - x_1x_3 + x_2 u \tag{52b}$$

$$\dot{x}_3 = u + 0.5x_2^2 + x_2 u \tag{52c}$$

Calculation of the coordinate transformation and feedback gives:

$$z_1 = x_1 - x_1x_3 \tag{53a}$$

$$z_2 = x_2 + 0.5x_1^2 - x_2x_3 \tag{53b}$$

$$z_3 = x_3 + x_1x_2 - x_1x_3 - x_3^2 \tag{53c}$$

$$v = x_1x_3 + 1.5x_2^2 - x_2x_3 + (1 - x_1 + x_2 - 2x_3)u. \tag{54}$$

With the above, we obtain the exact linearization (implying that the system (52) satisfies the Hunt–Su condition):

$$\dot{z}_1 = z_2 \tag{54a}$$

$$\dot{z}_2 = z_3 \tag{54b}$$

$$\dot{z}_3 = v. \tag{54c}$$

A simple feedback design $v = Kz$ that places the closed loop poles at locations -1, $-0.707 \pm 0.707j$ yields the gains as $k_1 = -1$, $k_2 = -2.4142$, $k_3 = -2.4142$. Using $v = K(x - \phi^{(2)}(x))$, (54), and (27), the feedback u is evaluated. The nonlinear input is then applied to the system (52). Note that this will introduce $O(x,u)^3$ terms into (52). In Figs. 3 through 8, we present simulation results and comparisons for the above linearization and feedback problem. In all plots, continuous lines represent the time response curves for a fictitious linear system equal to the linear part of (52) with feedback $u = Kx$ applied. The higher order linearization method of this paper for the nonlinear system (52) is compared against this exact linear model (dashed lines). A feedback design based on a first order approximation (i.e. the feedback $u = Kx$), and applied to the nonlinear model (52) is plotted with dotted lines. Figs. 3, 4, and 5 show the responses to a step input $x_2 = 0.4$. In Figs. 6, 7, and 8 the response curves for a step input $x_2 = -0.4$ is shown. The simulations demonstrate the advantage of the proposed nonlinear feedback. The time response of the nonlinear system with nonlinear controller is closer to the response of a linear system (specifically, the linear system obtained from the first order part of the vector field) than a control design based on a first order approximation. The effect and the improvement of this nonlinear control on the stability bounds of a nonlinear system is being investigated.

Conclusion

In this paper, we presented a method to solve the approximate linearization problem of nonlinear control systems. The problem is reduced to the solution of a set of linear equations as follows: First, the generalized homological equations are derived. By introducing an appropriate basis for expressing higher degree monomials in the vector field, a set of equations linear in the coefficients of the monomials are found. An exact solution to these set of equations is not always possible. A least square solution is proposed that minimizes, in a statistical sense as defined above, the error in the approximation.

We note that in the equivalent linear map, the case when *the nonlinear terms* to be cancelled are not in the range of the mapping exactly correspond to the violation of the integrability conditions in the Hunt–Su linearization theorem. In other words, the given nonlinear system in this case is not exactly linearizable. In the method developed here, we still find a "partially" linearizing solution to this problem. The least square solution *minimizes* precisely the error in such an approximation.

Especially for systems with higher dimensions and higher degrees of approximation, the dimension of the system of linear equations may become extremely large and difficult to solve. A computer program that automates the solutions is under development by the authors.

The multi–input case for the generalized homological equations is slightly more complicated to derive. Research is continuing for the description and solution of these equations in the most general input–output setting, and for an arbitrary degree of linearization.

References:

[1] Arnold, V.I., *Geometric Methods in the Theory of Ordinary Differential Equations*, Springer Verlag, New York, 1983.

[2] Brockett, R.W., "Feedback invariants for nonlinear systems," *IFAC*, v. 2, pp. 115-120, Helsinki, 1978.

[3] Brunovsky, P., "A classification of linear controllable systems," *Kybernetica cislo*, 3, pp. 173-188, 1970.

[4] Frezza, R., S. Karahan, A.J. Krener, M. Hubbard, "Application of an efficient nonlinear filter," 1987 International Symposium on the Mathematical Theory of Networks and Systems, June 17-20, Phoenix, AZ, 1987.

[5] Guckenheimer, J. and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer Verlag, New York, 1983.

[6] Guillemin, V.W., and S. Stenberg, "Remarks on a paper by Herman," *Trans. Amer. Math. Soc.*, v. 130, pp. 110-116, 1968.

[7] Hermann, R., "On the accessibility problem in control theory," in *Int. Symp. on Nonlinear Differential Equations and Nonlinear Mechanics*, New York, Academic Press, pp 325-332, 1963.

[8] Hermann, R., "The formal linearization of a semisimple Lie algebra of vector fields around a singular point," *Trans. Amer. Math. Soc.*, v. 130, pp. 105-109, 1968.

[9] Hermann, R., and A.J. Krener, "Nonlinear controllability and observability," *IEEE Trans. Automat. Contr.*, vol. AC-22, no. 5, pp.728-740, 1977.

[10] Hunt, L.R., and R. Su, "Linear equivalents of nonlinear time-varying system," *International Symposium on the Mathematical Theory of Netwoks and Systems*, Santa Monica, pp. 119-123, 1981.

[11] Hunt, L.R., R. Su, and G. Meyer, "Global transformations of nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. AC 28, pp. 24-31, 1983.

[12] Hunt, L.R., R. Su, and G. Meyer, "Design for multi-input nonlinear systems," in *Differential Geometric Control Theory*, R. W. Brockett, R. S. Millman and H. J. Sussmann, eds., Birkhauser, Boston, pp. 268-298, 1983.

[13] Jakubczyk, B. and W. Respondek, "On the linearization of control systems," *Bull. Acad. Polon. Sci., Ser. Sci. Math. Astronom. Phys.*, 28, pp.517-522, 1980.

[14] Karahan, S., "Higher degree linear approximations to nonlinear systems," Ph.D. dissertation, University of California, Davis, CA, 1988.

[15] Krener, A.J., "On the equivalence of control systems and the linearization of nonlinear systems," *SIAM J. Control*, vol. 11, pp. 670-676, 1973.

[16] Krener, A.J., "Approximate linearization by state feedback and coordinate change," *Systems and Control Letters*, vol. 5, pp.181-185, 1984.

[17] Krener, A.J., S. Karahan, M. Hubbard, and R. Frezza, "Higher order linear approximations to nonlinear systems," *Proceedings of 26th IEEE Conference on Decision and Control*, Los Angeles, CA, December 9-11, 1987.

[18] Krener, A.J., and A. Phelps, "Approximate normal forms for nonlinear observers," to appear.

[19] Poincaré, H., *Oeuvres*, Tome 1, Gauthier-Villars, Paris, 1928.

[20] Respondek, W., "Linearization, feedback and Lie brackets," *Proc. of Conf. on the Geometric Theory of Nonlinear Control Systems*, Wroclaw Technical University Press, Wroclaw, 1985.

[21] Sedwick, J.L., and D.L. Elliott, "Linearization of analytic vector fields in the transitive case," *J. Diff. Eq.*, v. 26, pp.370-390, 1977.

[22] Sommer, R., "Control design for multivariable non-linear time-varying systems," *Int. J. Control*, v. 31, pp. 883-891, 1980.

[23] Zeitz, M., "Canonical forms for nonlinear systems," *Proc. of Conf. on the Geometric Theory of Nonlinear Control Systems*, Wroclaw Technical University Press, Wroclaw, 1985.

[24] Zhevnin, A.A., and A.P. Krischenko, "Controllability of nonlinear systems and synthesis of control algorithms," *Soc. Phys. Dokl.* v. 26, n. 6, pp. 559-561, 1981.

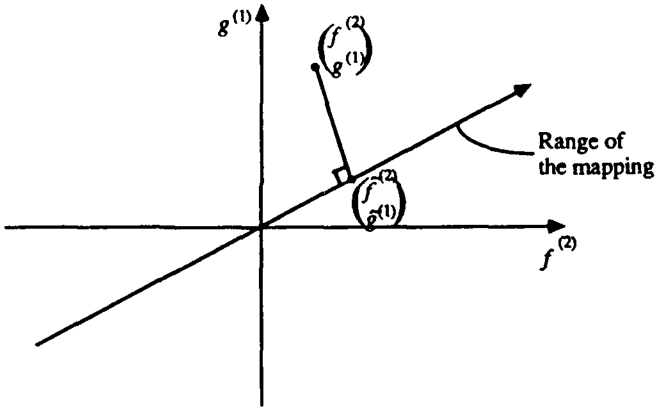

$g^{(1)}$
$\binom{f^{(2)}}{g^{(1)}}$
Range of the mapping
$\binom{\tilde{f}^{(2)}}{\tilde{g}^{(1)}}$
$f^{(2)}$

Fig.1 The range space of the mapping.

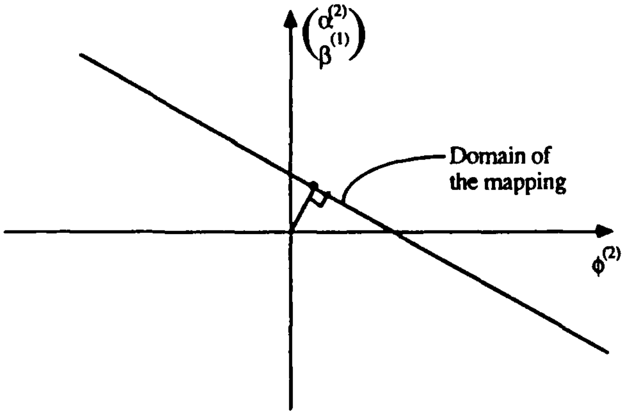


Fig. 2 Domain space of the mapping

LEGEND FOR FIGURES 3 TO 8:

(1) : ———— Fictitious linear model with linear feedbackcontrol.

(2) : — — — — Nonlinear system with nonlinear control (based on the method of this paper)

(3) : ·············· Nonlinear system with control based on first order approximation (same control as the fictitious linear model).

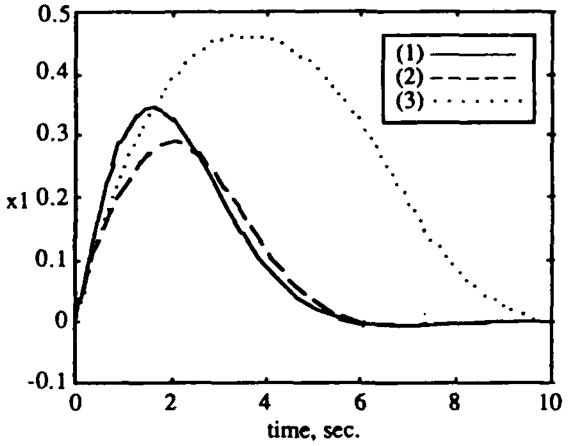


Fig. 3 Time domain response of state x_1; Simulation 1.

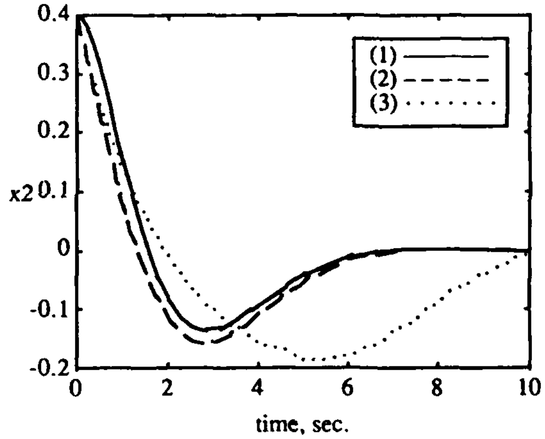


Fig. 4 Time domain response of state x_2; Simulation 1.

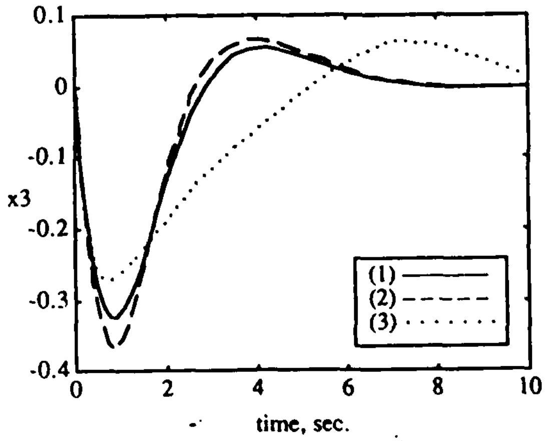


Fig. 5 Time domain response of state x_3; Simulation 1.

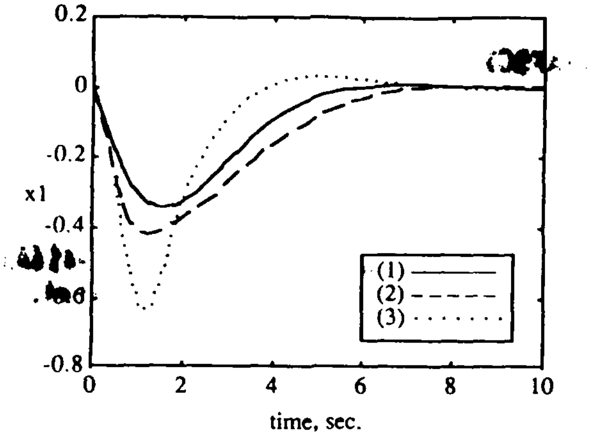


Fig. 6 Time domain response of state x_1; Simulation 2.

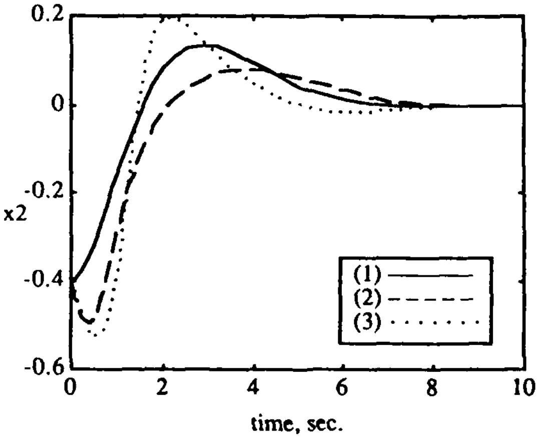


Fig. 7 Time domain response of state x_2; Simulation 2.

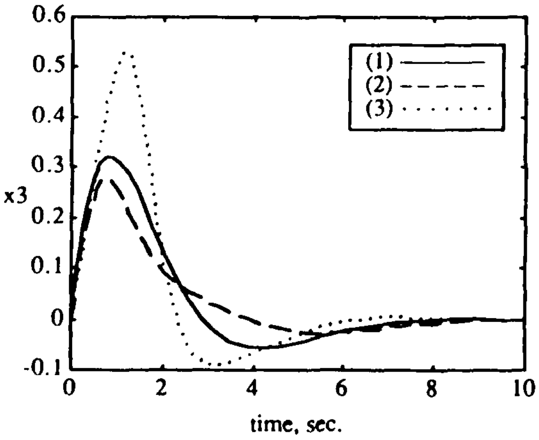


Fig. 8 Time domain response of state x_3; Simulation 2.

Nonlinear Controller Design via Approximate Normal Forms

Arthur J. Krener

Institute of Theoretical Dynamics

and

Department of Mathematics

University of California

Davis, CA 95616

Research supported in part by AFOSR–85–0267

1. Introduction. Over the past several years a group of faculty and graduate students at UC Davis have been developing a set of tools for the design of controllers and observers for nonlinear systems. Our approach has been based on normal forms and approximate normal forms for nonlinear systems. When a nonlinear system admits a normal form the design of a controller or observer is greatly simplified and standard linear design tools can be employed. The people that have been involved in this program are Mont Hubbard, Sinan Karahan, Andrew Phelps, Yi Zhu, Ruggero Frezza and myself. This work has been supported in part by AFOSR. In this paper I'll give an overview of our program.

2. Normal Forms. Following Kailath's terminology, [10], there are four normal forms for linear systems, i.e., controllable, observable, controller and observer form. The first two are relatively straightforward to obtain, provided the system is controllable or observable. However, the latter two are more useful in the design of stabilizing state feedback control laws and asymptotic state observers. If a linear system is both controllable and observable then it admits all four normal forms.

In [14] we discussed the nonlinear generalizations of the four linear normal forms. Unfortunately, even controllable and observable nonlinear systems do not admit all four nonlinear normal forms. A nonlinear system which admits controller normal form is sometimes said to be state feedback linearizable in the sense of Hunt–Su [8] and Jakubczyk–Respondek [9]. For a system in controller normal form, the design of a stabilizing state feedback control law is a straightforward task. However, most systems do not admit a controller normal form and even when one does, the transformation of a system into controller normal form involves solving a system of first order linear partial differential equations which can be quite difficult.

Similar remarks are even more appropriate for observer normal form. For a system in observer form, the design of an observer is a straightforward task. But very few systems admit such forms and the computation of observer normal form is, in general, extremely difficult.

For these reasons, we have introduced approximate versions of nonlinear controller and observer form [15, 16]. These may be thought of as finding systems nearby to the original which admit controller or observer form. The computation of such a system is relatively straightforward, and reduces to solving a set of linear equations. Unfortunately, these linear equations are not always solvable and they increase in size quite rapidly with the dimension of the system.

