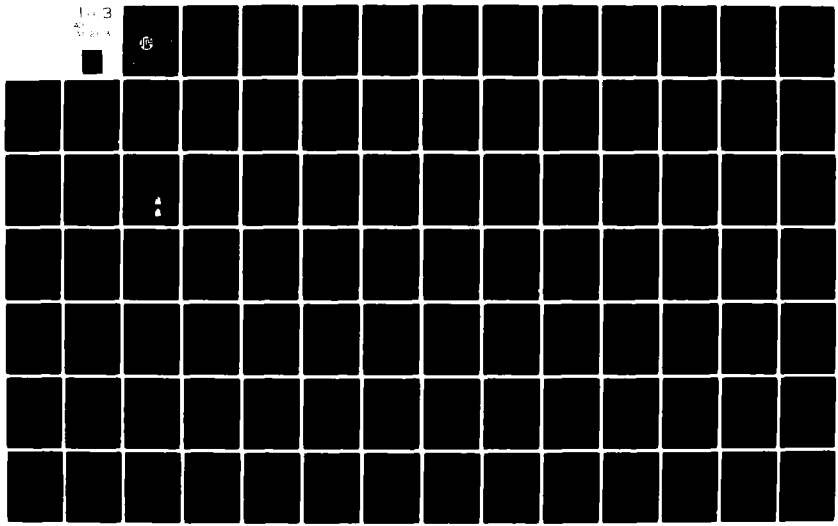
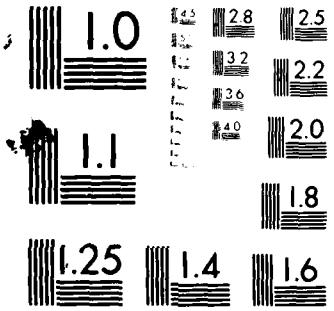
AD-A112 113 TEXAS TECH UNIV LUBBOCK INST FOR ELECTRONIC SCIENCE F/G 9/5
ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICES ELECTRONICS --ETC(U)
DEC 81 R SAEKS, L R HUNT, J MURRAY, J WALKUP N00014-76-C-1136

UNCLASSIFIED NL

1 OF 3
AD
A112113





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

(12)

AD A112113

ANNUAL REVIEW OF RESEARCH

under the

JOINT SERVICES ELECTRONICS PROGRAM

December 1981

DTIC
ELECTE
MAR 17 1982
S D
H

Institute for
Electronic Science

DTIC FILE COPY

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

TEXAS TECH UNIVERSITY
Lubbock, Texas 79409

82 03 17 12

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO. AD-A112 113

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

Annual Review of Research Under the Joint Services Electronics Program

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s) R. Saeks, L.R. Hunt, J. Murray, J. Walkup, and T.G. Newman

8. CONTRACT OR GRANT NUMBER(s)

N00014-76-C-1136

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Texas Tech University
Institute of Electronic Science
Lubbock, TX 79409

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE

December 1981

13. NUMBER OF PAGES

280

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

Office of Naval Research
800 N. Quincy Avenue
Arlington, VA.

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE - DISTRIBUTED UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DTIC
ELECTE
MAR 17 1982
H

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Fault Analysis, Control, Image Processing, Pointing and Tracking, Multidimensional Signal Processing.

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

This report represents the fifth year of research performed under the auspices of the Joint Services Electronics Program at Texas Tech University. The program is concentrated in the "information electronics" area and includes researchers from both the departments of Electrical Engineering and Mathematics. Specific work units deal with Feedback System Design, Nonlinear Control, Nonlinea Fault Analysis, Image Processing, and Pointing and Tracking.

Each work unit is represented in the report by a summary of the work perform ed during the past year, a list of publications and activities in the area, (over

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Ente

reprints of all papers which have been published during the past year, and abstracts of pending papers. In addition, the report includes lists of all grants and contracts administered by JSEP personnel, the department of Electrical Engineering and the Department of Mathematics; and a list of all publications prepared by JSEP personnel.

Accession For
NTIS GRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
By
Distribution/
Availability Codes
Dist Special
A

DTIC
COPY
INSPECTED

REVIEW OF RESEARCH

under the

JOINT SERVICES ELECTRONICS PROGRAM

at the

INSTITUTE FOR ELECTRONIC SCIENCE

TEXAS TECH UNIVERSITY

December 1981

Lubbock, Texas 79409

Abstract

This report represents the fifth year of research performed under the auspices of the Joint Services Electronics Program at Texas Tech University. The program is concentrated in the "information electronics" area and includes researchers from both the departments of Electrical Engineering and Mathematics. Specific work units deal with Feedback System Design, Nonlinear Control, Nonlinear Fault Analysis, Image Processing, and Pointing and Tracking.

Each work unit is represented in the report by a summary of the work performed during the past year, a list of publications and activities in the area, reprints of all papers which have been published during the past year, and abstracts of pending papers. In addition, the report includes lists of all grants and contracts administered by JSEP personnel, the department of Electrical Engineering and the Department of Mathematics; and a list of all publications prepared by JSEP personnel.

Contents

Significant Accomplishments Report..... 1

1. Feedback System Design, R. Saeks..... 5

Reprint of "Feedback System Design: The Tracking and Disturbance Rejection Problems"..... 11

Reprint of "Fractional Representation Approach to Adaptive Control"..... 29

Reprint of "Suboptimal Control with Optimal Quadratic Regulators"..... 33

Reprint of "Fractional Representation, Algebraic Geometry and the Simultaneous Stabilization Problem"..... 41

Abstracts for Pending Publications..... 43

2. Nonlinear Control, L.R. Hunt..... 57

Reprint of "Linear Equivalents of Nonlinear Time-Varying Systems"..... 63

Reprint of "The Poincare Lemma and Transformations of Nonlinear Systems"..... 71

Reprint of "Transforming Nonlinear Systems"..... 81

Reprint of "Global Mappings of Nonlinear Systems"..... 95

Reprint of "Local Transformations for Multi-input Nonlinear Systems"..... 103

Reprint of "Controllability of Nonlinear Hypersurface Systems"..... 111

Abstracts of Pending Publications..... 127

3. Nonlinear Fault Analysis, R. Saeks..... 153

Reprint of "Fault Diagnosis in Electronic Circuits"..... 159

Reprint of "Criteria for Analog Fault Diagnosis"..... 165

Reprint of "Analog Fault Diagnosis with Failure Bounds"..... 171

Reprint of "A Differential-Interpolative Approach to Analog Fault Simulation"..... 179

Abstracts of Pending Publications..... 185

PRECEDING PAGE BLANK-NOT FI

4. Multidimensional System Theory, J. Murray..... 197

Reprint of "A Time-Varying Approach to Two-Dimensional Digital Filtering"..... 201

Reprint of "Lumped-Distributed Networks and Differential-Delay Systems"..... 209

Abstracts of Pending Publications..... 223

5. Detection and Estimation in Imagery, J. Walkup..... 231

Reprint of "Estimation in signal-dependent film grain noise"..... 235

6. Pointing and Tracking, T. G. Newman..... 245

Reprint of "Adaptive Pattern Matching Using Control Theory in Lie Groups"..... 249

7. Director's Discretionary Fund, R. Saeks..... 257

Reprint of "Optimal Selection of IC Fabrication Parameters"..... 261

Abstracts of Pending Publications..... 269

Grants and Contracts Administered by JSEP Personnel..... 271

Grants and Contracts in Electrical Engineering..... 272

Grants and Contracts in Mathematics..... 275

Publications for JSEP Personnel..... 277

PRECEDING PAGE BLANK-NOT FILMED

Significant Accomplishments Report

A. Nonlinear Control

During the past year Professor L.R. Hunt and his students working jointly with researchers at NASA/AMES have developed an entirely new approach to nonlinear control system design problem. The key to the new approach is the formulation of an exact linearization theory which permits the nonlinear system under investigation to be transformed, without approximation, into an equivalent linear system to which classical design methodologies are applicable. This has, in turn, been achieved through the application of the powerful techniques of modern differential geometry through which it has been possible to convert the exact linearization concept from a theoretical abstraction into a viable design algorithm. Indeed, the algorithm has already been implemented at NASA/AMES in the design of a helicopter autopilot.

Although the exact linearization concept goes back to Poincare and has been investigated by a number of researchers during the past decade the class of transformations employed in this work has typically been limited to those which could be implemented by feedback. Unfortunately, this class of transformations has not proven to be amenable to mathematical analysis. To the contrary by adopting a larger class of transformations originally proposed by R. Su, Professor Hunt has been able to formulate a complete theory around the exact linearization concept. This includes precise necessary and sufficient conditions for a plant to admit an exact linearization and a partial differential equation whose solution defines the appropriate transformation with which to linearize a given system.

Although Dr. Hunt's research is extremely theoretical in nature his research has been closely coordinated with the autopilot design program at NASA/AMES. Indeed, AMES has already employed his work in the design of an experimental helicopter autopilot which is presently undergoing simulation and is expected to fly in the near future.

B. Nonlinear Fault Analysis

During the past year we completed work (we believe successfully) on a long standing JSEP work unit directed at the development of an algorithm for the solution of the analog fault diagnosis problem. Although many algorithms have been proposed over the years which were theoretically capable of determining the fault circuit components from external measurements the problem has been to find an algorithm which could do the job with limited computational resources and constraints on the number of measurements which can be made on the unit under test.

These problems have now been resolved via a new self-testing algorithm developed by Professor Saeks and his students. In essence, the algorithm exploits the fact that all of the system components do not fail simultaneously thereby permitting one to use a subset of the system components to test the remaining components. This, in turn, yields conditional test information which is valid if the given subset of components are, in fact, good. The results of several such conditional tests are then combined with the aid of an upper bound on the number of simultaneous failures to obtain the final diagnosis. Although the procedure is somewhat roundabout it meets most of the criteria which have been established over the years by researchers in the analog fault diagnosis area.

i). It is applicable to both linear and nonlinear systems modeled

in either the time or frequency domain.

ii). It can be used to locate multiple hard or soft faults.

iii). Finally, it is capable of locating failures in "replaceable Modules" such as an IC chip, a PC board, or a subsystem rather than discrete components.

As of the present time we have completed the algorithm development work related to the new algorithm while we are presently doing a preliminary investigation of the software engineering problems which must be resolved as a prerequisite to implementing the algorithm in a user oriented software code.

Texas Tech University | Institute for Electronic Science

Joint Services Electronics Program | Research Unit: 1

1. Title of Investigation: Feedback System Design
2. Senior Investigator: Richard Saeks Telephone: (806)-742-3528
3. JSEP Funds: $25,875
4. Other Funds:
5. Total Number of Professionals: PI 1 (1 mo.) RA 1 (1/2 time)
6. Summary:

Although control theorists have studied higher order and multivariate systems for more than a quarter of a century this research has had little impact on the DOD community in which single loop PI designs still predominate. Indeed, such controllers represent the physical limit of what can be achieved with the hydraulic and/or analog electronic hardware which is traditionally used to implement a control system. With the advent of the digital control computer, however, higher order and multivariate controllers have become a reality, wherein, one can implement any desired compensator design simply by burning the appropriate program into a ROM.

Given the renewed interest in higher order multivariate control brought about by the digital control computer the present work unit is directed toward the problem of developing an efficient algorithmic design procedure for linear multivariate control systems using frequency domain techniques. Our approach is based on a, now classical, result of Youla, Bongiorno and Jabr in which an explicit parameterization of the set of compensators which stabilize a given plant is formulated. Indeed, with the aid of a modified parameterization

PRECEDING PAGE BLANK-NOT FI

due to Desoer, Liu, Murray, and the author one can parameterize the set of compensators for a given plant in such a manner that the resultant feedback system gains are linear (actually affine) in the design parameter. This, in turn greatly simplifies the process of specializing the design to meet additional constraints and, as such, a powerful design theory has been obtained which includes the *trucking and disturbance rejection problems*, the *pole placement problem*, *robust design*, design with *stable or proper compensators*, and the *model matching problem* as well as the beginnings of a *simultaneous design* and *adaptive control theory*.

7. Publications and Activities:

A. Refereed Journal Articles

1. Saeks, R., and J. Murray, "Feedback System Design: The Tracking and Disturbance Rejection Problems," IEEE Trans. on Auto. Cont., Vol. AC-26, pp. 203-217, (1981).

2. Saeks, R., Murray, J., Chua, O., Karmokolias, C., and A. Iyer, "Feedback System Design: The Single Variate Case," Circuits, Systems, and Signal Processing, (to appear).

B. Conference Papers and Abstracts

1. Karmokolias, C., and R. Saeks, "A Fractional Representation Approach to Adaptive Control," Proc. of the IEEE Conf. on Decision and Control, Albuquerque, NM, Dec. 1980, pp. 272-273.

2. Saeks, R., and J. Murray, "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem," Proc. of the IEEE Inter. Symp. on Circuits and Systems, Chicago, April 1981, pp. 463-464.

3. Karmokolias, C., and R. Saeks, "Suboptimal Control with Optimal Quadratic Regulators," Proc. of the Conf. on Information Sciences Systems, Johns Hopkins Univ., April 1981, pp. 53-58.

4. Murray, J., and R. Saeks, "Simultaneous Design of Control Systems," Proc. of the IEEE Conf. on Decision and Control, San Diego, Dec. 1981, (to appear).

C. Preprints

1. Saeks, R., and J. Murray, "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem," (submitted for publication).

D. Dissertations and Theses

1. A. Iyer, Ph.D. Dissertation, (in preparation).

E. Conferences and Symposia

1. Saeks, R., and J. Murray, 1980 IEEE Conf. on Decision and Control, Albuquerque, NM, Dec. 1980.

2. Saeks, R., 1981 Joint Automatic Control Conf., Charlottesville, Va., June, 1981.

3. Saeks, R., IEE Conf. on Control and its Applications, Coventry, England, April, 1981.

4. Saeks, R., IEEE Inter. Symp. on Circuits and Systems, Chicago, Il. April 1979.

5. Saeks, R., and J. Murray, 1981 Texas Systems Workshop, Dallas, TX., April 1981.

6. Iyer, A., 24th Midwest Symp. on Circuits and Systems, Albuquerque, NM, June 1981.

F. Lectures

1. Saeks, R., NASA/AMES Research Center, Feb. 1981.

2. Saeks, R., University of Manchester Inst. of Science and Technology, April 1981.

3. Saeks, R., Cambridge Univ., April 1981.

FEEDBACK SYSTEM DESIGN: THE TRACKING AND DISTURBANCE REJECTION PROBLEMS

RICHARD SAEKS

AND

JOHN MURRAY

PRECEDING PAGE BLANK-NOT FILMED

Feedback System Design: The Tracking and Disturbance Rejection Problems

RICHARD SAEKS, FELLOW, IEEE, AND JOHN MURRAY, MEMBER, IEEE

Abstract—The problem of designing a compensator for a specified plant which simultaneously stabilizes the resultant feedback system and causes it to track a prescribed family of inputs and/or reject prescribed disturbances is considered. A set of linear design equations, in the space of stable systems, is formulated in a general linear systems setting and an explicit parameterization of the resultant solution space is obtained for a class of "generalized multivariate" systems. The theory is illustrated with several single and multivariate examples.

I. Introduction

THE FEEDBACK system design problem may naturally be subdivided into two tasks:

1) satisfaction of design constraints, and
2) "optimization" of system performance.

The first and foremost design constraint is stability. Additionally, system specifications may also include a tracking and/or disturbance rejection constraint. Once these constraints have been satisfied, the remaining design latitude may then be used to "optimize" the qualitative performance of the system: cost, energy consumption, overshoot, reliability, complexity, etc.

The purpose of the present paper is to report on the results of a research program whose goal is the formulation of a new algebraic fractional representation approach to the feedback system design problem. The key to this approach is a design philosophy pioneered by D. C. Youla, in which one parameterizes the entire solution space for the design problem, rather than simply constructing a single solution [15]–[17]. The approach is ideally suited for the feedback system design problem, wherein one can satisfy several design constraints by *sequentially reparameterizing* the controller as additional constraints are imposed with the system performance being "optimized" over the final parameterization. In [15] and [16] Youla, Bongiorno, and Jabr parameterized the stabilizing controllers for single-variate and multivariate feedback systems, respectively, and showed that the optimization of the system performance over the resultant parameteer space reduced to a standard Wiener-Hopf problem. More recently Youla has adopted a similar approach to the design of stochastic estimators in which the set of covariances which interpolate one's observations of a stochastic process are parameterized [17].

Manuscript received January 2, 1980; revised October 8, 1980. This work was supported in part by the Joint Services Electronics Program at Texas Technical University under ONR Contract 76-C-1136.

The authors are with the Department of Electrical Engineering, Texas Technical University, Lubbock, TX 79409.

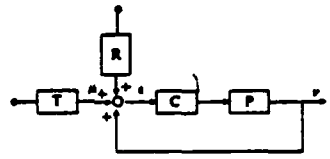


Fig. 1. Feedback system with reference generators.

The present research program began with the derivation of a simple algebraic proof for the Youla/Bongiorno/Jabr stabilization theory and the generalization of the theory to include a variety of "system placement" problems in a general linear system setting [6]. The spirit of this work was very close to that of several other authors in the area of "algebraic" control theory [1], [2], [4], [7], [12], [13]. Among the problems which have been studied by these authors are those of tracking and disturbance rejection. In the present paper, we study these problems in a general axiomatic setting. Two restrictions are made, however. First, we assume that the disturbances affect the system in a very simple way, and second, we assume that the denominators of the reference generators commute with everything. (See Section V for a more detailed explanation.) With this simplification, it is again possible to obtain a complete parameterization of all controllers which achieve the prescribed design contraints by sequentially reparameterizing the controllers as additional design constraints are imposed. Moreover, the feedback system gains which result from such a controller are linear (affine) in the resultant design parameter, thereby simplifying the second task referred to above, namely optimization.

Throughout the paper we will work with the feedback system of Fig. 1. Here, P is a given plant and C is the compensator to be designed while T and R are reference generators that model the inputs to be tracked and rejected, respectively. (The sense in which these generators model inputs will be made precise in Section III.) The problem of parameterizing the compensators that stabilize the feedback system, termed problem S, was the subject of [6] and will serve as the starting point for the present paper in which we investigate three additional design problems:

Problem ST: Parameterize the compensators that simultaneously stabilize the system and cause it to track the response of T.

Problem SR: Parameterize the compensators that simultaneously stabilize the system and cause it to reject the response of R.

PRECEDING PAGE BLANK-NOT FI

Problem STR: Parameterize the compensators that simultaneously stabilize the system, cause it to track the response of T, and reject the response of R.

Of course, the concepts of stability, tracking, and disturbance rejection will be made precise as required.

This, of course, is by no means the most general possible system. In particular, the disturbance is assumed merely to be added to the input, and the measured output is assumed to be the same as the regulated output. We remark in passing that the problem of rejecting an additive disturbance at the output is mathematically equivalent to the tracking problem.

In the remainder of this introduction our algebraic fractional representation theory is reviewed and the results on the stabilization problem, derived in [6], are summarized. In Section II a more powerful "doubly coprime" fractional representation theory is formulated and its properties developed. In the following section the concepts of asymptotic tracking and disturbance rejection are defined in our algebraic setting and theorems characterizing the stable feedback systems which track the response of T and/or reject the response of R are obtained. The fundamental feedback system design equations are derived in Section IV, wherein it is shown that problems ST, SR, and STR each reduce to the solution of a linear equation in the ring of stable operators. A necessary and sufficient condition for the existence of a solution of these equations and the desired parameterization of the solution space are obtained in a "generalized multivariate" class of systems in Section V. Finally, some examples of the theory are presented in Section VI.

Our algebraic theory [6] is set in a nest of rings, groups, and multiplicative structures

$$G \supset H \supset I \supset J.$$

Here, G is a ring with identity that represents the general class of systems with which we wish to work: rational matrices, continuous operators, a class of transcendental functions, etc. H is a subring of G containing the identity that models the systems, which are stable in some sense: poles in a prescribed region, transcendental functions with restricted singularities, stable operators, etc. Finally, I denotes the multiplicative set composed of elements of H that admit an inverse in G, while J denotes the multiplicative subgroup of I made up of elements that are invertible in H. Detailed examples of the axiomatic structure, $\{G, H, I, J\}$ were given in [6] and will not be repeated here.

We say that a plant P has a *right fractional representation* in $\{G, H, I, J\}$ if

$$P = p_r \tilde{p}_r^{-1} \tag{1.1}$$

where $p_r \in H$ and $\tilde{p}_r \in I$. Furthermore, we say that this representation is *right coprime* if there exists q_r and $\tilde{q}_r$ in H such that

$$q_r p_r + \tilde{q}_r \tilde{p}_r = 1. \tag{1.2}$$

This equality is equivalent to the classical coprimeness concept for rational functions and matrices while being well defined in our general ring theoretic setting. In particular, if G is the ring of rational functions and H is the ring of polynomials (e.g., in discrete-time systems), (1.2) implies that p_r and $\tilde{p}_r$ have no common zeros. If G is the ring of rational functions and H is the ring of exponentially stable rational functions, (1.2) implies that p_r and $\tilde{p}_r$ have no common right-half plane zeros.

Since the ring G is, in general, noncommutative, we also define a *left fractional representation* for P via the equality

$$P = \tilde{p}_l^{-1} p_l \tag{1.3}$$

for $p_l \in H$ and $\tilde{p}_l \in I$. Furthermore, we say that this representation is *left* coprime if there exists q_l and $\tilde{q}_l$ in H such that

$$p_l q_l + \tilde{p}_l \tilde{q}_l = 1. \tag{1.4}$$

Of course, in the classical case of a rational function or matrix these fractional representations are assured to exist. However, this is not the case in the general ring theoretic setting. Therefore, for distributed, time-varying and multidimensional systems, we assume that our plant admits such a representation as a prerequisite to the theory.

If one is given two fractional representations for a plant, say $P = x_r \tilde{x}_r^{-1}$ and $P = p_r \tilde{p}_r^{-1}$, where the second satisfies the coprimeness condition of (1.2) then the two representations differ by a *greatest right divisor* [6], $r \in H$, i.e.,

$$x_r = p_r r \tag{1.5}$$

and

$$\tilde{x}_r = \tilde{p}_r r. \tag{1.6}$$

To give an idea of the arguments used later, we prove this fact as follows. Since $\tilde{p}_r \in I$ the unique solution of (1.5) and (1.6) is $r = \tilde{p}_r^{-1} \tilde{x}_r$ and hence we must show that this $r \in H$. To this end we invoke (1.2) obtaining

$$\begin{aligned} r = \tilde{p}_r^{-1} \tilde{x}_r &= (q_r p_r + \tilde{q}_r \tilde{p}_r) \tilde{p}_r^{-1} \tilde{x}_r \\ &= q_r p_r \tilde{p}_r^{-1} \tilde{x}_r + \tilde{q}_r \tilde{x}_r = q_r x_r + \tilde{q}_r \tilde{x}_r \end{aligned}$$

which implies that $r \in H$ since it is expressed as the sum of products of elements of H. Similarly, if $P = \tilde{p}_l^{-1} p_l = \tilde{x}_l^{-1} x_l$, where $\tilde{p}_l^{-1} p_l$ is left coprime, then the two left fractional representations differ by a *greatest left* divisor [6] $l \in H$ such that

$$x_l = l p_l \tag{1.8}$$

and

$$\tilde{x}_l = l \tilde{p}_l. \tag{1.9}$$

Thus, our abstract fractional representation theory is quite similar to the classical theory for polynomials and rational functions even though it is set in an abstract ring and

includes distributed, time-varying, and multidimensional systems.

Since H represents a ring of stable systems we say that the feedback system of Fig. 1 is stable if all of its internal[1] and external gains are elements of H. By using the above described fractional representation theory a solution to problem S was obtained in [6]. In particular, if the plant P is given by $P=\tilde{p}_l^{-1}p_l$, then the compensator

$$C=(wp_l+\tilde{q}_r)^{-1}(-w\tilde{p}_l+q_r) \tag{1.10}$$

stabilizes the feedback system for any $w\in H$ such that $(wp_l+\tilde{q}_r)\in I$; and every stabilizing compensator is given by (1.10) for some $w\in H$. Moreover, when this compensator is used in the feedback system all of the usual gains of the closed-loop system are linear (actually affine) in w. Specifically, the system's input/output gain is given by

$$h_{rp}=-p_rw\tilde{p}_l+p_rq_r \tag{1.11}$$

while the system's input/error gain takes the form

$$h_{ep}=1+p_rw\tilde{p}_l-p_rq_r. \tag{1.12}$$

The remaining gains are also linear in w and are derived in [6].

We note that even though our approach to the feedback system design problem is formulated in a ring, no "ring theory" is employed. Indeed, the only mathematics required in the entire paper is addition, multiplication, subtraction, and inversion.

Finally, the above notation exemplifies a pattern that we will try to follow throughout the paper, namely, an input–output operator will be denoted by a single uppercase letter (P), its coprime factorizations will be denoted by the corresponding lowercase letter, with a tilde over the denominator, and subscript l or r for left or right (e.g., $p_r\tilde{p}_r^{-1}$), and last, the elements appearing in the coprimeness equation will be the next letter in the alphabet, with tilde and subscripts matching the elements which they multiply (e.g., $q_rp_r+\tilde{p}_r\tilde{q}_r=1$).

II. Doubly Coprime Fractional Representations

In the stabilization theory discussed in the previous section it was assumed that the plant admitted both left and right coprime fractional representations

$$P=p_r\tilde{p}_r^{-1}=\tilde{p}_l^{-1}p_l \tag{2.1}$$

for which there exist $q_r\in H$ and $\tilde{q}_r\in H$ such that

$$q_rp_r+\tilde{q}_r\tilde{p}_r=1 \tag{2.2}$$

and $q_l\in H$ and $\tilde{q}_l\in H$ such that

$$p_lq_l+\tilde{p}_l\tilde{q}_l=1. \tag{2.3}$$

Although this structure sufficed for the stabilization theory of problem S, one can, in fact, adopt a stronger structure without loss of generality [3].

Property 1: Assume that P admits both left and right coprime fractional representations such that (2.1) through (2.3) are satisfied. Then there exist q_l' and $\tilde{q}_l'$ in H such that

$$p_lq_l'+\tilde{p}_l\tilde{q}_l'=1 \tag{2.4}$$

and

$$\tilde{q}_rq_l'=q_r\tilde{q}_l'. \tag{2.5}$$

Proof: Recall that the inverse of a two by two matrix with elements in a noncommutative ring is given by

$$\begin{bmatrix} X & Y \\ Z & W \end{bmatrix}^{-1}=\begin{bmatrix} \Delta^{-1} & -\Delta^{-1}YW^{-1} \\ -W^{-1}Z\Delta^{-1} & W^{-1}+W^{-1}Z\Delta^{-1}YW^{-1} \end{bmatrix} \tag{2.6}$$

where $\Delta=X-YW^{-1}Z$, provided the indicated inverses exist. Applying this to the matrix

$$\begin{bmatrix} \tilde{q}_r & q_r \\ -p_l & \tilde{p}_l \end{bmatrix} \tag{2.7}$$

we find

$$\Delta^{-1}=\tilde{p}_r, \tag{2.8}$$

and, after some computation,

$$\begin{bmatrix} \tilde{q}_r & q_r \\ -p_l & \tilde{p}_l \end{bmatrix}^{-1}=\begin{bmatrix} \tilde{p}_r & -q_l' \\ p_r & \tilde{q}_l' \end{bmatrix} \tag{2.9}$$

where

$$q_l'=\tilde{p}_rq_r\tilde{q}_l+q_l-\tilde{p}_r\tilde{q}_rq_l \tag{2.10}$$

and

$$\tilde{q}_l'=\tilde{q}_l-p_rq_r\tilde{q}_l+p_r\tilde{q}_rq_l. \tag{2.11}$$

Equations (2.4) and (2.5) now follow immediately. ■

In essence, Property 1 implies that if a plant admits any pair of left and right coprime fractional representations then it also admits a stronger fractional representation in which the q's intertwine in a manner similar to the p's. We say that such a fractional representation is *doubly coprime* and (dropping the primes) we denote it by

$$\begin{bmatrix} \tilde{q}_r & q_r \\ -p_l & \tilde{p}_l \end{bmatrix}^{-1}=\begin{bmatrix} \tilde{p}_r & -q_l \\ p_r & \tilde{q}_l \end{bmatrix}. \tag{2.12}$$

Since a doubly coprime fractional representation exists whenever separate left and right coprime fractional representations exist we may work with a doubly coprime fractional representation without loss of generality. This in turn, means that rather than the three equalities (2.1)

[1] Our concept of stability here *requires* that all internal system gains are stable (i.e., in H) in addition to the gains observed between the actual inputs and outputs. These concepts are discussed in more detail in [6].

through (2.3) we have eight equalities with which to manipulate the feedback system equations. These are obtained by both premultiplying and postmultiplying the matrix of (2.12) by its inverse to compute the identity and take the form

$$p_l\tilde{p}_r = \tilde{p}_l p_r \tag{2.13}$$

$$\tilde{q}_r q_l = q_r \tilde{q}_l \tag{2.14}$$

$$\tilde{q}_r\tilde{p}_r + q_r p_r = 1 \tag{2.15}$$

$$p_l q_l + \tilde{p}_l\tilde{q}_l = 1 \tag{2.16}$$

$$\tilde{p}_r q_r = q_l \tilde{p}_l \tag{2.17}$$

$$p_r\tilde{q}_r = \tilde{q}_l p_l \tag{2.18}$$

$$\tilde{p}_r\tilde{q}_r + q_l p_l = 1 \tag{2.19}$$

and

$$p_r q_r + \tilde{q}_l\tilde{p}_l = 1. \tag{2.20}$$

Finally, we note that, if in the stabilization theory in [6] one begins with the above doubly coprime representation for the plant, the representation (1.10) for the controller is also doubly coprime: a simple calculation gives

$$\begin{bmatrix} \tilde{p}_l & p_l \\ w\tilde{p}_l - q_r & wp_l + \tilde{q}_r \end{bmatrix}^{-1} = \begin{bmatrix} p_r w + \tilde{q}_l & -p_r \\ -\tilde{p}_r w + q_l & \tilde{p}_r \end{bmatrix}.$$

III. Analysis

We say that the feedback system of Fig. 1 *tracks* T *if* $h_{\epsilon\mu}T \in H$. This definition may be justified by considering the case where H is the ring of exponentially stable rational functions. Here, if one lets μ be the impulse response of T, $\mu(s) = T(s)$ and

$$\mu(s) - \nu(s) = \epsilon(s) = h_{\epsilon\mu}(s)\mu(s) = h_{\epsilon\mu}(s)T(s). \tag{3.1}$$

Therefore, ν asymptotically tracks μ if and only if $h_{\epsilon\mu}(s)T(s)$ is exponentially stable, i.e., $h_{\epsilon\mu}T \in H$. Note that the impluse response of T may be unbounded (since T may be unstable) even though the input to the reference generator is bounded. Of course, the same intuition applies in a more general setting where the condition $h_{\epsilon\mu}T \in H$ implies that ν is asymptotic to μ in whatever sense the response of the systems represented by elements of H is asymptotic to zero.

Similarly, we say that the feedback system of Fig. 1 *rejects* R if $h_{\nu\mu}R \in H$. Once again if H is the ring of exponentially stable rational functions and μ is the impulse response of R, $\mu(s) = R(s)$ and

$$\nu(s) = h_{\nu\mu}(s)\mu(s) = h_{\nu\mu}(s)R(s), \tag{3.2}$$

which is asymptotic to zero if and only if $h_{\nu\mu}(s)R(s)$ is exponentially stable, i.e., $h_{\nu\mu}R \in H$. Thus, the system rejects R in the sense that its response to the impulse

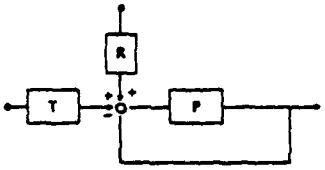


Fig. 2. Stable feedback system.

response of R is asymptotic to zero. See [1], [3], [4], and [7] for a fuller discussion of these concepts of tracking and rejection.

Now, consider the feedback system of Fig. 2. Adopting an argument similar to that employed by Francis [7] in a frequency domain setting, we obtain the following theorem, which characterizes the stable feedback systems which track T and/or reject R. For this purpose we assume that P, T, and R each admit left coprime fractional representations

$$P = \tilde{p}_l^{-1} p_l \tag{3.3}$$

$$T = \tilde{t}_l^{-1} t_l \tag{3.4}$$

and

$$R = \tilde{r}_l^{-1} r_l \tag{3.5}$$

where

$$p_l q_l + \tilde{p}_l\tilde{q}_l = 1 \tag{3.6}$$

$$t_l u_l + \tilde{t}_l\tilde{u}_l = 1 \tag{3.7}$$

and

$$r_l s_l + \tilde{r}_l\tilde{s}_l = 1. \tag{3.8}$$

Theorem 1: Assume that the feedback system of Fig. 2 is stable. Then, the system

1) tracks T if and only if $\tilde{t}_l$ is a right divisor of $\tilde{p}_l$, in the sense that $\tilde{p}_l\tilde{t}_l^{-1} \in H$;

2) rejects R if and only if $\tilde{r}_l$ is a right divisor of p_l, in the same sense.

Proof: We have

$$h_{\epsilon\mu}\tilde{t}_l^{-1} = h_{\epsilon\mu}\tilde{t}_l^{-1}(t_l u_l + \tilde{t}_l\tilde{u}_l) \tag{3.9}$$

$$= (h_{\epsilon\mu}\tilde{t}_l^{-1} t_l)u_l + h_{\epsilon\mu}\tilde{u}_l \tag{3.10}$$

and so $h_{\epsilon\mu}\tilde{t}_l^{-1} t_l \in H \Leftrightarrow h_{\epsilon\mu}\tilde{t}_l^{-1} \in H$. Now

$$h_{\epsilon\mu} = (\tilde{p}_l + p_l)^{-1}\tilde{p}_l \tag{3.11}$$

and since the system is stable, we know (from [6]) that $\tilde{p}_l + p_l \in J$.

Therefore,

$$h_{\epsilon\mu}\tilde{t}_l^{-1} \in H \Leftrightarrow (\tilde{p}_l + p_l)^{-1}\tilde{p}_l\tilde{t}_l^{-1} \in H$$

$$\Leftrightarrow \tilde{p}_l\tilde{t}_l^{-1} \in H, \tag{3.12}$$

and part 1) follows. Part 2) is proved similarly. ∎

IV. The Design Equations

We now turn our attention to the system design problems ST, SR, and STR. Since in each of these problems the compensator must stabilize the system in addition to causing it to track T and/or reject R, the theory of [6] implies that our compensator must take the form of (1.10) with $w \in H$ chosen to assure that $h_{\epsilon\mu}T \in H$ and/or $h_{r\mu}R \in H$. For this purpose we assume that the plant P admits a doubly coprime fractional representation (2.12) while T and R are assumed to admit left coprime fractional representations as in (3.4), (3.5), (3.7), and (3.8).

Theorem 2: For the feedback system of Fig. 1:

1) problem ST is soluble if and only if the equation

$$p_r w \tilde{p}_l + x \tilde{t}_l = p_r q_r - 1 \tag{4.1}$$

admits solutions w and x in H;

2) problem SR is soluble if and only if the equation

$$p_r w \tilde{p}_l + y \tilde{r}_l = p_r q_r \tag{4.2}$$

admits solutions w and y in H;

3) problem STR is soluble if and only if (4.1) and (4.2) admit simultaneous solutions w, x, and y in H.

Moreover, when a solution exists the set of compensators which solve the problem is given by

$$C = (wp_l + \tilde{q}_r)^{-1}(-w\tilde{p}_l + q_r) \tag{4.3}$$

where w is any solution of the appropriate equation(s), such that

$$wp_l + \tilde{q}_r \in I.$$

Proof: Since every stabilizing compensator is given by (1.10) for an arbitrary $w \in H$ with the resultant $h_{\epsilon\mu}$ given by (1.12) the w's that are both in H and simultaneously place

$$h_{\epsilon\mu}T = [1 + p_r w \tilde{p}_l - p_r q_r]\tilde{t}_l^{-1} t_l \tag{4.4}$$

in H define the class of compensators which solve problem ST. In the course of proving Theorem 1, we proved that $h_{\epsilon\mu}T \in H$ if and only if $h_{\epsilon\mu}\tilde{t}_l^{-1} \in H$. Now suppose that

$$h_{\epsilon\mu}\tilde{t}_l^{-1} = -x \in H. \tag{4.5}$$

Then,

$$1 + p_r w \tilde{p}_l - p_r q_r = -x\tilde{t}_l, \tag{4.6}$$

which is equivalent to (4.1).

Conversely, if (4.1) admits solutions w and x in H, then using this w in (1.10) yields a compensator which stabilizes the system with $h_{\epsilon\mu}\tilde{t}_l^{-1} = -x \in H$; that is, a compensator which solves the problem ST.

The proof of 2) is obtained by a parallel argument with $h_{\epsilon\mu}$ replaced by $h_{r\mu}$ and T replaced by R while 3) is obtained by recognizing that the solution of problem STR requires a simultaneous solution of problems ST and SR with the same compensator, hence the same w. ■

Following 14, we define p_r and $\tilde{t}_l$ to be *skew coprime* if there exists j and $\tilde{j}$ in H such that

$$p_r j + \tilde{j}\tilde{t}_l = 1.$$

We then can state the following corollary. (Compare [2], [4], [13]).

Corollary 1: For the feedback system of Fig. 1,

1) a necessary condition for problem ST to admit a solution is that p_r and $\tilde{t}_l$ are skew coprime;

2) a necessary condition for problem SR to admit a solution is that $\tilde{p}_l$ and $\tilde{r}_l$ are right coprime;

3) a necessary condition for problem STR to admit a solution is that p_r and $\tilde{t}_l$ are skew coprime, $\tilde{p}_l$ and $\tilde{r}_l$ are right coprime, and $\tilde{t}_l$ and $\tilde{r}_l$ are right coprime.

Proof: If problem ST admits a solution, (4.1) is satisfied for some w and x in H, hence, upon rearranging the terms in this equation, we have

$$p_r[q_r - w\tilde{p}_l] + [-x]\tilde{t}_l = p_r j + \tilde{j}\tilde{t} = 1 \tag{4.7}$$

where $j = q_r - w\tilde{p}_l$ and $\tilde{j} = -x$. To verify 2) we substitute (2.20) into (4.2) to obtain

$$p_r w \tilde{p}_l + y\tilde{r}_l = p_r q_r = 1 - \tilde{q}_l \tilde{p}_l \tag{4.8}$$

and rearrange terms to obtain

$$[p_r w + \tilde{q}_l]\tilde{p}_l + [y]\tilde{r}_l = 1 \tag{4.9}$$

as required. Finally, if problem STR admits a solution, (4.1) and (4.2) are satisfied with the same w, hence, so are (4.7) and (4.9). To obtain the final coprimeness condition we subtract (4.1) from (4.2), obtaining

$$[y]\tilde{r}_l + [-x]\tilde{t}_l = 1 \tag{4.10}$$

and thereby completing the proof. ■

Note that the coprimeness condition of (4.10) essentially says that one cannot simultaneously track and reject the same signal, and is therefore a natural auxiliary condition to guarantee the simultaneous solvability of the tracking and disturbance rejection problems.

Note also that (4.1) is linear in the unknowns. It follows that if this equation is solvable the solution space will be a linear manifold in H which can be represented in the form $w = Lo + d$ where L is an appropriate linear operator on H and $o \in H$ becomes our new design parameter. The compensators which satisfy the constraints of problem ST thus take the form

$$C = [(Lo + d)p_l + \tilde{q}_r]^{-1}[-(Lo + d)\tilde{p}_l + q_r] \tag{4.11}$$

and similarly for problems SR and STR. Moreover, upon substituting the expression $w = Lo + d$ into (1.11) and (1.12), we obtain expressions for the resultant feedback system gains that are linear in the new design parameter o. Thus, all of the properties associated with the solution of problem S, given in [6], are retained by the solutions of

problems ST, SR, and STR of the present paper. It remains, however, to derive explicit necessary and sufficient conditions for the existence of solutions to these problems and to formulate explicit expressions for the parameterization of the resultant solution space.

Finally, we note that one normally desires a proper controller C. The obvious approach is to try to find the set

$$\{w \in H \mid wp_l + \tilde{q}_r \in I\};$$

if one then takes G to be a ring of proper transfer operators, one has obtained all proper stabilizing controllers. Unfortunately, in the setting of general rings, the above set may be empty. Therefore, rather than impose the additional structure necessary for a general theory, we indicate what can be done in the classical multivariable case. The additional condition needed in this case is quite weak; we merely require that for any constant matrix A, there is a $w \in H$ such that

$$w(\infty) = A.$$

If we now take G to be the ring of proper transfer matrices, it is easy to see that the above set is nonempty, as follows.

Since $\tilde{p}_r \in I$, $\tilde{p}_r^{-1}(\infty)$ exists. Then, from (2.19),

$$(\tilde{p}_r^{-1}(\infty) q_l(\infty)) p_l(\infty) + \tilde{q}_r(\infty) = \tilde{p}_r^{-1}(\infty).$$

Now, if we let

$$H_\infty = \{w \in H \mid w(\infty) = \tilde{p}_r^{-1}(\infty) q_l(\infty)\}$$

it follows that, for all $w \in H_\infty$, $wp_l + \tilde{q}_r \in I$. In fact, it is easy to show that $w \in H_\infty$ parameterizes all strictly proper controllers. In general, one can show that it is possible to find a controller with $C(\infty) = A$ if and only if $(p_l(\infty)A + \tilde{p}_l(\infty))^{-1}$ exists, and that the set of all such stabilizing controllers is parameterized by the set of all $w \in H_A$, where

$$H_A = \{w \in H \mid w(\infty) = (q_r(\infty) - \tilde{q}_r(\infty)A)(\tilde{p}_l(\infty) + p_l(\infty)A)\}.$$

Similar considerations apply to the parameterizations occurring in subsequent sections of the paper.

V. The Multivariate Case

In this section we consider the solution of the feedback system design equations, (4.1) and (4.2), for a class of generalized *multivariate systems* which includes most time-invariant feedback systems encountered in engineering practice. To this end we let K denote a commutative ring of complex valued functions defined on a complex manifold $\hat{K}$ which includes the constant function, 1, and we let L denote the subring of K composed of functions which are analytic on a submanifold $\hat{L}$. Normally, $\hat{K}$ is the complex plane and $\hat{L}$ is a half-plane or disk with K representing any of the standard spaces of single-variate transfer functions and L representing the subspace of transfer functions which are stable in an appropriate sense. More generally, multidimensional systems are included in our theory with $\hat{K}$ taken to be C^n and $\hat{L}$ an appropriate polydisk.

Using the above function spaces we formulate an axiomatic structure $\{G, H, I, J\}$ by letting $G = K^{n \times n}$ be the set of n by n matrices whose entries are elements of K and $H = L^{n \times n}$ be the set of n by n matrices whose entries are elements of L. Of course, I and J may be constructed from G and H in a natural manner. Finally, we may embed L into $H = L^{n \times n}$ by identifying $l \in L$ with the n by n matrix $l1$. Under this embedding each $l \in L$ commutes with every $g \in G = K^{n \times n}$ and $h \in H = L^{n \times n}$.

Although we will normally work with general fractional representations in $\{G, H, I, J\}$, in order to parameterize our solution spaces we need to work with a fractional representation $M = m\tilde{m}^{-1} = \tilde{m}^{-1}m$ where $m \in L^{n \times n}$ and $\tilde{m} \in L$. Such fractional representations naturally model the situation where the denominator of a fractional representation is just the common denominator of each entry in an n by n matrix. Since L commutes with H, m and $\tilde{m}$ define both left and right fractional representations for M. Moreover, a standard function space argument using the usual rank coprimeness for matrices will reveal that $M = m\tilde{m}^{-1}$ is right coprime if, and only if, $M = \tilde{m}^{-1}m$ is left coprime. Therefore, we may say that m and $\tilde{m}$ are coprime without a qualifier and assume a doubly coprime representation

$$\begin{bmatrix} \tilde{n}_r & n_r \\ -m & \tilde{m} \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{m} & -n_l \\ m & \tilde{n}_l \end{bmatrix} \tag{5.1}$$

without loss of generality. Note, since $H = L^{n \times n}$ is noncommutative we may have $n_r \neq n_l$ and/or $\tilde{n}_r \neq \tilde{n}_l$.

In this generalized multivariate setting we would like to derive explicit solutions of the feedback system design problems ST, SR, and STR. To this end we assume that the plant P admits a doubly coprime fractional representation in our multivariate setting while the reference generators T and R are assumed to admit doubly coprime fractional representations

$$\begin{bmatrix} \tilde{u}_r & u_r \\ -t & \tilde{t} \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{t} & -u_l \\ t & \tilde{u}_l \end{bmatrix} \tag{5.2}$$

and

$$\begin{bmatrix} \tilde{s}_r & s_r \\ -r & \tilde{r} \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{r} & -s_l \\ r & \tilde{s}_l \end{bmatrix} \tag{5.3}$$

where $\tilde{t}$ and $\tilde{r}$ lie in L.

This assumption represents a restriction on our system to the effect that the coprime denominator of the transfer matrix T is a common denominator for the elements of T, and similarly for R [see (3)]. Since the signals to be tracked or rejected are by definition the response of these operators to a delta-function, our situation is clearly analogous to that in which the signals are generated as the

zero-input response of a set of ordinary differential equations.

A. Tracking

With the above restriction on the matrix T, we now proceed to parameterize the set of all controllers which solve the problem ST, i.e., those which simultaneously stabilize the system and cause it to track the impulse response of T. As in (6) our approach will be to find a particular solution to (4.1), and then to find the general solution of the corresponding homogeneous equation. Lastly, we show that these are the only solutions. In all three steps, extensive use is made of the various coprimeness conditions, and of the fact that $\tilde{t}$ commutes with everything.

First we observe that the tracking problem admits a solution if and only if p_r and $\tilde{t}$ are coprime [see (2), (4), (13)]. The necessity of this condition follows from the skew coprimeness condition of Corollary 1-1) since $\tilde{t} \in L$. Conversely, if p_r and $\tilde{t}$ are coprime, then by the discussion preceding equation (5.1), there exist j_l, j_r, $\tilde{j}_l$, and $\tilde{j}_r$ in H such that

$$\begin{bmatrix} \tilde{j}_r & j_r \\ -p_r & \tilde{t} \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{t} & -j_l \\ p_r & \tilde{j}_l \end{bmatrix}. \tag{5.4}$$

Now, by invoking (2.20) and the fact that $\tilde{t}$ commutes with H we have

$$\begin{aligned} p_r q_r - 1 &= -\tilde{q}_l \tilde{p}_l = -\left(p_r j_l + \tilde{t}\tilde{j}_l\right)\tilde{q}_l \tilde{p}_l \\ &= p_r\left[-j_l \tilde{q}_l\right]\tilde{p}_l + \left[-\tilde{j}_l \tilde{q}_l \tilde{p}_l\right]\tilde{t} \\ &= p_r\left[w_p\right]\tilde{p}_l + \left[x_p\right]\tilde{t} \end{aligned} \tag{5.5}$$

showing that

$$w_p = -j_l \tilde{q}_l \quad \text{and} \quad x_p = -\tilde{j}_l \tilde{q}_l \tilde{p}_l \tag{5.6}$$

are elements of H which satisfy (4.1).

To parameterize the solution space of (4.1) we assume that $E = \tilde{p}_l \tilde{t}^{-1}$ admits a doubly coprime fractional representation

$$\begin{bmatrix} \tilde{j}_r & f_r \\ -e_l & \tilde{e}_l \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{e}_r & -f_l \\ e_r & \tilde{f}_l \end{bmatrix}. \tag{5.7}$$

Then

$$\tilde{p}_l \tilde{t}^{-1} = \tilde{e}_l^{-1} e_l \tag{5.8}$$

and so

$$\tilde{e}_l \tilde{p}_l = e_l \tilde{t}. \tag{5.9}$$

Therefore, for any $v \in H = L^{n \times n}$,

$$p_r\left[v\tilde{e}_l\right]\tilde{p}_l + \left[-p_r v e_l\right]\tilde{t} = p_r v e_l \tilde{t} - p_r v e_l \tilde{t} = 0 \tag{5.10}$$

verifying that

$$w_h = v\tilde{e}_l \quad \text{and} \quad x_h = -p_r v e_l \tag{5.11}$$

are solutions to the homogeneous equation corresponding to (4.1).

To verify that (5.11) represents all homogeneous solutions to (4.1) we consider arbitrary homogeneous solutions w_h' and x_h' in H satisfying

$$p_r w_h' \tilde{p}_l + x_h' \tilde{t} = 0. \tag{5.12}$$

Now, let $v' = w_h' \tilde{e}_l^{-1}$, which yields

$$w_h' = v'\tilde{e}_l \tag{5.13}$$

and

$$x_h' = -p_r w_h' \tilde{p}_l \tilde{t}^{-1} = -p_r w_h' \tilde{e}_l^{-1} e_l = -p_r v' e_l \tag{5.14}$$

verifying that w_h' and x_h' are of the required form. It remains to show that $v' \in H$. To this end consider the string of equalities in which the coprimeness conditions of (5.4) and (5.7) are invoked along with the commutivity of $\tilde{t}$.

$$\begin{aligned} v' &= w_h' \tilde{e}_l^{-1} = w_h' \tilde{e}_l^{-1}\left(e_l f_l + \tilde{e}_l \tilde{f}_l\right) \\ &= w_h' \tilde{f}_l + \left(j_r p_r + \tilde{j}_r \tilde{t}\right) w_h' \tilde{e}_l^{-1} e_l f_l \\ &= w_h' \tilde{f}_l - j_r x_h' f_l + \tilde{j}_r \tilde{t} w_h' \tilde{e}_l^{-1} e_l f_l \\ &= w_h' \tilde{f}_l - j_r x_h' f_l + \tilde{j}_r w_h' \tilde{e}_l^{-1} e_l \tilde{t} \tilde{f}_l \\ &= w_h' \tilde{f}_l - j_r x_h' f_l + \tilde{j}_r w_h' \tilde{e}_l^{-1} \tilde{e}_l \tilde{p}_l f_l \\ &= w_h' \tilde{f}_l - j_r x_h' f_l + \tilde{j}_r w_h' \tilde{p}_l f_l \in H. \end{aligned} \tag{5.15}$$

It follows that (5.11) represents the set of all homogeneous solutions to (4.1) and the desired parameterization of the solution space is given by

$$w = -j_l \tilde{q}_l + v\tilde{e}_l \quad \text{and} \quad x = -\tilde{j}_l \tilde{q}_l \tilde{p}_l - p_r v e_l. \tag{5.16}$$

Summarizing the above development we have the following theorem.

Theorem 3: For the multivariate feedback system of Fig. 1 let P, T, and E be characterized by the doubly coprime fractional representations of (2.12), (5.2), and (5.7), with $\tilde{t} \in L$. Then *problem ST admits a solution if, and only if, p_r and $\tilde{t}$ are coprime*, in which case the set of compensators that satisfy the constraints of problem ST is given by

$$C = \left[\left(-j_l \tilde{q}_l + v\tilde{e}_l\right)p_l + \tilde{q}_r\right]^{-1}\left[-\left(-j_l \tilde{q}_l + v\tilde{e}_l\right)\tilde{p}_l + q_r\right] \tag{5.17}$$

where j_l and $\tilde{e}_l$ are defined by (5.4) and (5.7), respectively, and v is an arbitrary element of H such that the denominator of C is in I. Moreover, the feedback system gains resulting from the use of such a compensator are linear (affine) in the design parameter v. ■

Finally, since $\tilde{e}_l \in I$, considerations similar to those at the end of Section IV show that the set of proper controllers is nonempty.

B. Disturbance Rejection

The derivation of existence conditions for the solution of problem SR and the parameterization of the resultant solution space parallels the above derived solution to problem ST. For this reason only a partial sketch will be given here. First, we observe that the disturbance rejection problem admits a solution if, and only if, $\tilde{p}_l$ and $\tilde{r}$ are coprime. The necessity follows from condition 2) of Corollary 1. Conversely, if $\tilde{p}_l$ and $\tilde{r}$ are coprime there exist m_l, m_r, $\tilde{m}_l$, and $\tilde{m}_r$ in H such that

$$\begin{bmatrix} \tilde{m}_r & m_r \\ -\tilde{p}_l & \tilde{r} \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{r} & -m_l \\ \tilde{p}_l & \tilde{m}_l \end{bmatrix}. \tag{5.18}$$

Thus,

$$\begin{aligned} p_r q_r &= p_r q_r(\tilde{m}_r \tilde{r} + m_r \tilde{p}_l) \\ &= p_r[q_r m_r]\tilde{p}_l + [p_r q_r \tilde{m}_r]\tilde{r} \end{aligned} \tag{5.19}$$

verifying that

$$w_p = q_r m_r \quad \text{and} \quad y_p = p_r q_r \tilde{m}_r \tag{5.20}$$

satisfy (4.2).

Next we desire to parameterize the solution space of (4.2) when $\tilde{p}_l$ and $\tilde{r}$ are coprime. To this end we assume that $A = \tilde{r}^{-1} p_r$ admits a doubly coprime fractional representation

$$\begin{bmatrix} \tilde{b}_r & b_r \\ -a_l & \tilde{a}_l \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{a}_r & -b_l \\ a_r & \tilde{b}_l \end{bmatrix}. \tag{5.21}$$

Now, since

$$\tilde{r}^{-1} p_r = a_r \tilde{a}_r^{-1} \tag{5.22}$$

by an argument similar to that employed in the solution of the tracking problem we find that the homogeneous solutions to (4.2) are given by

$$w_h = \tilde{a}_r z \quad \text{and} \quad y_h = -a_r z \tilde{p}_l \tag{5.23}$$

and hence the desired parameterization of the solution space for 4.2 takes the form

$$w = q_r m_r + \tilde{a}_r z \tag{5.24}$$

and

$$y = p_r q_r \tilde{m}_r - a_r z \tilde{p} \tag{5.25}$$

where z is an arbitrary element of H.

Theorem 4: For the multivariate feedback system of Fig. 1 let P, R, and A be characterized by the doubly coprime fractional representations of (2.12), (5.3), and (5.21) with $\tilde{r} \in L$. Then, *problem SR admits a solution if, and only if, $\tilde{p}_l$ and $\tilde{r}$ are coprime*, in which case the set of compensators which satisfy the constraints of problem SR is given by

$$C = [(q_r m_r + \tilde{a}_r z)p_l + \tilde{q}_r]^{-1}[-(q_r m_r + \tilde{a}_r z)\tilde{p}_l + q_r] \tag{5.26}$$

where m_r and $\tilde{a}_r$ are defined by (5.18) and (5.21), respectively, and z is an arbitrary element of H such that the denominator of C is in I. Moreover, the feedback system gains resulting from the use of such a compensator are linear (affine) in the design parameter z. ■

Again, since $\tilde{a}_r \in I$ the considerations at the end of Section IV are applicable.