We start by introducing modified versions of controller and observer normal forms of the nonlinear system.

$$\dot{\xi} = f(\xi) + g(\xi)u \tag{2.1a}$$

$$y = h(\xi) \tag{2.1b}$$

$$\xi(0) \sim \xi^{o} = 0 \tag{2.1c}$$

around the nominal operating point ξ^{o}, which for convenience we assume to be 0. We assume $f(0) = 0$ and $h(0) = 0$. If this is not the case, then in many important cases it can be made so by a possibly time varying change of state and output coordinates. As is usual, the state ξ is n dimensional, the control u is m dimensional and the output y is p dimensional. It is relatively straightforward generalization to consider systems where y depends directly on u, as in

$$y = h(\xi) + k(\xi)\,u, \tag{2.1d}$$

however to simplify the exposition we shall not do so.

We are interested in studying (2.1) under the pseudogroup of state coordinate transformations around $\xi^{\circ} = 0$. In [14] we studied arbitrary change of coordinates and attempted to bring the system$^{\text{c}}$ into normal form based on prime systems. Such normal forms are closely related to Brunovsky form and its dual. In this article we shall restrict our attention to changes of state coordinates $x = x(\xi)$ whose Jacobian at $\xi^{\circ} = 0$ is the identity

$$\frac{\partial x}{\partial \xi}(0) = I.$$

Such transformations have two virtues. The first is that they leave invariant the first order linear approximation to (2.1),

$$\dot{z} = Az + Bu + 0(z,u)^2 \tag{2.2a}$$

$$y = Cz + 0(z)^2 \tag{2.2b}$$

$$z(t) = \xi(t) \tag{2.2c}$$

where

$$A = \frac{\partial f}{\partial \xi}(0) \tag{2.3a}$$

$$B = g(0) \tag{2.3b}$$

$$C = \frac{\partial h}{\partial \xi}(0) \tag{2.3c}$$

The second is that the nonlinear coordinates ξ and the normal form coordinates x agree to first order, (2.4a) $$\xi = x + \phi(x)$$

where

(2.4b) $$\phi(0) = 0, \ \frac{\partial\phi}{\partial x}(0) = 0.$$

Typically, the original coordinates in which the system is described have some natural meaning and the coordinates have different dimensions, e.g., distance, velocities, mass, etc. Property (2.4) means that at least to first order the normal form coordinates have the same dimensions and intuitive meanings as the natural coordinates.

The system (2.1) admits a modified controller form if there exists a change of state coordinates (2.4) which transforms (2.1) into

(2.5a) $$\dot{x} = Ax + Bu + B(\alpha(x) + \beta(x)u)$$

(2.5b) $$y = Cx + \gamma(x)$$

It follows from (2.4) that the nonlinear terms are quadratic or higher in (x,u), i.e.

(2.6a) $$\alpha(0) = 0, \ \frac{\partial\alpha}{\partial x}(0) = 0$$

(2.6b) $$\beta(0) = 0,$$

(2.6c) $$\gamma(0) = 0, \ \frac{\partial\gamma}{\partial x}(0) = 0.$$

We require that the $m \times m$ matrix $1 + \beta(x)$ be invertible for x of interest. These conditions (2.4) and (2.6) insure that A, B, C are given by (2.3). Hence the linear part of modified controller form of (2.1) is the same as the first order approximation (2.2) to (2.1).

The system (2.1) admits a modified observer form if there exists a change of state coordinates (2.4) which transforms (2.1) into

(2.7a) $\dot{x} = Ax + Bu + \alpha(\bar{y}) + \beta(\bar{y})u,$

(2.7b) $y = \bar{y} + \gamma(\bar{y}),$

(2.7c) $\bar{y} = Cx.$

It follows from (2.1) that the nonlinear terms are quadratic or higher in $\bar{y}$, i.e.

(2.8a) $\alpha(0) = 0,\ \frac{\partial\alpha}{\partial\bar{y}}(0) = 0$

(2.8b) $\beta(0) = 0,$

(2.8c) $\gamma(0) = 0,\ \frac{\partial\gamma}{\partial\bar{y}}(0) = 0.$

We require that the mapping (2.7b) be invertible between the y and $\bar{y}$ of interest. Once again the A, B, C of (2.7) are the same as those of (2.2, 3, 5).

Henceforth we shall drop the "modified" and refer to (2.5) and (2.7) as controller and observer forms. Of course generally it is not the same change of coordinates taking (2.1) to (2.5) and (2.7) and the α, β, γ are different. In particular, the dimensions and arguments of α, β, γ differ between (2.5) and (2.7). When necessary we shall use subscripts c and o to distinguish controller coordinates $x_c = \xi - \phi_c(x_c)$ and functions $\alpha_c(x_c)$, $\beta_c(x_c)$, $\gamma_c(x_c)$ from observer coordinates $x_o = \xi - \phi_o(x_o)$ and functions $\alpha_o(Cx_o)$, $\beta_o(Cx_o)$, $\gamma_o(Cx_o)$.

3. <u>Poincaré Linearization</u>. Henri Poincaré considered the problem of transforming a nonlinear vector field into a linear vector field by a change of coordinates around a critical point. We briefly describe his theory, a fuller description can be found in Guckenheimer

and Holmes [6] and Arnold [1].

We are given a single vector field

(3.1c) $\dot{\xi} = f(\xi)$

(3.1b) $f(0) = 0$

with a critical point at $\xi^{o} = 0$. We are interested in finding a change of coordinates (2.4) which transforms (3.1) into a linear vector field,

(3.2) $\dot{x} = Ax$

where A is given by (2.3a).

Poincaré noted that one could develop the change of coordinates term by term in homogeneous powers of x. At degree two we seek an n dimensional vector field $\phi^{(2)}(x)$ whose entries are homogeneous polynomials of degree 2 in x such that under the change of coordinates

(3.3) $\xi = x + \phi^{(2)}(x)$

the differential equation (3.1) is transformed to

(3.4) $\dot{x} = Ax + O(x)^3$

whose $O(x)^3$ denotes cubic and higher terms in x. Superscripts in parentheses will be used to indicate that the function is homogeneous of the degree of the superscript in its arguments. If we expand $f(\xi)$ in homogeneous powers of ξ, (3.5) $$f(\xi) = A\xi + f^{(2)}(\xi) + f^{(3)}(\xi) + \cdots$$

then (3.1) is transformed into (3.4) iff $\phi^{(2)}(\xi)$ satisfies the so called homological equation

(3.6a) $$[Ax, \phi^{(2)}(x)] = f^{(2)}(x)$$

where [,] is the Lie–Jacobi bracket

(3.6b) $$[Ax, \phi^{(2)}(x)] = \frac{\partial\phi^{(2)}}{\partial x}(x)\, Ax - A\phi^{(2)}(x)$$

It is straightforward to verify that $[Ax, \cdot]$ is a linear map from homogeneous vector fields of degree 2 into homogeneous vector fields of degree 2. Moreover the homogeneous n dimensional vector fields of degree 2 form a linear space of dimension $n^2(n+1)/2$. Hence (3.6a) is solvable for arbitrary $f^{(2)}$ iff zero is not an eigenvalue of the linear mapping defined by $[Ax, \cdot]$. Poincaré noted that the eigenvalues of this mapping are related to the eigenvalues of A in a simple fashion. To see why, suppose A is semisimple, i.e., there exists a basis $v^1,\dots,v^n$ of eigenvectors of A

(3.7a) $$Av^i = \lambda_i v^i$$

possibly over the complex numbers.

Let $w_1,\dots,w_n$ be a cobasis of left eigenvectors of A,

(3.7b) $$w_i A = \lambda_i w_i$$

Then the space of n vector fields homogeneous of degree 2 has as a basis

$$(3.8) \qquad \phi^k_{ij}(x) = v^k(w_i x)\,(w_j x)$$

when $1 \leq i \leq j \leq n$ and $1 \leq k \leq n$. A straightforward calculation yields

$$[Ax, \phi^k_{ij}(x)] = (\lambda_i + \lambda_j - \lambda_k)\, \phi^k_{ij}(x).$$

Hence the eigenvalues of $[Ax, \cdot]$ on vector fields homogeneous of degree 2 are

$$(3.9) \qquad \lambda_i + \lambda_j - \lambda_k$$

when $1 \leq i \leq n$ and $1 \leq k \leq n$.

Hence the homological equation (3.6a) is solvable if no expression of the form (3.9) is zero. Of course this is a sufficient but not necessary condition because a particular $f^{(2)}$ might well be the range of $[Ax, \cdot]$, e.g., $f^{(2)} = 0$.

If (3.6a) is solvable one can proceed to look for a transformation canceling the third degree terms in f,

$$(3.10) \qquad \xi = x + \phi^{(3)}(x)$$

and $[Ax, \cdot]$ is linear mapping of these vector fields homogeneous of degree 3 into themselves. The eigenvalues of this mapping are

$$(3.12) \qquad \lambda_i + \lambda_j + \lambda_k - \lambda_\ell$$

where $1 \leq i \leq j \leq k \leq n, 1 \leq \ell \leq n$.

Hence (3.11) is solvable for arbitrary $f^{(3)}$ iff none of (3.12) is zero. This generalizes to higher degree. If one of (3.9) or (3.12) or their generalization is zero then there is "resonance" and linearization is not always possible. We refer the reader to [1] and [6] for more details.

4. Approximate Controller Form. S. Karahan in his Ph.D. thesis [12] studied the application of Poincaré's method to finding controller forms and approximate controller forms. We give a brief description of his work.

One starts by expanding (2.1) into homogeneous powers of (ξ, u),

$$\dot{\xi} = A\xi + Bu + f^{(2)}(\xi) + g^{(1)}(\xi)\,u + \ldots \tag{4.1a}$$

$$y = C\xi + h^{(2)}(\xi) + \ldots \tag{4.1b}$$

One seeks a change of coordinates

$$\xi = x + \phi^{(2)}(x) \tag{4.2}$$

transforming (4.1) into approximate controller form

$$\dot{x} = Ax + Bu + B(\alpha^{(2)}(x) + \beta^{(1)}(x)u) + O(x,u)^3 \tag{4.3a}$$

$$y = Cx + \gamma^{(2)}(x) + O(x)^3 \tag{4.3b}$$

Following Poincaré, we see that this will happen iff

(4.4a) $[Ax, \phi^{(2)}(x)] + B\,\alpha^{(2)}(x) = f^{(2)}(x)$

(4.4b) $[Bu, \phi^{(2)}(x)] + B\beta^{(1)}(x)u = g^{(1)}(x)u$

where (4.4b) must hold for each constant u. We refer to these as the generalized homological equations. Like the homological equations, they are linear equations but they are generally not square. The space of unknown $\phi^{(2)}(x)$, $\alpha^{(2)}(x)$ and $\beta^{(1)}(x)$ is $n^2(n+1)/2 + mn(n+1)/2 + m^2n$ dimensional. The constraint space of $f^{(2)}$ and $g^{(1)}$ is $n^2(n+1)/2 + n^2m$ space. These dimensions agree iff $n = 2m+1$. Generally the map $\phi^{(2)}, \alpha^{(2)}, \beta^{(1)} \longmapsto f^{(2)}, g^{(1)}$ is not of full rank so it is not always solvable even when $n = 2m+1$.

Karahan has analyzed this mapping using a basis and cobasis related to the controllability matrix $(B, AB, ..., A^{n-1}B)$. We refer the reader to [12] for details.

Since the system (4.4a) is generally not solvable one is forced to seek approximate solutions. One way of doing so is to find $\bar{f}^{(2)}$ and $\bar{g}^{(1)}$ *in the range of the mapping* (4.2) which is closest in some least squares sense to the given $f^{(2)}$ and $g^{(1)}$. Moreover one would like to choose the smallest $\phi^{(2)}$, $\alpha^{(2)}$ and $\beta^{(1)}$ which maps into $\bar{f}^{(2)}$ and $\bar{g}^{(1)}$. Again we refer the reader to [12] for more details.

Before closing this section it should be mentioned how an approximate controller form (4.3) can be used to stabilize a nonlinear system (2.1) (or equivalently (4.1)) by state feedback. The standard approach is to approximate the nonlinear system to first order by (2.2), choose a stabilizing feedback law for (2.2), $u = Fz$, transform this back into original coordinates,

(4.5) $u = F\xi.$

Expressed in homogeneous terms the closed loop dynamics is

$$\dot{\xi} = (A + BF)\,\xi + f^{(2)}(\xi) + g^{(1)}(\xi)\,F\xi + O(\xi)^3 \tag{4.6}$$

and hence the system is locally stable around $\xi^o = 0$. Of course, if it is too far from $\xi^o = 0$, the quadratic and higher terms may drive it unstable.

In the normal form approach, we typically will use the same stabilizing state feedback gain F but to apply it to the second order linearization (4.3) rather than the first order linearization (2.2). The resulting feedback is

$$u + \alpha^{(2)}(x) + \beta^{(1)}(x)u = Fx \tag{4.7}$$

which results in x coordinates the closed loop system

$$\dot{x} = (A + BF)x + O(x)^3 \tag{4.8}$$

Generally speaking, it is better to implement the feedback in the original ξ coordinates taking advantage of the fact that the inverse to (4.2) is

$$x = \xi - \phi^{(2)}(\xi) + O(\xi)^3 \tag{4.9}$$

Neglecting higher than quadratic terms we obtain from (4.7) the feedback

$$u = F\xi - (F\phi^{(2)}(\xi) + \alpha^{(2)}(\xi) + \beta^{(1)}(\xi)\,F\xi) + O(\xi)^3. \tag{4.10}$$

Note that to first degree the standard feedback (4.5) and the feedback (4.9) agree. However, the second degree terms of (4.10) cancel the second degree terms of (4.6) to obtain in x coordinates (4.8). One expects that (4.10) is asymptotically stabilizing over a larger neighborhood of ξ^o then (4.5).

Of course one can also seek a higher degree approximate controller form. The dimensions of the homological and generalized homological equations grow exponentially in the degree of the approximation. Hence this approach may not be recommended. It might be more efficient and effective to find approximate controller forms of degree two around several operating points rather than an approximate controller form of degree three around a single point.

5. Approximate Observer Form. The work I'm about to describe is joint with Andrew Phelps. We seek a change of coordinates of the form (4.2) which transforms (4.1) into approximate observer form

$$\text{(5.1a)} \qquad \dot{x} = Ax + Bu + \alpha^{(2)}(\bar{y}) + \beta^{(1)}(\bar{y})u + O(x,u)^3$$

$$\text{(5.1b)} \qquad y = Cx + \gamma^{(2)}(\bar{y}) + O(x)^3$$

$$\text{(5.1c)} \qquad \bar{y} = Cx.$$

As before this is possible iff we can solve another set of generalized homological equations

$$\text{(5.2a)} \qquad [Ax, \phi^{(2)}(x)] + \alpha^{(2)}(Cx) = f^{(2)}(x)$$

$$\text{(5.2b)} \qquad [Bu, \phi^{(2)}(x)] + \beta^{(1)}(Cx)u = g^{(1)}(x)u$$

$$\text{(5.2c)} \qquad \gamma^{(2)}(Cx) - C\phi^{(2)}(x) = h^{(2)}(x)$$

As before (5.2b) must hold for each constant u.

These equations are a linear mapping from the space of functions $\phi^{(2)}(x)$, $\alpha^{(2)}(Cx)$, $\beta^{(1)}(Cx)$, $\gamma^{(2)}(Cx)$ to the space of functions $f^{(2)}(x)$, $g^{(1)}(x)$, $h^{(2)}(x)$. The dimension of the domain is $n^2(n+1)/2 + np(p+1)/2 + m\,2p + p^2(p+1)/2$ and that of the range is $n^2(n+1)/2 + mn(n+1)/2 + pn(n+1)/2$. In general, these equations (5.2) are not solvable so as before one must seek a least squares solution. We shall report on that in more detail at another time.

If (2.1) (equivalently (4.1)) can be transformed to approximate observer form then it is easy to construct an observer. We choose H so that A + HC is sufficiently stable. An approximation $\hat{x}(t)$ to $x(t)$ is defined to evolve according to

$$(5.3) \qquad \begin{aligned} \dot{\hat{x}} &= (A + HC)\,\hat{x} + Bu - H(y - \gamma^{(2)}(y)) \\ &\quad + \alpha^{(2)}(y-\gamma^{(2)}(y)) + \beta^{(2)}(y-\gamma^{(2)}(y))u \end{aligned}$$

then the error $\tilde{x}(t) = x(t) - \tilde{x}(t)$
satisfies

$$(5.4) \qquad \dot{\tilde{x}} = (A + HC)\,\tilde{x} + O(x, \hat{x}, u)^3.$$

Hence if the initial error is not too large and u is also not too large, we can expect $\tilde{x}(t) \to 0$ as $t \to \infty$.

Of course, it is preferable to implement the observer in natural coordinates so we transform (5.3) using $\hat{\xi} = \hat{x} + \phi^{(2)}(\hat{x})$ to obtain

(5.5a) $$\dot{\xi} = A\hat{\xi} + Bu + H(\hat{y} - y) + f^{(2)}(\hat{\xi}) + g^{(1)}(\hat{\xi})\, u$$
$$+ \alpha^{(2)}(y) - \alpha^{(2)}(\hat{y}) + (\beta^{(2)}(y) - \beta^{(2)}(\hat{y}))u$$
$$+ H(\gamma^{(2)}(y) - \gamma^{(2)}(\hat{y})) + \frac{\partial\phi^{(2)}}{\partial\xi}\,(\hat{\xi})\, H(\hat{y} - y)$$
$$+ O(\xi,\, \hat{\xi},\, u)^3$$

(5.5b) $$\hat{y} = C\hat{\xi} + h^{(2)}(\hat{\xi})$$

(5.5c) $$\hat{\xi}(0) = \xi^{o} = 0.$$

Notice that the linear part of (5.5) is the observer for (2.1) one would obtain from the linear approximation (2.2), namely

(5.6a) $$\dot{\hat{z}} = A\hat{z} + Bu + H(\hat{y} - y)$$

(5.6b) $$\hat{y} = C\hat{z}$$

(5.6c) $$\hat{z}(0) = \xi^{o} = 0.$$

The error $\tilde{z} = \xi - \hat{z}$ between (2.1) and (5.6) satisfies

(5.7a) $$\dot{\tilde{z}} = (A + HC)\,\tilde{z} + O(\xi,\, \hat{\xi},\, u)^2$$

while the error of the observer (5.5) expressed in $\tilde{x}$ coordinates satisfies (5.4). Hence one expects (5.5) to perform better as an observer for (2.1) over a larger operating range.

As with the state feedback (4.10), the second degree terms of the observer (5.5) are a correction to the standard linear observer for the quadratic nonlinearations of the original system. In implementations one would replace the state ξ in the state feedback control law (4.9) with the estimate $\hat{\xi}$ from (5.5).

One can continue this process and look for a third degree change of coordinates which transforms the system into approximate observer form where the error terms are $O(\xi, u)^3$. One obtains in this further third order correction to the state feedback (4.10) and observer (5.5). Viewed in this light, we see that the approximate normal form approach allows us to start with a standard linear design based on the linear approximation (2.2) and build in a succession of higher degree corrections to overcome the nonlineararities of (4.1). Throughout we can keep the same feedback gain K and observer gain H, and these can be chosen by standard linear design techniques applied to the linear approximation (2.2).

6. <u>Coprime Factorizations</u>. This work is joint with Yi Zhu [19]. Suppose we have a system in controller normal form

(6.1a) $$\dot{x}_c = Ax_c + Bu + B(\alpha_c(x_c) + \beta_c(x_c)u)$$

(6.1b) $$y = Cx_c + \gamma_c(x_c)$$

(6.1c) $$x_c(0) = 0$$

where the c–subscripts indicate coordinates and functions associated to controller normal form. We view (6.1) as defining an input/output map

(6.2a) $$G: u(\cdot) \longmapsto y(\cdot)$$

from functions $u(t)$ to $y(t)$ for $t \geq 0$.

We seek a right factorization of G

$$G = N \circ M^{-1}$$

where N and M are input/output maps

(6.2b) $M\colon v(\cdot) \longmapsto u(\cdot)$

(6.2c) $N\colon v(\cdot) \longmapsto y(\cdot),$

M is invertible and o denotes composition. There is a large and growing literature on coprime factorization of both linear and nonlinear systems. A sampling is [2–5, 7, 10, 11, 13, 17, 18, 20–26]. In particular our approach follows [3, 4].

To describe the input/output maps M and N we shall use a state space realization. In particular we define M to be the input/output map of

(6.3a) $\dot{\xi}_c = (A + BF)\, \xi_c + Bv$

(6.3b) $\alpha_c(\xi_c) + (1 + \beta_c(\xi_c))u = F\xi_c + v$

(6.3c) $\xi_c(0) = 0$

where (6.3b) defines u as a function of ξ_c and v.

We consider the composition $N = G \circ M$, this is realized by the 2n dimensional system (6.1, 3) described in ξ_c, x_c coordinates. Let $e = x_c - \xi_c$ then

$$(6.4) \qquad \dot{e} = Ae + B(-F\xi_c - v + \alpha_c(x_c) + (1 + \beta_c(x_c))(1 + \beta_c(\xi_c))^{-1}(F\xi_c + v - \alpha_c(\xi_c)))$$

If $e(t) = 0$ then $\dot{e}(t) = 0$. Since $e(0) = 0$ we conclude that $e(t) = 0$ for all $t \geq 0$. In other words, the realization (6.1, 3) of N is not controllable because $e(t)$ is unaffected by the input $v(t)$.

A controllable realization of N is

$$(6.5a) \qquad \dot{\zeta}_c = (A + BF)\zeta_c + Bv$$

$$(6.5b) \qquad y = C\zeta_c + \gamma_c(\zeta_c)$$

$$(6.5c) \qquad \zeta_c(0) = 0$$

Hence we conclude that $G = N \circ M^{-1}$ where N and M are realized by (6.5) and (6.3). Notice that M is invertible since $(1 + \beta_c)$ is invertible by assumption.

Notice also that if (A, B) is a controllable pair then we can choose F so that (6.3) and (6.5) are stable systems. Hence we have factored G over the ring of stable nonlinear systems. We are being deliberately vague about the precise definition of a stable nonlinear system. It is clear that (6.3, 5) are "stable" under any reasonable definition.

Of course, we are interested in coprime factorizations over the ring of stable nonlinear systems. Again we should not try to make this concept precise but following Hammer [7] and others we shall say that $G = N \circ M^{-1}$ is a coprime factorization if there exists $\tilde{P}$, the input/output map of a stable system,

$$(6.6a) \qquad \tilde{P}: \binom{u}{y} \longmapsto w$$

such that the composition

(6.6b) $\tilde{P} \circ \binom{M}{N}: v \longmapsto \binom{u}{y} \longmapsto w$

is the identity, $w = v$.

The input/output map of $\binom{M}{N}$ can be realized by an n dimensional system

(6.7a) $\dot{\xi}_c = (A + BF)\, \xi_c + Bv$

(6.7b) $\alpha_c(\xi_c) + (1 + \beta_c(\xi_c))u = F\xi_c + v$

(6.7c) $y = C\xi_c + \gamma_c(\xi_c)$

(6.7d) $\xi_c(0) = 0$

A left inverse of (6.7) is

(6.8a) $\dot{z}_c = Az_c + Bu + B\,(\alpha_c(z_c) + \beta_c(z_c)\, u)$

(6.8b) $w = \alpha_c(z_c) + (1 + \beta_c(z_c))u - Fz_c$

(6.8c) $z_c(0) = 0$

If $e = \xi_c - z_c$ then

$$\dot{e} = Ae + B(\alpha_c(\xi_c) - \alpha_c(z_c) + (\beta_c(\xi_c) - \beta_c(z_c))u)$$

If $e(t) = 0$ the $\dot{e}(t) = 0$ and since $e(0) = 0$ it follows that $e(t) = 0$ for all $t \geq 0$. If $e(t) = \xi_c(t) - z_c(t) = 0$ then $w(t) = v(t)$ so (6.8) inverts (6.7).

However we do not know that (6.8) is stable. To insure the stability of (6.8), we must add to (6.8a) an extra term. This term must stabilize (6.8) and also must be zero when $\xi_c = z_c$ so that (6.8) remains a left inverse of (6.7). How do we find such a term?

Notice that the dynamics (6.8a) is the same as the dynamics of the original system (6.1a) and notice that the other output y of (6.7) does not appear in (6.8). Perhaps we can inject y into (6.8a) to stabilize it? This is more or less equivalent to asking whether output injection can be used to stabilize the original system (6.1). This is always possible for systems in observer form, hence we assume that there exists a change of coordinates

$$x_c = x_o + \phi(x_o) \tag{6.9}$$

satisfying (2.4b) transforming (6.1) into observer form

$$\dot{x}_o = Ax_o + Bu + \alpha_o(Cx_o) + \beta_o(Cx_o)u \tag{6.10a}$$

$$y = Cx_o + \gamma_o(Cx_o) \tag{6.10b}$$

$$x_o(0) = 0 \tag{6.10c}$$

Suppose we consider a similar change of coordinates for (6.8)

$$z_c = z_o + \phi(z_o) \tag{6.11}$$

to obtain

$$\dot{z}_o = Az_o + Bu + \alpha_o(Cz_o) + \beta_o(Cz_o)u \tag{6.12a}$$

$$w = \alpha_o(z_o + \phi(z_o)) + (1 + \beta_c(z_o + \phi(z_o)))u - F(z_o + \phi(z_o)) \tag{6.12b}$$

(6.12c) $z_0(0) = 0.$

We add to (6.12a) the term

(6.13a) $$\alpha_0(\bar{y}) - \alpha_0(Cz_0) + (\beta_0(\bar{y}) - \beta_0(Cz_0))\, u + H(Cz_0 - \bar{y})$$

to obtain

(6.12aa) $$\dot{z}_0 = (A + HC)\, z_0 + Bu + \alpha_0(\bar{y}) + \beta_0\, (\bar{y})\, u - Hy$$

where $\bar{y}$ is a function of y of (6.7c) defined by

(6.13b) $$y = \bar{y} + \gamma_0(\bar{y}) = C\xi_0 + \gamma_0(C\xi_0)$$

and ξ_0 is the state of (6.7) in observer coordinates

(6.13c) $$\xi_c = \xi_0 + \phi(\xi_0)$$

Notice that (6.13a) is zero whenever $\xi_0 = z_0$, hence the input/output map $\tilde{P}$ of the (6.12aa, b, c) is also an inverse of (6.7). Also, if (C,A) is an observable pair then we can choose H so that (6.12aa, b, c) is stable.

In summary, we have shown that if a nonlinear system admits both controller and observer form than its input/output map G can be factored into the composition $N \circ M^{-1}$ of input/output maps of stable sy^tems N and M. Moreover this composition is coprime in the sense that the input/output map $\binom{M}{N}$ has a left inverse $\tilde{P}$ which is realized by a stable system.

We have not presented this as a theorem because we are reluctant at this point in time to give formal definitions of coprimeness and stability for nonlinear systems. However the above development is very analogous to the linear theory [3, 4]. See also Hammer [7]

Unfortunately the analogy is not so straightforward for left copiime factorizations. The theory of left coprime factorizations for nonlinear systems has some substantial differences with the linear theory.

We start with a system in observer form (6.10) realizing an input/output map G. We define another input output map

(6.14) $$\tilde{M}: \binom{u}{y} \longmapsto w,$$

by

(6.15a) $$\dot{\xi}_0 = (A + HC)\xi_0 - H\bar{y} + \alpha_0(\bar{y}) + \beta_0(\bar{y})\, u$$

where $\bar{y}$ is an invertible function of the input y defined by

(6.15b) $$y = \bar{y} + \gamma_0(\bar{y})$$

and the output is

(6.15c) $$w = -C\xi_0 + \bar{y}$$

(6.15d) $$\xi_0(0) = 0$$

Consider the serial connection of (6.10) and (6.15), this is not a realization of the $\tilde{M} \circ G$ but it is a realization of $\tilde{N} = \tilde{M} \circ \binom{I}{G}$. (This is the first important difference with the linear theory). If we define $\xi_o = x_o - \xi_o$ then $\tilde{N}$ is realized by

(6.16a) $\qquad \dot{\zeta}_o = (A + HC)\zeta_o + Bu$

(6.16b) $\qquad w = C\zeta_o$

(6.16c) $\qquad \zeta_o(0) = 0$

because in ξ_o, x_o coordinates for (6.10, 15) only the ξ_o coordinates are observable from the output w. We consider $\tilde{N}$, $\tilde{M}$ as a left factorization of G, although it is really a left factorization of $\binom{I}{G}$ in the sense that

(6.17) $\qquad \tilde{M} \circ \binom{I}{G} = \tilde{N}$

Notice that we cannot compose this on the left with $\tilde{M}^{-1}$ since $\tilde{M}$ is not invertible as a mapping from $\binom{u}{y}$ to w.

Perhaps the best way of viewing the situation is

(6.18a) $\qquad \begin{bmatrix} I & 0 \\ 0 & \tilde{M} \end{bmatrix} \circ \begin{bmatrix} I \\ G \end{bmatrix} = \begin{bmatrix} I \\ \tilde{N} \end{bmatrix}$

or

(6.18b) $\qquad \begin{bmatrix} I \\ G \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \tilde{M} \end{bmatrix}^{-1} \circ \begin{bmatrix} I \\ \tilde{N} \end{bmatrix}$

The matrix notation is somewhat misleading because $\tilde{M}$ depends on both u and y.

In any case, if (C, A) is an observable pair then both (6.15) and (6.16) can be made stable by proper choice of H. In particular, the nonlinearities in (6.15a) are memoryless functions of the inputs u and y hence (6.15) is BIBO stable.

Next we address the coprimeness of the above factorizatioi..

We consider the input/output map

(6.19a) $$[-\tilde{N}, \tilde{M}]: \binom{u}{y} \longmapsto w$$

where again the matrix notation is somewhat misleading since both u and y are inputs to $\tilde{M}$, i.e.,

(6.19b) $$w = -\tilde{N}(u) + \tilde{M}\binom{u}{y}.$$

This input/output map can be realized by an n *dimensional system*

(6.20a) $$\dot{\xi}_o = (A + HC)\xi_o + Bu + \alpha(\bar{y}) + \beta_o(\bar{y})u - H\bar{y}$$

where $\bar{y}$ is an invertible function of the input y defined by

(6.20b) $$y = \bar{y} + \gamma_o(\bar{y})$$

and the output w is given by

(6.20c) $$w = -C\xi_o + \bar{y}$$

We wish to find a input/output map P realized by a stable system so that P is a right inverse of $[-\tilde{N}, \tilde{M}]$,

(6.21a) $\quad P: v \longmapsto \binom{u}{y}$

(6.21b) $\quad [-\tilde{N}, \tilde{M}] \circ P: v \longmapsto w = v.$

We start by constructing an inverse for (6.20),

(6.22a) $\quad \dot{z}_0 = Az_0 - Hv + Bu + \alpha_0(\bar{y}) + \beta_0(\bar{y})u$

(6.22b) $\quad \bar{y} = Cz_0 + v$

(6.22c) $\quad y = \bar{y} + \gamma_0(\bar{y})$

(6.22d) $\quad u = ?$

(6.22e) $\quad z_0(0) = 0$

We leave unspecified for the moment the output u which also appears in the dynamics (6.22a). Notice that if $e = \xi_0 - z_0$ is the error between the states of (6.20) and (6.22) then $\dot{e} = 0$ whenever $e = 0$. Since $e(0) = 0$ we conclude that $e(t) = 0$ for all $t \geq 0$ and so by (6.20c) and (6.22b) we have $w(t) = v(t)$. In other words (6.22) is a right inverse of (6.20).

What about the stability of (6.22)? We would like to choose the output u in such a way that (6.22a) is stable in some sense. If we ignore the $-Hv$ term of (6.22a) this looks like the original system is observer form. This is not exactly true because $\bar{y}$ is defined by (6.22b) with v present. Suppose the original system can be transformed into controller form (6.1) by a change of coordinates (6.9). If we apply a similar change of coordinates (6.11) to (6.22) we obtain

(6.23a) $$\dot{z}_c = Az_c + Bu + B(\alpha_c(z_c) + \beta_c(z_c)u) - Hv - \frac{\partial\phi}{\partial z_o}(Hv) + (1 + \frac{\partial\phi}{\partial z_o})(\alpha_o(Cz_o + v) - \alpha_o(Cz_o) + (\beta_o(Cz_o + v) - \beta_o(Cz_o))u)$$

Suppose we choose an F such that $(A + BF)$ is stable and define u by

(6.22dd) $$\alpha_c(z_c) + \beta_c(z_c)\,u = Fz_c.$$

When the input $v = 0$, (6.23a) becomes

(6.23b) $$\dot{z}_c = (A + BF)\,z_c.$$

Unfortunately we cannot conclude that (6.23a) is BIBO stable since the input v is multiplied by a function of the state.

We conclude by noting that a "nonlinear Bezout identity" holds for the above. In other words beside $\tilde{P}$ being a left inverse (6.6b) for $\binom{M}{N}$ and P a right inverse (6.21b) for $[-\tilde{N}, \tilde{M}]$, it is also true that

(6.24a) $$[-\tilde{N}, \tilde{M}] \circ \binom{M}{N}: v \longmapsto \binom{u}{y} \longmapsto w = 0$$

and

(6.24b) $$\tilde{P} \circ P: v \longmapsto \binom{u}{y} \longmapsto w = 0$$

In abuse of notation we summarize these equations by

$$(6.25) \qquad \begin{bmatrix} \tilde{P} \\ [-\tilde{N}, \;\; \tilde{M}] \end{bmatrix} \circ \begin{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} & P \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

The verification of (6.24) is straightforward.

From the work of Doyle [3] Francis [4] and others the, the existence of a nonlinear Bezout identity suggests that it might be possible to develop a nonlinear version of Youla's Q parameterization of all stable and stabilizing controller of a linear system. This generalization would apply to those nonlinear systems which admit both controller and observer form. This class is very thin, but perhaps such a result could be extended approximately to those systems that approximately admit controller and observer form. Work in these areas is continuing.

7. Concluding Remarks. We have briefly described an approach to nonlinear compensator design based on nonlinear normal forms and approximately normal forms. This approach is being pursued by a group of researchers at U.C. Davis with support from AFOSR. The principle advantage of the normal forms approach is that to a large extent it reduces problems in nonlinear design to problems in linear design. We are developing software tools which utilize this approach as a compliment to existing linear design software so that these linear design packages can be used for nonlinear systems that admit at least approximate normal forms.

References

[1] ARNOLD, V.I., Geometrical Methods in the Theory of Ordinary Differential Equations, Springer Verlag, NY, 1983.

[2] C. DESOER, R.W. LIU, J. MURRAY AND R. SAEKS, Feedback System Design: the Fractional Representation Approach, IEEE Trans. Automat. Control, AC–25 (1980), pp. 399–412.

[3] J. DOYLE, Lecture Notes in Advances in Multivariable Control, Office of Naval Research/Honeywell Workshop, Minneapolis, MN, 1984.