C. Simultaneous Tracking and Disturbance Rejection

To obtain a solution of problem STR in our multivariate setting we must find a simultaneous solution to (4.1) and (4.2), using the same w, hence, the same compensator. Since we already have a complete parameterization of the solution spaces for these equations taken individually, the solution of the simultaneous problem reduces to finding values for the arbitrary parameters, v and z, which yield the same w. In short, upon combining (5.16) and (5.25), the solution to problem STR may be obtained by solving the linear equation

$$q_r m_r + \tilde{a}_r z = -j_l \tilde{q}_l + v \tilde{e}_l \tag{5.27}$$

for z and v in H. From corollary 1-3) we have the necessary condition that p_r and $\tilde{t}$, $\tilde{p}_l$ and $\tilde{r}$, and $\tilde{t}$ and $\tilde{r}$ all be coprime. Moreover, since $\tilde{t}$ and $\tilde{r}$ are both in L, the coprimeness condition for $\tilde{t}$ and $\tilde{r}$ may be formulated in terms of elements $\hat{t}$ and $\hat{r}$ in L as follows.

Lemma 1: Assume that $\tilde{t}$ and $\tilde{r}$ are right coprime in the sense that there exist $\hat{t}_r$ and $\hat{r}_r$ in H such that $\hat{t}_r \tilde{t} + \hat{r}_r \tilde{r} = 1$. Then there exist $\hat{t}$ and $\hat{r}$ in L such that

$$\tilde{t}\hat{t} + \tilde{r}\hat{r} = 1. \tag{5.28}$$

Proof: Recalling that $\hat{t}_r$ and $\hat{r}_r$ are matrices, we let $\hat{t}$ and $\hat{r}$ be the (1, 1) entries in $\hat{t}_r$ and $\hat{r}_r$, respectively. ■

Recall that $\tilde{e}_l^{-1} e_l$ is a left coprime fractional representation of $\tilde{p}_l \tilde{t}^{-1} = \tilde{t}^{-1} \tilde{p}_l$; it follows that $\tilde{t} = k_l \tilde{e}_l$ for some $k_l \in H$. Similarly, $\tilde{r} = \tilde{a}_r n_r$ for some $n_r \in H$. We use this to verify the sufficiency of the three coprimeness conditions as follows.

We rewrite (5.27) in the form

$$v \tilde{e}_l - \tilde{a}_r z = q_r m_r + j_l \tilde{q}_l. \tag{5.29}$$

Now, starting with (5.28) and invoking the fact that L commutes with G we obtain

$$\begin{aligned} [q_r m_r + j_l \tilde{q}_l] &= [q_r m_r + j_l \tilde{q}_l]\tilde{t}\hat{t} + \tilde{r}\hat{r}[q_r m_r + j_l \tilde{q}_l] \\ &= [q_r m_r + j_l \tilde{q}_l]\hat{t} k_l \tilde{e}_l + \tilde{a}_r n_r \hat{r}[q_r m_r + j_l \tilde{q}_l] \\ &= v_p \tilde{e}_l - \tilde{a}_r z_p \end{aligned} \tag{5.30}$$

where

$$v_p = [q_r m_r + j_l \tilde{q}_l]\hat{t} k_l \quad \text{and} \quad z_p = -n_r \hat{r}[q_r m_r + j_l \tilde{q}_l] \tag{5.31}$$

verifying that our simultaneous equations admit a solution.

To parameterize the resultant solution space, we observe that

$$[\tilde{a}_r d]\tilde{e}_l - \tilde{a}_r[d\tilde{e}_l] = 0 \tag{5.32}$$

implying that

$$v_h = \tilde{a}_r d \quad \text{and} \quad z_h = d\tilde{e}_l \tag{5.33}$$

are homogeneous solutions to (5.27) for all $d \in H$. To verify that (5.33) represents all homogeneous solutions to (5.27) we consider an arbitrary pair of solutions v'_h and z'_h in H such that

$$v'_h \tilde{e}_l - \tilde{a}_r z'_h = 0. \tag{5.34}$$

We now let $d' = \tilde{a}_r^{-1} v'_h$, in which case

$$v'_h = \tilde{a}_r d' \tag{5.35}$$

and

$$z'_h = \tilde{a}_r^{-1} v'_h \tilde{e}_l = d'\tilde{e}_l, \tag{5.36}$$

verifying that v'_h and z'_h are the required form. It remains only to show that $d' \in H$ for which purpose we invoke the various coprimeness conditions and the commutativity of L with G as follows.

$$\begin{aligned}
d' &= \tilde{a}_r^{-1} v'_h = (b_r a_r + \tilde{b}_r \tilde{a}_r)\tilde{a}_r^{-1} v'_h \\
&= \tilde{b}_r v'_h + b_r a_r \tilde{a}_r^{-1} v'_h (\tilde{t}\tilde{t} + \hat{r}\hat{r}) \\
&= \tilde{b}_r v'_h + b_r a_r \tilde{a}_r^{-1} \hat{r} v'_h \hat{r} \\
&\quad + b_r a_r \tilde{a}_r^{-1} v'_h (e_l f_l + \tilde{e}_l \tilde{f}_l)\tilde{t}\tilde{t} \\
&= \tilde{b}_r v'_h + b_r a_r \tilde{a}_r^{-1} \tilde{a}_r n_r v'_h \hat{r} \\
&\quad + b_r a_r \tilde{a}_r^{-1} v'_h \tilde{e}_l \tilde{f}_l \tilde{t}\tilde{t} + b_r a_r \tilde{a}_r^{-1} v'_h e_l \tilde{t} \tilde{f}_l \hat{t} \\
&= \tilde{b}_r v'_h + b_r a_r n_r v'_h \hat{r} + b_r a_r z'_h \tilde{f}_l \tilde{t}\tilde{t} \\
&\quad + b_r a_r \tilde{a}_r^{-1} v'_h \tilde{e}_l \tilde{p}_l f_l \hat{t} \\
&= \tilde{b}_r v'_h + b_r a_r n_r v'_h \hat{r} + b_r a_r z'_h \tilde{f}_l \tilde{t}\tilde{t} \\
&\quad + b_r a_r z'_h \tilde{p}_l f_l \hat{t} \in H.
\end{aligned} \tag{5.37}$$

Consistent with the above, the required parameterization for the solution space of (5.29) is given by

$$v = [q_r m_r + j_l \tilde{q}_l]\hat{t} k_l + \tilde{a}_r d$$

and

$$z = -n_r \hat{r}[q_r m_r + j_l \tilde{q}_l] + d\tilde{e}_l \tag{5.38}$$

which upon substitution into (5.16) or (5.25) yields

$$\begin{aligned}
w &= q_r m_r + \tilde{a}_r\{-n_r \hat{r}[q_r m_r + j_l \tilde{q}_l] + d\tilde{e}_l\} \\
&= q_r m_r - \tilde{r}\hat{r}[q_r m_r + j_l \tilde{q}_l] + \tilde{a}_r d\tilde{e}_l \\
&= \tilde{t}\tilde{t}(q_r m_r) - \tilde{r}\hat{r}(j_l \tilde{q}_l) + \tilde{a}_r d\tilde{e}_l.
\end{aligned} \tag{5.39}$$

We have thus proven the following.

Theorem 5: For the multivariate feedback system of Fig. 1 let P, T, R, E, and A be characterized by the doubly coprime fractional representations of (2.12), (5.2), (5.3), (5.7), and (5.21) with $\hat{r}$ and $\hat{t}$ in L. Then, *problem STR* admits a solution if and only if p_r and $\tilde{t}$, $\tilde{p}_l$ and $\tilde{r}$, and $\tilde{t}$ and $\tilde{r}$ are coprime, *in which case the set of compensators which satisfy the constraints of problem STR is given by*

$$\begin{aligned}
C = &\left\{\left[\tilde{t}\tilde{t}(q_r m_r) - \tilde{r}\hat{r}(j_l \tilde{q}_l) + \tilde{a}_r d\tilde{e}_l\right] p_l + \tilde{q}_r\right\}^{-1} \\
&\cdot \left\{-\left[\tilde{t}\tilde{t}(q_r m_r) - \tilde{r}\hat{r}(j_l \tilde{q}_l) + \tilde{a}_r d\tilde{e}_l\right]\tilde{p}_l + q_r\right\}
\end{aligned} \tag{5.40}$$

where j_l, $\tilde{e}_l$, m_r, $\tilde{a}_r$, $\tilde{t}$, and $\hat{r}$ are defined by (5.4), (5.7), (5.18), (5.21), and (5.28), respectively, and d is an arbitrary element of H such that the denominator of C is in I. Moreover, the feedback system gains resulting from the use of such a compensator are linear (affine) in the design parameter d. ∎

Again, the considerations at the end of Section IV are applicable.

Although we have gone through some rather complex derivations in the preceding it should be noted that the only mathematics employed are addition, multiplication, subtraction, and inversion. Moreover, the results of these derivations are given in the form of explicit expressions for the compensators that satisfy the constraints of the three design problems.

Although the expression (5.40) in particular looks extremely complicated, most of the terms occurring in it are fixed transfer matrices; it can be rewritten in the form

$$C = [(g_0 + \tilde{a}_r d\tilde{e}_l)p_l + \tilde{q}_r]^{-1}[-(g_0 + \tilde{a}_r d\tilde{e}_l)\tilde{p}_l + q_r]$$

or even in the form

$$C = (g_1 d g_2 + g_3)^{-1}(g_4 d g_5 + g_6)$$

where all of the g_i are fixed transfer matrices.

VI. Examples

To minimize repetition, our first two examples are continuations of examples begun in [6], where the coprime fractional representations employed below are computed.

A. A Single-Variate Tracking and Disturbance Rejection Problem

Although the single-variate case is already well understood, the theory is most readily illustrated in this case, and hence we begin with a single variate example. Here, the plant is taken to be

$$\begin{aligned}
P(s) &= \left[\frac{(s+1)}{(s^2-4)}\right] = \left[\frac{(s+1)}{(s+2)^2}\right]\left[\frac{(s-2)}{(s+2)}\right]^{-1} \\
&= p(s)\tilde{p}(s)^{-1}
\end{aligned} \tag{6.1}$$

where

$$\begin{aligned}
&\left[\frac{16}{3}\right]\left[\frac{(s+1)}{(s+2)^2}\right] + \left[\frac{(s+2/3)}{(s+2)}\right]\left[\frac{(s-2)}{(s+2)}\right] \\
&\qquad = q(s)p(s) + \tilde{q}(s)\tilde{p}(s) = 1.
\end{aligned} \tag{6.2}$$

Of course, since the single-variate problem is commutative these q's and p's yield a doubly coprime fractional representation in $\{G, H, I, J\}$. Here G is taken to be the ring of proper rational functions and H is the ring of rational functions whose poles have a real part which is less than -1. Thus, the theory will yield "strongly stable" systems in the sense that their poles will be bounded away from the imaginary axis. Moreover, since the theory uses the same stability concept for the tracking and disturbance rejection problems as for stabilization, the resultant solutions to these problems will be "strongly asymptotic" in the same sense.

Now, let us consider the problem of tracking a step function. Here we let

$$T(s)=\left[\frac{1}{s}\right]=\left[\frac{1}{(s+2)}\right]\left[\frac{s}{(s+2)}\right]^{-1}$$
$$=t(s)\tilde{t}(s)^{-1} \tag{6.3}$$

where

$$\left[\frac{4(s+1)}{(s+2)}\right]\left[\frac{1}{(s+2)}\right]+\left[\frac{s}{(s+2)}\right]\left[\frac{s}{(s+2)}\right]$$
$$=u(s)t(s)+\tilde{u}(s)\tilde{t}(s)=1. \tag{6.4}$$

Moreover,

$$[4]\left[\frac{(s+1)}{(s+2)^2}\right]+\left[\frac{s}{(s+2)}\right]\left[\frac{s}{(s+2)}\right]$$
$$=j(s)p(s)+\tilde{j}(s)\tilde{t}(s)=1 \tag{6.5}$$

verifying that $p(s)$ and $\tilde{t}(s)$ are coprime and, hence, assuring the existence of a solution to our tracking problem. To compute the solution we define

$$E(s)=\tilde{p}(s)\tilde{t}(s)^{-1}=\left[\frac{(s-2)}{(s+2)}\right]\left[\frac{s}{(s+2)}\right]^{-1}$$
$$=e(s)\tilde{e}(s)^{-1} \tag{6.6}$$

where

$$[-1]\left[\frac{(s-2)}{(s+2)}\right]+[2]\left[\frac{s}{(s+2)}\right]$$
$$=f(s)e(s)+\tilde{f}(s)\tilde{e}(s)=1. \tag{6.7}$$

From 5.16 we then have

$$w(s)=-j(s)\tilde{q}(s)+v(s)\tilde{e}(s)$$
$$=\left[-4\frac{(s+2/3)}{(s+2)}\right]+\left[\frac{s}{(s+2)}\right]v(s). \tag{6.8}$$

Substituting this value of $w(s)$ into (1.11) and (1.12) we obtain the system gains

$$h_{y_p}(s)=\left[\frac{(s+1)}{(s+2)^4}\right]\left[\left(\frac{28}{3}s^2+16s+16\right)-(s-2)sv(s)\right] \tag{6.9}$$

and

$$h_{e_p}(s)=\left[\frac{s}{(s+2)^4}\right]\left[s(s-2)(s+2/3)+(s+1)(s-2)v(s)\right] \tag{6.10}$$

in terms of the design parameter $v(s)$. Note the zero at zero in $h_{e_p}(s)$ assures that the system will track a step function as required.

Now, let us consider an alternative tracking problem where we are required to track $e^{2t}U(t)$. That is,

$$T(s)=\left[\frac{1}{(s-2)}\right]=\left[\frac{1}{(s+2)}\right]\left[\frac{(s-2)}{(s+2)}\right]^{-1}$$
$$=t(s)\tilde{t}(s)^{-1} \tag{6.11}$$

where

$$[4]\left[\frac{1}{(s+2)}\right]+[1]\left[\frac{(s-2)}{(s+2)}\right]$$
$$=u(s)t(j)+\tilde{u}(s)\tilde{t}(s) \tag{6.12}$$

and

$$\left[\frac{16}{3}\right]\left[\frac{(s+1)}{(s+2)^2}\right]+\left[\frac{(s+2/3)}{(s+2)}\right]\left[\frac{(s-2)}{(s+2)}\right]$$
$$=j(s)p(s)+\tilde{j}(s)\tilde{t}(s)=1. \tag{6.13}$$

Thus, $p(s)$ and $\tilde{t}(s)$ are coprime and we may proceed to construct the desired set of compensators. For this purpose we define a system $E(s)$ by

$$E(s)=\tilde{p}(s)\tilde{t}(s)^{-1}=\left[\frac{(s-2)}{(s+2)}\right]\left[\frac{(s-2)}{(s+2)}\right]^{-1}$$
$$=1=[1][1]^{-1}=e(s)\tilde{e}(s)^{-1} \tag{6.14}$$

where

$$=[0][1]+[1][1]=f(s)e(s)+\tilde{f}(s)\tilde{e}(s)=1. \tag{6.15}$$

Note that in this example $\tilde{p}(s)$ and $\tilde{t}(s)$ are not coprime. This is, however, not necessary as long as we can construct a coprime fractional representation for their ratio, $E(s)$. Finally, substitution into (5.16) yields

$$w(s)=-\left[\frac{16(s+2/3)}{3(s+2)}\right]+v(s) \tag{6.16}$$

while

$$h_{y_p}(s)=\left[\frac{(s+1)}{(s+2)^4}\right]\left[\frac{32}{3}\left(s^2+\frac{4}{3}s+\frac{4}{3}\right)-(s-2)(s+2)v(s)\right] \tag{6.17}$$

and

$$h_{e_p}(s)=\frac{(s-2)}{9(s+2)^4}\cdot\left[(9s^3-6s^2-20s-8)+9(s+1)(s+2)v(s)\right]. \quad (6.18)$$

As before, the $(s-2)$ factor in the numerator of the $h_{e_p}(s)$ verifies the tracking property. Also note that as $v(s)$ spans the set of "strongly stable" functions so does $w(s)$. Thus, every compensator that stabilizes the system in our sense also solves the tracking problem. Although redundant, the extra term in (6.16) complicates our expression for $C(s)$, $h_{y_p}(s)$ and $h_{e_p}(s)$ and should be eliminated if possible. In our theory this would be achieved by choosing more opportune fractional representations for the various functions with which we deal.

Now, let us consider the problem of rejecting a step function for which we let

$$R(s)=\left[\frac{1}{s}\right]=\left[\frac{1}{(s+2)}\right]\left[\frac{s}{(s+2)}\right]^{-1}=r(s)\tilde{r}(s)^{-1} \quad (6.19)$$

where

$$\left[\frac{4(s+1)}{(s+2)}\right]\left[\frac{1}{(s+2)}\right]+\left[\frac{s}{(s+2)}\right]\left[\frac{s}{(s+2)}\right]=s(s)r(s)+\tilde{s}(s)\tilde{r}(s)=1. \quad (6.20)$$

Here,

$$[2]\left[\frac{s}{(s+2)}\right]+[-1]\left[\frac{(s-2)}{(s+2)}\right]=\tilde{m}(s)\tilde{r}(s)+m(s)\tilde{p}(s)=1 \quad (6.21)$$

showing that $\tilde{p}(s)$ and $\tilde{r}(s)$ are coprime. Next we compute

$$A(s)=p(s)\tilde{r}(s)^{-1}=\left[\frac{(s+1)}{(s+2)^2}\right]\left[\frac{s}{(s+2)}\right]^{-1}=a(s)\tilde{a}(s)^{-1} \quad (6.22)$$

and

$$[4]\left[\frac{(s+1)}{(s+2)^2}\right]+\left[\frac{s}{(s+2)}\right]\left[\frac{s}{(s+2)}\right]=b(s)a(s)+\tilde{b}(s)\tilde{a}(s)=1 \quad (6.2$$

obtaining

$$w(s)=q(s)m(s)+\tilde{a}(s)z(s)=\left[-\frac{16}{3}\right]+\left[\frac{s}{(s+2)}\right]z(s) \quad (6.24)$$

$$h_{y_p}(s)=\frac{s(s+1)}{(s+2)^4}\left[\frac{32}{3}(s+2)-(s-2)z(s)\right] \quad (6.25)$$

and

$$h_{e_p}(s)=\frac{(s-2)}{3(s+2)^4}\left[(s+2)(3s^2-8s-12)+3s(s+1)z(s)\right]. \quad (6.26)$$

Here, the factor of s in the numerator of $h_{y_p}(s)$ indicates that the feedback system will, indeed, reject step functions. Moreover, as we have previously indicated, any stabilizing controller tracks $e^{2t}U(t)$ as is indicated by the $(s-2)$ factor in the numerator of $h_{e_p}(s)$. Interestingly, if we apply our theory to solve this simultaneous tracking and rejection problem we obtain an equivalent, but more complex, parameterization of the desired feedback systems. To this end we let

$$[2]\left[\frac{s}{(s+2)}\right]+[-1]\left[\frac{(s-2)}{(s+2)}\right]=\hat{r}(s)\tilde{r}(s)+\hat{t}(s)\tilde{t}(s)=1 \quad (6.27)$$

and obtain

$$w(s)=\tilde{t}(s)\hat{t}(s)q(s)m(s)-\tilde{r}(s)\hat{r}(s)j(s)\tilde{q}(s)+\tilde{a}(s)d(s)\tilde{e}(s)=\left[\frac{1}{(s+2)^2}\right]\left[-\frac{16}{3}\left(s^2+\frac{4}{3}s+4\right)+s(s+2)d(s)\right] \quad (6.28)$$

$$h_{y_p}(s)=\frac{(s+1)s}{(s+2)^5}\left[\frac{32}{3}\left(s^2+\frac{8}{3}s+\frac{20}{3}-(s-2)(s+2)d(s)\right]\right. \quad (6.29)$$

and

$$h_{e_p}(s)=\left[\frac{(s-2)}{9(s+2)^5}\right]\left[(9s^4+12s^3+32s^2-112s-144)+9s(s+1)(s+2)d(s)\right]. \quad (6.30)$$

Equations (6.28) through (6.30) represent a complete parameterization of the simultaneous solutions to the tracking and disturbance rejection problem where $d(s)$ is a design parameteer which may be chosen to "optimize" some other aspect of the feedback system design. For instance, if we would like to create an additional zero of $h_{y_p}(s)$ at $s=1$ we let $d\simeq-36.74$ and obtain

$$H_{y_p}(s)=\frac{s(s+1)}{27(s+2)^5}(1280s^2+768s-2048). \quad (6.31)$$

B. A Multivariate Lumped-Distributed Disturbance Rejection Problem

We now consider a lumped/distributed multivariate plant

$$P(s)=\begin{bmatrix}\frac{e^{-1/s}}{(s+1)} & \frac{(s-1)}{(s+1)}\\ 0 & \frac{1}{(s-1)}\end{bmatrix}. \quad (6.32)$$

which is included in our axiomatic theory by taking G to be the ring of 2 by 2 matrices whose entries are L_∞ functions on the imaginary axis and H to be the subring of 2 by 2 matrices whose entries are H_∞ functions on the imaginary axis. We would like to design a compensator for this plant which will simultaneously stabilize the feedback system and cause it to reject the sinusoidal input $\sin(t)U(t)$. To this end we obtain a doubly coprime fractional representation

$$\begin{bmatrix} \tilde{q}_r & q_r \\ -p_l & \tilde{p}_l \end{bmatrix}^{-1} = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 \\ \hline -\dfrac{e^{1/s}}{(s+1)} & -\dfrac{(s-1)}{(s+1)} & 1 & 0 \\ 0 & -\dfrac{1}{(s+1)} & 0 & \dfrac{(s-1)}{(s+1)} \end{array}\right]^{-1}$$

$$= \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & \dfrac{(s-1)}{(s+1)} & 0 & -2 \\ \hline \dfrac{e^{-1/s}}{(s+1)} & \dfrac{(s-1)^2}{(s+1)^2} & 1 & \dfrac{2(s-1)}{(s+1)} \\ 0 & \dfrac{1}{(s+1)} & 0 & 1 \end{array}\right]$$

$$= \begin{bmatrix} \tilde{p}_r & -q_l \\ p_r & \tilde{q}_l \end{bmatrix} \tag{6.33}$$

and we let

$$R(s) = \begin{bmatrix} \dfrac{1}{(s^2+1)} & 0 \\ 0 & \dfrac{1}{(s^2+1)} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{1}{(s+1)^2} & 0 \\ 0 & \dfrac{1}{(s+1)^2} \end{bmatrix} \begin{bmatrix} \dfrac{(s^2+1)}{(s+1)^2} & 0 \\ 0 & \dfrac{(s^2+1)}{(s+1)^2} \end{bmatrix}^{-1}$$

$$= r(s)\tilde{r}(s)^{-1} \tag{6.34}$$

where

Of course, since the r's and s's are all commutative (6.35) defines a doubly coprime fractional representation for $R(s)$.

To solve the disturbance rejection problem we must now verify that $\tilde{p}_l$ and $\tilde{r}$ are coprime, for which we have

$$\begin{bmatrix} 1 & 0 \\ 0 & \dfrac{2}{(s+1)^2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{(s-1)}{(s+1)} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \dfrac{(s+3)}{(s+1)} \end{bmatrix} \begin{bmatrix} \dfrac{(s^2+1)}{(s+1)^2} & 0 \\ 0 & \dfrac{(s^2+1)}{(s+1)^2} \end{bmatrix}$$

$$= m(s)\tilde{p}_l(s) + \tilde{m}(s)\tilde{r}(s) = 1. \tag{6.36}$$

As before, these matrices are all commutative and thus define a doubly coprime fractional representation for our problem. We then define

$$A(s) = p_r(s)\tilde{r}(s)^{-1} = \begin{bmatrix} \dfrac{e^{-1/s}(s+1)}{(s^2+1)} & \dfrac{(s-1)^2}{(s^2+1)} \\ 0 & \dfrac{(s+1)}{(s^2+1)} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{e^{-1/s}}{(s+1)} & \dfrac{(s-1)^2}{(s+1)^2} \\ 0 & \dfrac{1}{(s+1)} \end{bmatrix} \begin{bmatrix} \dfrac{(s^2+1)}{(s+1)^2} & 0 \\ 0 & \dfrac{(s^2+1)}{(s+1)^2} \end{bmatrix}^{-1}$$

$$= a_r(s)\tilde{a}_r(s)^{-1} \tag{6.37}$$

which is right coprime. Of course, we can also formulate a left coprime representation for $A(s)$ along with the appropriate b's required to construct a doubly coprime fractional representation. For the present purpose, however, all that is required is $\tilde{a}_r(s)$ and hence we will not derive the remaining a's and b's. Substituting q_r, m, and $\tilde{a}_r$ into (5.25) now yields the required set of w's for the solution of problem SR for our lumped-distributed multi-

$$\begin{bmatrix} \dfrac{2(s-1)}{(s+1)} & 0 \\ 0 & \dfrac{2(s-1)}{(s+1)} \end{bmatrix} \begin{bmatrix} \dfrac{1}{(s+1)^2} & 0 \\ 0 & \dfrac{1}{(s+1)^2} \end{bmatrix} + \begin{bmatrix} \dfrac{(s+3)}{(s+1)} & 0 \\ 0 & \dfrac{(s+3)}{(s+1)} \end{bmatrix} \begin{bmatrix} \dfrac{(s^2+1)}{(s+1)^2} & 0 \\ 0 & \dfrac{(s^2+1)}{(s+1)^2} \end{bmatrix}$$

$$= s(s)r(s) + \tilde{s}(s)\tilde{r}(s) = 1. \tag{6.35}$$

variate system.

$$w(s)=q_r(s)m(s)+\bar{a}_r(s)z(s)$$

$$=\begin{bmatrix} 0 & 0 \\ 0 & \dfrac{4}{(s+1)^2} \end{bmatrix}+\begin{bmatrix} \dfrac{(s^2+1)}{(s+1)^2} & 0 \\ 0 & \dfrac{(s^2+1)}{(s+1)^2} \end{bmatrix} z(s) \tag{6.38}$$

where $z(s)$ is an arbitrary stable matrix. Finally, substitution of this $w(s)$ into 1.10 yields the required compensator while the input/output gain for the resultant feedback system takes the form

$$h_{r\mu}(s)=\begin{bmatrix} 0 & \dfrac{2(s^2+1)(s+3)(s-1)^2}{(s+1)^5} \\ 0 & \dfrac{2(s^2+1)(s+3)}{(s+1)^4} \end{bmatrix}$$

$$-\begin{bmatrix} \dfrac{e^{-1/s}}{(s+1)} & \dfrac{(s-1)^2}{(s+1)^2} \\ 0 & \dfrac{1}{(s+1)} \end{bmatrix}\begin{bmatrix} z_{11}(s) & z_{12}(s) \\ z_{21}(s) & z_{22}(s) \end{bmatrix}$$

$$\cdot\begin{bmatrix} \dfrac{(s^2+1)}{(s+1)^2} & 0 \\ 0 & \dfrac{(s^2+1)(s-1)}{(s+1)^3} \end{bmatrix}. \tag{6.39}$$

Note that the factor (s^2+1), in the numerator of $h_{r\mu}(s)$ implies that the feedback system will reject the required sinusoid, while choosing $z_{21}(s)=0$ will preserve the desirable triangular nature of the plant.