[4] B. FRANCIS, A Course in H_∞ Control Theory, Lecture Notes in Control and information science, Vol. 88, Springer–Verlag, 1987.

[5] B. FRANCIS AND J. DOYLE, Linear Control Theory with an H_∞ Optimality Criterion, SIAM J. Control and Optimization, Vol 25, No. 4, July 1987, pp. 815–844.

[6] J. GUCKENHEIMER AND P. HOLMES, Nonlinear Oscillations, Dynamical Systems and Bifurcation of Vector Fields, Springer Verlag, NY, 1983.

[7] J. HAMMER, Fraction Representations of Nonlinear Systems: A Simplified Approach, Int. J. of Control, Vol. 46, No. 2, August 1987, pp. 455–472.

[8] L.R. HUNT AND R. SU, Linear equivalents of nonlinear time varying systems, Proc. MTNS, Santa Monica pp. 119–123, 1981.

[9] B. JAKUBCZYK AND W. RESPONDEK, On the linearization of control systems, Bull. Acad. Polon. Sci., Ser. Math. Astron. Phys. Vol 28 pp. 517–522, 1980.

[10] T. KAILATH, Linear Systems, Prentice Hall, Englewood Cliffs, 1980.

[11] M. VIDYASAGAR, Control System Synthesis: A Factorization Approach, MIT Press, 1985.

[12] S. KARAHAN, Higher order linear approximation to nonlinear systems, Ph. D., Mechanical Engineering, University of California, Davis, 1988.

[13] P. KHARGONEKAR AND E. SONTAG, On the relation between stable matrix factorizations and regulable realizations of linear systems over rings. IEEE Trnas. on Auto. Contr. Vol. 27, 1982, pp. 627–636.

[14] A.J. KRENER, Normal forms for linear and nonlinear systems. In Differential Geometry, the Interface between Pure and Applied Mathematics, M. Luksik, C. Martin and W. Shadwick, eds. Contempary Mathematics V. 68, American Mathematical Society, Providence, 157–189, 1986.

[15] A.J. KRENER, S. KARAHAN, M. HUBBARD AND R. FREZZA, Higher order linear approximations to nonlinear control systems. Proceedings, IEEE Conf. on Decision and Control, Los Angeles, pp. 519–523, 1987.

[16] A. PHELPS AND A.J. KRENER, Computation of observer normal form using MACSYMA. In Nonlinear Dynamics and Control, C. Byrnes, C. Martin and R. Saeks, eds. North Holland, 1988.

[17] E. SONTAG, Smooth stabilization implies coprime factorization, IEEE Trans. Auto. Control, to appear, 1989.

[18] E. SONTAG, Nonlinear control via equilinearization, Proc. IEEE Conf. on Decision and Control, Los Angeles, pp. 1363–1367, 1987.

[19] YI ZHU, Masters Thesis, Applied Mathematics, University of California, Davis, 1988.

[20] C.A. Desoer, C.A. Lin, "Nonlinear Unity Feedback Systems and Q–Parametrization," Analysis and Optimization of Systems, A. Bensoussan and J.L. Lions ed., *Lecture Notes in Control and Information Sciences*, 62, Springer Verlag, 1984.

[21] C.A. Desoer, M.G. Kabuli, "Nonlinear Plants, Factorizations and Stable Feedback Systems," *Proceedings of 26th IEEE Conference on Decision and Control*, Los Angeles, pp. 155–156, December 1987.

[22] C.A. Desoer, M.G. Kabuli, "Stabilization and Robustness of Nonlinear Unity–Feedback System: Factorization Approach," *International Journal of Control*, vol. 47, no. 4, pp. 1133–1148, 1988.

[23] C.A. Desoer, M.G. Kabuli, "Right Factorization of a Class of Time–Varying Nonlinear Systems," *IEEE Transactions on Automatic Control*, vol. 33, no. 8, pp. 755–757, August 1988.

[24] J. Hammer, "Non–Linear Systems, Stabilization, and Coprimeness," *International Journal of Control*, vol. 42, no. 1, pp. 1–20, 1985.

[25] C.N. Nett, C.A. Jacobson, M.J. Balas, "A Connection Between State–Space and Doubly Coprime Fractional Representations," *IEEE Transaction on Automatic Control*, vol. 29, no. 9, pp. 831–832, September 1984.

[26] D.C. Youla, H.A. Jabr and J.J. Bongiorno, Jr., "Modern Wiener–Hopf Design of Optimal Controllers, Part II: The Multivariable Case," *IEEE Transactions on Automatic Control*, vol. 21, pp. 319–338, June 1976.

Presented at the Imperial College Workshop on Applied Stochastic
Analysis April 1989, London. (Proceedings to be published in
Stochastics and Stochastics Reports)

AFOSR-TR- 89-1689

A LOCAL CHARACTERIZATION OF RECIPROCAL DIFFUSIONS

J.M.C. CLARK*

Department of Mathematics
University of California Davis

A reciprocal diffusion is a 'thinned' diffusion in the following intuitive sense: view the paths of a Markov diffusion on [0,1] as the shoots in a dense thicket of highly contorted shoots, all cut to a unit height (height is the independent parameter). A reciprocal diffusion is a diffusion thicket that has been thinned out just sufficiently so that the tops and bases of the shoots conform to an artificially imposed joint distribution. In particular a pinned diffusion is a reciprocal diffusion thinned so that the shoots start and end at single points (a pointed 'bush'). Much of the structure of the original distribution of the diffusion will be present in all its thinnings. The aim of this paper is to give a local characterization of this common 'reciprocal' structure.

Krener (1988a) has made a fundamental study of the infinitesimal properties of reciprocal diffusions. The problem of local characterization is one of a number of intriguing questions that were only partially answered in his paper. It turns out that the reciprocal structure of a Gaussian reciprocal diffusion has a beautifully simple characterization in terms of a linear self–adjoint second order differential operator (Frezza, Krener, Levy, 1989). Our concern is with the nonGaussian case, which has to be treated by different methods.

To proceed it is necessary to introduce some notation. Let Ω be the space of $\mathbb{R}^n$–valued continuous functions on [0,1], $x_t(\omega)$ the coordinate map $\omega(t)$ and $\mathcal{F}$ the Borel σ–field generated by the topology of uniform convergence. Suppose $X = (\{x_t : 0 \leq t \leq 1\}, \Omega, \mathcal{F}, \mathbb{P})$ is a Markov diffusion

* This research was supported by the National Science Foundation under Grant DMS 8601635 and by the United States Air Force under Grant AFOSR [illegible]

with continuous strictly positive initial and transition probability densities $p_0(x)$ and $p_{sx}(t,y)$, $0 \leq s < t \leq 1$. It is a straightforward matter to show that the probability law $\mathbb{P}$ of x has a unique factorization into the initial distribution $p_0(x)\,dx$ the final conditional distribution $p_{0x}(1,y)\,dy$ and a probability kernel $(x,y) \to \mathbb{P}_{0x}^{1y}$ on $\mathcal{F}$, narrowly continuous as a function of (x,y), that is the law of X conditional on $x_0 = x$, $x_1 = y$.

Now suppose $k(x,y)$ is a strictly positive continuous probability density on $\mathbb{R}^n \times \mathbb{R}^n$ and let $\mathbb{Q}$ be the modified law on $\mathcal{F}$: $\mathbb{Q}(A)$: $= \int \mathbb{P}_{0x}^{1y}(A)\, k(x,y)\, dxdy$. Then we shall call the process $Y = (\{x_t\}, \mathbb{Q})$, with law $\mathbb{Q}$ on $(\Omega, \mathcal{F})$, *a reciprocal diffusion governed by a Markov diffusion* X. If the marginal density k is replaced by a distribution $\pi(dx, dy)$ we could create other reciprocal diffusions from X with laws singular to $\mathbb{P}$; for instance, the pinnings of X at $t = 0$ and 1, but to keep the development simple we restrict our attention to governed reciprocal diffusions. Markov diffusions are trivially reciprocal and two diffusions governing the same reciprocal diffusion govern each other; in fact it follows from Theorem 3.1 of (Jamison, 1974) that the ratio of their end–point densities must take the form $h_0(x)\, h_1(y)$ for some continuous positive functions h_0 and h_1. Reciprocal diffusions, on the other hand, are not generally Markov. For instance, if Y has the law of $\{Z + B_t$: $0 \leq t \leq 1\}$, where $\{B_t\}$ is a Brownian bridge with $B_0 = B_1 = 0$, and Z is an independent nondegenerate random variable, then Y is a reciprocal diffusion (governed by a Brownian motion) that is not Markov. However, reciprocal diffusions inherit many of the properties of Markov diffusions. For $0 \leq r \leq t \leq 1$, let $\mathcal{I}_r^t$ denote the *interior* σ–field $\sigma\{x_s: r \leq s \leq t\}$ and $\mathcal{E}_r^t$ the *exterior* σ–field $\sigma\{x_s: 0 \leq s \leq r$ or $t \leq s \leq 1\}$ in $\mathcal{F}$. Then a reciprocal diffusion Y is a Markov field on the line in the sense that for any $0 \leq r \leq t \leq 1$, conditioned on x_r and x_t, elements of $\mathcal{I}_r^t$ are independent of elements of $\mathcal{E}_r^t$. Also for any $0 \leq s \leq 1$ the pinned process $Y^x = (\{x_t\}, \mathbb{Q}(\cdot \mid x_s = x))$ is a Markov diffusion. The first property is the definition of the reciprocal process introduced by Bernstein (1932) to give a probabilistic interpretation of Schrödinger's equation; the second the definition of a reciprocal diffusion in the sense of Krener (1988b). A detailed study of these and other properties can be found in Jamison (1974, 75). A reciprocal process possesses a 'governing' Markov process provided some technical 'end–point' conditions are satisfied. The writing of a governing diffusion into the definition adopted here simply allows us to avoid these technicalities.

The kernel $(x,y) \to \mathbb{P}^{1y}_{0x}$ can be thought of as the *reciprocal structure* that is common to the Markov diffusion X and all its subservient reciprocal diffusions. Similarly for $s < t$ the unique narrowly continuous probability kernel on $\mathscr{F}^t_s$ given by $\mathbb{P}^{ty}_{sx}(A) = \mathbb{P}(A | x_s = x, x_t = y)$ represents the common reciprocal structure on the interval $[s,t]$.

We can now define the notion of a set of local reciprocal characteristics. Let $\mathscr{R}$ denote the set of all probability laws of reciprocal diffusions. We shall say that an assignment of a continuous function $\rho^{\mathbb{Q}}(\cdot,\cdot)$ on $[0,1] \times \mathbb{R}^n$ to every law $\mathbb{Q}$ in $\mathscr{R}$ is a *reciprocal invariant on* $[s,t,]$ if the restriction of $\rho^{\mathbb{Q}}(u,\cdot)$ to the interval $s \leq u \leq t$ is the same for all $\mathbb{Q}$ with a common reciprocal structure on $[s,t]$ and that $\rho^{\cdot}$ is a *local reciprocal invariant* if it is a reciprocal invariant on $[s,t]$ for all $0 \leq s < t \leq 1$. Finally we shall say that a family $\{\rho^{\cdot}_1, \rho^{\cdot}_2, \cdots, \rho^{\cdot}_n\}$ of local reciprocal invariants is a *set of local reciprocal characteristics* if it uniquely determines the reciprocal structure on $[0,1]$; that is, given a set of functions $\{r_1, r_2, \cdots, r_n\}$ within the joint range of $\{\rho^{\cdot}_1, \cdots, \rho^{\cdot}_n\}$ those laws in the inverse image $\{\mathbb{Q}: \rho^{\mathbb{Q}}_i = r_i, i = 1, \cdots, n\}$ possess a common reciprocal structure $\mathbb{P}^{1\cdot}_{0\cdot}$.

The next section presents a theorem showing how a set of local reciprocal characteristics for a reciprocal diffusion can be constructed by a differential transformation of its *semimartingale* local characteristics.

The final section presents two formulae that give a probabilistic interpretation of local reciprocal characteristics. The first is a curiously factored form of Girsanov's formula that is applicable to pinned diffusions.

The second requires some preliminary explanation. Krener (1988b) has introduced a number of interrelated Feller like postulates, that are satisfied by reciprocal diffusions and that display the infinitesimal form of particular conditional amounts of their first and second differences. The diffusion matrix $\underset{\sim}{a}(t,x)$ (that is, the rate of change of the matrix quadratic variation process) plays a central role. One important postulate states that for constant $x, v \in \mathbb{R}^n$, the limit*

$$\lim_{h\to 0} \frac{1}{h^2} E[x_{t+h} - 2x_t + x_{t-h} | x_{t+h} = x + vh, x_{t-h} = x - vh]$$
$$= f(t,x) + g(t,x)v \qquad (1)$$

for some vector and matrix functions f and g (when expressed in Riemann normal coordinates with respect to the Riemann tensor a^{-1}). Another is the corresponding expression (1') for the (matrix) second moment of the second difference; here the right member is $2\,a(t,x)$.

Our second formula (Theorem 3) can be regarded as an integrated non–asymptotic variation of (1) for constant diffusion matrices a. Together with the matrix a its coefficients form a set of local reciprocal characteristics. So, at least in this integrated form, Krener's postulate (1), together with the matrix a, characterize the reciprocal structure. The general case with non–constant a will be treated elsewhere.

My thanks are due to A.J. Krener who besides offering me hospitality and support at U.C. Davis, largely provided the inspiration for this work, and to B.C. Levy and R. Frezza for educating me in the ways of Gaussian reciprocal processes.

* In (1988a) Krener uses a slightly different conditioning in his postulates: $E[\,\cdot\,| x_{t-h} + x_{t+h} = 2x]$. This introduces quantities such as Nelson's current velocity that are not generally reciprocal invariants.

A LOCAL RECIPROCAL CHARACTERIZATION.

From now on we use the summation convention. Using the notation of the previous section, suppose $x = (\{x_t\}: 0 \leq t \leq 1\} \Omega, \mathcal{F}, \mathbb{P})$ (or $(\{x_t\}, \mathbb{P})$ for short) is a Markov diffusion in $\mathbb{R}^n$ with diffusion matrix $a(t,x)$ and drift vector $b(t,x)$. For simplicity we assume that the components of a and b are C_b^∞, (that is, they are bounded and have bounded derivatives of all order) and that a is uniformly elliptic. Then x possesses a strictly positive transition density $p_{sx}(t,y)$. It is also a continuous semimartingale on $\{\mathcal{F}_0^t\}$ and has the decomposition

$$dx_t = b(t, x_t)\, dt + dN_t,$$

where $\{N_t\}$ is a continuous martingale on $\{\mathcal{F}_0^t\}$ diffusing at the rate $a(t,x_t)$; that is, its (matrix) quadratic variation is of the form $\int_0^t a(s, x_s)ds$.

Now suppose $Y = (\{x_t\}, \mathbb{Q})$ is a reciprocal diffusion governed by X. Suppose k is its positive continuous end–point density, and k_0 its initial density. Let $h_x(1,z)$ be the conditional end–point density relative to $p_{0x}(1,z)$ given by

$$h_x(1,z) = k(x,z)/(k_0(x)\, p_{0x}(1,z))$$

then $h_x(1,\cdot)$ is positive and continuous and possesses a 'space–time harmonic extension' with respect to x given by:

$$h_x(t,y) = \int h_x(1,z)\, p_{ty}(1,z)\, dz, \text{ for } 0 \leq t \leq 1.$$

Let D_0 denote $\partial/\partial t$ and D_i denote $\partial/\partial y^i$. By Ito's formula h_x satisfies (for fixed x, with the summation convention in force)

$$D_0 h_x + b^i D_i h_x + a^{ij} D_i D_j h_x = 0$$

on $[0,1] \times \mathbb{R}^n$. Then $M_t := h_{x_0}(t, x_t)$ is the martingale $E_{\mathbb{P}}[d\mathbb{Q}/d\mathbb{P} \,|\, \mathcal{F}_0^t]$ and has the representation

$$dM_t^x = M_t^x \cdot \Sigma_i D_i(\log h_{x_0}(t, x_t))\, dN_t.$$

A 'Girsanov' argument then leads to the representation: On $(\Omega, \mathcal{F}, \mathbb{Q})$,

$$dx_t = c_{x_0}(t, x_t)\, dt + d\tilde{N}_t,$$

where $c_x^i(t,y) = b^i(t,y) + a^{ij} D_j \log h_x(t,y)$ and where $\tilde{N}_t$ is a continuous martingale on $\{\mathcal{F}_0^t\}$, again diffusing at rate $a(t, x_t)$. The full details of the argument can be found in Theorem 2 (and its proof) in Jamison (1975). Notice that $c_{\cdot}^i(\cdot,\cdot)$ is continuous and uniquely defined. The pair a and $c_{\cdot}$ can be thought of as the local semimartingale characteristics of Y with respect to the filtration $\{\mathcal{F}_0^t\}$. We can obtain a different characterization by conditioning at $t = 1$; that is, with respect to the augmented filtration $\{\mathcal{G}_t^1\}$, then

$$dx_t = e_{x_1}(t, x_t)dt + d(\text{martingale})$$

where the martingale (on $\{\mathcal{G}_t^1\}$) still diffuses at rate $a(t, x_t)$ and e_z takes the form: for $0 \le t < 1$

$$e_z^i(t,y) = b^i(t,y) + a^{ij}(t,y)\, D_j \log p_{t,y}(1,z)$$

The argument is almost the same, though now $e_z(t,y)$ becomes singular at $t = 1$.

Similar but more complicated characterizations can be obtained for conditionings at intermediate times. Now let $\alpha_{..}$ be the inverse of the matrix a. Introduce the differential operator

$$\mathrm{Re}(a,b)_i = D_0(\alpha_{ij}\, b^j) + \tfrac{1}{2} D_i(b^j\, \alpha_{jk}\, b^k + a^{jk}\, D_j(\alpha_{k\ell}\, b^\ell)),\ i = 1, \cdots, n,$$

mapping a C^∞ matrix–vector pair into a C^∞ vector field. We can now make the following statement.

Theorem 1 Let X be a Markov diffusion with a continuous positive initial density, a diffusion matrix a(t,y) and drift vector b(t,y). Let Y be a governed reciprocal diffusion with semimartingale characteristics $\bar{a}(t,y)$ and $c_x(t,y)$. We assume that the components of a, b, $\bar{a}$ and $c_.$ are C_b^∞ and that a and $\bar{a}$ are uniformly elliptic. Let $0 \leq s < u \leq 1$. Then X and Y possess the same reciprocal structure on [s,u] if and only if the following functions of (t,y) are equal on $[s,u] \times \mathbb{R}^n$

i) $a = \bar{a}$

ii) For all $x \in \mathbb{R}^n$, i, j = 1, 2, $\cdots$ n.

$$D_i(\alpha_{jk}\, b^k) - D_j(\alpha_{ik}\, b^k) = D_i(\bar{\alpha}_{jk}\, c_x^k) - D_j(\bar{\alpha}_{ik}\, c_x^k)$$

iii) For all $x \in \mathbb{R}^n$ i = 1, 2, $\cdots$ n.

$$\mathrm{Re}(a,b)_i = \mathrm{Re}(\bar{a}, c_x)_i$$

Remark 1) This result is close to being a corollary of a remarkable short–time asymptotic expansion of Krener (1988b) for the reciprocal transition density $\mathbb{P}(x_t \in dy \,|\, x_{t-h} = x,\, x_{t+h} = z)$ of a Markov diffusion that parallels the corresponding expansions of Azencott (1981) and Molchanov (1975) for Markov transition densities. However, the expansion is quite complicated and I have not used it in the proof as the following argument is more direct.

2) The theorem also holds if the c_x is replaced by the augmented semimartingale characteristics $c_z(t,y)$ provided all the functions are

restricted to $0 \leq t < 1$.

3) *For simplicity let* β_i *be the function* $\alpha_{ij}\, b^j$ for $i = 1, \cdots, n$ and let β_0 be the function $-1/2(b^j\, \alpha_{jk}\, b^k + a^{jk}\, D_j(\alpha_{k\ell}\, b^\ell))$. Then the left member of ii) is $D_i\beta_j - D_j\beta_i$ and that of iii) is $D_0\beta_i - D_i\beta_0$. Then the theorem states that the $n(n+1)/2$ functions $D_i\beta_j - D_j\beta_i$, $i, j = 0, 1, \cdots, n$, $i < j$, and the $n(n+1)/2$ components of the matrix a (or more pedantically, their assignments to the probability law of X) form a set of local reciprocal characteristics. Clearly these characteristics are not independent; in fact, in differential–geometric terms, the $D_i\beta_j - D_j\beta_i$ are the components of the exterior derivative of the 1–form $(\beta_0, \beta_1, \cdots, \beta_n)$.

Proof of the 'only if' part. Without loss of generality we may take $s = 0$, $u = 1$. Assume that X and Y possess the same kernel $\mathbb{P}^{1y}_{0x}$ on $\mathscr{S}$. Then Y is governed by X and by the previous argument it follows that, for some positive continuous X–space time harmonic function $h_x(t,y)$, $\bar{a}(t,y) = a(t,y)$ and

$$c^i_x(t,y) = b^i(t,y) + a^{ij}(t,y)\, D_j \log h_x(t,y)$$

Let β_i, $i = 0, 1, \cdots, n$ be the terms introduced in Remark 3 and let $\bar\beta_i$ be the corresponding terms (note $a = \bar{a}$).

$$\bar\beta_i = \alpha_{ij}\, c^j_x \; i = 1, \cdots n, \quad \bar\beta_0 = -\tfrac{1}{2}(c^j_x\, \alpha_{jk}\, c^k_x + a^{jk}\, D_j(\alpha_{k\ell}\, c^\ell_x))$$

(From now on we generally omit the subscript x). Then $\bar\beta_i - \beta_i = D_i \log h_x$ for $i = 1, \cdots n$. If we can show that in addition $\bar\beta_0 - \beta_0 = D_0 \log h_x$ then $\bar\beta_0 - \beta_0$ is a gradient vector and, by Remark 3, ii) and iii) will follow immediately. From (1) $\ell = \log h_x$ satisfies

$$\begin{aligned} D_0\ell &= -b^i D_i\ell \,\bar{t}\, \tfrac{1}{2} a^{ij}(D_iD_j\ell + D_i\ell D_j\ell) \\ &= -a^{ij}\,\beta_j(\overline{\beta}_i - \beta_i) - \tfrac{1}{2} a^{ij}(D_i(\overline{\beta}_j - \beta_j) + (\overline{\beta}_i - \beta_i)(\overline{\beta}_j - \beta_j)) \\ &= -\tfrac{1}{2} a^{ij}[\overline{\beta}_i\beta_j - \beta_i\beta_j + D_i(\overline{\beta}_j - \beta_j)) \\ &= (\overline{\beta}_0 - \beta_0) \end{aligned}$$

So $\overline{\beta}_. - \beta_.$ is a gradient on $[0,1] \times \mathbb{R}^n$ and the proof is complete.

<u>The 'if' part</u>. Assume the identities hold for the pairs (a,b) and $(\bar{a}, c_.)$. To show that X and Y have the same reciprocal structure it suffices to show that (1) holds for some X–space–time harmonic function $h_x(t,y)$ with $h_x(0,x) = 1$. Now for each x the vector function $(\overline{\beta}_{x,i}(t,y) - \beta_i(t,y))$ is by ii) and iii) the gradient of some smooth function ϕ on $[0,1] \times \mathbb{R}^n$. Then it follows from a reversal of the derivation above and the definition of $\overline{\beta}_{x,0}$ and β_0 that

$$D_0\,\phi_x + b^i\,D_i\,\phi_x + \tfrac{1}{2} a^{ij}(D_iD_j\phi_x + D_i\phi_x D_j\phi_x) = 0.$$

Set $h_x(t,y) = \exp[\phi_x(t,y) - \phi_x(0,x)]$. Then h_x is a positive smooth X–space–time harmonic function with the properties required.

PROBABILISTIC INTERPRETATIONS

The previous theorem is purely analytical in nature and sheds no light on the probabilistic eaning of local reciprocal characteristics. In this section we give two probabilistic formulae in which they occur naturally. They are formulated for Markov diffusions. Theorem 1 makes it clear how they can be extended to non–Markovian reciprocal diffusions.

Suppose $X = (\{x_t\}, \mathbb{P})$ with diffusion a and drift b is the Markov diffusion satisfying the conditions of Theorem 1. Let $\overline{X} = (\{x_t\}. \bar{\mathbb{P}})$ be the Markov diffusion with diffusion a and zero drift. $p_{0x}(t,y)$. $\bar{p}_{0x}(t,y)$ are the corresponding transition densities. Let $\mathbb{P}_{0x}$. $\mathbb{P}^{1z}_{0x}$. $\bar{\mathbb{P}}_{0x}$ and $\bar{\mathbb{P}}^{1z}_{0x}$

respectively denote narrowly continuous versions of $\mathbb{P}(\cdot \mid x_0 = x)$,

$\mathbb{P}(\cdot \mid x_0 = x, x_1 = z)$, etc. $\beta_0, \beta_1, \cdots \beta_n$ **are defined as before.** The first formula is a factored version of Girsanov's formula for $\mathbb{P}^{1z}_{0x}$.

Theorem 2 There is a random variable S, defined a.s. $\mathbb{P}^{1z}_{0x}$–uniquely for all $x, z \in \mathbb{R}^n$, and a function k continuous on $\mathbb{R}^{2n}$ such that

$$S = \exp [\tfrac{1}{2} \int_0^1 \int_0^1 (D_\mu \beta_\nu(\epsilon t, \epsilon x_t) - D_\nu \beta_\mu(\epsilon t, \epsilon x_t)) \epsilon d\epsilon (x_t^\mu \partial x_t^\nu - x_t^\nu \partial x_t^\mu)$$

and such that for almost all z (Lebesgue).

$$\frac{d\mathbb{P}^{1z}_{0x}}{d\mathbb{P}^{1z}_{0x}} (\omega) = k(x,z)\, S(\omega), \qquad \text{a.s. } \mathbb{P}^{1z}_{0x}$$

where $\int \cdots \partial x_t^\nu$ denotes a Stratonovich integral (convention of Rogers–Williams (1987)), the summation indices μ, ν range over $0, 1, 2, \cdots, n$, and $x_t^0 = t$.

Remark 1 The integrand of the double integral in S is precisely the "exact 2–form" made up of the reciprocal characteristics given by Theorem 1. Notice that S does not depend on x and z.

2) There are many similar expressions; this is just one of the simplest. The double integral is given by a Stokes formula on the fan–shaped surface obtained by subtending the path $\{t, x_t \colon 0 \le t \le 1\}$ at the origin. Any surface stretched between the path and a piecewise smooth curve beginning at x_0 and ending at x_1, and depending only on x_0 and x_1 could also be used.

3) It is highly likely that the phrase "for almost all z" can be dropped, but I have not proved this.

Proof Let G be the Girsanov formula for $d\,\mathbb{P}_{0x}/d\,\mathbb{P}_{0x}$; in terms of the β_i, $i = 1, \cdots, n$, this is

$$G = \exp\,[\int_0^1 \beta_i\, dx_t^i - \frac{1}{2}\int_0^1 \beta_i\, a^{ij}\, \beta_j\, dt$$

($\beta_i = \beta_i(t, x_t)$, etc.,). G and the other stochastic integrals that we introduce can be so chosen that they are defined a.s. $\mathbb{P}^{1z}_{0x}$—uniquely for all $x, z \in \mathbb{R}^n$. Now transform G into Stratonovich form and than expand the resulting line integral by a stochastic form of Stoke's theorem. These are familiar steps in stochastic differential geometry; see for instance Meyer (1982, p. 202) for the former and Bismut (1981, p. 208) for the latter. So

$$\begin{aligned} G &= \exp[\int_0^1 \beta_i\, \partial x_t^i - \frac{1}{2}\int_0^1 (\,a^{ij}\, D_j\, \beta_i + \beta_i\, a^{ij}\, \beta_j)dt] \\ &= \exp \int_0^1 \beta_\mu\, \partial x_t^\mu \end{aligned}$$

Now

$$\int_0^1 \epsilon\beta_\mu\,(\epsilon t,\, \epsilon x_t)\, \partial x_t^\mu \Big|_{\epsilon=1} = \int_0^1\int_0^1 (\beta_\mu(\epsilon t,\, \epsilon x_t) + \epsilon\, D_\nu \beta_\mu(\epsilon t,\, \epsilon x_t) x_t^\nu) d\epsilon\, \partial x_t^\mu$$

Making use of a stochastic Fubini's theorem (Ikeda and Watanabe 1981) we have

$$\begin{aligned} \int_0^1\int_0^1 \beta_\mu\, d\epsilon\, \partial x_t^\mu = \int_0^1 &[\beta_\mu\,(\epsilon,\, \epsilon x_1)\, x_1^\mu - \beta_\mu(0,\, \epsilon x_0) x_0^\mu \\ &- \epsilon \int_0^1 D_\nu\, \beta_\mu(\epsilon t,\, \epsilon x_t)\, x_t^\mu\, \partial x_t^\nu] d\epsilon \end{aligned}$$

and on combining these we find

$$\int_0^1 \beta_\mu (t, x_t)\partial x_t^\mu = \int_0^1 (\beta_\mu(\epsilon, \epsilon x_1)\, x_1^\mu - \beta_\mu(0, \epsilon x_0)x_0^\mu)\, d\epsilon$$
$$+ \int_0^1 \int_0^1 D_\nu \beta_\mu(\epsilon t, \epsilon x_t)\, (x_t^\nu\, \partial x_t^\mu - x_t^\mu\, \partial x_t^\nu)\epsilon d\epsilon$$
$$= I(x_0, x_1) + J \quad (= \log G)$$

On rearranging the double integral we find $S = e^J$. For all $A \epsilon \mathscr{F}$, $B \epsilon \mathscr{B}(\mathbb{R}^n)$, we have the identities

$$\int_B \mathbb{P}_{0x}^{1z}(A)\, p_{0x}(1,z)dz = \mathbb{P}_{0x}(A \cap x_1^{-1} B)$$
$$= \int_A 1_{x^{-1}B}\, G\, d\, \bar{\mathbb{P}}_{0x} = \int_B \int_A G\, d\, \bar{\mathbb{P}}_{0x}^{1z}\, \bar{p}_{0x}(1,z)\, dz$$
$$= \int_B [\int_A G\, d\bar{\mathbb{P}}_{0x}^{1z}\, k_0(x,z)] p_{0x}(1,z)\, dz.$$

where $k_0 = \bar{p}_{0x}(1,z)/p_{0x}(1,z)$. Hence for almost all z, and all $A \epsilon \mathscr{F}$,

$$\mathbb{P}_{0x}^{1z}(A) = \int_A G\, d\bar{\mathbb{P}}_{0x}^{1z}\, k_0(x,z),$$

and so

$$\frac{d\mathbb{P}_{0x}^{1z}}{d\bar{\mathbb{P}}_{0x}^{1z}} = G\, k_0(.\,,z) = S\, e^{I(x_0,\, x_1)} k_0(x,z)$$

Note that $x_0 = x$ and $x_1 = z$ a.s. $\bar{\mathbb{P}}_{0x}^{1z}$. Setting $k(x,z) = e^{I(x,z)}\, k_0(x,z)$ then gives us the result.

Krener's postulate (1) is quite subtly defined in that the conditioning variables slide, as the limit is taken, along geodesics of the Riemannian metric α_{ij}. The following formula can be thought of as an integral variation of (1) conditioned more conventionally on the members of a partially nested family of σ-fields. It is stated without proof for the case

with constant diffusion matrix a. A full treatment of the general case, which requires the introduction of the stochastic differential geometric concepts of stochastic development and parallel translation, will be presented elsewhere.

It is convenient to introduce the following definition. Given the set–up $(\Omega, \mathscr{F}, \{\mathscr{F}_s^t\}, \{\mathscr{E}_s^t\}, \mathbb{P})$ we shall say that a continuous process Z_t is a *free motion* if for all $0 \leq s < t < u \leq 1$

$$E[Z_t - \frac{u - t}{u - s} Z_s - \frac{t - s}{u - s} Z_u \mid \mathscr{E}_s^u] = 0$$

If we set $r = t - h$, $u = t + h$ then we see that this definition essentially says that the mean 'acceleration' of Z_t (conditioned on $\mathscr{E}_{t-h}^{t+h}$) is zero. (cf. martingales; for these the predicted 'velocity' is zero) For example, if X is a Brownian motion, then $Z_t = x_t + k(x_0, x_1)t$ is a free motion. If a free motion Z is adapted to $\{\mathscr{E}_t^1\}$ then it is a continuous semimartingale; I have not explored its other properties. However, in the following particular case the martingale part of the free motion is a Brownian motion. If X is Gaussian Z can be given a more precise description (see Frezza, Krener, Levy 1989).

<u>Theorem 3</u> With the definition of Theorem 1, if $[a^{ij}]$ is a constant ma.rix, then the process

$$Z_t^i = x_t^i - a^{ij} \int_0^t \int_0^s [(D_k \beta_j - D_j \beta_k)\partial x_r^k + (D_0 \beta_j - D_j \beta_0) \, dr] \, ds \quad (2)$$

is a free motion, where $D_k\beta_j = D_k\beta_j(r, x_r)$, etc., and j, k are summed over 1, 2, ···, n. Alternatively, for any $0 \leq s < t < u \leq 1$

$$E[x_t^i - \frac{u - t}{u - s} x_s^i - \frac{t - s}{u - s} x_u^i$$
$$+ a^{ij} \int_s^u \gamma_{su}^t(r) \, [(D_k\beta_j - D_j\beta_k)\partial x_r^k + (D_0\beta_j - D_j\beta_0)dr] \mid \mathscr{E}_s^u] = 0$$

where $$\gamma^t_{su}(r) = \frac{(u - t)(r - s)}{u - s} \quad \text{for } r \leq t$$
$$= \frac{(u - r)(t - s)}{(u - s)} \quad \text{for } r > t.$$

Remarks 1) Following Krener's point of view, we can formally but usefully rewrite the integral equation (2) as a second order stochastic differential equation.

$$\partial^2 x_t = a^{ij}(D_k\beta_j - D_j\beta_k)\, \partial x^k_t\, dt + a^{ij}(D_0\beta_j - D_j\beta_0)\, dt^2 + \partial^2 Z_t \tag{3}$$

where $\partial^2 x_t$ etc. is to be thought of as an infinitesimal central second difference, with the usual proviso that (3) is no more than a mnemonic for (2).

2) A comparison with Krener's postulate (1) suggests that the coefficient of $\partial x dt$ in (3) should be g in (1) and that of dt^2 should be f. This is certainly so for Gaussian X, but for nonGaussian processes the agreement has still to be verified.