C. A Periodically Varying Discrete-Time Tracking Problem

Although time-varying systems are not traditionally viewed as multivariate, the class of periodically varying discrete time systems admits a frequency domain theory which closely resembles the classical multivariate theory. In fact, we can apply the results of Section V to this class of time-varying systems. Although it is not well known in this system theory community, the frequency domain theory for periodically varying discrete-time systems has been rediscovered by a number of researchers over the past quarter century in one form or another [5], [8]–[11]. The basic theory, however, always employs an n by n matrix of z-transforms to model a single-variate system of period n. For the present example we will take $n=2$ in which case a single-variate system is modeled by a transfer function in

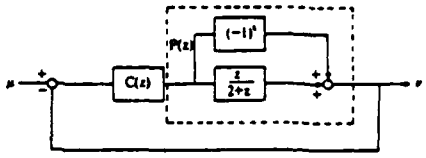


Fig. 3. Periodically varying feedback system.

the ring G, composed of 2 by 2 matrices of rational functions which have the form

$$P(z)=\begin{bmatrix} P_1(z^2)+zP_2(z^2) & P_3(z^2)-zP_4(z^2) \\ P_3(z^2)+zP_4(z^2) & P_1(z^2)-zP_2(z^2) \end{bmatrix} \tag{6.40}$$

where P_i, $i=1, 2, 3, 4$, are rational functions. As usual, the stable systems H are taken to be the subring of G composed of functions which are analytic inside the unit circle of the complex plane. All of the constituent components of a discrete-time system of period 2 can be modeled by such matrices and the usual operational calculus for interconnected systems remains valid [5], [8]–[11]. For instance, a constant scale factor with gain k is modeled by the matrix

$$P(z)=\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} \tag{6.41}$$

while a single-variate time-invariant component is modeled by

$$P(z)=\begin{bmatrix} T(z) & 0 \\ 0 & T(z)_* \end{bmatrix} \tag{6.42}$$

where $T(z)$ is the usual transfer function for the component and $T(z)_*=T(-z)$. Finally, the periodically varying multiplier defined by

$$y_k=(-1)^k u_k \tag{6.43}$$

is modeled by

$$P(z)=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{6.44}$$

Now, consider the periodically varying feedback system of Fig. 3. Here, the plant is a time-invariant system with a parallel time-varying gain and it is thus modeled by the transfer function matrix

$$P(z)=\begin{bmatrix} \dfrac{z}{(2+z)} & 1 \\ 1 & \dfrac{-z}{(2-z)} \end{bmatrix}. \tag{6.45}$$

Since this system is stable (its poles are located at $z=\pm 2$), we may obtain a doubly coprime representation for $P(z)$ by letting $p(z)=P(z)$, $\bar{p}(z)=1$, $q(z)=0$, and $\bar{q}(z)=1$. Since these matrices are all mutually commutative $p(z)$

may serve for both $p_r(z)$ and $p_l(z)$ and similarly for the remaining p's and q's.

Let us now consider the problem of designing a compensator for this plant which will cause it to track a prescribed input without destabilizing the system. For this purpose we take

$$T(z)=\begin{bmatrix} \dfrac{1}{(1-z^2)^2} & 0 \\ 0 & \dfrac{1}{(1-z^2)^2} \end{bmatrix}, \tag{6.46}$$

which generates an input which oscillates between zero and a ramp. Here $T(z)$ is unstable (since it has poles at ± 1) but its inverse is stable. Therefore, we may form a doubly coprime fractional representation for $T(z)$ with $t(z)=1$,

$$\tilde{t}(z)=\begin{bmatrix} (1-z^2)^2 & 0 \\ 0 & (1-z^2)^2 \end{bmatrix} \tag{6.47}$$

$s(z)=1$, and $\tilde{s}(z)=0$. Now

$$\begin{bmatrix} \dfrac{z}{(2+z)} & 1 \\ 1 & \dfrac{-z}{(2-z)} \end{bmatrix} \cdot \begin{bmatrix} \dfrac{z(15-6z^2)}{4(4-z^2)(2-z)} & \dfrac{(15-6z^2)}{4(4-z^2)} \\ \dfrac{(15-6z^2)}{4(4-z^2)} & \dfrac{-z(15-6z^2)}{4(4-z^2)(2+z)} \end{bmatrix} + \begin{bmatrix} (1-z^2)^2 & 0 \\ 0 & (1-z^2)^2 \end{bmatrix} \begin{bmatrix} \dfrac{1}{(4-z^2)^2} & 0 \\ 0 & \dfrac{1}{(4-z^2)^2} \end{bmatrix}$$

$$=p(z)j_l(z)+\tilde{t}(z)\tilde{j}_l(z)=1 \tag{6.48}$$

and hence $p(z)$ and $\tilde{t}(z)$ are coprime, verifying the existence of a solution to our tracking problem. The final step required to obtain the desired compensator is to compute

$$E(z)=\tilde{p}(z)\tilde{t}(z)^{-1}=\begin{bmatrix} \dfrac{1}{(1-z^2)^2} & 0 \\ 0 & \dfrac{1}{(1-z^2)^2} \end{bmatrix}. \tag{6.49}$$

Since $E(z)$ coincides with $T(z)$ we may let $e(z)=t(z)$, $\tilde{e}(z)=\tilde{t}(z)$, $f(z)=s(z)$, and $\tilde{f}(z)=\tilde{s}(z)$ define our doubly coprime fractional representation for $E(z)$. Substituting into (5.16) we obtain

$$w(z)=-\begin{bmatrix} \dfrac{z(15-6z^2)}{4(4-z^2)(2-z)} & \dfrac{(15-6z^2)}{4(4-z^2)} \\ \dfrac{(15-6z^2)}{4(4-z^2)} & \dfrac{-z(15-6z^2)}{4(4-z^2)(2+z)} \end{bmatrix} + \begin{bmatrix} v_1(z^2)+zv_2(z^2) & v_3(z^2)-zv_4(z^2) \\ v_3(z^2)+zv_4(z^2) & v_1(z^2)-zv_2(z^2) \end{bmatrix} \cdot \begin{bmatrix} (1-z^2)^2 & 0 \\ 0 & (1-z^2)^2 \end{bmatrix}, \tag{6.50}$$

which defines all compensators which satisfy the constraints of problem ST where the v_i are arbitrary stable rational functions. To construct an example of a compensator we take $v_i=0$, $i=1,2,3,4$, which yields a $w(z)$, which is just the first term in (6.50). Substituting into (1.10) we obtain the compensator

$$C(z)=\begin{bmatrix} \dfrac{z}{(2-z)} & 1 \\ 1 & \dfrac{-z}{(2+z)} \end{bmatrix} \frac{(15-6z^2)(4-z^2)}{4(1-z^2)^2}. \tag{6.51}$$

Interestingly, this time-varying compensator perfectly cancels the time-variation in the plant yielding a time-invariant open-loop gain and time-invariant feedback system gains in the form

$$h_{rp}(z)=\begin{bmatrix} \dfrac{(15-6z^2)}{(4-z^2)^2} & 0 \\ 0 & \dfrac{(15-6z^2)}{(4-z^2)^2} \end{bmatrix} \tag{6.52}$$

$$h_{lp}(z)=\begin{bmatrix} \dfrac{(1-z^2)^2}{(4-z^2)^2} & 0 \\ 0 & \dfrac{(1-z^2)^2}{(4-z^2)^2} \end{bmatrix}. \tag{6.53}$$

Here, the $(1-z^2)^2$ factor in $h_{lp}(z)$ indicates that the

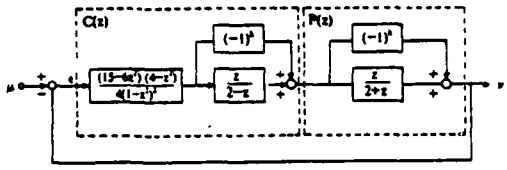


Fig. 4. Periodically varying feedback system that tracks $1/(1-z^2)^2$.

required tracking property is attained while the system gains are clearly stable. The resultant feedback system is sketched in Fig. 4. Although it is not clear from the figure that the open loop gain for this system is time-invariant, this fact can be verified by computing $P(z)C(z)$ and showing that it has the form of (6.42). Alternatively, an opportune rearrangement of the subsystems in Fig. 4 will lead to an equivalent time-invariant system.

VII. Conclusion

Although the above derivations were occasionally complex, it is important to note that the mathematics employed was quite elementary. Indeed, at no time have we used any mathematical techniques which are more sophisticated than addition, multiplication, subtraction, and inversion. Moreover, a single proof technique has been employed throughout the present paper as well as in [6]. First, one formulates a design equation which is linear and characterized by two unknowns in the ring of stable systems. A particular solution for this equation is constructed in terms of a specified coprimeness condition. A homogeneous solution is formulated in terms of an arbitrary stable design parameter. Finally, a coprimeness condition is used to verify that all homogeneous solutions have been obtained and the desired parameterization of the solution space is formulated in terms of the specified particular and homogeneous solutions. We believe that this algebraic formulation and solution technique is fundamental to the feedback system design problem and we are presently investigating several additional applications thereof.

Although our design equations, formulated in Theorem 2, are applicable to arbitrary linear systems, the solutions, formulated in Section VI, are restricted to multivariate systems with $\bar{t}$ and $\bar{r}$ in L. An inspection of the proof of Theorem IV will, however, reveal that the multivariate assumption was not used in the proof of the existence of a solution to problem *SR* (although it was used in the parameterization of the solution space and in the derivations of the solutions to problem *ST* and *STR*). We thus have the following.

Theorem 6: For the feedback system of Fig. 1 let P and R be characterized by doubly coprime fractional representations. Then *problem SR admits a solution if, and only if,* p_l *and* r_l are right coprime. (No commutativity assumption is necessary.)

References

[1] P. J. Antsaklis and J. B. Pearson, "Stabilization and regularization in linear multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 928–930, Oct. 1978.

[2] G. Bengtsson, "Output regulation and internal models—A frequency domain approach," *Automatica*, vol. 13, pp. 333–345, 1977.

[3] F. M. Callier and C. A. Desoer, "Stabilization, tracking, and disturbance rejection in linear multivariable distributed systems," in *Proc. 17th IEEE Decision Contr. Conf.*, San Diego, CA, Jan. 1979, pp. 513–514. Also, University of California, Berkeley, Tech. Memo. UCB/ERL M78/83, Dec. 1978.

[4] L. Cheng and J. B. Pearson, "Frequency domain synthesis of multivariable linear regulators," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 3–15, Feb. 1978.

[5] J. H. Davis, "Stability conditions derived from spectral theory: Discrete systems with periodic feedback," *SIAM J. Contr.*, vol. 10, pp. 1–13, 1972.

[6] C. A. Desoer, R.-W. Liu, J. J. Murray, and R. Saeks, "Feedback system design: The fractional approach to analysis and synthesis," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 401–412, 1980.

[7] B. A. Francis, "The multivariable servomechanism problem from the input-output viewpoint," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 322–328, June 1977.

[8] E. I. Jury and F. J. Mullin, "The analysis of sampled-data control systems with periodically time-varying sampling rate," *IRE Trans. Automat. Contr.*, vol. AC-4, pp. 15–21, 1959.

[9] E. I. Jury, "A note on multirate sampled-data systems," *IEEE Trans. Automat. Contr.*, vol. AC-12, pp. 319–320, June 1967.

[10] G. M. Kranc, "Input-output analysis of multirate feedback systems," *IRE Trans. Automat. Contr.*, vol. AC-3, pp. 21–28, 1957.

[11] R. A. Meyer and C. S. Burrus, "A unified analysis of multirate and periodically time-varying digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 162–168, 1975.

[12] L. Pernebo, "Algebraic control theory for linear multivariable systems," Ph.D. dissertation, Dep. Automat. Contr., Lund Inst. Technol., Lund, Sweden, May 1978.

[13] W. A. Wolovich, "Output regulation and tracking in linear multivariable systems," *Proc. CDC*, San Diego, CA, Jan. 1979.

[14] ——, "Skew prime polynomial matrices," *IEEE Trans. Automat. Cont.*, vol. AC-23, pp. 880–887, 1978.

[15] D. C. Youla, J. J. Bongiorno, and H. A. Jabr, "Modern Wiener-Hopf design of optimal controllers—Part I," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 3–15, 1976.

[16] ——, "Modern Wiener-Hopf design of optimal controllers—Part II," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 319–338, 1976.

[17] D. C. Youla, "The FEE: A new tunable high-resolution spectral estimator Part I—Background theory and derivation," Polytechnic Inst. of New York, Brooklyn, NY, unpublished notes, 1979.

Richard Saeks (S'59–M'65–SM'74–F'77) was born in Chicago, IL, in 1941. He received the B.S. degree in 1964, the M.S. degree in 1965, and the Ph.D. degree in 1967 from Northwestern University, Evanston, IL, Colorado State University, Fort Collins, and Cornell University, Ithaca, NY, respectively, all in electrical engineering.

He is presently Paul Whitfield Horn Professor of Electrical Engineering and Mathematics at Texas Tech University, Lubbock, where he is involved in teaching and research in the areas of fault analysis, circuit theory, and mathematical system theory.

Dr. Saeks is a member of the American Mathematical Society, the Society for Industrial and Applied Mathematics, and Sigma Xi.

John Murray (M'78) was born in Galway, Ireland, on August 8, 1947. He received the B.Sc. and M.Sc. degrees from University College, Cork, Ireland, in 1969 and 1970, respectively, and the Ph.D. degree from the University of Notre Dame, Notre Dame, IN, in 1974, all in mathematics.

He is currently with the Department of Electrical Engineering, Texas Tech University, Lubbock. His principal research interests are in the areas of several complex variables, multidimensional system theory, and time-varying systems.

A FRACTIONAL REPRESENTATION APPROACH TO ADAPTIVE CONTROL

C. KARMOKOLIAS

AND

R. SAEKS

PRECEDING PAGE BLANK-NOT FI

A FRACTIONAL REPRESENTATION APPROACH TO ADAPTIVE CONTROL†

C. Karmokolias and R. Saeks
Department of Electrical Engineering
Texas Tech University
Lubbock, Texas 79409

The main problem of adaptive control theory is to design a system S which is capable of automatically adjusting the generated control input to the plant P. Such adjustments may be necessary for a variety of reasons, such as insufficient knowledge about the plant, plant perturbations, etc. A multitude of adaptive control techniques have been proposed through the years. A characteristic shared by all of them is the presence of some means of identifying the unknown or perturbed plant. Of course, the design of such a mechanism, termed here the identifier, is an important question in its own right. The design, however, of an adaptive controller is heavily influenced by the particular technique used to generate the control and it therefore inherits the technique's features.

A recent advance in control theory is an approach to feedback control based upon the representation of the plant as the ratio of two operators, both of which belong to an operator ring H. (Ref). A brief overview of the approach is as follows. Consider the following ring structure R

$$R = (G,H,I,J) \tag{1.1}$$

where G is a not necessarily commutative ring with identity representing the general class of systems of interest. The subring H also contains the identity and represents the class of systems which in some sense are stable. I is the set of elements in H which admit an inverse in G and J the set of elements in H which admit an inverse in H. As shown in (Ref),

$$G \supset H \supset I \supset J \tag{1.2}$$

A plant P is said to have a doubly coprime fractional representation if for $(N_r,N_l,U_r,U_l,V_r,V_l) \subset H$ and $(D_r,D_l) \subset I$

$$P=N_rD_r^{-1}=D_l^{-1}N_l \tag{1.3}$$

$$U_rN_r+V_rD_r=1 \tag{1.4}$$

$$N_lU_l+D_lV_l=1 \tag{1.5}$$

The aim now is to design a system S so that the system's input-output map h is placed in H. Consider the system shown in Fig. 1.1 and assume that P has a doubly coprime fractional representation.

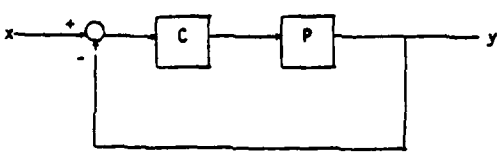


Fig. 1.1. A feedback control system.

For any arbitrary w, let the compensator C be defined as

$$C=(wN_l+V_r)^{-1}(-wD_l+U_r). \tag{1.6}$$

It was shown that if $w \subset H$, then the input-output map h also belongs to H and

$$h=N_r(-wD_l+U_r). \tag{1.7}$$

An important element of the approach is that it provides a complete characterization of the set of compensators which place h in the ring H. It is therefore desirable to investigate the conditions under which fractionally represented feedback systems can be adaptively controlled.

Suppose then that either in the limit as $t \to \infty$, or for all times $t \geq t_0$, an input-output map H in H is desired; in other words, suppose that, with the appropriate time interpretation, it is required that

$$h=H. \tag{1.8}$$

Clearly, there exists a choice of three independent variables, namely w, U_r and V_r, to satisfy two linear equations. The decision was made to consider w as a parameter in H. Thus the problem can in general be stated as seeking the particular coprimeness operator pair U_r,V_r which for a given w in H simultaneously satisfies Eq.s 1.4 and 1.8.

The two main problems to be addressed here are the acquisition and the plant-follower. In the former, the linear, time-invariant plant P is assumed to be insufficiently specified at the initial time t_0. The intention is to provide a feedback system S which consists of an identifier ID and an adaptor AD as shown in Fig. 1.2. The identifier provides the adaptor with estimates p(t) of the plant P such that $\lim_{t\to\infty} p(t) = P$. Then, using these estimates, the adaptor

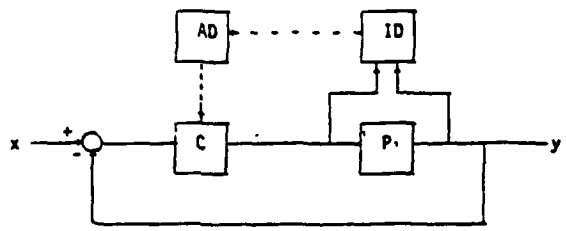


Fig. 1.2. An adaptive control system.

provides the compensator with an operator pair $(u_r(t),v_r(t))$ such that the required coprimeness pair

† This research supported in part by the Joint Services Electronics Program of Texas Tech University under ONR Contract 76-C-1136

PRECEDING PAGE BLANK-NOT FILMED

$\{U_r, V_r\}$ is obtained in the limit. The first task is to delineate the class of plants for which such a system is possible. This can be done by deriving the necessary and sufficient conditions for a solution to exist under the assumption of instantaneous identification, (i.e., a perfect identifier). Then it would remain to show that in the non-ideal case the solution can be attained adaptively. In other words it would be required that Eq.s 1.4 and 1.8 are satisfied in the limit.

In the plant-follower problem the linear plant P is perfectly known at the initial time t_0, but it undergoes perturbations thereafter. The intention is to provide the compensator with an operator pair $\{U_r(t), V_r(t)\}$ such that the systems input-output map remains invariant under the plant's perturbations. In other words Eq.s 1.4 and 1.8 are to be satisfied at every point in time. Again the class of plants for which a solution exists is delineated under the perfect identifier assumption. In the non-ideal case it is desirable to examine the extent to which the input-output map is perturbed due to the plant perturbations.

As always, stability is a question of paramount importance. A consequence of the fractional representation approach is the fact that a system is stable in the sense of H whenever the system's input-output map is time-invariant and the coprimeness operators belong to H. This is exploited in the ideal case of both problems. But, whereas, in the acquisition problem the derived stability conditions are time-independent and hence easy to check a priori, in the plant-follower they are time-dependent and thus the task of verifying whether they hold or not is considerably harder. However, the problem is by-passed by showing that in this case the question of the coprimeness operators belonging to H is equivalent to the classical question of stability in the sense of H of a system with time-invariant feedforward path and memoryless, time-varying feedback path. In the adaptive case of the plant-follower problem stability is resolved by a similar criterion applied to the entire adaptive acquisition problem, the fact that the input-output map converges to a time-invariant element of H suggests that the system is stable as long as the map remains bounded. It is shown that for uniform asymptotic stability this is in fact the case as long as a sufficiently "good" identifier is used. (The quality of the identifier is also shown to be related to the robustness of the adaptive plant-follower system).

The requirement to control the entire input-output map restricts the application to a class of plants which, for all practical purposes, is only slightly larger than the miniphase case. But if a less restrictive requirement is imposed the class becomes considerably larger. The point is demonstrated by the pole positioning problem for plants represented as rational functions (not necessarily proper). It is shown that the problem is equivalent to solving a linear, algebraic equation. Furthermore, a solution to the equation is shown to exist provided that the number of poles to be positioned is sufficiently large. In terms of adaptive control, the equation must be solved repeatedly in time by any of the available methods, (e.g. a continuation algorithm).

SUBOPTIMAL CONTROL WITH OPTIMAL QUADRATIC REGULATORS

C. KARMOKOLIAS

AND

R. SAEKS

SUBOPTIMAL CONTROL WITH OPTIMAL QUADRATIC REGULATORS*

C. Karmokolias** R. Saeks
Texas Tech University

Abstract

A new approach to the design of suboptimal controllers for constrained, nonlinear, decentralized, and non-quadratic systems is presented. Here, one designs a quadratic regulator for an idealized system but chooses the weighting matrices for the regulator to optimize its performance as a controller of the actual system relative to a prescribed (not necessarily quadratic) performance measure. The approach is illustrated via several examples.

I. Introduction

Historically, the control system designer has been faced with the dilemma: "Should he work with an idealized model of a system which is amenable to simple solutions or a "real world" model which may be intractable?" The former approach is epitomized by the LQG school wherein a highly idealized model of a real world system yields an easily implementable analytic control theory (11,12). Alternatively, a more realistic model may be employed in conjunction with a nonlinear programming algorithm at the cost of a more complex implementation and increased computer requirements (2,3).

The purpose of the present paper is to describe an intermediate approach to the control system design problem, (4,5,13) wherein one designs a quadratic regulator for an idealized system but chooses the weighting matrices for the regulator to optimize its performance as a controller of the actual system relative to a prescribed (not necessarily quadratic) *performance measure*. The advantage of such an approach is that the resultant regulator has the same "ease of implementation" and most of the "stability characteristics" associated with the classical LQG problem. The disadvantage is that the system performance is suboptimal. Computationally, the process does not require any more effort than required for a nonlinear optimization. Moreover, the on-line computation needed to implement the system is greatly reduced from that which would be required for an optimal non-linear controller.

The basic steps required for our proposed design procedure are:

i. Approximate the given (nonlinear, interconnected, or constrained) system by a linear state model.

ii. Design a regulator for the approximate state model which minimizes a quadratic performance measure with weighting matrices Q and R.

iii. Evaluate the performance of the actual system using the above quadratic control strategy.

iv. Optimize the performance of the actual system under such a control strategy as a function of the weighting matrices Q and R.

As is the case with any "real world" design algorithm, its effectiveness can be measured only by its performance in engineering practice. As such, the remainder of the paper is devoted to a series of examples in which the above-described design procedure is applied in a variety of settings and compared with existing design procedures. The examples include three tracking systems, a case of control under input constraints and a case of decentralized control.

II. Tracking Systems

To illustrate the design procedure we begin with a simple first order tracking problem. Here the system dynamics are

$$\dot{x}(t) = -x(t) + u(t) \; ; \; x(t_0) = x_0 \qquad 2.1$$

The optimal input $u^*(t)$ minimizing the quadratic index

$$J_x = \int_0^1 [qx^2(t) + u^2(t)]\, dt \qquad 2.2$$

is

$$u^*(t) = -P(q,t)\, x(t) \qquad 2.3$$

*This research supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.
**Presently with the Rearfott Div., Singer

PRECEDING PAGE BLANK-NOT FI

where $P(q,t)$ is the Riccati equation solution. Hence the optimal state trajectory is given by

$$x^*(t) = x_o e^{-(t + \int_0^t P(q,\sigma)\, d\sigma)} \qquad 2.4$$

We now seek the weight q^* which minimizes a given performance index J_Σ.

2.5

<u>Case 1</u>: $J_\Sigma = \int_0^1 \{[1-x^*(t)]^2 + [u^*(t)]^2\}\, dt$

The trajectories required by J_Σ and typical regulator trajectories are shown in Fig. 2-1. Obviously q^* depends upon x_o. On the other hand, finding q^* given an x_o is not always trivial (refer to Fig. 2-2).

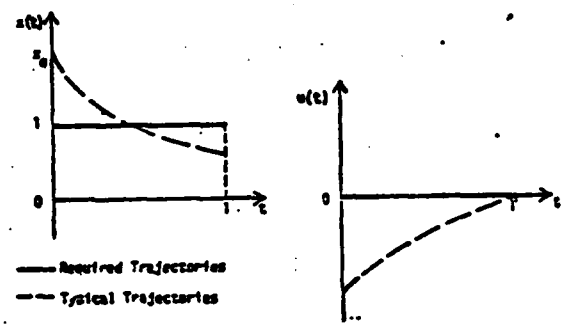


Fig. 2-1 Required and Typical Performance—Case 1

As $x_o \to \pm\infty$, $q^* \to +1$ because for x_o being very large, $(x(t)-1)^2 \simeq x^2(t)$ and thus equal importance is placed upon the state and the input term of J_Σ. As x_o goes to zero from above,

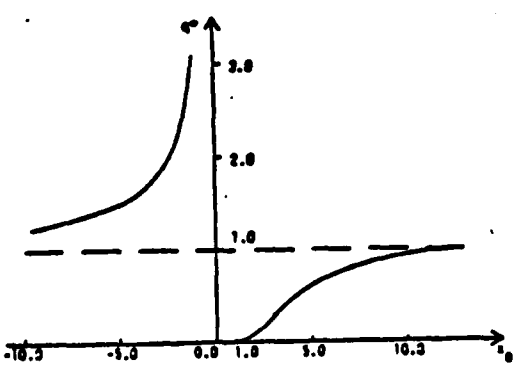


Fig. 2-2 Variation of q^* with x_o, Case 1

the state trajectory approximates its requirement better and better. Thus the regulator is "instructed" to emphasize the input cost by decreasing q^* which then causes the input to decrease and thus approximate its requirement better. Then, since a scalar stable system with a quadratic regulator is always bounded by its natural response, which in turn is always bounded by the initial condition, the best approximation to the $x=1$ requirement that the system could ever achieve for $0<x_o\leq 1$ is its own natural response. Thus for this range $q^*=0$. For $x_o=0$ q^* is indeterminate since both state and input are identically zero. Once x_o becomes slightly negative, the regulator is instructed to drive the state to zero as fast as possible because the state can now only add to the error. However, as x_o becomes more and more negative, the effort in driving the state to zero becomes significant also and hence q^* approaches unity. Table 2-1 shows that the present approach holds an advantage over the competitive approach with an optimally chosen constant gain h.

Table 2-1 Time-Invariant Versus Time-Variant Optimal Designs—Case 1

x_o	h^*	q^*	$J_\Sigma(x_o,h^*)$	$J_\Sigma(x_o,q^*)$	% Difference
1.0	1.0	0.0	0.0	0.16809	—
2.0	-0.0330	0.10	0.19873	0.19820	0.27
4.0	-0.1755	0.50	2.64189	2.62500	0.64
50.0	-0.3205	0.99	914.92200	911.40000	0.39

<u>Case 2</u>:

Required and typical trajectories are shown in Fig. 2-3 whereas the dependence of q^* upon x_o is shown in Fig. 2-4. Once again, to benchmark our results, Table 2-2 compares the present approach to optimally choosing a time invariant gain h. A noted improvement is obtained for $x_o=1$, the advantage becoming less profound as x_o gets larger.