REFERENCES

Azencott, R. (1984) Densities des diffusions en temps petit: develepements asymptotiques. *Seminaire de Probabilité XVIII 1982/83, Lect. Notes in Math* 1059 Berlin: Springer–Verlag 402–498.

Bernstein, S. (1932) Sur Les liaisons entre les grandeurs aléatoires. *Verh. des Intern. Mathematikerkongr.* I Zurich.

Bismut J.–M. (1981) *Mechaniqué Aleatoire Lect. Notes in Math.* 866. Berlin: Springer Verlag.

Frezza, R., Krener, A.J., Levy, B., (1989) The self adjointness of second order models of Gaussian reciprocal processes. Preprint.

Jamison, B. (1974) Reciprocal processes *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 30 65–81.

Jamison, B. (1975) The Markov processes of Schrödinger *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 32 323–331.

Krener, A.J. (1988a) Reciprocal diffusions and stochastic differential equations of second order. *Stochastics* 24 393–422.

Krener, A.J. (1988b) unpublished report.

Meyer, P.A. (1982) Geometrie Differentelle Stochastique (bis) *Seminaire de Probabilitẽs XVI 1980/81 Supplement: Geometrie Differentielle Stochastique. Lect. Notes in Math.* 921 Springer Verlag, 165–207.

Molchanov, S.A. (1975) Diffusion processes and Riemannian geometry. *Russ. Math Surveys* 30 (1) 1–63, from *Uspektic Mat. Nank.* 30 (1) 3–59.

Rogers, L.C.G., Williams, D. (1987) *Diffusions, Markov Processes and Martingales, Volume 2. Ito Calculus.* Chichester: John Wiley and Sons.

The Fractional Representations of a Class of Nonlinear Systems

Arthur Krener, Yi Zhu

Department of Mathematics

University of California at Davis

Davis, California 95616, USA

Abstract Right and left coprime fractional representations are shown to exist for a special class of nonlinear systems which have both controller and observer forms. Furthermore, a generalized Bezout identity is given for this class of nonlinear systems.

1.Introduction The purpose of this paper is to obtain the right and the left coprime factorizations of a class of nonlinear systems which have both controller and observer forms and to prove that a generalized Bezout identity holds for this class of nonlinear systems.

The fractional representation method of linear systems can be traced back to the early 1970's when Rosenbrock[16] used transfer matrix to study multi-input/multi-output systems. Since then, this method has been studied intensively by many researchers[4], [7], [11]. Desoer[3] generalized the general concept of coprimeness to a ring. Recently, Hammer [9] used the coprime factorization approach to tackle the discrete nonlinear systems, and there is a growing interest in applying the coprime factorizations approach to study nonlinear systems.

In this paper, we use the state space descriptions as realizations of a nonlinear mapping so that we can work on both state space and input-output mapping just as one does in the linear situation. Our approach is based on the nonlinear normal form theory of Krener[13]. Using the controller form of a nonlinear system, it is possible to design a nonlinear feedback controller. The given system can be right factorized into a composition of a stable postprocessor and an inverse of a stable preprocessor. The right coprimeness concept is based on this right factorization. If we combine the postprocessor and the preprocessor, they form a higher dimensional system. The existence of a stable left inverse of this higher order system is our definition of right coprimeness.

This follows Hammer[9]. In a similar way, we start from the observer form of a nonlinear system used to design a nonlinear asymptotic observer. A modification of this gives a way to left factorize the given system into a composition of an inverse of a stable postprocessor and a stable preprocessor. If we put them together to form a higher dimensional system, the existence of a stable right inverse of this higher order system is our definition of the left coprimeness. We shall give the rigorous definitions later on. It turns out that for those nonlinear systems which have both controller and observer forms the development of right coprime factorizations is a rather straightforward generalization of the linear theory but the development of left coprime factorizations differs with substantially from the linear situation. The nonlinear right and left coprime factorizations satisfy a generalized Bezout identity.

This paper is organized as following. We first talk about two normal forms of nonlinear systems briefly, and then discuss right and left factorizations. Finally, the generalized Bezout identity is given.

2.Normal forms This work is based on nonlinear normal form theory. A complete discussion of nonlinear normal form theory is beyond of the scope of this paper. We only describe those aspects of controller form and observer form which we are going to use in this paper. A more detailed discussion can be found in Krener's paper[13].

Consider the following nonlinear system

(2.1a) $$\dot{\xi} = f(\xi) + g(\xi)u$$

(2.1b) $$y = h(\xi)$$

(2.1c) $$\xi(0) = \xi^0.$$

where f: $R^n \longrightarrow R^n$, g: $R^n \longrightarrow R^{m\times n}$, h: $R^n \longrightarrow R^p$ are all smooth(C^∞) functions. Assume that f(0) = 0 and h(0) = 0. We seek a local change of coordinates $x = x(\xi)$ under which (2.1) has a simpler form in some neighborhood of the norminal $\xi^o = 0$. In this paper we shall restrict our attention to the changes of coordinates whose Jacobian at the $\xi^o = 0$ is the identity

$$\frac{\partial x}{\partial \xi}(0) = I.$$

Such changes of coordinates can be written

(2.2) $$\xi = x + \phi(x),$$

where

$$\phi(0) = 0, \qquad \frac{\partial \phi}{\partial x}(0) = 0.$$

They leave invariant the first order linear approximation to (2.1)

(2.3a) $$\dot{z} = Az + Bu + O(z,u)^2$$

(2.3b) $$y = Cz + O(z)^2$$

(2.3c) $$z(0) = \xi(0)$$

where

(2.4) $$A = \frac{\partial f}{\partial \xi}(0), \quad B = g(0), \quad C = \frac{\partial h}{\partial \xi}(0)$$

Controller form

We say system (2.1) admits a *controller form* if there is a change of coordinates (2.2) and a controllable pair (A,B) such that (1) can be transformed into the following form

(2.5a) $$\dot{x} = Ax + Bu + B\alpha_c(x) + B\beta_c(x)u$$

(2.5b) $$y = Cx + \gamma_c(x)$$

where

$$\alpha_c(x) = O(x)^2, \quad \beta_c(x) = O(x), \quad \gamma_c(x) = O(x)^2,$$

and A and B are constant matrices defined in (2.4). $\alpha_c(x)$, $\beta_c(x)$, and $\gamma_c(x)$ are all arbitrary smooth matrix-valued functions of dimension m×1, m×m, and p×1 respectively. Furthermore the matrix I + $\beta_c(x)$ must be invertible at each of interest.

Controller form is useful in designing nonlinear feedback controllers. If we let

$$\bar{u} = u + \alpha_c(x) + \beta_c(x)u,$$

then (2.5a) becomes a linear system

$$\dot{x} = Ax + B\bar{u}.$$

Since (A,B) is a controllable pair so that we can find a matrix F such that A+BF is stable. If we apply the feedback $\bar{u} = Fx + v$ or equivalently

$$(2.6) \qquad u = [\, I + \beta_c(x)]^{-1}(\, Fx - \alpha_c(x) + v\,),$$

then the closed-loop system is a stable system

$$(2.7) \qquad \dot{x} = (\, A + BF)x\,.$$

The nonlinear feedback control law (2.6) is used to cancel the nonlinearity of the given system. We see from it that the invertibility of $I + \beta_c(x)$ is necessary to guarantee the existence of the nonlinear feedback (6).

Observer form

The system (2.1) is said to admit an *observer form* if there is a change of coordinates (2.2) and an observable pair (C,A) such that (1) can be transformed into the following form

$$(2.8a) \qquad \dot{x} = Ax + Bu + \alpha_0(\bar{y}) + \beta_0(\bar{y})u$$

$$(2.8b) \qquad y = \bar{y} + \gamma_0(\bar{y})$$

$$(2.8c) \qquad \bar{y} = Cx.$$

where

$$\alpha_0(x) = O(x)^2, \quad \beta_0(x) = O(x), \quad \gamma_0(x) = O(x)^2,$$

and A and C are constant matrices defined in (2.4). $\alpha_0(x)$, $\beta_0(x)$, and $\gamma_0(x)$ are smooth matrix-valued functions with dimensions n×1, n×m, and p×1. They are arbitrary except $x + \gamma_0(x)$ must be locally invertible function. Notice that the change of coordinates is typically different from the one in controller form.

Observer form is useful in designing asymptotic observers. Now introduce an observer

$$(2.9) \qquad \dot{\hat{x}} = A\hat{x} + \alpha_0(\bar{y}) + \beta_0(\bar{y})u - H(\bar{y} - C\hat{x}).$$

If we denote $\tilde{x} = x - \hat{x}$, then the error satisfies

$$(2.10) \qquad \dot{\tilde{x}} = (\, A + HC)\tilde{x}\,.$$

Since the pair (A,C) is observable, we can choose a matrix H such that A+HC is a stable matrix, and then (9) is an asymptotic observer of (8).

If a nonlinear system can be transformed into controller form and observer form (typically in two different coordinate systems), then we can design an observer-based controller just like we do for linear systems.

The question here to ask is that under what conditions the nonlinear system (2.1) can be transformed into controller form (5) and observer form (8), and for what conditions the pairs (A,B) and (C,A) of the linearization are controllable and observable pairs respectively. The answer is the nonlinear system (1) must be controllable and observable in the nonlinear sense and satisfy certain additional conditions. Again readers who are interested in the details should refer to[13].

3.Right fractional description Suppose we have a nonlinear system in controller normal form

(3.1a) $\dot{x}_c = Ax_c + Bu + B(\alpha_c(x_c) + \beta_c(x_c)u)$

(3.1b) $y = Cx_c + \gamma_c(x_c)$

(3.1c) $x_c(0) = x_c^0$

where the c-subscripts indicate coordinates and functions associated to controller normal form. We view (3.1) as defining an input/output map

(3.2a) $G\colon\ u(\cdot) \mapsto y(\cdot)$

from functions u(t) to y(t) for $t \geq 0$. We seek a right factorization of G

$$G = N \circ M^{-1}$$

where N and M are input/output maps

(3.2b) $M\colon v(\cdot) \mapsto u(\cdot)$

(3.2c) $N\colon v(\cdot) \mapsto y(\cdot)$,

M is invertible and $\circ$ denotes composition. Among others, Khargonekar and Sontag[12], Doyle[5] Francis[6,7], Sontag[12] have treated such factorization of linear systems and Hammer[9] has discussed similar ideas for nonlinear discrete systems. We shall follow these authors, particularly [5,7].

To describe the input/output maps **M** and **N** we shall use a state space realization. In particular we define **M** to be the input/output map of

(3.3a) $$\dot{\xi}_c = (A + BF)\,\xi_c + Bv$$

(3.3b) $$\alpha_c(\xi_c) + (I + \beta_c(\xi_c))u = F\xi_c + v$$

(3.3c) $$\xi_c(0) = 0$$

where (3.3b) defines u as a function of ξ_c and v.

We consider the decomposition **N** = **G**∘**M**, which is realized by the 2n dimensional system (3.1,3.3) described in ξ_c, x_c coordinates. Let $e = x_c - \xi_c$ then

(3.4) $$\dot{e} = Ae + B\big(-F\xi_c - v + \alpha_c(x_c) + (I + \beta_c(x_c))(I + \beta_c(\xi_c))^{-1}(F\xi_c - \alpha_c(\xi_c) + v)\big)$$

If e(t) = 0 then $\dot{e}(t) = 0$. Since e(0) = 0 we conclude that e(t) = 0 for t ≥ 0. In other words, the realization ( ) of **N** is not controllable because e(t) is unaffected by the input v(t).

A controllable realization of **N** is

(3.5a) $$\dot{\zeta}_c = (A + BF)\zeta_c + Bv$$

(3.5b) $$y = C\zeta_c + \gamma_c(\zeta_c).$$

(3.5c) $$\zeta_c(0) = 0$$

Hence we conclude that **G** = **N**∘**M**$^{-1}$ where **N** and **M** are realized by (3.5) and (3.3). Notice that **M** is invertible since $(I + \beta_c)$ is invertible by assumption.

Notice also that if (A,B) is a controllable pair then we can choose F so that (3.3) and (3.5) are stable systems. Hence we have factored **G** over the ring of stable nonlinear systems. We are being deliberately vague about the precise definition of a stable nonlinear system. It is clear that (3.3,3.5) are "stable" under any reasonable definition.

Of course, we are interested in coprime factorizations over the ring of stable nonlinear systems. Again we shall not try to make this concept precise but following Hammer[9] and others we shall say that **G** = **N**∘**M**$^{-1}$ is a coprime factorization if there exists $\tilde{\mathbf{P}}$, the input/outpuı map of a stable system,

(3.6a) $$\tilde{\mathbf{P}}: \begin{bmatrix} u \\ y \end{bmatrix} \mapsto w$$

such that the composition

(3.6b) $$\tilde{P}\circ\begin{bmatrix}M\\N\end{bmatrix} : v \mapsto \begin{bmatrix}u\\y\end{bmatrix} \mapsto w$$

is the identity, w = v.

The input/output map of $\begin{bmatrix}M\\N\end{bmatrix}$ can be realized by an n dimensional system

(3.7a) $$\dot{\xi}_c = (A + BF)\,\xi_c + Bv$$

(3.7b) $$\alpha_c(\xi_c) + (I + \beta_c(\xi_c))u = F\xi_c + v$$

(3.7c) $$y = C\xi_c + \gamma_c(\xi_c)$$

(3.7d) $$\xi_c(0) = 0.$$

One left inverse of (3.7) is realized by

(3.8a) $$\dot{z}_c = Az_c + Bu + B(\alpha_c(z_c) + \beta_c(z_c)u)$$

(3.8b) $$w = \alpha_c(z_c) + (I + \beta_c(z_c))u - Fz_c$$

(3.8c) $$z(0) = 0.$$

If $e = \xi_c - z_c$ then

$$\dot{e} = Ae + B\big(\alpha_c(\zeta_c) - \alpha_c(x_c) + (\beta_c(\zeta_c) - \beta_c(x_c))u\big)$$

If e(t) = 0 then $\dot{e}(t) = 0$ and since e(0) = 0 it follows that e(t) =0 for all $t \geq 0$. If $e(t) = \xi_c(t) - z_c(t)$ = 0 then w(t) = v(t) for all $t \geq 0$ so (3.8) is inverts (3.7).

However we do not know that (3.8) is stable. To insure the stability of (3.8), we must add to (3.8a) an extra term. This term must stabilize (3.8) and must be zero when $\xi_c = z_c$ so that (3.8) remains a left inverse of (3.7). How do we find such a term?

Notice that the dynamics (3.8a) is the same as the dynamics of the original system (3.1a) and notice that the other y of (3.7) does not appear in (3.8). Perhaps we can inject y into (3.8a) to stabilize it? This is more or less equivalent to asking whether output injection can be used to stabilize the original system (3.1). This is always possible for systems in observer form, hence we assume that there exists a change of coordinates

(3.9) $$x_c = x_o + \phi_{co}(x_o)$$

satisfying (2.2) transforming (3.1) into the observer form

(3.10a) $\dot{x}_o = Ax_o + Bu + \alpha_o(Cx_o) + \beta_o(Cx_o)u$

(3.10b) $y = Cx_o + \gamma_o(Cx_o)$

(3.10c) $x_o(0) = x_o^{\,0},$

Suppose we consider a similar change of coordinates for (3.8)

(3.11) $z_c = z_o + \phi_{co}(z_o)$

to obtain

(3.12a) $\dot{z}_o = Az_o + Bu + \alpha_o(Cz_o) + \beta_o(Cz_o)u$

(3.12b) $w = \alpha_c(z_o + \phi_{co}(z_o)) + (I + \beta_c(z_o + \phi_{co}(z_o)))u - F(z_o + \phi_{co}(z_o))$

(3.12c) $z_o(0) = 0$

We add to (3.12a) the term

(3.13a) $\alpha_o(\bar{y}) - \alpha_o(Cz_o) + (\beta_o(\bar{y}) - \beta_o(Cz_o))u + H(Cz_o - y)$

on the right hand side of (3.14a) to obtain

(3.12aa) $\dot{z}_o = (A + HC)z_o + Bu + \alpha_o(\bar{y}) + \beta_o(\bar{y})u - H\bar{y}$

where $\bar{y}$ is a function of y of (3.7c) defined by

(3.13b) $y = \bar{y} + \gamma_o(\bar{y}) = C\xi_o + \gamma_o(C\xi_o)$

and ξ_o is the state of (3.7) in observer coordinates

(3.13c) $\xi_c = \xi_o + \phi(\xi_o)$

Notice that (3.13a) is zero whenever $\xi_o = z_o$, hence the input/output map $\widetilde{\mathbf{P}}$ of the (3.12aa, b, c) is also an inverse of (3.7). Also, if (C,A) is an observable pair then we can choose H so that (3.12aa, b, c) is stable.

Now we summarize the analysis into the following right factorization theorem.

Theorem If the nonlinear system (2.1) admits controller form (3.1) and observer form (3.10), then there exist stable mappings **M**: $v \longmapsto u$ and **N**: $v \longmapsto y$, where **M** is invertible, such that the input/output mapping **G** defined by the system can be factorized as $\mathbf{G}=\mathbf{N}\circ\mathbf{M}^{-1}$. And furthermore, **M** and **N** are right coprime, that is, the mapping $\begin{bmatrix}\mathbf{M}\\ \mathbf{N}\end{bmatrix}: v \longmapsto \begin{bmatrix}u\\ y\end{bmatrix}$ has a stable left inverse $\widetilde{\mathbf{P}}$: $\begin{bmatrix}u\\ y\end{bmatrix} \longmapsto v$ such that the composition $\widetilde{\mathbf{P}}\circ\begin{bmatrix}\mathbf{M}\\ \mathbf{N}\end{bmatrix}$ is identity mapping.