2.6

$$J_\Sigma = \int_0^1 \{[1 - t - x^*(t)]^2 + [u^*(t)]^2\}\, dt$$

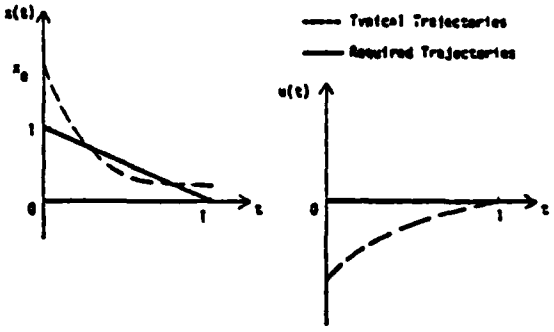


Fig. 2-3 Required and Typical Performance—Case 2

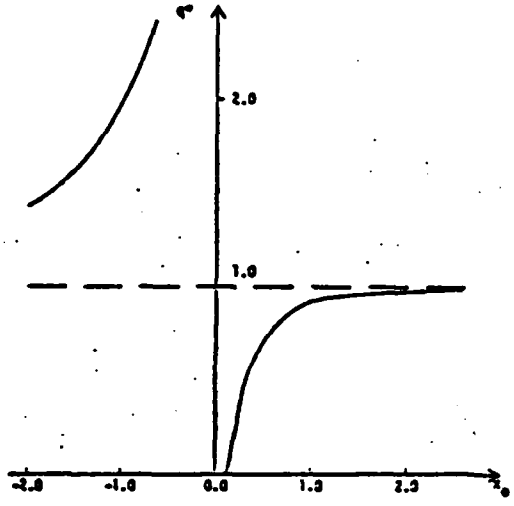


Fig. 2-4 Variation of q^* with x_0—Case 2

Table 2-2 Time Invariant Versus Time Variant Optimal Designs—Case 2

x_0	h^*	q^*	$J(x_0,h^*)$	$J(x_0,q^*)$	% Difference
1.0	-0.3688	0.24	0.05619	0.02649	112.22
2.0	-0.3645	0.59	0.55001	0.51680	6.43
4.0	-0.3472	0.80	3.86142	3.81600	1.34
50.0	-0.3350	0.96	936.06000	929.76000	0.68

As a variation on the tracking problem, consider the design of a pulse forming system whose output is required to approximate a triangular pulse. Specifically assume that $x_2(t)$ is to approximate the pulse shown in Fig. 2-5 where

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.08 & 0.05 \\ 0.05 & -0.08 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0.05 \\ 0.08 \end{bmatrix} u(t). \qquad 2.7$$

with $x_2(0)=0$ and $x_1(0)$ variable. Our goal is to design a regulator for the index

$$J_s = \int_0^{20} [x^t(t)Qx(t) + u^2(t)]\, dt \qquad 2.8$$

so that the measure

$$J_z = \int_0^{10} (|x_2(t) - 0.02t| + 0.001|u(t)|)\, dt + \int_{10}^{20} (|x_2(t) + 0.02t - 0.4| + 0.001|u(t)|)\, dt \qquad 2.9$$

is minimized over the diagonal matrix Q. The optimal designs are plotted against $x_1(0)$ in Fig. 2-6. Optimal trajectories of $x_2(t)$

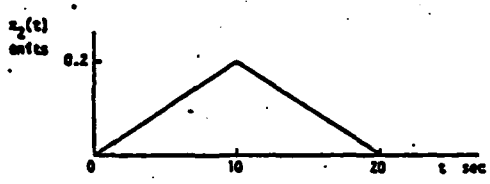
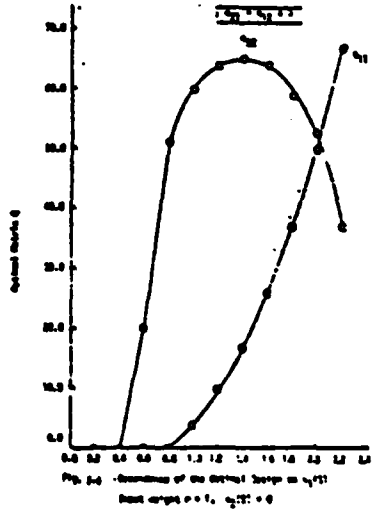
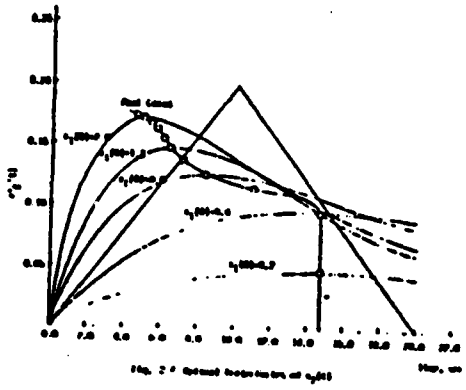


Fig. 2-5 Desired Triangular Pulse





are shown in Fig. 2-7. It is perhaps of interest to note that as $x_1(0)\to\infty$, the control is dominated by Q_{11} causing $x_2(t)$ to be regulated through its derivative.

III. An Aircraft Landing Problem

A four state model

$$\dot{x}(t)=Fx(t) + Gu(t) \qquad 3.1$$

for the landing of an aircraft was presented in (1,2) and was used in (3) to treat the same problem with dynamic programming. Although somewhat simplified, the model is adequate for illustrating our technique.

The states are defined in terms of the aircraft's coordinates and angles as (refer to Fig. 3-1).

$$x(t)=(\dot{\theta}(t),\theta(t),\dot{h}(t),h(t)), \qquad 3.2$$

whereas the input u(t) is the elevator deflection angle $\delta_e(t)$ which is mechanically constrained between -35° and 15°. Consistent with our suboptimal design approach we will initially neglect this limiting effect and design a quadratic regulator for the linear system but we will choose the weighting matrices so that the behavior of the actual nonlinear system is optimal relative to a performance index J . This index is chosen to simultaneously achieve a safe and comfortable landing. The complexity of the resulting J is such that rather than giving an explicit description here we will only list some of the factors entering its formulation. The interested reader is referred to (4,5) for the details.

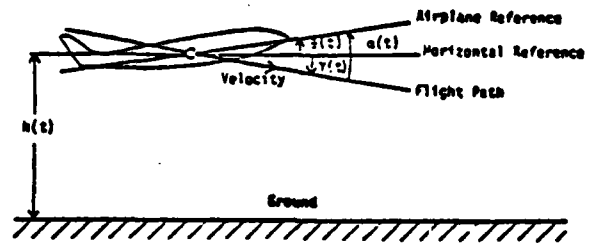


Fig. 3-1 Aircraft Coordinates and Angles

Safety Conditions

(1) The angle of attack $\alpha(t)$ must be less than about 18° and not considerably negative.

(2) At touchdown the airplane must be on the runway and within a distance d_s from the runway's start.

(3) At touchdown the rate of descent h(T) must be between -1.75 ft/sec and 0.0 ft/sec.

(4) At touchdown the pitch angle $\theta(T)$ must be within 0° and 10°.

Comfort Conditions

(1) Avoid all accelerations.

(2) Avoid a "hard" landing.

(3) Avoid a negative pitch angle.

The underlying assumption is that safety takes precedence over comfort and thus should any of the safety conditions be violated, $J_1=\infty$. It is noted that the selection of a matrix Q to satisfy these conditions is not a trivial task.

It was shown in (4) that stability considerations for the nonlinear model lead to the establishment of an upper bound for Q. In particular, if F is invertible and P(Q,t) the Riccati equation solution, the upper bound on Q is established by

$$R=G^TP(Q)F^{-1}G \qquad 3.3$$

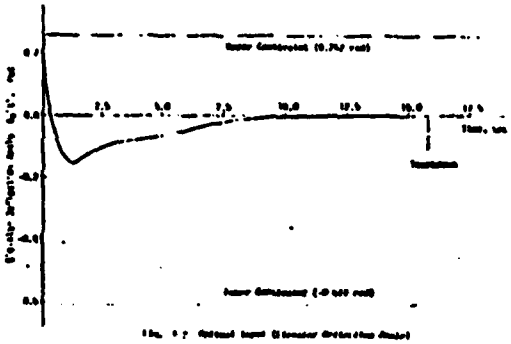
where $P(Q)=\lim P(Q,t)$. The landing was simulated on an IBM-370 and for an initial state

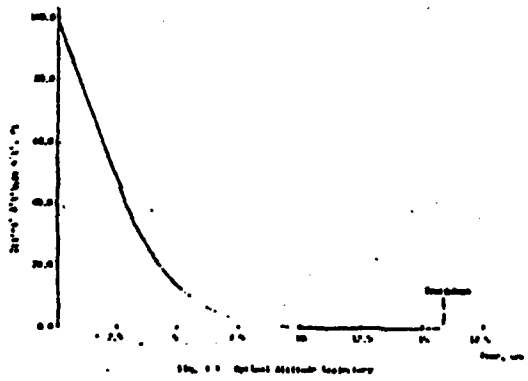
$$x(0)=(0.0,-0.0181,-20.0,100.0) \qquad 3.4$$

the optimal matrix Q was found to be

$$\mathrm{diag}\,Q^*=(21.0,21.0,0.0016,0.00047). \qquad 3.5$$

The resultant optimal input and elevation trajectories are shown in Figs. 3-2 and 3-3.





IV. Decentralized Control

The dimension of a large dynamical system is often so large that control on a global scale is prohibitive due to excessive requirements on computer storage or time. In such cases it may be desirable to consider the system as a collection of subsystems $S_i, i=1,2,\ldots,N$ and control each subsystem separately. Several schemes have been suggested varying from ignoring the subsystem interconnections to providing a separate control to neutralize their effect (6,7,8,9,10).

As an alternative we adopt our suboptimal approach to control system design. Here the given interconnected system is initially approximated by a decoupled system for which decoupled quadratic regulators are designed. These regulators are then employed to control the coupled system with the weighting matrices being chosen to optimize the performance of the resultant coupled system relative to some performance measure J_Σ.

To illustrate the approach we consider the system

$$\dot{x}(t) = Fx(t) + Gu(t) + Hx(t);\ x(t_0) = x_0 \qquad 4.1$$

where F and G are block diagonal matrices characterizing the decoupled system components and H is a coupling matrix.

For our example we take

$$F = \left[\begin{array}{cc|cc|cc} -0.10 & 0.05 & & & & \\ 0.05 & -0.10 & & 0 & & 0 \\ \hline & 0 & -0.10 & 0.05 & & 0 \\ & & 0.05 & -0.8 & & \\ \hline & 0 & & 0 & -0.01 & 0.0 \\ & & & & 0.05 & -0.08 \end{array}\right] \qquad 4.2$$

$$G = \left[\begin{array}{c|c|c} 0.10 & 0 & 0 \\ 0.0 & & \\ \hline 0 & 0.10 & 0 \\ & 0.0 & \\ \hline 0 & 0 & 0.05 \\ & & 0.01 \end{array}\right] \qquad 4.3$$

and

$$H = \left[\begin{array}{cc|cc|cc} & 0 & 0.0 & 0.0 & 0.01 & 0.0 \\ & & 0.0 & 0.01 & 0.0 & 0.0 \\ \hline 0.0 & 0.0 & & 0 & 0.0 & 0.0 \\ 0.01 & 0.0 & & & 0.0 & 0.0 \\ \hline 0.0 & 0.0 & 0.0 & 0.0 & & 0 \\ 0.0 & 0.0 & 0.01 & 0.0 & & \end{array}\right] \qquad 4.4$$

We now approximate the given system by the decoupled system

$$\dot{x}(t) = Fx(t) + Gu(t) \qquad 4.5$$

and design a decoupled regulator which minimizes

$$J_s = \int_0^{10} [x^t(t)Qx(t) + u^t(t)u(t)]\ dt \qquad 4.6$$

where Q is a block diagonal matrix. This regulator is then used to control the coupled system of Eq. 4-1 with Q chosen to minimize the global performance measure

$$J_\Sigma = \int_0^{10} [x'(t)\ x(t) + u'(t)\ u(t)]\ dt \qquad 4.7$$

The resultant optimal Q was found to be

$$Q^* = \left[\begin{array}{cc|cc|cc} 0.5 & 0.0 & & & & \\ 0.0 & 0.5 & & 0 & & 0 \\ \hline & & 0.5 & 0.0 & & \\ & 0 & 0.0 & 0.0 & & 0 \\ \hline & & & & 0.0 & 0.0 \\ & 0 & & 0 & 0.0 & 1.5 \end{array}\right] \quad 4.8$$

It is interesting to compare the above described approach to the so-called "naive" scheme where the couplings between subsystems are totally ignored. (To the contrary, we design decoupled regulators for each subsystem but choose the weight matrices through a minimization of J where all the couplings are included). In this particular example the naive approach yields a J =6.590, whereas our algorithm yields J =6.349, a 3.8 percent improvement.

V. Conclusions

Our purpose in the preceding has been the formulation of a new approach to the suboptimal control of nonlinear, constrained, decentralized and non-quadratic systems. In essence we restrict ourselves to a subclass of controllers by assuming that our controller will take the form of a quadratic regulator for a linear approximation of the given system. We then choose a particular regulator so as to optimize the performance of the given system within our class. The advantage of the approach is the relative ease in implementing the resultant control strategy and the stability inherent in the use of a quadratic regulator. The disadvantage is the suboptimality of the result and the computational effort required to choose the optimal weighting matrices.

VI. References

(1) C.W.Merriam, F.J.Ellert,"Synthesis of Feedback Controls Using Optimization Theory--An Example,"IEEE Trans. Auto. Control, pp. 89-103, 1963.

(2) D.E.Kirk,Optimal Control Theory,An Introduction. Prentice-Hall,Inc: Englewood Cliffs, New Jersey. 1970.

(3) C.W.Merriam. Optimization Theory and the Design of Feedback Control Systems. McGraw-Hill: New York. 1964.

(4) C.Karmokolias, Suboptimal Control With Optimal Quadratic Regulators, Ph.D. Dissertation, Texas Tech Univ., 1979.

(5) C.Karmokolias and R. Saeks, "Suboptimal Design of an Aircraft Landing System", Unpublished Notes, Texas Tech Univ., 1979.

(6) M.D.Mesarovic, D.Macko, Y.Takahara, "Theory of Hierarchical Multilevel Systems," Academic, New York. 1970.

(7) F.N. Bailey, F.C. Wang, "Decentralized Control Strategies for Linear Systems," Proc. Sixth Asilomar Conf. Circuits and Systems, pp. 370-4, November 1972.

(8) T.Ishimatsu, A.Mohri, M.Takata, "Optimization of Weakly-Coupled Subsystems by a Two-Level Method," Int. J. Control, Vol. 22, No. 6, pp. 877-82, December 1975.

(9) D.D. Siljak, M.K. Sundareshan, "A Multilevel Optimization of Large Scale Dynamic Systems," IEEE Trans. Auto. Control, Vol. AC-21, No. 1, pp. 79-84, February 1976.

(10) D.D. Siljak. Large-Scale Dynamic Systems, Stability and Structure. North-Holland: New York. 1978.

(11) A.P. Sage. Optimum Systems Control. Prentice-Hall, Inc: Englewood Cliffs, N. J. 1968.

(12) B.D.O.Anderson, J.B. Moore, Linear Optimal Control. Prentice Hall: New Jersey. 1971.

(13) C.Karmokolias, R.Saeks, "Optimal Selection of Weighting Matrices in Kalman Regulators," Proc. 21st Midwest Symposium on Circ. and Syst., pp. 72-6, August 1978.

FRACTIONAL REPRESENTATION, ALGEBRAIC GEOMETRY AND THE SIMULTANEOUS STABILIZATION PROBLEM

R. SAEKS

AND

J. MURRAY

1931 IEEE INTERNATIONAL SYMPOSIUM ON
CIRCUITS AND SYSTEMS PROCEEDINGS

RADISON CHICAGO HOTEL, CHICAGO, IL., APRIL 27-29, 1931

FRACTIONAL REPRESENTATION, ALGEBRAIC GEOMETRY, AND THE SIMULTANEOUS STABILIZATION PROBLEM *

R. Saeks and J. Murray
Department of Electrical Engineering
Texas Tech University
Lubbock, Texas 79409

ABSTRACT

An explicit relationship between the fractional representation approach to feedback system design and the algebro-geometric approach to system theory is formulated and used to derive a global solution to the feedback system design problem. These techniques are then applied to the simultaneous stabilization problem yielding a natural geometric criterion for a set of plants to be simultaneously stabilized by a single compensator.

I. SUMMARY

Classically, in control theory one is given a plant and desires to design a control system around this plant which meets certain design specifications. In fact, however, a "real world" plant is never known exactly and, as such, a realistic design must simultaneously meet specifications over an entire range of plants which (hopefully) include the actual plant. The simplest form of the resultant *simultaneous design problem* is the *robust design problem* wherein one desires to meet the design specifications in an ϵ-ball around a prescribed nominal plant. Although this is satisfactory for dealing with modeling errors it cannot cope with plants containing unknown parameters and/or plants characterized by multiple modes of operation. For instance, the dynamics of an airplane or rocket vary widely with altitude while the dynamics of an electric motor change with speed and load. To cope with these problems we must formulate a *simultaneous design theory* in which one designs a control system to simultaneously meet specifications over a prescribed set of plants. Of course, the set of plants may be taken to be a ball in which case the classical robustness theory is replicated. Alternatively, one may choose to work with a set of plants in which one or more parameters vary over a prescribed range and/or a discrete set of plants; say the dynamics of a two speed motor in its high and low speed settings.

The simultaneous design concept is possibly best illustrated in the 1st order case, wherein a simple geometric solution suggests itself. Assume that our plants are of the form

$$p(s) = \frac{A}{s + B} \qquad (1.1)$$

and we desire to design a stable feedback system using a proportional compensator with gain t. This results in a system with characteristic function

$$d(s) = s + (B + tA) \qquad (1.2)$$

and, as such, the feedback system will be stable if and only if $B + tA > 0$. Here, for a given compensator, t, the feedback system will be stable if and only if the point (A,B) lies above the line with slope 1/t as shown in figure 1a. As such, if we want to simultaneously stabilize an entire set of plants their representations on the A-B plane must all lie above a line through the origin. For instance, the set of plants indicated by the hatched region in figure 1b. can be simultaneously stabilized (by a compensator with gain $-\frac{1}{2}$) while the set of plants shown in figure 1c. cannot be simultaneously stabilized since they subtend an angle greater than 180 degrees on the A-B plane. Similarly, the set of plants shown in figure 1d. cannot be simultaneously stabilized since they cross the negative A-axis.

The example suggests two alternative criteria for the simultaneous stabilization problem. One may adopt an algebraic criterion to the effect that

$$B + tA > 0 \qquad (1.3)$$

for each plant in the prescribed set and some t. While such a test is definative it is local in nature allowing one to test for stabilizability on a plant by plant basis but yielding no global criterion with which to characterize a set of plants which is simultaneously stabilizable. To the contrary one may adopt a global geometric viewpoint to the effect that a prescribed set of plants is simultaneously stabalizable if and only if it is contained in an appropriate half-plane. The goal of the present paper is the formulation of a similar geometric criterion for the simultaneous stabilization problem applicable to general linear systems.

The starting point for our theory is the ring theoretic fractional representation theory introduced by the authors in a series of recent papers in which the set of stabilizing compensators for a given plant are parameterized.[1,4] Indeed, with

* This research supported in part by the Joint Services Electronics Program at Texas Tech Univ. under ONR Contract 76-C-1136.

PRECEDING PAGE BLANK-NOT FI

minor modifications one can invoke the same theory to parameterize the set of plants which are stabilized by a given compensator. This, in turn, yields an immediate algebraic criterion for the simultaneous stabilization problem. The resultant criterion is, however, local in nature just as that of equation 1.3. The desired global criterion for simultaneous stabilization can, however, be obtained if one first translates the fractional representation theory into an appropriate geometric setting.

Indeed, the appropriate geometric setting proves to be just the Grassmannian first introduced into the system theory literature by Hermann and Martin.[2,3] Unlike their frequency domain formulation, however, we obtain the Grassmannian directly from the ring theoretic fractional representation previously employed by the authors. Indeed, the Grassmannian is obtained simply by factoring out the non-uniqueness inherent in the fractional representation theory. As such, in addition to formulating the global theory necessary for our study of the simultaneous stabilization problem the geometric approach yields new insight into the relationship between the fractional representation theory (which we identify with the elements of a general linear group) and the system itself (which we identify with the elements of a Grassmannian).

II. REFERENCES

1. Desoer, C.A., Liu, R.-w., Murray, J.J., and R. Saeks, "Feedback System Design: The Fractional Representation Approach to Analysis and Synthesis", IEEE Trans. on Auto. Cont., Vol. AC-25, pp. 401-412, (1980).

2. Hermann, R., and C. Martin "Applications of Algebraic Geometry to Linear System Theory", IEEE Trans. on Auto. Cont., Vol. AC-22, pp. 19-25, (1977).

3. Herman, R., and C. Martin "Applications of Algebraic Geometry to System Theory: The MacMillan Degree and Kronecker Indices as Topological and Holomorphic Invariants", SIAM Jour. on Cont., Vol. 16, pp. 743-755, (1978).

4. Saeks, R., and J.J. Murray, "Feedback System Design: The Tracking and Disturbance Rejection Problems", IEEE Trans. on Auto. Cont., (to appear).

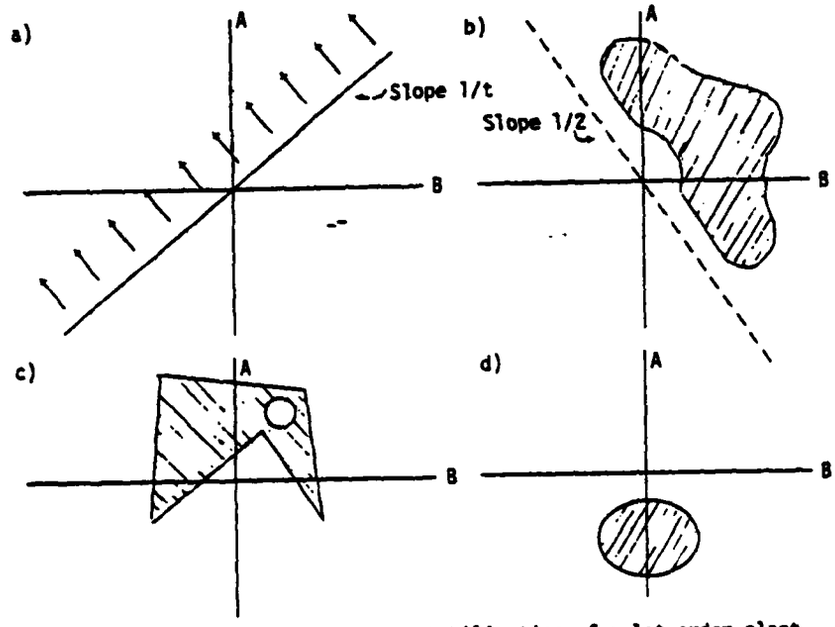


Figure 1. Simultaneous stabilization of a 1st order plant with a proportional compensator.

ABSTRACT OF

FEEDBACK SYSTEM DESIGN: THE SINGLE-VARIATE CASE

R. SAEKS, J. MURRAY, O. CHUA,
C. KARMOKOLIAS AND A. IYER

Abstract

A recently developed algebraic approach to the feedback system design problem is reviewed via the derivation of the theory in the single-variate case. This allows the simple algebraic nature of the theory to be brought to the fore while simultaneously minimizing the complexities of the presentation. Rather than simply giving a single solution to the prescribed design problem we endeavor to give a complete parameterization of the set of compensators which meet specifications. Although this might at first seem to complicate our theory it, in fact, opens the way for a sequential approach to the design problem in which one parameterizes the set of compensators which meet one specification and then characterizes the subset of those compensators which meet the second spec., etc. etc. Specific problems investigated include feedback system stabilization, the tracking and disturbance rejection problem, robust design, transfer function design, pole placement simultaneous stabilization, and stable stabilization.

PRECEDING PAGE BLANK-NOT FILMED

ABSTRACT OF

FRACTIONAL REPRESENTATION, ALGEBRAIC GEOMETRY AND THE SIMULTANEOUS STABILIZATION PROBLEM

R. SAEKS AND J. MURRAY

PRECEDING PAGE BLANK-NOT FILMED

Abstract

An explicit relationship between the fractional representation approach feedback system design and the algebro-geometric approach to system theory is formulated and used to derive a global solution to the feedback system problem. These techniques are then applied to the simultaneous stabilization problem yielding a natural geometric criterion for a set of plants to be simultaneously stabilized by a single compensator.

PRECEDING PAGE BLANK-NOT FI

ABSTRACT OF

SIMULTANEOUS DESIGN OF CONTROL SYSTEMS

R. SAEKS AND J. MURRAY

PRECEDING PAGE BLANK-NOT FILMED

Abstract

The problem of designing a feedback controller which stabilizes a number of plants simultaneously is discussed from the fractional representation point of view. An abstract solution of this general simultaneous stabilization problem is presented, and an elementary, explicit criterion is given for the simultaneous stabilizability of two systems. Finally, some examples and counter examples are presented, and some open problems are discussed.

PRECEDING PAGE BLANK-NOT FILMED

NONLINEAR CONTROL

L.R. HUNT

PRECEDING PAGE BLANK-NOT FILMED

Texas Tech University | Institute for Electronic Sciences

Joint Services Electronics Program | Research Unit: 2

1. Title of Investigation: Nonlinear Control
2. Senior Investigator: L. Roberts Hunt Telephone:
3. JSEP Funds: Current $25,875
4. Other Funds: Current $52,090*
5. Total Number of Professionals: PI's 2 (3 mo.) RA's
6. Summary:

Although *linearizations* have long been employed in the analysis and design of nonlinear control systems their applicability is severely limited by the approximate nature of the concept. To the contrary, if one could formulate an *exact transformation of a nonlinear system into a linear system* the established techniques for linear system analysis and design could be applied to the nonlinear problem. The goal of the present work unit is the formulation of such an *exact linearization theory* via the *differential geometric techniques* previously employed by the senior investigator in his investigation of the *controllability*, *observability*, and *stabilizability* characteristics of a nonlinear system.

Although the exact linearization problem goes back to Poincare and has been studied by a number of system theorists over the past decade with the aid of a generalized class of transformations introduced by Su we have been able to formulate readily testable necessary and sufficient conditions for the solution of the exact linearization problem. Moreover, when these conditions are satisfied the required transformation is given by the

*NASA Grant in support of Professor Hunt's leave of absence at NASA/AMES during the 1981/1982 academic year.

PRECEDING PAGE BLANK-NOT FI

solution of appropriate partial differential equations.

Although one might expect that the set of nonlinear systems which admit an exact linearization would be quite thin; and, indeed, this is the case; in practice, we have found that many "real world" systems either satisfy the required conditions and/or are approximable by such systems. Indeed, the exact linearization concept has been successfully implemented at NASA/AMES on several autopilot systems.

7. Publications and Activities:

A. Refereed Journal Articles

1. Hunt, L.R., "Controllability of Nonlinear Hypersurface Systems", in Algebraic and Geometric Methods of Linear System Theory (eds. C.I. Byrnes and C.F. Martin), Providence, AMS, 1980, pp. 209-224.

2. Hunt, L.R., "N-dimensional Controllability with n-1 Controls", IEEE Trans. on Auto. Cont., (to appear).

3. Hunt, L.R. "Sufficient Conditions for Controllability", IEEE Trans. on Circuits and Systems, (to appear).

B. Conference Papers and Abstracts

1. Hunt, L.R., and R. Su, "Local Transformations for Multi-input Nonlinear Systems", Proc. of the Joint Auto. Cont. Conf., Charlottesville, June 1981, paper FA3B.

2. Hunt, L.R., and R. Su, "Global Mappings of Nonlinear Systems", Proc. of the Joint Auto, Cont. Conf., Charlottesville, June 1981, paper FA3C.

3. Hunt, L.R., and R. Su, "Linear Equivalents of Nonlinear Time-Varying Systems", Proc. of the Inter. Symp. on the Mathematics of Networks and Systems, Santa Monica, Aug. 1981, pp. 119-123.

4. Hunt, L.R., and R. Su, "Poincare Lemma and Transformations of Nonlinear Systems", Proc. of the Inter. Symp. on the Mathematics of Networks and Systems", Santa Monica, Aug. 1981, pp. 111-118.

5. Hunt, L.R., and R. Su, "Transforming Nonlinear Systems", Proc. of the 24th Midwest Symp. on Circuits and Systems, Albuquerque, June 1981, pp. 341-345.

6. Hunt, L.R., Meyer, G., and R. Su, "Transformations of Nonhomogeneous Nonlinear Systems", *Proc. of the 19th Allerton Conf.* on Communications, Control and Computing, Oct 1981, (to appear)

7. Hunt, L.R., and R. Su, "Control of Nonlinear Time-Varying Systems", Proc. of the 20th IEEE Conf. on Decision and Control, Dec. 1981, (to appear).