4.Left fractional description In linear system theory, observability is dual to controllability. Unfortunately the analogy is not so straightforward for coprime factorizations. The theory of right coprime factorization of nonlinear systems is very similar to the theory for linear systems but theory of left coprime factorizations for nonlinear systems has some substantial differences with the linear theory.

We start with a system in observer form (3.10) realizing an input/output map G. We define another input output map

(4.1) $$\tilde{M}:\begin{bmatrix}u\\y\end{bmatrix}\longmapsto w$$

by

(4.2a) $$\dot{\xi}_o = (A + HC)\xi_o - H\bar{y} + \alpha_o(\bar{y}) + \beta_o(\bar{y})u$$

where $\bar{y}$ is an invertible function of the input y defined by

(4.2b) $$y = \bar{y} + \gamma_o(\bar{y})$$

and the output is

(4.2c) $$w = -C\xi_o + \bar{y}$$

(4.2d) $$\xi_o(0) = 0$$

Consider the serial connection of (3.10) and (4.2), this is not a realization of the $\tilde{M}\circ G$ but it is a realization of $\tilde{N} = \tilde{M}\circ\begin{bmatrix}I\\G\end{bmatrix}$. (This is the first important difference with the linear theory). If we define $\zeta_o = x_o - \xi_o$, then $\tilde{N}$ is realized by

(4.3a) $$\dot{\zeta}_o = (A + HC)\zeta_o + Bu$$

(4.3b) $$w = C\zeta_o$$

(4.3c) $$\zeta_o(0) = 0$$

because in ξ_o, x_o coordinates for (3.10, 4.2) only the ξ_o are observable from the output w. We consider $\tilde{N}$, $\tilde{M}$ as a left factorization of G, although it is really a left factorization of $\begin{bmatrix}I\\G\end{bmatrix}$ in the sense that

(4.4) $$\tilde{M}\circ\begin{bmatrix}I\\G\end{bmatrix} = \tilde{N}$$

Notice that we cannot compose this on the left with $\tilde{M}^{-1}$ since $\tilde{M}$ is not invertible as a mapping from $\begin{bmatrix} u \\ y \end{bmatrix}$ to w.

Perhaps the best way of viewing the situation is

(4.5a) $$\begin{bmatrix} I & 0 \\ 0 & \tilde{M} \end{bmatrix} \circ \begin{bmatrix} I \\ G \end{bmatrix} = \begin{bmatrix} I \\ \tilde{N} \end{bmatrix}$$

or

(4.5b) $$\begin{bmatrix} I \\ G \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \tilde{M} \end{bmatrix}^{-1} \circ \begin{bmatrix} I \\ \tilde{N} \end{bmatrix}$$

The matrix notation is somewhat misleading because $\tilde{M}$ depends on both u and y.

In any case, if (C,A) is an observable pair then (4.2) and (4.3) can be made stable by proper choice of H. In particular, the nonlinearities in (4.2) are momeryless functions of the inputs u and y hence (4.2) is BIBO stable.

Next we address the input/output map

(4.6a) $$\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix} : \begin{bmatrix} u \\ y \end{bmatrix} \longmapsto w$$

where again the matrix notation is somewhat misleading since both u and y are inputs to $\tilde{M}$, i.e.,

(4.6b) $$w = -\tilde{N}(u) + \tilde{M} \begin{bmatrix} u \\ y \end{bmatrix}.$$

This input/output map can be realized by an n dimensional system

(4.7a) $$\dot{\xi}_o = (A + HC)\xi_o + \alpha_o(\bar{y}) + Bu + \beta_o(\bar{y})u - H\bar{y}$$

where $\bar{y}$ is an invertible function of the input y defined by

(4.7b) $$y = \bar{y} + \gamma_o(\bar{y})$$

and the output w is given by

(4.7c) $$w = -C\xi_o + \bar{y}.$$

We wish to find an input/output map **P** realized by a stable system so that **P** is a right invertible of $\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix}$,

(4.8a) $$\mathbf{P} : v \longmapsto \begin{bmatrix} u \\ y \end{bmatrix}$$

(4.8b) $$\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix} \circ \mathbf{P} : v \mapsto w = v.$$

We start by constructing an inverse for (4.7), (4.9a) $$\dot{z}_o = Az_o - Hv + Bu + \alpha_o(\bar{y}) + \beta_o(\bar{y})u$$

(4.9b) $$\bar{y} = Cz_o + v$$

(4.9c) $$y = \bar{y} + \gamma_o(\bar{y})$$

(4.9d) $$u = ?$$

(4.9e) $$z_o(0) = 0$$

We leave the u, which also appears in the dynamics (4.9a),unspecified for the moment. Notice that if $e = \xi_o - z_o$ is the error between the states of (4.7) and (4.22) then $\dot{e} = 0$ whenever $e = 0$. Since $e(0) = 0$ we conclude that $e(t) = 0$ for all $t \geq 0$ and so by (4.7c) and (4.9b) we have $w(t) = v(t)$. In other words (4.9) is a right inverse of (4.7).

What about stability of (4.9)? We would like to choose the output u in such a way that (4.9a) is stable in some sense. If we ignore the $-Hv$ term of (4.9a) this looks like the original system is in observer form. This is not exactly true because $\bar{y}$ is defined by (4.9b) by a change of coordinates (3.9). If we apply a similar change of coordinates (3.11) to (4.9) we obtain

(4.10a) $$\dot{z}_c = Az_c + Bu + B(\alpha_c(z_c) + \beta_c(z_c)u) - (I + \frac{\partial\phi_{co}}{\partial z_o}(z_o))(Hv) + \\ (I + \frac{\partial\phi_{co}}{\partial z_o}(z_o))\{[\alpha_o(Cz_o + v) - \alpha_o(Cz_o)] + [\beta_o(Cz_o + v) - \beta_o(Cz_o)]u\}$$

Suppose we choose an F such that (A + BF) is stable and define u by

(4.9dd) $$\alpha_c(z_c) + (I + \beta_c(z_c))u = Fz_c .$$

The u in (4.9d) is chosen in (4.9dd).

When the input v = 0, (4.10a) becomes

(4.10b) $$\dot{z}_c = (A + BF)z_c .$$

We view (4.10a),(4.10b),(4.9b) and (4.9c) as a realization of $\mathbf{P} : v \longmapsto \begin{bmatrix} u \\ y \end{bmatrix}$, then $\mathbf{P}$ is a right inverse of $\begin{bmatrix} -\tilde{\mathbf{N}}, \tilde{\mathbf{M}} \end{bmatrix}$.

We summarize the above analysis as the following theorem.

Theorem If nonlinear system (2.1) admits observer and controller forms, then there exist stable mappings $\tilde{\mathbf{M}}: \begin{bmatrix} u \\ y \end{bmatrix} \longmapsto v$ and $\tilde{\mathbf{N}}: u \longmapsto v$ such that $\tilde{\mathbf{M}} \circ \begin{bmatrix} I \\ G \end{bmatrix} = \tilde{\mathbf{N}}$. Furthermore $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{N}}$ are left coprime under the sense that the left combined mapping $\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix} : \begin{bmatrix} u \\ y \end{bmatrix} \longmapsto v$ has a right stable inverse P: $v \longmapsto \begin{bmatrix} u \\ y \end{bmatrix}$ such that the composition $\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix} \circ P$ is identity .

5.Bezout identity We conclude by noting that a "nonlinear Bezout identity" holds for the above. In other words beside $\tilde{P}$ being a left inverse (3.6b) for $\begin{bmatrix} M \\ N \end{bmatrix}$ and P a right inverse (4.9b) for $\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix}$, it is also true that

(5.1) $$\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix} \circ \begin{bmatrix} M \\ N \end{bmatrix} : v \mapsto \begin{bmatrix} u \\ y \end{bmatrix} \mapsto w = 0$$

and

(5.2) $$\tilde{P} \circ P : v \mapsto w = 0.$$

We summarize these equations by (5.3) in the following theorem.

Theorem (nonlinear Bezout identity) The systems defined in last two theorems $\begin{bmatrix} M \\ N \end{bmatrix}$, $\tilde{P}$, $\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix}$, and P satisfy the following Bezout identity:

(5.3) $$\begin{bmatrix} \tilde{P} \\ -\tilde{N}, \tilde{M} \end{bmatrix} \circ \begin{bmatrix} M & P \\ N & \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

where $\begin{bmatrix} M \\ N \end{bmatrix}$, $\tilde{P}$, $\begin{bmatrix} -\tilde{N}, \tilde{M} \end{bmatrix}$, P have realizations as discussed.

6.Conclusion We have briefly described an approach to nonlinear factorizations based on nonlinear normal forms. The research is just the beginning. Many concepts are not rigorously defined. We hope that this attempt will give the nonlinear system study a push in this direction.

References

[1] P. Antsaklis, Some Relations Satisfied by Prime Polynomial Matrices and Their Role in Linear Multivariable System Theory,IEEE Trans. on Aut. Control, Vol.AC-24, NO.4, August 1979.

[2] C. Desoer and M. Kabuli, Right Factorization of a Class of Time Varying Nonlinear Systems, IEEE Trans. on Aut. Control, Vol.33, NO.8 pp755-757, August 1988.

[3] C. Desoer, R.W. Liu, J. Murray and R. Saeks, Feedback System Design: the Fractional Representation Approach, IEEE Trans. Automat. Control, AC-25 (1980), pp.399-412.

[4] C. Desoer and M. Vidyasagar, Feedback Systems: Input-Output Properties, Academic Press, 1975

[5] J. Doyle, Lecture Notes in Advances in Multivariable Control, Office of Naval Research/ Honeywell Workshop, Minneapolis, MN, 1984.

[6] B. Francis, A Course in H_∞ Control Theory, Lecture Notes in Control and Information Science, Vol. 88, Springer-Verlag, 1987.

[7] B. Francis and J. Doyle, Linear Control Theory with an H_∞ Optimality Criterion, SIAM J. Control and Optimization, Vol 25, No.4,July 1987, pp. 815-844.

[8] J. Hammer, Nonlinear Systems, Stabilization and Coprimeness, Int.J.Control,42,pp1-20, 1985

[9] J. Hammer, Fraction Representations of Nonlinear Systems: A Simplified Approach, Int. J.of control, Vol.46, No.2, August 1987, pp. 455-472.

[10] A. Isidori, Nonlinear Control Systems: An Introduction, Lecture Notes in Control and Information Sciences, Vol.72, Springer-Verlag, Berlin, 1985.

[11] T. Kailath, Linear Systems, Prentice Hall, Englewood Cliffs, 1980.

[12] P. Khargonekar and E. Sontag, On the Relation between Stable Matrix Factorizations and Regulable Realizations of Linear Systems over Rings.

[13] A. Krener, Normal Forms for Linear and Nonlinear Systems, Contemporary mathematics, Vol.68, 1987

[14] A. Krener and W. Respondek, Nonlinear Observers with Linearizable error dynamics,SIAM J. Control and Optimization, Vol.23, No.2, March 1985.

[15] A. Krener, Nonlinear Controller Design via Approximate Normal Forms, SIAM Anual Conf. Minneapolis, 1988

[16] T.T.Tay and J.B.Moore, Left Coprime Factorizations and a class of Stabilizing controllers for Nonlinear Systems, IEEE Decision and Control 1988, pp449-454

[17] H. Rosenbrock, State space and multivariable theory, Nelson, London, 1970.

[18] M.S.Verma, Coprime Factorizations of Nonlinear and Time-varying Systems, IEEE Decision and Control 1988, pp444-448

[19] M. Vidyasagar, Control System Synthesis: A Factorization Approach, MIT Press, 1985.

[20] D. Youla, H.Jabr and J.Bongiorno,Jr., Modern Weiner-Hopf Design of Optimal Controllers: Part II, IEEE Trans. Aut. Control, AC21, pp.319-338,1976.

[21] Y. Zhu, MS Thesis, App. Math, UC Davis, 1988

OBSERVATION OF A RIGID BODY FROM MEASUREMENT OF A PRINCIPLE AXIS*

Wei Kang and Arthur J. Krener

Department of Mathematics
University of California
Davis, CA95616

ABSTRACT

The spacecraft attitude control has been studied before, see [3]. In this note, we give a method of observing the attitude of a freely rotating spacecraft by measuring one of its principal axis. This problem is solved in §2 and §3. In §4, a method of determining the angular velocity by the trajectory of one of its coordinates is given.

1. INTRODUCTION

In the following, we consider a freely rotating rigid body with no externel torques acting on it. Let $\{e_1,e_2,e_3\}$ be a set of orthonormal axis fixed in the spacecraft, with the origin at the center of mass and each axis parallel to one of the principal axis. A second frame $\{r_1,r_2,r_3\}$ is an inertially fixed basis. In page 445 of [2], Symon describes the motion of a freely rotating body as follows. Fix an ellipsoid on the spacecraft, which can be represented as

$$\{\sum_{i=1}^{3} x_i e_i \mid \sum_{i=1}^{3} I_i x_i^2 = 1\}$$

It is called the inertia ellipsoid. There is a fixed plane, P, which is called the invariant plane. As the spacecraft is rotating freely, one can imagine that the inertia ellipsoid is fixed on the spacecraft and it is rolling on the invariant plane without slipping and its center is fixed at the origin, see figure 1.

Suppose $\mathbf{r}$ is the vector from the origin to the point of contact between the ellipsoid and the invariant plane. Then, the angular velocity ω satisfies

$$\omega = b\mathbf{r}$$

where b is a constant.The following equations describe the evolution of the spacecraft's attitude

$$I_1\dot\omega_1 + (I_3 - I_2)\omega_2\omega_3 = 0$$

$$I_2\dot\omega_2 + (I_1 - I_3)\omega_1\omega_3 = 0$$

$$I_3\dot\omega_3 + (I_2 - I_1)\omega_1\omega_2 = 0$$

$$\dot\phi = \frac{\sin\psi}{\sin\theta}\omega_1 + \frac{\cos\psi}{\sin\theta}\omega_2 \qquad (1)$$

$$\dot\psi = \frac{\sin\psi\cos\theta}{\sin\theta}\omega_1 - \frac{\cos\psi\cos\theta}{\sin\theta}\omega_2 + \omega_3$$

$$\dot\theta = \cos\psi\omega_1 - \sin\psi\omega_2$$

where (ϕ, ψ, θ) are Euler angles (see [3]),and

$$\omega = \sum_{i=1}^{3} \omega_i e_i$$

is the angular velocity. The inertia tensor is

$$I = \begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix}$$

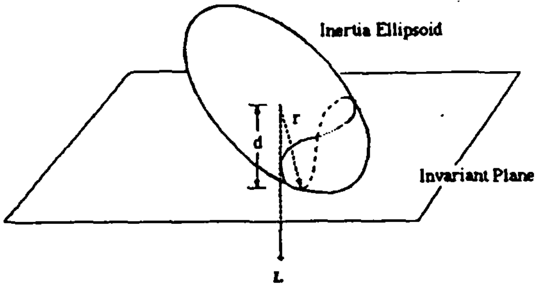


Figure 1. The inertia ellipsoid rolls on the invariant plane

2. OBSERVABILITY

Suppose the output of system (1) is the position of the third principal axis, $e_3(t)$, in inertial coordinates. We consider the observability of the model, i.e, whether the complete motion can be determined from the time history of $e_3(t)$ in inertial coordinates and the equations of motion (1). Rewrite the system as follows:

$$\dot\omega_1 = \alpha\omega_2\omega_3$$

$$\dot\omega_2 = \beta\omega_1\omega_3 \qquad (2\text{-}1)$$

$$\dot\omega_3 = \gamma\omega_1\omega_2$$

$$\begin{bmatrix} \dot\phi \\ \dot\psi \\ \dot\theta \end{bmatrix} = \begin{bmatrix} \frac{\sin\psi}{\sin\theta} & \frac{\cos\psi}{\sin\theta} & 0 \\ -\sin\psi\,\mathrm{ctg}\theta & -\cos\psi\,\mathrm{ctg}\theta & 1 \\ \cos\psi & -\sin\psi & 0 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} \qquad (2\text{-}2)$$

The inertial coordinates of $e_3(t)$ are the observation and are given by

$$y = \begin{bmatrix} \sqrt{1 - \cos^2\theta}\sin\phi \\ -\sqrt{1 - \cos^2\theta}\cos\phi \\ \cos\theta \end{bmatrix} = \begin{bmatrix} h_1(\theta,\phi) \\ h_2(\theta,\phi) \\ h3(\theta,\phi) \end{bmatrix} \qquad (2\text{-}3)$$

Here

$$\alpha = \frac{I_2 - I_3}{I_1}, \quad \beta = \frac{I_3 - I_1}{I_2}, \quad \gamma = \frac{I_1 - I_2}{I_3}$$

Theorem 1.The system (2) is observable iff $\alpha+\beta \neq 0$ and $(\beta+1)\omega_1^2 + (\alpha-1)\omega_2^2 \neq 0$.