C. Preprints

1. Hunt, L.R., and R. Su, "Global Transformations of Nonlinear Systems", (submitted for publication).

2. Hunt, L.R., and R. Su, "Multi-input Nonlinear Systems", (submitted for publication).

D. Dissertations and Theses

1. H. Ford, Ph.D., Dissertation, (in preparation).

E. Conferences and Symposia

1. Hunt, L.R., Joint Auto. Cont. Conf., Charlottesville, June 1981.

2. Hunt, L.R., 24th Midwest Symp. on Circuits and Systems, Albuquerque, June 1981.

3. Hunt, L.R., Inter. Symp. on the Mathematics of Networks and Systems, Santa Monica, Aug. 1981.

4. Hunt, L.R., 19th Allerton Conf. on Communications, Control and Computing, Urbana, Oct. 1981.

LINEAR EQUIVALENTS OF NONLINEAR TIME-VARYING SYSTEMS

L. R. HUNT AND RENJENG SU

AMES RESEARCH CENTER, NASA
MOFFETT FIELD, CALIFORNIA

PRECEDING PAGE BLANK-NOT FI

LINEAR EQUIVALENTS OF NONLINEAR TIME-VARYING SYSTEMS

L. R. Hunt* and Renjeng Su**

Ames Research Center, NASA
Moffett Field, California.

Abstract

Recent results have shown that a single-input time-invariant system of the form

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

can locally be transformed into a series of integrators if and only if (1) the vector fields $g, (ad^1 f,g), \ldots, (ad^{n-2} f,g)$ are involutive, and (2) $g, (ad^1 f,g), \ldots, (ad^{n-1} f,g)$ are linearly independent in a neighborhood of the origin in $\mathbb{R}^n$. This result is generalized to time-varying systems. A parallel theorem is obtained in terms of the time-varying version Lie derivative Γ.

1. INTRODUCTION

This paper is concerned with the problem of equivalence of nonlinear systems and a particular linear system, that is, a series of integrators. Expressed in state space form it is

$$\dot{x}_1 = x_2,\ \dot{x}_2 = x_3,\ \ldots,\ \dot{x}_{n-1} = x_n,\ \dot{x}_n = u,$$

where $x_1, x_2, \ldots, x_n$ are the state variables and u the control. We shall call this system the <u>canonical linear system</u> and denote it by Σ_0.

In the past this problem has been studied by Meyer and Cicolani[1] and by Brockett[2]. They obtain two different (but intersecting) classes of nonlinear systems which can be transformed into the canonical linear system Σ_0.

In reference 3, Su points out the difference between the equivalence relations defined by Meyer and Brockett. Considering the class of nonlinear systems with scalar input of the form $\dot{x} = f(x,u)$, and using the equivalence relation defined by Meyer, necessary and sufficient conditions for a nonlinear system to be equivalent to the system Σ_0 are given. This is the largest equivalence class in terms of state coordinates change and feedback, and it properly contains the results of Meyer and Brockett.

Later in reference 4, the authors extended that result to a sufficient theorem on the problem of global equivalence. In another paper[5] (submitted to this conference) they also show an interesting connection between the renowned Poincaré lemma and the construction of the desired transformation.

*Researcher supported by Ames Research Center, NASA, under the Intergovernmental Personnel Agreement (IPA) Program and the Joint Services Electronics Program at Texas Tech University under Office of Naval Research (ONR) Contract 76-C-1136.

**Research Associate of National Research Council at Ames Research Center, NASA.

PRECEDING PAGE BLANK-NOT FILMED

This paper is devoted to generalizing the previous results from reference 3 to the time-varying case. In section 2 we define the equivalence relation; the main results are stated in section 3. Because the arguments in reference 3 can be applied here with only slight modification, most of the proofs of the theorems are omitted.

2. EQUIVALENCE RELATION

We consider the scalar input, time-varying systems on $\mathbb{R}^n$ of the form

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

where the origin is an equilibrium state for any time t when $u = 0$.

A <u>$\mathcal{F}$-transformation</u> is a map $T: \mathbb{R}^{n+2} \to \mathbb{R}^{n+1}$ such that for any $t \in \mathbb{R}$ the restricted map $T(\cdot,t): \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is a diffeomorphism. For each system $\Sigma: \dot{x} = f(x,u,t)$ the combined vector (x,u) of the state and the control is considered as an element of the space $\mathbb{R}^{n+1}$. With the $\mathcal{F}$-transformations we can define a concept of system equivalence.

A system $\Sigma_1: \dot{x} = f(x,u,t)$ is said to be <u>$\mathcal{F}$-related</u> to a system $\Sigma_2: \dot{y} = g(y,v,t)$ if there is a $\mathcal{F}$-transformation T that carries each state trajectory of Σ_1 into a state trajectory of Σ_2; that is, setting $T(\phi(t;x_0,u),u(t),t) = (y(t),v(t))$, we have

$$\frac{dy}{dt}(t) = g(y(t),v(t),t),$$

where $\phi(t;x_0,u)$ is the state trajectory of Σ_1 with respect to the initial condition x_0 and control $u(t)$. The following important observation of this relation is parallel to that in reference 3.

Proposition 1: A system $\Sigma_1: \dot{x} = f(x,u,t)$ is $\mathcal{F}$-related to a system $\Sigma_2: \dot{y} = g(y,v,t)$ if and only if there is a transformation $T = (T_1, \ldots, T_n, T_{n+1})$ such that T has the property

$$\frac{\partial T_1}{\partial u} = \ldots = \frac{\partial T_n}{\partial u} = 0, \text{ and } \frac{\partial T_{n+1}}{\partial u} \neq 0$$

and satisfies the following system of partial differential equations:

$$\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} f_j + \frac{\partial T_i}{\partial t} = g_i(T_1, \ldots, T_n, T_{n+1}, t) \qquad i = 1, \ldots, n,$$

where $f = (f_1, \ldots, f_n)$ and $g = (g_1, \ldots, g_n)$.

Based on this observation one can prove that the $\mathcal{F}$-relation is indeed an equivalence relation among the systems. Therefore, we are justified in saying that two systems are <u>$\mathcal{F}$-equivalent</u>. In the next section we shall characterize the equivalence class that contains the canonical linear system Σ_0.

3. MAIN RESULTS

First we introduce our notations. Since only the local theory is attempted here, everything on $\mathbb{R}^n$ will be defined locally near the origin.

We assume the reader is familiar with the basic definitions of vector fields and one-forms on $\mathbb{R}^n$. The vector fields and the one-forms in this paper may be varying in time, namely, their coefficients may be functions of time when they are expressed in a fixed coordinate frame.

For each smooth scalar function ζ on $\mathbb{R}^n$ <u>the differential operator d</u> maps ζ into a one-form $d\zeta$ defined by $(\partial\zeta/\partial x_1)dx_1 + \ldots + (\partial\zeta/\partial x_n)dx_n$. For a vector field $f = f_1(\partial/\partial x_1) + \ldots + f_n(\partial/\partial x_n)$ and a one-form $\omega = \omega_1\, dx_1 + \ldots + \omega_n\, dx_n$, <u>the dual product</u> of ω and f, denoted by $\langle\omega,f\rangle$, is a scalar function defined by $\omega_1 f_1 + \ldots + \omega_n f_n$.

In the course of the development, several types of derivatives with respect to a vector field will be used. We state the definitions with a given vector field f.

(1) For a scalar function ζ on $\mathbb{R}^n$,
$L^0 f(\zeta) = \zeta$, $L^1 f(\zeta) = \langle d\zeta, f\rangle$, and
$L^n f(\zeta) = L^1 f(L^{n-1} f(\zeta))$.

(2) For a one-form ω on $\mathbb{R}^n$,
$\mathcal{L}^0 f(\omega) = \omega$, $\mathcal{L}^1 f = d\langle \omega, f\rangle$, and
$\mathcal{L}^n f(\omega) = \mathcal{L}^1 f(\mathcal{L}^{n-1} f(\omega))$.

For a time-varying vector field g, we have two types of derivatives:

(3) $(ad^0 f, g) = g$,
$(ad^1 f, g) = (\partial f/\partial x) g - (\partial g/\partial x) f$, and
$(ad^n f, g) = (ad^1 f, (ad^{n-1} f, g))$.

(4) $(\Gamma^0 f, g) = g$,
$(\Gamma^1 f, g) = (ad^1 f, g) - \partial g/\partial t$, and
$(\Gamma^n f, g) = (\Gamma^1 f, (\Gamma^{n-1} f, g))$.

An important formula involving the derivatives of types (1), (2), and (3) is

$$\langle \mathcal{L}^1 f(\omega), g\rangle = \langle \omega, (ad^1 f, g)\rangle + L^1 f(\langle \omega, g\rangle). \quad (1)$$

Now we are ready to study the problem of characterizing those systems that are $\mathcal{F}$-equivalent to the system Σ_0.

If a system Σ: $\dot{x} = f(x,u,t)$ is $\mathcal{F}$-equivalent to the system Σ_0 with a transformation $T = (T_1, T_2, \ldots, T_n, T_{n+1})$, then, from Proposition 1 we have

$$\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} f_j(x,u,t) + \frac{\partial T_i}{\partial t}(x,t) = T_{i+1}(x,t) \qquad i = 1, \ldots, n-1 \quad (2)$$

$$\sum_{j=1}^{n} \frac{\partial T_n}{\partial x_j} f_j(x,u,t) + \frac{\partial T_n}{\partial t}(x,t) = T_{n+1}(x,u,t) \quad (3)$$

An observation similar to one in reference 3 is given as follows. The $n-1$ equations in (2) say that for each fixed pair of state and time (x,t), the n components $f_1, \ldots, f_n$ of f, considered as variables of u, satisfy $n-1$ linear equations with constant coefficients. This leads to the following theorem.

Theorem 1: If a system $\dot{x} = f(x,u,t)$ is $\mathcal{F}$-equivalent to Σ_0, then f can be expressed in the form

$$f(x,u,t) = f(x,t) + g(x,t) \cdot \phi(x,u,t)$$

for some vector fields f and g and some scalar function ϕ with $f(0,t) = 0$, $\phi(0,0,t) = 0$, and $\partial\phi/\partial u \neq 0$ for any t.

From now on we will only examine systems of the form $\dot{x} = f(x,t) + g(x,t)u$; this will not result in any loss of generality (see ref. 3). A system will also be represented by a pair (f,g) of time-varying vector fields.

From equations (2) and (3), a system (f,g) is $\mathcal{F}$-equivalent to the system Σ_0 if and only if there is a transformation $T = (T_1, \ldots, T_n, T_{n+1})$ such that the following equations hold.

$$\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j}(f_j(x,t) + g_j(x,t)u) + \frac{\partial T_i}{\partial t}(x,t) = T_{i+1}(x,t), \quad (4)$$

for $i = 1, \ldots, n-1$, and

$$\sum_{j=1}^{n} \frac{\partial T_n}{\partial x_j}(f_j(x,t) + g_j(x,t)u) + \frac{\partial T_n}{\partial t}(x,t) = T_{n+1}(x,u,t). \quad (5)$$

In the terminology introduced early in this section, a transformation T satisfies the system of equations (4) and (5) if and only if T satisfies

$$\langle dT_i, g\rangle = 0, \qquad i = 1, \ldots, n-1, \quad (6)$$

$$\langle dT_i, f\rangle + \frac{\partial T_i}{\partial t} = T_{i+1} \qquad i = 1, \ldots, n-1, \quad (7)$$

$$\langle dT_n, f + gu\rangle + \frac{\partial T_n}{\partial t} = T_{n+1}, \quad (8)$$

$$\langle dT_n, g\rangle \neq 0. \quad (9)$$

Considering the case $i = 2$, equations (6) and (7) state that

$$\langle dT_2, g\rangle = 0, \quad (10)$$

$$\langle dT_2, f\rangle + \frac{\partial T_1}{\partial t} = T_2. \quad (11)$$

Letting $dT_1 = \omega$ and substituting (11) into (10), we obtain

$$\langle \mathscr{L}^1 f(\omega), g \rangle + \left\langle d\left(\frac{\partial T_1}{\partial t}\right), g \right\rangle = 0 \qquad (12)$$

By formula (1)

$$\langle \mathscr{L}^1 f(\omega), g \rangle = \langle \omega, (\mathrm{ad}^1 f, g) \rangle + L^1 f(\langle \omega, g \rangle)$$

and $\langle \omega, g \rangle = 0$, we have $\langle \mathscr{L}^1 f(\omega), g \rangle = \langle \omega, (\mathrm{ad}^1 f, g) \rangle$. Because the operators d and $\partial/\partial t$ commute, we have $d(\partial T_1/\partial t) = \partial/\partial t(dT_1) = \partial\omega/\partial t$. The second term in equation (12) then is $\langle \partial\omega/\partial t, g \rangle$, and the fact that $\langle \omega, g \rangle = 0$ implies $(\partial\langle \omega, g \rangle)/\partial t = 0$. By the chain rule,

$$\frac{\partial\langle \omega, g \rangle}{\partial t} = \left\langle \frac{\partial\omega}{\partial t}, g \right\rangle + \left\langle \omega, \frac{\partial g}{\partial t} \right\rangle,$$

which implies

$$\left\langle \frac{\partial\omega}{\partial t}, g \right\rangle = -\left\langle \omega, \frac{\partial g}{\partial t} \right\rangle.$$

Summing up these observations, equation (12) becomes

$$\langle \omega, (\mathrm{ad}^1 f, g) \rangle - \left\langle \omega, \frac{\partial g}{\partial t} \right\rangle = 0.$$

In terms of the operator Γ, it can be expressed as $\langle \omega, (\Gamma^1 f, g) \rangle = 0$. Similar computations change equations (6) and (9) into, respectively,

$$\langle dT_1, (\Gamma^i f, g) \rangle = 0, \qquad i = 0, 1, \ldots, n-2 \qquad (13)$$

and

$$\langle dT_1, (\Gamma^{n-1} f, g) \rangle \neq 0. \qquad (14)$$

It then can be readily checked that the existence of a transformation T satisfying equations (6)-(9) is equivalent to the existence of a scalar function T_1 such that equations (13) and (14) hold. Except for the use of the time-varying version of the Lie derivative, this result is exactly parallel to that obtained in the time-invariant case in reference 3.

Now we are ready to state the main theorem which characterizes the equivalence class in which we are interested. The proof is omitted because of its similarity to that given in reference 3. We remark that the time-varying version of the Frobenius theorem plays a crucial role.

Theorem 2: A time-varying system

$$\Sigma: \dot{x} = f(x,t) + g(x,t)\phi(x,u,t)$$

(as defined in Theorem 1) is $\mathscr{F}$-equivalent to the system Σ_0 if and only if

(1) The vector fields g, $(\Gamma^1 f, g), \ldots, (\Gamma^{n-1} f, g)$ span $\mathbb{R}^n$ at the origin at any time t, and

(2) The vector fields g, $(\Gamma^1 f, g), \ldots, (\Gamma^{n-2} f, g)$ are involutive near the origin for any t.

4. CONCLUDING REMARKS

The conditions in Theorem 2 describe a special $\mathscr{F}$-equivalence class of time-varying systems. For a nonlinear system in this equivalence class we can solve the system of partial differential equations (13) and (14) for a transformation that will turn the system into a series of integrators; that is, the system Σ_0. The system Σ_0 is not only linear but also time invariant. Notice that the transformation is varying in time in order to "cancel" the time dependency of the system. It is not surprising that a time-varying version of Lie derivative Γ has to be used to generalize the result in reference 3. This operator Γ in its linear version has previously appeared in the literature; for example, see Hermes.[6]

5. REFERENCES

1. Meyer, G.; and Cicolani, L.: A Formal Structure for Advanced Automatic Flight Control Design-System Concepts and Flight Evaluations. AGARDograph on Theory and Applications of Optimal Control in Aerospace Systems, P. Kant, ed., 1980.

2. Brockett, R. W.: Feedback Invariants for Nonlinear Systems. IFAC Congress, Helsinki, 1978.

3. Su, R.: On the Linear Equivalents of Nonlinear Systems. Submitted to Systems and Control Letters.

4. Hunt, L. R.; and Su, R.: Global Transformations of Nonlinear Systems. Submitted to 1981 Joint Automatic Control Conference.

5. Hunt, L. R.; and Su, R.: The Poincaré Lemma and Transformations of Nonlinear Systems. Submitted to 1981 International Symposium on the Mathematical Theory of Networks and Systems.

6. Hermes, H.: On Local and Global Controllability. SIAM J. Control, vol. 12, no. 2, 1974.

6. BIOGRAPHIES

Renjeng Su was born in China in 1950. He received a B.S. in chemical engineering from Chen-Kung University in Taiwan in 1972. His M.S. and D.Sc. are both in systems science and mathematics from Washington University at St. Louis, Mo., in 1980. Currently, he is a research associate of the National Research Council at Ames Research Center, NASA, Moffett Field, Calif.

L. R. Hunt was born in Shreveport, La., on Dec. 5, 1942. He received his B.S. degree in 1964 from Baylor University and his Ph.D. degree in 1970 from Rice University, both in mathematics. He is currently on leave from Texas Tech University, working at NASA Ames Research Center, Moffett Field, Calif.

Dr. Hunt's research interests are nonlinear systems and control, several complex variables, and partial differential equations. He is a member of the Institute of Electrical and Electronics Engineers, the American Mathematical Society, and the Society for Industrial and Applied Mathematics.

THE POINCARE LEMMA AND TRANSFORMATIONS OF NONLINEAR SYSTEMS

L. R. HUNT AND RENJENG SU

AMES RESEARCH CENTER, NASA
MOFFETT FIELD, CALIFORNIA

PRECEDING PAGE BLANK-NOT FILMED

THE POINCARÉ LEMMA AND TRANSFORMATIONS OF NONLINEAR SYSTEMS

L. R. Hunt* and Renjeng Su**

Ames Research Center, NASA
Moffett Field, California

Abstract

Recent results have classified those nonlinear systems that can be transformed to linear systems. For single-input systems of the form

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t))$$

we assume that the vector fields $g, [f,g], \ldots, (ad^{n-2}f,g)$ are involutive and that $g, [f,g], \ldots, (ad^{n-1}f,g)$ are linearly independent in a neighborhood of the origin in $\mathbb{R}^n$. It is shown that the transformation mapping this system to a linear system exists by virtue of the famous Poincaré lemma from differential geometry.

1. INTRODUCTION

Because of the extensive literature concerning time-invariant linear systems, it is interesting to characterize the nonlinear systems that can be transformed to these linear systems. In this paper we consider single-input, time-invariant nonlinear systems of the form

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)) \tag{1}$$

where f and g are $\mathscr{C}^\infty$ complete vector fields on an open set in $\mathbb{R}^n$ containing the origin and $f(0) = 0$. The $\mathscr{C}^\infty$ transformations of interest to us are $T = (T_1, T_2, \ldots, T_{n+1})$, which map $(x_1, \ldots, x_n, u)$ space to $(T_1, T_2, \ldots, T_{n+1})$ space so that the system (1) is mapped to the linear system

$$\left.\begin{aligned} \dot{T}_1 &= T_2 \\ \dot{T}_2 &= T_3 \\ &\vdots \\ \dot{T}_n &= T_{n+1} \end{aligned}\right\} \tag{2}$$

If we think of T_{n+1} as being the control in equation (2), this system is in integrator form with $T_1, T_2, \ldots, T_n$ being the space variables. We want (i) T to have a nonsingular Jacobian matrix, (ii) $T(0) = 0$, (iii) $T_1, T_2, \ldots, T_n$ to be functions of $x_1, x_2, \ldots, x_n$ only and to have a nonsingular Jacobian matrix in these variables, and (iv) T_{n+1} to be a function of $x_1, \ldots, x_n, u$ which can be inverted as a function of u. In addition, we can also ask where the T transformation is a diffeomorphism.

*Researcher supported by Ames Research Center, NASA, under the Intergovernmental Personnel Agreement Program and the Joint Services Electronics Program at Texas Tech University under Office of Naval Research Contract 76-C-1136.

**National Research Council Research Associate at Ames Research Center, NASA.

PRECEDING PAGE BLANK-NOT FI

Necessary and sufficient conditions for the local existence of such a transformation are given in reference 1, and a constructive proof of the transformation in addition to global results is found in reference 2. The purpose of this paper is to show that the existence of such a mapping depends on the application of the Poincaré lemma from differential geometry.

Several authors have examined the equivalence of nonlinear systems and linear systems under various assumptions. Meyer and Cicolani [3],[4] considered the block-triangular nonlinear (possibly time-varying) systems; Krener[5] gave conditions for a nonlinear system to be transformed to a linear system under state space coordinate changes; and Brockett[6] studied the equivalence of nonlinear and linear systems under coordinate changes and additive feedback. The transformation theory developed in references 1 and 2 contains the results from the authors just mentioned.

In section 2 of this paper we give definitions and study the system of linear partial differential equations from reference 1 that determines the existence of a transformation of the type that is of interest to us. Section 3 contains examples, the construction of the transformation, and several applications of the Poincaré lemma.

2. DEFINITIONS AND TECHNIQUES

We give basic definitions and examine the technique in reference 1 that proves the existence of the transformation.

Letting X and Y be vector fields on $\mathbb{R}^n$ (or on an open subset of $\mathbb{R}^n$), we define the <u>Lie bracket</u> of X and Y

$$[X,Y] = \frac{\partial Y}{\partial x} X - \frac{\partial X}{\partial x} Y ,$$

where $\partial Y/\partial x$ and $\partial X/\partial x$ are $n \times n$ Jacobian matrices. Successive Lie brackets like $[X,[X,Y]]$, $[Y,[X,Y]]$, etc., can be introduced, and we set

$$\begin{aligned}
(\mathrm{ad}^0 X,Y) &= Y , \\
(\mathrm{ad}^1 X,Y) &= [X,Y] , \\
(\mathrm{ad}^2 X,Y) &= [X,[X,Y]] , \\
&\vdots \\
(\mathrm{ad}^k X,Y) &= [X,(\mathrm{ad}^{k-1} X,Y)] .
\end{aligned}$$

A set of $\mathscr{C}^\infty$ vector fields $\{X_1, \ldots, X_r\}$ on $\mathbb{R}^n$ is <u>involutive</u> if there exist $\mathscr{C}^\infty$ functions $\gamma_{ijk}(x)$ such that

$$[X_i,X_j](x) = \sum_{k=1}^{r} \gamma_{ijk}(x) X_k(x) ,$$

$$1 \le i,j \le r, \quad i \ne j .$$

The classical Frobenius theorem states that given a point $x_0 \in \mathbb{R}^n$ and an involutive set $\{X_1, \ldots, X_r\}$ of linearly independent vector fields on $\mathbb{R}^n$, there is a unique maximal r dimensional $\mathscr{C}^\infty$ submanifold S of $\mathbb{R}^n$ containing x_0 so that the tangent space to S at each $x \in S$ is the space spanned by $X_1, \ldots, X_r$ at x. We say that S is the <u>integral manifold</u> of $X_1, \ldots, X_r$ through x_0.

A $\mathscr{C}^\infty$ n dimensional manifold M is called (<u>smoothly</u>) <u>contractible</u> to a point $x_0 \in M$ if there is a $\mathscr{C}^\infty$ function

$$H: \; M \times [0,1] \to M$$

such that

$$\begin{aligned}
H(x,1) &= x \quad \text{for } x \in M \\
H(x,0) &= x_0 .
\end{aligned}$$

Of course, $\mathbb{R}^n$ is smoothly contractible to the origin, and a <u>star-shaped region</u> with respect to a point x_0 is contractible to that point (see ref. 7).

Let ω be a k-form on M, $1 \le k \le n-1$, and let d be the standard differential operator mapping ω to a $k+1$ form (in our theory ω is a one-form). A k-form ω is called <u>closed</u> if $d\omega = 0$ and <u>exact</u>

if $\omega = d\eta$ for some $k-1$ form η on M. This leads to a famous result of Poincaré, which we state for $\mathscr{C}^\infty$ forms.

Lemma 1: If M is smoothly contractible to a point $x_0 \in M$, then every closed form ω on M is exact.

In later applications of this result M is an open neighborhood of the origin in $\mathbb{R}^n$. Recall that we are interested in a transformation $T = (T_1, T_2, \ldots, T_{n+1})$ which takes the system

$$\dot{x} = f + ug \tag{1}$$

to the system

$$\left.\begin{aligned} \dot{T}_1 &= T_2 \\ \dot{T}_2 &= T_3 \\ &\vdots \\ \dot{T}_n &= T_{n+1} \end{aligned}\right\} \tag{2}$$

so that the conditions (i) through (iv) (specified in the Introduction) hold. By design, this mapping will be a diffeomorphism near the origin; conditions under which we have a global diffeomorphism are discussed in reference 2. Necessary and sufficient conditions for the local existence of such a map are that (a) the matrix $\{g, [f,g], \ldots, (ad^{n-1}f,g)\}$ has rank n in some neighborhood of the origin in $(x_1, x_2, \ldots, x_n)$ space, and (b) the set of vector fields $g, [f,g], \ldots, (ad^{n-2}f,g)$ is involutive in some neighborhood of the origin in the same space. By condition (a) above, $g, [f,g], \ldots, (ad^{n-2}f,g)$ are linearly independent, and the Frobenius theorem implies the existence of a $\mathscr{C}^\infty$ $n-1$ dimensional integral manifold of $g, [f,g], \ldots, (ad^{n-2}f,g)$, using condition (b) above.

From reference 1 we know that a transformation $T = (T_1, T_2, \ldots, T_n)$ must satisfy the system of partial differential equations

$$\begin{aligned} &\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} g_j = 0, \quad i = 1, \ldots, n-1, \\ &\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} f_j = T_{i+1}, \quad i = 1, \ldots, n-1, \end{aligned} \tag{3}$$

and

$$\sum_{j=1}^{n} \frac{\partial T_n}{\partial x_j} (f_j + ug_j) = T_{n+1}.$$

Solving this system of equations is shown to be equivalent to finding a $\mathscr{C}^\infty$ function T_1 such that

$$\left.\begin{aligned} &\langle dT_1, (ad^k f,g)\rangle = 0, \quad k = 0, 1, \ldots, n-2 \\ &\langle dT_1, (ad^{n-1}f,g)\rangle \neq 0, \end{aligned}\right\} \tag{4}$$

where $\langle .,. \rangle$ denotes the duality of one-forms and vector fields. Thus a transformation satisfying conditions (i) through (iv) (specified in the Introduction) exists if we can find such a T_1 that vanishes at the origin.

3. GENERAL RESULTS

We show that the Poincaré lemma can be used to discover all of the $\mathscr{C}^\infty$ functions T_1 satisfying equations (4) under the assumptions that the matrix $\{g, [f,g], \ldots, (ad^{n-1}f,g)\}$ has rank n and the set $g, [f,g], \ldots, (ad^{n-2}f,g)$ is involutive. The transformations $T = (T_1, T_2, \ldots, T_{n+1})$ with conditions (i) through (iv) (from the Introduction) holding can be found by applying equations (3).

Example 1: We examine the nonlinear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ 0 \end{bmatrix} + u \begin{bmatrix} 0 \\ 1 \end{bmatrix} = f(x) + ug(x)$$

on $\mathbb{R}^2$. Computing the first Lie bracket we have

$$[f,g] = -\begin{bmatrix}\cos x_2 \\ 0\end{bmatrix},$$

and g and $[f,g]$ are linearly independent on $\{(x_1,x_2): -\pi/2 < x_2 < \pi/2\} = U$. Certainly the vector field $g = \begin{bmatrix}0\\1\end{bmatrix}$ is nonvanishing, so there does exist a transformation (T_1,T_2,T_3) with (x_1,x_2) in a neighborhood of the origin in $\mathbb{R}^2$ [(T_1,T_2) is a diffeomorphism for a sufficiently small neighborhood].

We show that the Poincaré lemma implies the existence of a transformation with (x_1,x_2) in U. We need to demonstrate the existence of a $\mathscr{C}^\infty$ function T_1 satisfying equations (4). Consider the one-form $\omega = 1\,dx_1 + 0\,dx_2$ on U. Since it is closed and U is contractible to the origin, there is a $\mathscr{C}^\infty$ function T_1 satisfying $dT_1 = \omega$ by Poincaré. Now $\langle dT_1,g\rangle = 0$, and since g and $[f,g]$ are linearly independent on U and $dT_1 \neq 0$, $\langle dT_1,[f,g]\rangle \neq 0$. The transformation T exists for $(x_1,x_2)\varepsilon U$, and taking T_1 with $T_1(0) = 0$ conditions i) through iv) hold as desired. Such a transformation is $(T_1,T_2,T_3) = (x_1,\sin x_2,(\cos x_2)u)$, which is one-to-one on U.

<u>Example 2</u>: Consider the system on $\mathbb{R}^2$

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}f_1(x)\\ f_2(x)\end{bmatrix} + u\begin{bmatrix}g_1(x)\\ g_2(x)\end{bmatrix} = f + ug\,.$$

Assume that there exists a smoothly contractible neighborhood U of the origin in (x_1,x_2) space on which

1. g is nonzero and $\partial g_2/\partial x_2 = -(\partial g_1/\partial x_1)$
2. g and $[f,g]$ are linearly independent.

The one-form $\omega = g_2\,dx_1 - g_1\,dx_2$ is closed on U ($d\omega = (\partial g_2/\partial x_2 + \partial g_1/\partial x_1)dx_2 \wedge dx_1$)), so there exists a $\mathscr{C}^\infty$ function T_1 [we take $T_1(0) = 0$] so that $dT_1 = \omega$. Again $\langle dT_1,g\rangle = 0$ and $\langle dT_1,[f,g]\rangle \neq 0$ by design.

<u>Example 3</u>: Suppose we have

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}f_1(x)\\ f_2(x)\end{bmatrix} + \begin{bmatrix}g_1(x)\\ g_2(x)\end{bmatrix} = f + ug$$

on $\mathbb{R}^2$. Assume there exists a smoothly contractible neighborhood U of the origin in (x_1,x_2) space on which

1. g is nonzero
2. g and $[f,g]$ are linearly independent
3. $[[f,g],g] = \alpha g$ for some $\mathscr{C}^\infty$ function α.

Assumption (3) above is due to Brockett[6] and is one of his conditions for a transformation (using coordinate changes and additive feedback) to a linear system. In reference 1 it is shown that a transformation of the type discussed in this paper exists in a neighborhood of the origin under assumptions (1) and (2) only. It is interesting to note that Example 1 does not satisfy assumption (3).

We take the one-form

$$\omega = \frac{g_2}{\det C}\,dx_1 - \frac{g_1}{\det C}\,dx_2\,,$$

where C is the 2×2 matrix with columns g and $[f,g]$, which is nonsingular on U. This form is closed if and only if

$$\frac{\partial\left(\frac{g_2}{\det C}\right)}{\partial x_2} = -\frac{\partial\left(\frac{g_1}{\det C}\right)}{\partial x_1}$$

Letting $[f,g] = h$, $[h,g] = \alpha g$ gives

$$\frac{\partial g_1}{\partial x_1}h_1 + \frac{\partial g_1}{\partial x_2}h_2 - \frac{\partial h_1}{\partial x_1}g_1 - \frac{\partial h_1}{\partial x_2}g_2 = \alpha g_1$$

and

$$\frac{\partial g_2}{\partial x_1}h_1 + \frac{\partial g_2}{\partial x_2}h_2 - \frac{\partial h_2}{\partial x_1}g_1 - \frac{\partial h_2}{\partial x_2}g_2 = \alpha g_2\,.$$

Substituting these last two equations into

$$\frac{\partial\left(\frac{g_2}{\det C}\right)}{\partial x_2} + \frac{\partial\left(\frac{g_1}{\det C}\right)}{\partial x_1}$$

we get 0; that is, assumption (3) implies that the one-form ω is closed on U. By the Poincaré lemma there exists a

$\mathscr{C}^\infty$ function T_1 on U with $dT_1 = \omega$. The facts that $\langle dT_1, \omega \rangle = 0$ and $\langle dT_1, [f,g] \rangle \neq 0$ follow easily.

We consider the general case for equations (1) and suppose that the rank assumption on the matrix $\{g, [f,g], \ldots, (ad^{n-1}f,g)\}$ and the involutive assumption on $g, [f,g], \ldots, (ad^{n-2}f,g)$ hold on some open neighborhood of the origin in $(x_1, x_2, \ldots, x_n)$ space. We seek a solution T_1 of equations (4). It is important to remember that first-order linear partial differential equations are solved by reducing them to systems of ordinary differential equations. At each stage of our constructive procedure we give an equation for T_1 to satisfy, but we wait until a construction is completed and let the Poincaré lemma give a solution T_1 to all of these equations. Since f and g are complete and $g, [f,g], \ldots, (ad^{n-1}f,g)$ are linearly independent, we know that $[f,g], \ldots, (ad^{n-1}f,g)$ are complete.

For all $s \in \mathbb{R}$ we solve the system

$$\frac{dx}{ds} = (ad^{n-1}f,g)$$

with initial conditions $x(0) = 0$ to obtain the unique integral curve $x(s)$ of $(ad^{n-1}f,g)$ through the origin in $(x_1, x_2, \ldots, x_n)$ space. The partial differential equation $\langle dT_1, g \rangle = 0$ is solved by considering for all $t_1 \in \mathbb{R}$ the system

$$\frac{dx}{dt_1} = g \ , \quad \frac{dT_1}{dt_1} = 0 \ .$$

We denote by $x(s,t_1)$ the solution of the first system $dx/dt_1 = g$ with initial conditions $x(s,0) = x(s)$.

We then examine for all $t_2 \in \mathbb{R}$ the system

$$\frac{dx}{dt_2} = [f,g] \ , \quad \frac{dT_1}{dt_2} = 0 \ .$$

In this way we deal with the partial differential equation $\langle dT_1, [f,g] \rangle = 0$. The solution of $dx/dt_2 = [f,g]$ with initial conditions $x(s,t_1,0) = x(s,t_1)$ is $x(s,t_1,t_2)$. Continuing this process, the last step involves the system

$$\frac{dx}{dt_{n-1}} = (ad^{n-2}f,g) \ , \quad \frac{dT_1}{dt_{n-1}} = 0 \ ,$$

which is associated with the partial differential equation $\langle dT_1, (ad^{n-2}f,g) \rangle = 0$. By $x(s, t_1, \ldots, t_{n-1})$ we denote the integral curves of the vector field $(ad^{n-2}f,g)$ with $s, t_1, \ldots, t_{n-2}$ behaving like parameters.

Since $g, [f,g], \ldots, (ad^{n-1}f,g)$ are linearly independent, the matrix

$$\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t_1} & \cdots & \frac{\partial x_1}{\partial t_{n-1}} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t_1} & \cdots & \frac{\partial x_2}{\partial t_{n-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial s} & \frac{\partial x_n}{\partial t_1} & \cdots & \frac{\partial x_n}{\partial t_{n-1}} \end{bmatrix}$$

called the <u>noncharacteristic matrix</u>, is nonsingular. Thus $s, t_1, \ldots, t_{n-1}$ can be solved for $x_1, x_2, \ldots, x_n$ in a neighborhood of the origin. We denote the map

$$(s,t_1,\ldots,t_{n-1}) \to (x_1(s,t_1,\ldots,t_{n-1}), x_2(s,t_1,\ldots,t_{n-1}), \ldots, x_n(s,t_1,\ldots,t_{n-1}))$$

by F and note that its Jacobian matrix is the noncharacteristic matrix.

We are now ready to prove our result concerning mappings of nonlinear systems to linear systems.

<u>Theorem 1</u>: Let U be an open subset in $(x_1, x_2, \ldots, x_n)$ space containing the origin and suppose that

(a) $g, [f,g], \ldots, (ad^{n-1}f,g)$ are linearly independent on U, and

(b) $g, [f,g], \ldots, (ad^{n-2}f,g)$ are involutive on U. Also assume the mapping F is one-to-one and that the noncharacteristic matrix is nonsingular on an open subset V, which is smoothly contractible to the origin in $(s, t_1, \ldots, t_{n-1})$ space and with $U \subset F(V)$. Then there exists a transformation $T = (T_1, T_2, \ldots, T_{n+1}): U \to \mathbb{R}^{n+1}$ taking system (1) to system (2) so that (i) T has a nonsingular Jacobian matrix xith respect to $(x_1, \ldots, x_n, u)$, (ii) $T(0) = 0$, (iii) $T_1, T_2, \ldots, T_n$ are independent of u and have a nonsingular $n \times n$ Jacobian matrix, and (iv) T_{n+1} is a function of $(x_1, \ldots, x_n, u)$ which can be inverted as a function of u.

<u>Proof</u>: Recall that we need to find a $\mathscr{C}^\infty$ function T_1 that vanishes at the origin in $(x_1, x_2, \ldots, x_n)$ space and satisfies equations (4). We have constructed the function $F(s, t_1, \ldots, t_{n-1})$ whose Jacobian matrix is the noncharacteristic matrix.

Let $\phi(s)$ by any $\mathscr{C}^\infty$ function of s which does not vanish on V (think of ϕ as defined on V because of its independence in $(t_1, t_2, \ldots, t_{n-1})$). Consider the one-form $\omega = \phi(s)ds + 0\, dt_1 + \ldots + 0\, dt_{n-1}$ on V. Since it is obviously closed on V there exists a $\mathscr{C}^\infty$ function $T_1(s)$ satisfying $dT_1 = \omega$ such that $T_1(0) = 0$.

Certainly

$$\frac{\partial T_1}{\partial t_1} = 0, \frac{\partial T_1}{\partial t_2} = 0, \ldots, \frac{\partial T_1}{\partial t_{n-1}} = 0$$

on V, but how does this relate to the equations $\langle dT_1, g\rangle = 0$, $\langle dT_1, [f,g]\rangle = 0, \ldots \langle dT_1, (ad^{n-2}f,g)\rangle = 0$? Since $g, [f,g], \ldots, (ad^{n-2}f,g)$ are involutive on U, the Frobenius theorem and our construction of the map F imply that for each fixed s we have an $n-1$ dimensional integral manifold of $g, [f,g], \ldots, (ad^{n-2}f,g)$ as $t_1, t_2, \ldots, t_{n-1}$ vary. Now, T_1 is constant on each integral manifold and must solve $\langle dT_1, g\rangle = 0$, $\langle dT_1, [f,g]\rangle = 0, \ldots, \langle dT_1, (ad^{n-2}f,g)\rangle = 0$ as functions of $s, t_1, \ldots, t_{n-1}$. Our assumptions on F and the noncharacteristic matrix imply that we can solve for $(s, t_1, \ldots, t_{n-1})$ as functions of $(x_1, x_2, \ldots, x_n)$ and hence the desired partial differential equations are solved on U.

Since $g, [f,g], \ldots, (ad^{n-1}f,g)$ are linearly independent on U and

$$dT_1 = \frac{\partial T_1}{\partial s}\frac{\partial s}{\partial x_1} dx_1 + \frac{\partial T_1}{\partial s}\frac{\partial s}{\partial x_2} dx_2 + \ldots + \frac{\partial T_1}{\partial s}\frac{\partial s}{\partial x_n} dx_n = \phi(s)ds$$

is nonzero on U, the equation $\langle dT_1, (ad^{n-1}f,g)\rangle \neq 0$ holds on the set U.

We conclude with the following example.

<u>Example 4</u>: We take the nonlinear system

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}\sin x_2\\ \sin x_3\\ 0\end{bmatrix} + u\begin{bmatrix}0\\0\\1\end{bmatrix} = f(x(t)) + ug(x(t))$$

on $\mathbb{R}^3$. Computing Lie brackets we have

$$[f,g] = -\begin{bmatrix}0\\ \cos x_3\\ 0\end{bmatrix},$$

$$(ad^2f,g) = \begin{bmatrix}\cos x_2 \cos x_3\\ 0\\ 0\end{bmatrix},$$

$$[g,[f,g]] = \begin{bmatrix}0\\ \sin x_3\\ 0\end{bmatrix}.$$

Thus conditions (a) and (b) of Theorem 1 are fulfilled on the open set

$$U = \left\{(x_1,x_2,x_3) \in \mathbb{R}^3 : -\frac{\pi}{2} < x_2 < \frac{\pi}{2}, -\frac{\pi}{2} < x_3 < \frac{\pi}{2}\right\}.$$

The solution of

$$\frac{dx}{ds} = (ad^2 f, g) \ , \quad x(0) = 0$$

is $x_1(s) = s$, $x_2(s) = 0$, and $x_3(s) = 0$. The system

$$\frac{dx}{dt_1} = g$$

with initial conditions $x_1(s,0) = s$, $x_2(s,0) = 0$, and $x_3(s,0) = 0$ has the three tuple solution $x_1(s,t_1) = s$, $x_2(s,t_1) = 0$, and $x_3(s,t_1) = t_1$. Similarly, for

$$\frac{dx}{dt_2} = [f,g] \ ,$$

satisfying $x_1(s,t_1,0) = s$, $x_2(s,t_1,0) = 0$, and $x_3(s,t_1,0) = t_1$, we find $x_1(s,t_1,t_2) = x_1 = s$, $x_2(s,t_1,t_2) = x_2 = (-\cos t_1)t_2$, and $x_3(s,t_1,t_2) = x_3 = t_1$. The mapping $F = (s,(-\cos t_1)t_2,t_1)$ is one-to-one on the smoothly contractible to $(0,0,0)$ set

$$V = \left\{(s,t_1,t_2) \in \mathbb{R}^3 : -\frac{\pi}{2} < t_1 < \frac{\pi}{2}\right\} \ ,$$

and the noncharacteristic matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & (\sin t_1)t_2 & (-\cos t_1) \\ 0 & 1 & 0 \end{bmatrix}$$

is nonsingular on this set. Solving for (s,t_1,t_2) as functions of (x_1,x_2,x_3) we find $s = x_1$, $t_1 = x_3$, and $t_2 = -(x_2/\cos x_3)$, and these are defined on U (i.e., $U \subset F(V)$).

Hence by Theorem 1, a transformation with the desired properties exists. Taking $T_1 = s$, one such transformation is

$$T_1 = x_1 \ ,$$

$$T_2 = \sin x_2 \ ,$$

$$T_3 = \cos x_2 \sin x_3 \ ,$$

$$T_4 = -\sin x_2 \sin^2 x_3 + (\cos x_2 \cos x_3)u \ .$$

Other choices of T_1 yield similar transformations.

Roger Brockett presented results concerning mappings of nonlinear systems to linear systems and provided a construction similar to that preceding Theorem 1 at a 1978 CBMS conference held at the University of California, Davis.

4. REFERENCES

1. Su, R.: On the Linear Equivalents of Nonlinear Systems. Submitted to Systems and Control Letters.
2. Hunt, L. R.; and Su, R.: Global Transformations of Nonlinear Systems. Submitted to IEEE Trans. on Autom. Contr.
3. Meyer, G.; and Cicolani, L.: A Formal Structure for Advanced Automatic Flight Control Systems. NASA TN D-7940, 1975.
4. Meyer, G.; and Cicolani, L.: Application of Nonlinear System Inverses to Automatic Flight Control Design-System Concepts and Flight Evaluations. AGARDograph on Theory and Applications of Optimal Control in Aerospace Systems, P. Kant, ed., 1980.
5. Krener, A. J.: On the Equivalence of Control Systems and the Linearization of Nonlinear Systems. SIAM J. Control, Vol. 11, 1973, pp. 670-676.
6. Brockett, R. W.: Feedback Invariants for Nonlinear Systems. IFAC Congress, Helsinki, 1978.
7. Spivak, M.: Differential Geometry. Vol. I. Publish or Perish, Inc., Berkeley, Calif., 1970.

5. BIOGRAPHIES

L. R. Hunt was born in Shreveport, La., on Dec. 5, 1942. He received his B.S. degree in 1964 from Baylor University and his Ph.D. degree in 1970 from Rice University, both in mathematics. He is currently on leave from Texas Tech University, working at Ames Research Center, NASA, Moffett Field, Calif.

Dr. Hunt's research interests are nonlinear systems and control, several complex variables, and partial differential equations. He is a member of the Institute of Electrical and Electronics Engineers, the American Mathematical Society, and the Society for Industrial and Applied Mathematics.

Renjeng Su was born in China in 1950. He received a B.S. in chemical engineering from Chen-Kung University in Taiwan in 1972. His M.S. and D.Sc. are both in systems science and mathematics from Washington University at St. Louis, Mo., in 1980. Currently, he is a research associate of National Research Council at Ames Research Center, NASA, Moffett Field, Calif.

TRANSFORMING NONLINEAR SYSTEMS

L. R. HUNT AND RENJENG SU

AMES RESEARCH CENTER, NASA
MOFFETT FIELD, CALIFORNIA

TRANSFORMING NONLINEAR SYSTEMS*

L. R. Hunt And Renjeng Su†

ABSTRACT

We consider nonlinear systems of the form

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t))$$

where $f, g_1, \ldots, g_m$ are C^∞ complete vector fields on $\mathbb{R}^n$ and $f(0)=0$. Because of the amount of literature devoted to the study of linear time invariant systems, it is reasonable to ask necessary and sufficient conditions for the above system to be transformable to a linear system. Using such a transformation we could construct a regulator for the nonlinear system by building one for the linear system (G. Meyer has done this in his study of automatic flight control). It is the purpose of this paper to find conditions (depending on Lie brackets) for a transformation to exist. Basically, we choose a canonical form for a linear time invariant system and investigate the possible mapping of our nonlinear system to that canonical form.

I. Introduction

Suppose we have a nonlinear plant that we are to control to perform some task. For example, an aircraft which is designed to automatically fly

*Research supported by NASA Ames Research Center under the IPA Program and the Joint Services Electronics Program at Texas Tech University under ONR Contract N00014-76-C-1126.
†Research Associate of National Research Council at Ames Research Center.

PRECEDING PAGE BLANK-NOT FILMED

a designated path despite modelling errors and disturbances (see [1], [2], [3]). We consider the mathematics associated with such a problem.

Assume the dynamics of our plant are described by the system

$$\dot{x}(t) = h(x,t,u), \tag{1}$$

where $x \in \mathbb{R}^n$ and h is a complete C^∞ vector field. Our research involves multi-input and time varying systems, but to save notation we emphasize the single input and time invariant system

$$\dot{x}(t) = h(x,u), \tag{2}$$

and mention the more general results.

Since the design theory for controllable linear time invariant systems is treated in the literature, if our nonlinear system is equivalent (say using coordinate changes and feedback) to a controllable linear system, then we can use this fact in our control problem.

Thus we are interested in characterizing those nonlinear systems which are transformable to controllable linear systems. The transformations we consider are maps $T = (T_1, T_2, \cdots, T_{n+1})$ which take $V \times \mathbb{R}$ (with variables $(x_1, x_2, \cdots, x_n, u)$), where V is an open neighborhood of the origin in state space, onto an open set in $\mathbb{R}^{n+1}$ ($(T_1, T_2, \cdots, T_{n+1})$ space) containing the origin, so that the following properties hold:

i) $T(0) = 0$

ii) $T_1, T_2, \cdots, T_n$ are functions of $x_1, x_2, \cdots, x_n$ only and have a nonsingular Jacobian matrix on V

iii) T_{n+1} is a function of $x_1, x_2, \cdots, x_n, u$ for which $\frac{\partial T_{n+1}}{\partial u}$ is nonzero on $V \times \mathbb{R}$

iv) $T_1, T_2, \cdots, T_{n+1}$ satisfy

$$(3) \qquad \begin{aligned} \dot{T}_1 &= T_2 \\ \dot{T}_2 &= T_3 \\ &\vdots \\ \dot{T}_n &= T_{n+1} \end{aligned} \qquad \text{or} \qquad \begin{bmatrix} \dot{T}_1 \\ \dot{T}_2 \\ \vdots \\ \dot{T}_n \end{bmatrix} = \begin{bmatrix} T_2 \\ T_3 \\ \vdots \\ 0 \end{bmatrix} + T_{n+1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

that is, $T_1, T_2, \cdots, T_n$ are the state variables and T_{n+1} the control for our controllable linear system

v) $T = (T_1, T_2, \cdots, T_{n+1})$ is one-to-one on $V \times \mathbb{R}$.

Results in this paper are local (near the origin in state space), but global theorems have been proved in [4].

It is shown in [5] that if the system (2) is transformable to the system (3), then it can be "reduced" to

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)), \tag{4}$$

where f and g are C^∞ vector fields on V, and we assume $f(0) = 0$. Hence we wish to map system (4) to system (3).

For related results concerning the transformations from nonlinear systems to linear systems we refer to the research of Krener [6], Brockett [7], Meyer [3], Jakubczyk and Respondek [8], and the authors [4],[5],[9], [10],[11].

In section II we give basic definitions and the partial differential equations we must solve to build a transformation. Section III contains the main result, a constructive proof of a transformation, and comments about the more general theory.

II. Definitions and Preliminaries

If f and g are C^∞ vector fields on an open set in $\mathbb{R}^n$, we define the <u>Lie bracket</u> of f and g

$$[f,g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g,$$

where $\frac{\partial g}{\partial x}$ and $\frac{\partial f}{\partial x}$ are $n \times n$ Jacobian matrices. We let

$$(ad^0 f,g) = g$$

$$(ad^1 f,g) = [f,g]$$

$$(ad^2 f,g) = [f,[f,g]]$$

$$\vdots$$

$$(ad^k f,g) = [f,(ad^{k-1} f,g)].$$

For h a C^∞ function on V we define the <u>Lie derivative of h with respect to f</u> as

$$L_f(h) = \langle dh,f\rangle,$$

with dh being the gradient and $\langle\cdot,\cdot\rangle$ being the duality between one forms and vector fields. If w is a C^∞ one form, then we have the <u>Lie derivative of w with respect to f</u>

$$L_f(w) = \left(\frac{\partial w^*}{\partial x} f\right)^* + w\frac{\partial f}{\partial x},$$

where * denotes transpose and $\frac{\partial w^*}{\partial x}$ and $\frac{\partial f}{\partial x}$ are Jacobian matrices.

A relation between the three types of Lie derivatives just defined is

$$L_f\langle w,g\rangle = \langle L_f(w),g\rangle + \langle w,[f,g]\rangle, \tag{5}$$

where g is a C^∞ vector field and f and w are as before.

In [5] it is proved we can transform system (4) to system (3) by $T = (T_1,T_2,\cdots,T_{n+1})$ if and only if $T_1,T_2,\cdots,T_{n+1}$ have linearly independent gradients and satisfy

$$\begin{aligned} &\langle dT_i,g\rangle = 0,\ i=1,2,\cdots,n-1 \\ &\langle dT_i,f\rangle = L_f(T_i) = T_{i+1},\ i=1,2,\cdots,n-1 \\ &\langle dT_n,f+ug\rangle = L_{f+ug}(T_n) = T_{n+1}\ . \end{aligned} \tag{6}$$

Using the formula (5) repeatedly these equations become

$$\begin{aligned} &\langle dT_1,(ad^k f,g)\rangle = 0,\ k=0,1,\ldots,n-2 \\ &(-1)^{n-1}u\langle dT_1,(ad^{n-1}f,g)\rangle + \langle dT_n,f\rangle = T_{n+1}. \end{aligned} \tag{7}$$

Hence we have a desired transformation T if and only if we can find a solution T_1 of

$$\begin{aligned} &\langle dT_1,(ad^k f,g)\rangle = 0,\ k=0,1,\cdots,n-2 \\ &\langle dT_1,(ad^{n-1}f,g)\rangle \neq 0 \end{aligned} \tag{8}$$

which vanishes at the origin. The remaining coordinate functions $T_2,T_3,\cdots,T_n$ are easily derived from (6).

III. Construction of a Transformation

A collection of C^∞ vector fields $a_1(x), a_2(x), \cdots, a_r(x)$ of $\mathbb{R}^n$ is called <u>involutive</u> if functions $\gamma_{ijk}(x)$ exist so that

$$[a_i, a_j](x) = \sum_{k=1}^{r} \gamma_{ijk}(x) a_k(x), \qquad 1 \le i, j \le r, i \ne j.$$

Our main result depends on the assumptions

a) the set $\{g, [f,g], \cdots, (ad^{n-1} f, g)\}$ spans an n dimensional space, and

b) the set $\{g, [f,g], \cdots, (ad^{n-2} f, g)\}$ is involutive.

<u>Theorem 3.1</u> Conditions a) and b) hold for $(x_1, x_2, \cdots, x_n)$ in some neighborhood of the origin in $\mathbb{R}^n$ if and only if there is an open set $V \subset \mathbb{R}^n$ containing the origin and a transformation $T = (T_1, T_2, \cdots, T_{n+1}) : V \times \mathbb{R} \to \mathbb{R}^{n+1}$ such that properties i) through v) hold.

This theorem is proved in [5] , and we illustrate a method for constructing such a transformation. For a real parameter $s_1 \in \mathbb{R}$ we solve

$$\frac{dx(s_1)}{ds_1} = (ad^{n-1} f, g)$$

with initial conditions $x(0) = 0$. Then we consider for $s_2 \in \mathbb{R}$ the system

$$\frac{dx(s_1, s_2)}{ds_2} = (ad^{n-2} f, g)$$

satisfying $x(s_1, 0) = x(s_1)$. Repeating the process n-2 more times we arrive at the final system

$$\frac{dx(s_1, s_2, \ldots, s_n)}{ds_n} = g$$

with initial conditions $x(s_1,s_2,\cdots,s_{n-1},0) = x(s_1,s_2,\cdots,s_{n-1})$. We have a map

$$(s_1,s_2,\cdots,s_n) \to (x_1(s_1,s_2,\cdots,s_n),\ x_2(s_1,s_2,\cdots,s_n),\cdots,x_n(s_1,s_2,\cdots,s_n))$$

with a Jacobian matrix at $(0,0,\cdots,0)$ equal to the matrix with columns $(ad^{n-1}f,g),(ad^{n-2}f,g),\cdots,[f,g],g$. Since this last matrix is nonsingular at the origin by condition a) we can solve for $(s_1,s_2,\cdots,s_n)$ as functions of $(x_1,x_2,\cdots,x_n)$ in an open neighborhood of the origin using the inverse function theorem.

If we can find a function T_1 which solves (8) as a function of $s_1,s_2,\cdots,s_n$, then we know T_1 as a function of $x_1,x_2,\cdots,x_n$. Because $g,[f,g],\cdots,(ad^{n-2}f,g)$ are involutive, as we fix s_1 and let $s_2,s_3,\cdots,s_n$ vary we get an integral manifold of this involutive set by the famous Frobenius Theorem. Hence if we choose $T_1 = s_1$, then we have a solution of (8). As was mentioned earlier $T_2,T_3,\cdots,T_{n+1}$ are found by differentiation.

An illustration of this technique is given in the following example.

<u>Example 3.2</u> Consider the nonlinear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ x_3 - x_1^{10} \\ 0 \end{bmatrix} + u \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = f(x(t)) + u(t)g(x(t)))$$

for $(x_1,x_2,x_3) \in V$ with

$$V = \{(x_1,x_2,x_3) : -\frac{\pi}{2} < x_2 < \frac{\pi}{2}\}.$$

Computing we find

$$[f,g] = -\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad (ad^2 f,g) = \begin{bmatrix} \cos x_2 \\ 0 \\ 0 \end{bmatrix}.$$

and conditions a) and b) of Theorem 3.1 hold on V. Solving

$$\frac{dx}{ds_1} = (ad^2 f,g)$$

$$\frac{dx}{ds_2} = [f,g]$$

$$\frac{dx}{ds_3} = g$$

in order and with the correct initial conditions we have $x_1 = s_1$, $x_2 = -s_2$, $x_3 = s_3$. Thus our transformation $T = (T_1,T_2,T_3,T_4)$ is

$$T_1 = x_1$$

$$T_2 = \sin x_2$$

$$T_3 = (\cos x_2)(x_3 - x_1^{10})$$

$$T_4 = (\cos x_2)(-10x_1^9)\sin x_2 + (-\sin x_2)(x_3 - x_1^{10})^2 + (\cos x_2)u.$$

It is interesting to remark how these transformations for (4) are used in practice. We want to choose the control u to drive the plant. If all state variables are available to us (estimates can be used if they are not), we map to the linear system and choose our control T_{n+1}. Then to

find u we just have to solve the equation

$$(-1)^{n-1}u < dT_1,(ad^{n-1}f,g) > + < dT_n,f > = T_{n+1}$$

which is linear in u.