* Research supported in part by AFOSR 85-0267

Proof: To prove this system is observable, using the method of [1], we need to find the dimension of the distribution $\mathcal{C}(r)$ generated by $\{dy\ , L_F dy, ..., L_F^{r-1}dy\}$. Because

$$h_3 = \pm \sqrt{1 - h_1^2 + h_2^2}$$

the dimension of $\mathcal{C}(r)$ can be determined by:

$$dh_1, L_F dh_1, ..., L_F^{r-1}dh_1$$

$$dh_2, L_F dh_2, ..., L_F^{r-1}dh_2$$

But $L_F^k dh_i$, i=1,2, are complicated. To determine the dimension, we make a change of coordinates in the output space. Let

$$A = \begin{bmatrix} \frac{\partial h_1}{\partial \phi} & \frac{\partial h_1}{\partial \theta} \\ \frac{\partial h_2}{\partial \phi} & \frac{\partial h_2}{\partial \theta} \end{bmatrix}$$

Then, det(A) = $\cos\theta\sin\theta$; and A is nonsingular whenever $\theta \neq \frac{k\pi}{2}$ .The angle θ depends on the choice of the inertially fixed basis. We can chose suitable basis to avoid the case $\theta = \frac{k\pi}{2}$ in a local neighborhood.Therefore, by change of coordinates in the output space we can take θ and ϕ the output..The observability of (2) is equivalent to the observability with respect to the output

$$y_1 = \begin{bmatrix} \theta \\ \phi \end{bmatrix}$$

By calculation, we know that

$$L_F\theta = \omega_1\cos\psi - \omega_2\sin\psi$$

$$L_F\phi = (\omega_1\sin\psi + \omega_2\cos\psi)\frac{1}{\sin\theta}$$

$$L_F^2\theta = \sin\theta\cos\theta(L_F\phi)^2 - \omega_3\sin\theta(L_F\phi)$$

$$+ \alpha\omega_2\omega_3\cos\psi - \beta\omega_1\omega_3\sin\psi$$

$$L_F^2\phi = \text{ctg}\theta(L_F\phi)(L_F\theta) + \frac{\cos\theta}{\sin\theta}(L_F\theta L_F\phi)$$

$$+ \omega_3\frac{L_F\theta}{\sin\theta} + \frac{\sin\psi}{\sin\theta}\alpha\omega_2\omega_3 + \frac{\cos\psi}{\sin\theta}\beta\omega_1\omega_3$$

In $dL_F^2\phi$ and $dL_F^2\theta$, all the terms containing $d\theta$, $d\phi$, $dL_F\phi$ or $dL_F\theta$ can be cancelled, the dimension of $\mathcal{C}(3)$ is the same as the rank of the following matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\omega_1 S\psi - \omega_2 C\psi & -C\psi & -S\psi & 0 \\ 0 & 0 & \omega_1 C\psi - \omega_2 S\psi & S\psi & C\psi & 0 \\ 0 & 0 & -\alpha\omega_2\omega_3 S\psi - \beta\omega_1\omega_3 C\psi & -\beta\omega_3 S\psi & \alpha\omega_3 C\psi & (\alpha-1)\omega_2 C\psi - (\beta+1)\omega_1 S\psi \\ 0 & 0 & \alpha\omega_2\omega_3 C\psi - \beta\omega_1\omega_3 S\psi & \beta\omega_3 C\psi & \alpha\omega_3 S\psi & (\alpha-1)\omega_2 S\psi + (\beta+1)\omega_1 C\psi \end{bmatrix}$$

where $C\psi = \cos\psi$ and $S\psi = \sin\psi$.

The determinant of this matrix is

$$[(\beta + 1)\omega_1^2 + (\alpha - 1)\omega_2^2](\alpha + \beta)$$

Therefore, this is not zero iff $\mathcal{C}(3)$ has full dimension.

Remark 1: $(\beta + 1)\omega_1^2 + (\alpha - 1)\omega_2^2 \neq 0$ implies that $\omega_1$and ω_2 can not be zero at the same time. If $\omega_1 = \omega_2 = 0$, then the spacecraft turns around e_3, the output is a constant vector and it is impossible to determine e_1, e_2 from the trajectory of $e_3(t)$.

Remark 2: The condition $\alpha + \beta \neq 0$ implies $I_1 \neq I_2$. If $I_1 = I_2$, then the spacecraft is symmetric with respect to e_3, We can not tell the difference between e_1 and e_2. Moreover, e_1 and e_2 are not uniquely defined. So it is impossible to determine the position of e_1, e_2.

3. ATTITUDE DETERMINATION

In this section, we assume that $I_1 > I_2 > I_3$ and $\omega_3 \neq 0$. If $\omega_3 = 0$, then $\omega_1 \neq 0$ (see [3]). So, from the similar method in this section, we can determine the attitude by measuring e_1.

From [3]and the introduction, we know that the motion of the spacecraft is totally determined by the following three constants.

(1) The direction of L, which is the normal vector of the invariant plane P.

(2) The distance, d, from the origin to the tangent plane P.

(3) The energy T.

As the inertia ellipsoid rolls on P, the vector e_3 turns around the axis L. Suppose that the coordinate of $e_3(t)$ in the inertially fixed basis is (x(t), y(t), z(t)), it is expressed by Euler angles in (2). Imagine that the curve described by $e_3(t)$ has mass with the density a constant 1. Then L is a vector passing through the center of this mass. So the coordinates of L in the inertially fixed basis are

$$x_L = \frac{\int_0^{s_0} x(s)ds}{s_0},\ y_L = \frac{\int_0^{s_0} y(s)ds}{s_0},\ z_L = \frac{\int_0^{s_0} z(s)ds}{s_0} \qquad (3)$$

Where s is the length of the curve described by $e_3(t)$ at time t. The number s_0 is the length of the smallest closed curve described by $e_3(t)$ if $e_3(t)$ moves in periodic, If it is not periodic, we must take the limit of these integrals as the length s_0 goes to ∞.

To find d, we consider $L \cdot e_3$. Let's take the inertia ellipsoid as

$$I_1 r_1^2 + I_2 r_2^2 + I_3 r_3^2 = 1 \qquad (4)$$

So, the vector L is parallel to

$$I_1 r_1 e_1 + I_2 r_2 e_2 + I_3 r_3 e_3$$

where $r = \sum_{i=1}^{3} r_i e_i$ is the vector from the origin to the point of contact between the ellipsoid and the invariant plane. Therefore

$$L = \frac{\sum_{i=1}^{3} I_i r_i e_i}{\sqrt{\sum_{i=1}^{3} I_i^2 r_i^2}} \qquad (5)$$

and

$$d = L \cdot r = \frac{I_1 r_1^2 + I_2 r_2^2 + I_3 r_3^2}{\sqrt{\sum_{i=1}^{3} I_i^2 r_i^2}} = \frac{1}{\sqrt{\sum_{i=1}^{3} I_i^2 r_i^2}}$$

So

$$e_i \cdot L = \frac{I_i r_i}{\sqrt{\sum_{i=1}^{3} I_i^2 r_i^2}} = r_i I_i d \qquad (6)$$

The function $|e_3 \cdot L|$ has its maximum value iff $|r_3|$ has its maximum value.From the first three equations of system (2), we can easily prove

$$\frac{\omega_2^2}{\beta} - \frac{\omega_3^2}{\gamma} = \text{constant}$$

So $(\omega_2(t),\omega_3(t))$ describes an ellipse. In [2], it was proved that $\mathbf{r} = b\omega$, therefore (r_2,r_3) *is also on an ellipse. The function* $|r_3(t)|$ has the maximum value implies r_2=0. The equation (6) implies that $|r_3(t)|$ has its maximum value iff $|L \cdot e_3|$ has its maximum value. Denote this maximum value of $|L \cdot e_3|$ by A. So $|L \cdot e_3|=A$ implies r_2=0. From (4) and (5), we obtain

$$I_1 r_1^2 + I_3 r_3^2 = 1 \qquad (7)$$

$$L = \frac{I_1 r_1 e_1 + I_3 r_3 e_3}{\sqrt{I_1^2 r_1^2 + I_3^2 r_3^2}} \qquad (8)$$

so, L is in the e_1, e_3 plane.

$$L \cdot e_1 = \pm\sqrt{1 - (Le_3)^2} = \pm\sqrt{1 - A^2} \qquad (9)$$

The equation (6) and (9) imply

$$r_1 I_1 d = \pm\sqrt{1-A^2}$$

$$r_3 I_3 d = A$$

Therefore

$$r_1 = \pm \frac{\sqrt{1-A^2}}{I_1 d}, \quad r_2 = \frac{A}{I_3 d} \qquad (10)$$

Substitute (10) to (7), we have

$$\frac{1 - A^2}{I_1 d^2} + \frac{A^2}{I_2 d^2} = 1$$

From this, we obtain

$$d = \sqrt{\frac{1 - A^2}{I_1} + \frac{A^2}{I_3}} \qquad (11)$$

Now, we try to determine the energy T by the frequency of e_3. Suppose

$$X = (x_1(t), x_2(t)\, x_3(t))$$

is the solution of

$$\dot{x}_1 = \alpha x_2 x_3$$
$$\dot{x}_2 = \beta x_1 x_3$$
$$\dot{x}_3 = \gamma x_1 x_2$$

such that the initial condition is on the inertia ellipsoid and

$$d = \frac{1}{\sqrt{\sum_{i=1}^{3} x_i^2 I_i^2}}$$

Its period is a_0. Consider

$$\omega_i = \lambda x_i(\lambda t)$$

It can be proved that ω_1, ω_2 and ω_3 satisfy the first three equations in (2).The energy

$$T = \frac{1}{2}\,\omega I \omega = \frac{\lambda^2}{2}(I_1 x_1^2 + I_2 x_2^2 + I_3 x_3^2) = \frac{\lambda^2}{2}$$

The period if ω is $\frac{a_0}{\lambda}$. Suppose the period of ω_3 is a, then

$$2a = \frac{a_0}{\lambda} = \frac{a_0}{\sqrt{2T}}$$

So

$$T = \frac{a_0^2}{8a^2} \qquad (12)$$

Here, a is the period of ω_3, which is unknown. But we proved that

$$e_3 \cdot L = r_3 I_3 d$$

$$r_3 = b\omega_3$$

where b is some constant. So, a is also the period of $e_3 \cdot L$.

Therefore, we can use (3) to determine L, (11) to determine d and (12) to determine T. From the proof, we could see that the center of the curve described by $e_3(t)$ is L, the amplitude of $|e_3 \cdot L|$ determines d and the frequency of $|e_3 \cdot L|$ determines the energy T.

4. ANGULAR VELOCITY OBSERVATION AND THE OBSERVER NORMAL FORM

In this section, we study the observability of the following system:

$$\begin{aligned}\dot{\omega}_1 &= \alpha\omega_2\omega_3\\ \dot{\omega}_2 &= \beta\omega_1\omega_3\\ \dot{\omega}_3 &= \gamma\omega_1\omega_2\\ y &= \omega_1\end{aligned} \qquad (13)$$

This is a subsystem in the spacecraft attitude problem which is related to the angular velocity and the energy. In this section, we assume $I_1 > I_2 > I_3$ or $I_1 < I_2 < I_3$.

Theorem 2. If $I_3 \neq I_2$ $(\alpha \neq 0)$, $\omega_1 \neq 0$, $\omega_2 + \omega_3 \neq 0$, then system (13) is observable.

Proof: The following relations can be easily proved.

$$L_F d\omega_1 = \alpha\omega_3 d\omega_2 + \alpha\omega_2 d\omega_3$$

$$L_F^2 d\omega_1 = \alpha\beta\omega_3^2 d\omega_1 + 2\alpha\beta\omega_1\omega_3 d\omega_3$$
$$+ \alpha\gamma\omega_2^2 d\omega_1 + 2\alpha\gamma\omega_1\omega_2 d\omega_2$$

Therefore, the dimension of the distribution generated by $d\omega_1$, $L_F d\omega_1$, $L_F^2 d\omega_1$ is the same as the rank of the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \alpha\omega_3 & \alpha\omega_2 \\ 0 & 2\alpha\gamma\omega_1\omega_2 & 2\alpha\beta\omega_1\omega_3 \end{bmatrix}$$

Its determinant is

$$2\alpha^2\omega_1(\beta\omega_3^2 - \gamma\omega_2^2)$$

Because $I_1 > I_2 > I_3$ or $I_1 < I_2 < I_3$, we know that β and γ have different signs. So, the distribution has dimension 3 whenever $\alpha \neq 0$, $\omega_1 \neq 0$, $\omega_2^2 + \omega_3^2 \neq 0$. The theorem follows.

Remark: In the remarks after theorem 1, we explained why the condition $I_1 \neq I_2$ and $\omega_1^2 + \omega_2^2 \neq 0$ arise. This can also be used to explain the condition on I_2, I_3, ω_2 and ω_3 in theorem 2. The condition $\omega_1 \neq 0$ is necessary. From the following discussion, we can see that if $\omega_1 = 0$, then ω_2, ω_3 can not be estimated by $y = \omega_1$.

Theorem 3. Under the same hypotheses as theorem 2, then

$$\omega_2^2 = \frac{\beta}{\alpha}\omega_1^2 - \beta\cdot\alpha\cdot\max\{\omega_1^2\} \qquad (14)$$

$$\omega_3^2 = \frac{\gamma}{\alpha}\omega_1^2 - \gamma\cdot\alpha\cdot\max\{\omega_1^2\} \qquad (15)$$

Proof: From (13), we have

$$\frac{1}{\alpha}\frac{d\omega_1^2}{dt} = \frac{1}{\beta}\frac{d\omega_2^2}{dt} = \frac{1}{\gamma}\frac{d\omega_3^2}{dt}$$

Therefore,

$$\frac{\omega_2^2}{\beta} = \frac{\omega_1^2}{\alpha} + c_1$$

$$\frac{\omega_3^2}{\gamma} = \frac{\omega_1^2}{\alpha} + c_2$$

Because $I_1 > I_2 > I_3$ or $I_1 < I_2 < I_3$, the constants α and β have different signs, the constants α and γ have the signs.

$$\frac{\omega_2^2}{\beta} - \frac{\omega_1^2}{\alpha} = c_1$$

is an ellipse. So $(\max\{\omega_1\},0)$ is on the ellipse. So

$$c_1 = -\alpha\cdot\max\{\omega_1^2\}$$

Since

$$\frac{\omega_3^2}{\gamma} - \frac{\omega_1^2}{\alpha} = c_2$$

is a hyperbola, $\omega_1 \neq 0$ means that ω_1^2 takes its minimum value if $\omega_3 = 0$. So

$$c_2 = -\alpha\cdot\min\{\omega_1^2\}$$

Therefore, the formulas in theorem 3 are proved.

In the following, we are going to find a kind of change of coordinates so that (13) can be transformed to observer normal form, i.e, we want to change (13) and make it look like

$$\dot{x} = Ax + f(y,u)$$
$$y = Cx$$

where (C,A) is an observable pair.

In [1], this method is discussed in detail. In example 7.3 of [1], the author proved a necessary and sufficient condition for a system like (13) to be transformed to observer form. Unfortuately, it can be proved that system (13) does not satisfy this condition. Therefore, we have to think about this problem from another point of view.

In theorem 4, we will find a family of changes of coordinates $x = x(\omega,c)$ such that for each output trajectory, there is a constant c_0 so that $x = x(\omega,c_0)$ transforms (13) to an observer normal form. $x(\omega,c)$ and the observer normal forms are continuous with respect to c.

Theorem 4. Under the same hypotheses as theorem 2, we define

$$x_1 = \omega_1$$

$$x_2 = \alpha\omega_2\omega_3$$

$$x_3 = \alpha\beta\omega_1\omega_3^2 + \alpha\gamma\omega_1\omega_2^2 \qquad (16)$$
$$- \frac{2}{3}(\beta\gamma + \beta^2 + \gamma^2)y^3 - \alpha\beta\gamma c y$$

$$c = -\alpha\{\max(y^2) + \min(y^2)\}$$

Then $x(\omega(t),c)$ satisfies

$$
\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= x_3 + \frac{2}{3}(\beta\gamma + \beta^2 + \gamma^2)y^3 + \alpha\beta\gamma c y \qquad (17) \\
\dot{x}_3 &= 0 \\
y &= x_1
\end{aligned}
$$

Proof: substitute (16) into (17), and use (14), (15).

In this note, we assume the spacecraft moves freely. An obvious question is, how to observe the attitude when the system has nonzero input ? This is an important open question.

Another interesting problem is, for what range of the output, system (13) can be estimated by theorem 4 without changing the parameter c.

The results in this note are applicable to the rigid body problem, but most recent spacecraft research is directed towards large flexible space structures and the models are much more complicated. However, the rigid dynamics are still interesting and important.

REFERENCES

[1] Arthur J. Krener and Witold Respondek, "Nonlinear Observers With Linearization Error Dynamics," SIAM J. Control And Optimization, Vol. 23, No. 2, 1985, pp.197.

[2] K.R. Symon, "Mechanics," Addison-Wesly Publishing Company,1961.

[3] P.E. Crouch, "*Spacecraft Attitude Control and Stabilization* Applications of Geometric Control Theory to Rigid Body Models," IEEE Trans. Automat. Control, AC-29, 1984, pp. 321.

[4] Arthur J. Krener and R. Hermann, " Nonlinear Observability and Controllability." IEEE Trans. Automat. Contr., vol. AC-22, pp. 728-740.

[5] G. Meyer, " On the use of Euler's theorem on rotations for the synthesis of attitude control systems," Ames Res. Cen., Moffet Field, CA, NASA Tech. Note NASA TN. D-3643, 1966.

[6] G. Mayer, "Design and global analysis of spacecraft attitude control systems," Ames Res. Cen. Moffet Field. CA, NASA Tech. Note NASA TR. R-361, 1971.

[7] J. Baillieul, " Controllability and Observability of Polynomial Dynamical System," Nonlinear Anal., Theory, Methods and Appl., vol. 5. no. 5, pp. 543-552, 1981.