If we consider the multi-input system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t)g_i(x(t)) \tag{9}$$

then for our target controllable linear system we choose a Brunovsky [12] canonical form associated with a set of Kronecker indices $\kappa_1,\kappa_2,\cdots,\kappa_m$. A transformation $T = (T_1,T_2,\cdots,T_{n+m})$ exists if and only if the following conditions are satisfied for all $(x_1,x_2,\cdots,x_n)$ near the origin. The set

$$C = \{g_1,[f,g_1],\cdots,(ad^{\kappa_1-1}f,g_1),g_2,[f,g_2],\cdots,(ad^{\kappa_2-1}f,g_2),\cdots,$$
$$g_m,[f,g_m],\cdots,(ad^{\kappa_m-1}f,g_m)\}$$

spans an n dimensional space, the sets $C_j \cap C$ with

$$C_j = \{g_1,[f,g_1],\cdots,(ad^{\kappa_j-2}f,g_1),g_2,[f,g_2],\cdots,(ad^{\kappa_j-2}f,g_2),\cdots,$$
$$g_m,[f,g_m],\cdots,(ad^{\kappa_j-2}f,g_m)\}$$

are involutive, and the span of each C_j equals the span of $C_j \cap C$ for $j = 1,2,\cdots,m$. This result is proved in [9], and a discussion [11] of this topic was presented at the recent JACC meeting.

For time varying nonlinear systems one must replace the Lie bracket $[\cdot,\cdot]$ with a time varying Lie derivative (see [10], [13], and [14]). The time variable t appears as a parameter in constructing the transformation,

and we map to a controllable linear time invariant system.

In system (4) we assume that $f(0) = 0$. Current research is in progress for which the origin may not be an equilibrium point of the vector field f.

The authors wish to thank George Meyer for valuable conversations.

REFERENCES

[1] G. Meyer and L. Cicolani, A formal structure for advanced flight control systems, NASA TND-7940, 1975.

[2] G. Meyer and L. Cicolani, Applications of nonlinear system inverses to automatic flight control design-system concepts and flight evaluations, AGARDograph on Theory and Applications of Optimal Control in Aerospace Systems, P. Kant, ed., 1980.

[3] G. Meyer, The design of exact nonlinear model followers, 1981 Joint Automatic Control Conference, to appear.

[4] L. R. Hunt and R. Su, Global transformations of nonlinear systems, submitted.

[5] R. Su, On the linear equivalents of nonlinear systems, submitted.

[6] A. J. Krener, On the equivalence of control systems and linearization of nonlinear systems, SIAM J. Control 11(1973), 670-676.

[7] R.W. Brockett, Feedback invariants for nonlinear systems, IFAC Congress, Helsinki, 1978.

[8] B. Jakubczyk and W. Respondek, On linearization of control systems, Bull. Acad. Polon. Sci., Ser. Sci. Math. Astronom. Phys., to appear.

[9] L. R. Hunt and R. Su, Multi-input nonlinear systems, submitted.

[10] L. R. Hunt, G. Meyer, and R. Su, Time varying systems, in preparation.

[11] L. R. Hunt and R. Su, Local transformations for multi-input nonlinear systems, 1981 Joint Automatic Control Conference, to appear.

[12] P. Brunovsky, A classification of linear controllable systems, Kibernetica (Praha)6(1970), 173-187.

[13] L. R. Hunt and R. Su, Linear Equivalents of time varying nonlinear systems, 1981 International Symposium on Mathematical Theory of Networks and Systems, to appear.

[14] L. R. Hunt and R. Su, Control of nonlinear time varying systems, submitted.

L. R. Hunt and Renjeng Su
MS 210-3
NASA Ames Research Center
Moffett Field, CA 94035
415-965-5453

Hunt is on leave from
Department of Mathematics
Texas Tech University
Lubbock, TX 79409

GLOBAL MAPPINGS OF NONLINEAR SYSTEMS

L. R. HUNT AND RENJENG SU

AMES RESEARCH CENTER, NASA
MOFFETT FIELD, CALIFORNIA

Global Mappings of Nonlinear Systems

L. R. Hunt* and Renjeng Su

Ames Research Center, NASA
Moffett Field, California 94035

If f and g are complete $\mathscr{C}^\infty$ vector fields on $\mathbb{R}^n$ we examine the nonlinear system

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

We find sufficient conditions for the existence of a global transformation defined for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ which takes our nonlinear system to a linear system. A constructive proof is given of the transformation by solving a system of partial differential equations. We require that the vector fields $g, [f,g], \ldots, (ad^{n-1}f, g)$ span an n-dimensional space at each point of $\mathbb{R}^n$, that the set $\{g, [f,g], \ldots, (ad^{n-2}f, g)\}$ is involutive on $\mathbb{R}^n$, and that the noncharacteristic matrix defined in our construction of the transformation satisfies the assumptions of various global inverse function theorems.

*On leave from Department of Mathematics, Texas Tech University, Lubbock, Texas, 79409.

PRECEDING PAGE BLANK-NOT FI

NOMENCLATURE

$\mathscr{C}^\infty$ infinitely differentiable functions

$\mathbb{R}^n$ n dimensional real Euclidean space

u_i controls

$\dot{x}$ time derivative of x

I. INTRODUCTION

Suppose we have the nonlinear system

$$\dot{x}(t) = f(x,t) + \sum_{i=1}^{m} u_i(t) g_i(x,t) , \qquad (1)$$

where $f, g_1, \ldots, g_m$ are complete (no finite escape time) $\mathscr{C}^\infty$ vector fields on $\mathbb{R}^n$ and $f(0,t) = 0$ for each t. Recent results have contained necessary and sufficient conditions for there to be a local transformation of system (1) to a controllable time invariant linear system. We combine these results with certain versions of the global inverse function theorem to yield conditions under which a global transformation exists. To simplify notation in this paper we restrict to the single input time invariant (autonomous) system

$$\dot{x}(t) = f[x(t)] + u(t) g[x(t)] , \qquad (2)$$

but our theory will generalize to the system (1).

In this single input case we map to the linear system in integrator form

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{y}_{n-1} \\ \dot{y}_n \end{bmatrix} = \begin{bmatrix} y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \\ 0 \end{bmatrix} + v \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} . \qquad (3)$$

If we were considering the general system (1) then we attempt to map to a linear system in its Brunovsky (3) canonical form associated with the Kronecker indices.

We indicate the mappings of interest to us. A $\mathscr{C}^\infty$ transformation $T = (T_1, T_2, \ldots, T_{n+1})$ takes $\mathbb{R}^{n+1}$ ($(x_1, x_2, \ldots, x_n, u)$ space) to $\mathbb{R}^{n+1}$ ($(T_1, T_2, \ldots, T_n, T_{n+1}) = (y_1, y_2, \ldots, y_n, v)$ space) so that T has the following properties:

i) $T(0) = 0$,

ii) $T_1, T_2, \ldots, T_n$ are functions of $x_1, x_2, \ldots, x_n$ only and have a nonsingular Jacobian matrix on $\mathbb{R}^n$,

iii) T_{n+1} is a function of $x_1, x_2, \ldots, x_n, u$ which can be inverted in terms of u for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$,

iv) $T_1, T_2, \ldots, T_n$ are the state variables and T_{n+1} the control for our linear system (3),

v) $T = (T_1, T_2, \ldots, T_{n+1})$ is a one-to-one mapping of $\mathbb{R}^{n+1}$ to $\mathbb{R}^{n+1}$ taking system (2) to system (3).

We denote by $[f,g]$ the Lie brackets of our vector fields f and g, $(ad^1 f, g) = [f,[f,g]], \ldots,$ $(ad^k f, g) = [f,(ad^{k-1} f, g)]$. If the set $\{g, [f,g], \ldots, (ad^{n-1} f, g)\}$ spans an n-dimensional space at the origin in $\mathbb{R}^n$ and the set $\{g, [f,g], \ldots, (ad^{n-2} f, g)\}$ is involutive for all points near the origin, then system (2) can be

transformed to system (3) for all $(x_1, x_2, \ldots, x_n)$ in some open neighborhood of the origin. If these conditions on $\{g, [f,g], \ldots, (ad^{n-1}f, g)\}$ and $\{g, [f,g], \ldots, (ad^{n-2}f, g)\}$ hold at every point in R^n, we give a construction of a transformation $T = (T_1, T_2, \ldots, T_{n+1})$. In the process of building T we introduce the noncharacteristic matrix (named for the noncharacteristic condition in partial differential equations). If this matrix satisfies the ratio condition on R^n, or other conditions for which we can apply a global inverse function theorem, the transformation T exists for all $(x_1, x_2, \ldots, x_n) \in R^n$ and satisfies properties i) through v) listed above. For other results concerning the transformation from nonlinear systems to linear systems we refer to the work of Krener (13), Brockett (2), Meyer and Cicolani (14), (15), and Jakubczyk and Respondek (11), and the authors (7, 8, 9, 10). Our results depend on the local theory developed in (18). Details, proofs, and other theories concerning global transformations appear in (7).

Classical results involving global inverse function theorems are found in the work of Hadamard (4), (5), and (6). Other interesting theories in this direction are attributed to Palais (16), Berger and Berger (1), Wu and Desoer (19), Kou, Elliot, and Tarn (12), and Sandberg (17).

Section II contains definitions and the system of partial differential equations from (18) that we must solve in order to find a transformation of the desired type. In section III we state our main result, give an outline of its proof, and present several examples of its application.

II. DEFINITIONS

For $\mathscr{C}^\infty$ vector fields f and g on R^n (or generally on a differentiable manifold), we define the Lie bracket

$$[f,g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g$$

with $\partial g/\partial x$ and $\partial f/\partial x$ denoting $n \times n$ Jacobian matrices. We also have

$$\begin{aligned}
(ad^0 f, g) &= g \\
(ad^1 f, g) &= [f,g] \\
(ad^2 f, g) &= [f,[f,g]] \\
&\vdots \\
(ad^k f, g) &= [f, (ad^{k-1} f, g)] .
\end{aligned}$$

A collection of $\mathscr{C}^\infty$ vector fields $f_1, f_2, \ldots, f_r$ on R^n is called involutive if functions $a_{ijk}(x)$ exist such that

$$[f_i, f_j](x) = \sum_{k=1}^{r} a_{ijk}(x) f_k(x), \quad 1 \le i,j \le r, \quad i \ne j .$$

Given a point $x_0 \in R^n$ and an involutive set $\{f_1, f_2, \ldots, f_r\}$ of vector fields on R^n, then there exists a unique maximal r-dimensional $\mathscr{C}^\infty$ submanifold S of R^n containing x_0 so that the tangent space to S at each point $x \in S$ is the space spanned by $f_1(x), f_2(x), \ldots, f_r(x)$. In this case S is the integral manifold of $f_1, f_2, \ldots, f_r$ through x_0 and exists by the Frobenius Theorem.

Let f be a $\mathscr{C}^\infty$ vector field and h a $\mathscr{C}^\infty$ function. Then the Lie derivative of h with respect to f is

$$L_f(h) = \langle dh, f \rangle$$

where dh is the gradient and $\langle \cdot, \cdot \rangle$ is the duality between one forms and vector fields. If

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ \cdot \\ f_n \end{bmatrix}$$

then

$$\langle dh, f \rangle = \frac{\partial h}{\partial x_1} f_1 + \frac{\partial h}{\partial x_2} f_2 + \ldots + \frac{\partial h}{\partial x_n} f_n .$$

For ω a $\mathscr{C}^\infty$ one form on R^n, we have the Lie derivative of ω with respect to f:

$$L_f(\omega) = \left(\frac{\partial \omega^*}{\partial x} f \right)^* + \omega \frac{\partial f}{\partial x} ,$$

where $*$ denotes the transpose and $\partial \omega^*/\partial x$ and $\partial f/\partial x$ are Jacobian matrices.

Our three Lie derivatives are related by the formula

$$L_f \langle \omega, g \rangle = \langle L_f(\omega), g \rangle + \langle \omega, [f,g] \rangle . \tag{4}$$

Recall that we want a transformation that maps our system (2) to system (3), which we rewrite in the $T = (T_1, T_2, \ldots, T_{n+1})$ coordinates (for R^{n+1}) as

$$\begin{aligned}
\dot{T}_1 &= T_2 \\
\dot{T}_2 &= T_3 \\
&\vdots \\
\dot{T}_n &= T_{n+1} .
\end{aligned}$$

From (18), necessary and sufficient conditions for the existence of a transformation for $(x_1, x_2, \ldots, x_n)$ in some neighborhood of the origin in R^n are:

a) The set $\{g, [f,g], \ldots, (ad^{n-1}f, g)\}$ spans an n-dimensional space, and
b) The set $\{g, [f,g], \ldots, (ad^{n-2}f, g)\}$ is involutive,

both in a neighborhood of the origin. A transformation T must satisfy the partial differential equations.

$$\left.\begin{aligned}
&\langle dT_i, g \rangle = 0, \quad i = 1, 2, \ldots, n-1 \\
&\langle dT_i, f \rangle = L_f(T_i) = T_{i+1}, \quad i = 1, 2, \ldots, n-1 \\
&\langle dT_n, f + ug \rangle = L_{f+ug}(T_n) = T_{n+1}
\end{aligned}\right\} \tag{5}$$

as shown in (18).

Using the formula (4) these equations become

$$\left.\begin{aligned}
&\langle dT_i, (ad^k f, g) \rangle = 0, \quad k = 0, 1, \ldots, n-2 \\
&\langle dT_n, f + ug \rangle = T_{n+1}
\end{aligned}\right\} \tag{6}$$

The second equation in (6) is the same as

$$(-1)^{n-1}\langle dT_1,(ad^{n-1}f,g)\rangle u + \langle dT_n,f\rangle = T_{n+1} .$$

Thus to find a transformation so that conditions i) through v) as given in the introduction hold, we must find a solution T_1 of

$$\left.\begin{aligned} &\langle dT_1,(ad^k f,g)\rangle = 0, \quad k = 0,1, \ldots, n-2 \\ &\langle dT_1,(ad^{n-1}f,g)\rangle \neq 0 \end{aligned}\right\} \quad (7)$$

that vanishes at the origin. The functions $T_2,T_3, \ldots, T_{n+1}$ are then easily derived from equations (5)

The local theory gives us a transformation which is applicable to a neighborhood of the origin in R^n, but gives us no idea of the size of the neighborhood. Next we construct a solution T_1 and introduce conditions under which the transformation $(T_1,T_2, \ldots, T_{n+1})$ is global.

III. CONSTRUCTION OF THE MAPPING

We begin this section by building a function T_1 which satisfies equations (7). The first $(n-1)$ equations are partial differential equations that are solved by reducing to ordinary differential equations.

Parameters $t_1,t_2, \ldots, t_{n-1}$ are introduced as follows. For all $t_1 \epsilon R$ we solve

$$\frac{dx}{dt_1} = (ad^{n-1}f,g)$$

with initial conditions $x(0) = 0$ to find the integral curve of $(ad^{n-1}f,g)$ through the origin. For every $t_2 \epsilon R$ we examine

$$\frac{dx}{dt_2} = (ad^{n-2}f,g)$$

satisfying $x(t_1,0) = \cancel{0}\ x(t_1)$. Continuing, we solve for all $t_3 \epsilon R$ the equation

$$\frac{dx}{dt_3} = (ad^{n-3}f,g)$$

with $x(t_1,t_2,0) = x(t_1,t_2)$. This argument is repeated until the final step is reached, solving

$$\frac{dx}{dt_n} = g$$

with the initial conditions

$$x(t_1,t_2, \ldots, t_{n-1},0) = x(t_1,t_2, \ldots, t_{n-1}) .$$

Thus we have a function

$$(t_1,t_2, \ldots,t_n) \to (x_1(t_1,t_2, \ldots,t_n),x_2(t_1,t_2, \ldots,t_n), \ldots,x_n(t_1,t_2, \ldots,t_n))$$

which has as its Jacobian matrix the <u>noncharacteristic matrix</u>

$$\begin{bmatrix} \frac{\partial x_1}{\partial t_1} & \frac{\partial x_1}{\partial t_2} & \cdots & \frac{\partial x_1}{\partial t_n} \\ \frac{\partial x_2}{\partial t_1} & \frac{\partial x_2}{\partial t_2} & & \frac{\partial x_2}{\partial t_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial t_1} & \frac{\partial x_n}{\partial t_2} & \cdots & \frac{\partial x_n}{\partial t_n} \end{bmatrix} \quad (8)$$

By design this matrix evaluated at the origin in R^n is the same as the matrix with columns $(ad^{n-1}f,g),(ad^{n-2}f,g), \ldots, g$ evaluated at $(0,0, \ldots, 0)$. This last matrix is nonsingular, so by the inverse function theorem there is an open neighborhood of the origin on which we can solve for $t_1,t_2, \ldots, t_n$ as functions of $x_1,x_2, \ldots, x_n$.

If we can find a solution T_1 of equation (7) as a function of $t_1,t_2, \ldots, t_n$ we have a solution in the x variables also.

We now use our assumption that $g,[f,g], \ldots, (ad^{n-2}f,g)$ are involutive. The Frobenius Theorem tells us that if we fix the parameter t_1 and let $t_2,t_3, \ldots, t_n$ vary we get an integral manifold of $g,[f,g], \ldots, (ad^{n-2}f,g)$. An obvious choice (our transformation is not unique) for T_1 is $T_1 = t_1$. This function is constant on each integral manifold of $g,[f,g], \ldots, (ad^{n-2}f,g)$ and

$$\frac{\partial T_1}{\partial t_2} = \frac{\partial T_1}{\partial t_3} = \cdots = \frac{\partial T_1}{\partial t_n} = 0 .$$

Hence the first $(n-1)$ equations in equations (7) hold.

Suppose $\langle dT_1,(ad^{n-1}f,g)\rangle = 0$. Since $T_1 = t_1$ we have $dT_1 \neq 0$, and $(ad^{n-1}f,g) \neq 0$ at the origin implies that $(ad^{n-1}f,g)$ is tangent to the integral manifold of $g,[f,g], \ldots, (ad^{n-2}f,g)$ at $(0,0, \ldots, 0)$, a contradiction to the assumption that $g,[f,g], \ldots, (ad^{n-1}f,g)$ are linearly independent there.

Hence there is an open neighborhood of the origin on which conditions i) through v) hold. Note that $T_1 = t_1$ and the mapping which has the noncharacteristic matrix as its Jacobian matrix maps the origin to the origin, implying that $T(0) = 0$.

The function T_1 satisfies

$$\langle dT_1,(ad^k f,g)\rangle = 0, \quad k = 0,1, \ldots, n-2$$

and

$$\frac{\partial T_1}{\partial t_2} = \frac{\partial T_1}{\partial t_3} = \cdots = \frac{\partial T_1}{\partial t_n} = 0 .$$

Likewise T_2 is a solution of

$$\langle dT_2,(ad^k f,g)\rangle = 0, \quad k = 0,1, \ldots, n-2$$

(use formula (4)), and

$$\frac{\partial T_2}{\partial t_3} = \frac{\partial T_2}{\partial t_4} = \cdots = \frac{\partial T_2}{\partial t_n} = 0 .$$

Continuing in this way we have T_3 independent of $t_4, t_5, \ldots, t_n$, and in general T_i does not depend on $t_{i+1}, t_{i+2}, \ldots, t_n$ for $i \le i \le n-1$. Since $T_1, T_2, \ldots, T_n$ are linearly independent and their Jacobian matrix with respect to $t_1, t_2, \ldots, t_n$ has all entries above the diagonal zero, the mapping $(T_1, T_2, \ldots, T_n)$ is one-to-one as a function in the t variables.

Hence, wherever assumptions a) and b) hold, the only possible obstruction to constructing a transformation $T = (T_1, T_2, \ldots, T_{n+1})$ with the desired properties is contained in the noncharacteristic matrix. Our global results depend on this matrix.

The first theorem that we need is found in (12).

Theorem 3.1

Suppose there is a map $H: \mathbb{R}^n \to \mathbb{R}^n$ which is differentiable with Jacobian matrix $J(x)$. If there exists a constant $c > 0$ such that the absolute values of the leading principal minors $\Delta_1, \Delta_2, \ldots, \Delta_n$ of $J(x)$ satisfy

$$|\Delta_1| \ge c, \quad \frac{|\Delta_2|}{|\Delta_1|} \ge c, \quad \ldots, \quad \frac{|\Delta_n|}{|\Delta_{n-1}|} \ge c$$

for all $x \in \mathbb{R}^n$, then H is one-to-one from $\mathbb{R}^n$ onto $\mathbb{R}^n$.

The condition stated on the absolute values of leading principal minors is called the ratio condition. Our first result is proved by applying Theorem 3.1, the construction of our transformation, and the comments made during and after the construction.

Theorem 3.2

For system (2) assume that the set $\{g, [f,g], \ldots, (ad^{n-1}f, g)\}$ spans an n-dimensional space at each point of $\mathbb{R}^n$ and the set $\{g, [f,g], \ldots, (ad^{n-2}f, g)\}$ is involutive on $\mathbb{R}^n$. If the noncharacteristic matrix (8) satisfies the ratio condition on $\mathbb{R}^n$, then there exists a $\mathcal{C}^\infty$ transformation $T = (T_1, T_2, \ldots, T_{n+1})$ with the following properties.

i) $T(0) = 0$,

ii) $T_1, T_2, \ldots, T_n$ are functions of $x_1, x_2, \ldots, x_n$ only and have a nonsingular Jacobian matrix on $\mathbb{R}^n$,

iii) T_{n+1} is a function $x_1, x_2, \ldots, x_n, u$ which can be inverted in terms of u for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$,

iv) $T_1, T_2, \ldots, T_n$ are the state variables and T_{n+1} the control for our linear system (3),

v) $T = (T_1, T_2, \ldots, T_{n+1})$ is a one-to-one mapping of $\mathbb{R}^{n+1}$ to $\mathbb{R}^{n+1}$ taking system (2) to system (3).

Example 3.3

We take the nonlinear system on $\mathbb{R}^2$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} x_1^2 + e^{x_2} + x_2 \\ x_1^2 \end{bmatrix} + u \begin{bmatrix} 0 \\ 1 \end{bmatrix} = f(x(t)) + ug(x(t)).$$

Our first Lie bracket

$$[f,g] = \begin{bmatrix} -(e^{x_2} + 1) \\ 0 \end{bmatrix}$$

is linearly independent from g on $\mathbb{R}^2$, and the involutive assumption on g is trivial for two dimensions.

Solving $dx_1/dt_1 = -(e^{x_2} + 1)$, $dx_2/dt_1 = 0$ with $x_1(0) = 0$ and $x_2(0) = 0$ we have $x_1 = -2t_1$ and $x_2 = 0$. Examining $dx_1/dt_2 = 0$ and $dx_2/dt_2 = 1$ with initial conditions $x_1(t_1, 0) = -2t_1$, and $x_2(t_1, 0) = 0$ we find $x_1 = -2t_1$, and $x_2 = t_2$. In this case the noncharacteristic matrix is

$$\begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix},$$

which fulfills the ratio conditions with $c = 1/2$. Hence our transformation

$$(T_1, T_2, T_3) = \Big(-\frac{x_1}{2}, -\frac{1}{2}\Big(\frac{1}{2} x_1^2 + e^{x_2} + x_2\Big) + \frac{1}{2} x_1, -\frac{1}{2} x_1 \Big(\frac{1}{2} x_1^2 + e^{x_2} + x_2\Big) + \frac{1}{2} x_1 - \frac{1}{2}(e^{x_2} + 1)(x_1^2 + u) \Big)$$

is defined on all of $\mathbb{R}^2$ and has properties i) through v).

The proof of the following corollary depends on results like Theorem 3.1 from (12).

Corollary 3.4

Suppose $\{g, [f,g], \ldots, (ad^{n-1}f, g)\}$ spans an n-dimensional space on $\mathbb{R}^n$ and $\{g, [f,g], \ldots, (ad^{n-2}f, g)\}$ is involutive there. If the assumption on the ratio condition for the noncharacteristic matrix in Theorem 3.1 is replaced by any of the following hypotheses, then the conclusions of that theorem remain valid.

1) There exists an $n \times n$ nonsingular constant matrix A such that A multiplied on the right by the noncharacteristic matrix satisfies the ratio condition on $\mathbb{R}^n$.

2) The noncharacteristic matrix (with a possible premultiplication by an $n \times n$ nonsingular constant matrix A) is positive definite on $\mathbb{R}^n$.

3) The determinant of the noncharacteristic matrix is positive on $\mathbb{R}^n$ and the sum of the noncharacteristic matrix and its adjoint has non-negative principal minors for all $x \in \mathbb{R}^n$. In this case, the noncharacteristic matrix may be premultiplied by an $n \times n$ nonsingular constant matrix A as before.

Often we cannot construct a transformation on all of $\mathbb{R}^n$ but are able to do so on some fixed open subset U of $\mathbb{R}^n$. In this direction we consider the following example.

Example 3.5

Let

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ \sin x_3 \\ 0 \end{bmatrix} + u \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = f(x(t)) + u(t) g(x(t))$$

and

$$U = \Big\{ (x_1, x_2, x_3) \in \mathbb{R}^3 : -\frac{\pi}{2} < x_2, x_3 < \frac{\pi}{2} \Big\}.$$

Our Lie brackets are

$$[f,g] = \begin{bmatrix} 0 \\ -\cos x_3 \\ 0 \end{bmatrix}$$

and

$$[f,[f,g]] = (ad^2 f,g) = \begin{bmatrix} \cos x_2 \cos x_3 \\ 0 \\ 0 \end{bmatrix},$$

which together with g span a three-dimensional space on U. Also

$$[g,[f,g]] = \begin{bmatrix} 0 \\ \sin x_3 \\ 0 \end{bmatrix}$$

implying that g and $[f,g]$ are involutive on U. Solving

$$\frac{dx}{dt_1} = (ad^2 f,g), \quad \frac{dx}{dt_2} = [f,g], \quad \frac{dx}{dt_3} = g$$

in order and with the proper initial conditions we find that $x_1(t_1,t_2,t_3) = t_1, x_2(t_1,t_2,t_3) = -(\cos t_2)t_3$, and $x_3(t_1,t_2,t_3) = t_2$. Solving for t_1,t_2,t_3 as functions of x_1,x_2,x_3 we obtain $t_1 = x_1, t_2 = x_3$, and $t_3 = x_2/(-\cos x_3)$, all valid on U. The transformation

$$T_1 = x_1$$

$$T_2 = \sin x_2$$

$$T_3 = \cos x_2 \sin x_3$$

$$T_4 = -\sin x_2 \sin^2 x_3 + (\cos x_2 \cos x_3)u$$

holds for all (x_1,x_2,x_3) in U.

CONCLUSIONS

We have given a theorem and its corollary in which we state conditions for having a global transformation of a nonlinear system to a linear system. In doing so we gave an explicit method for constructing such a transformation and introduced the important noncharacteristic matrix. Examples that illustrated our technique were presented.

REFERENCES

1 Berger, M. S. and Berger, M. S., Perspectives in Nonlinearity, Benjamin, New York, 1968.

2 Brockett, R. W., "Feedback Invariants for Nonlinear Systems," IFAC Congress, Helsinki, 1978.

3 Brunovsky, P., "A Classification of Linear Controllable Systems," Kibernetika (Praha), vol. 6, 1970, pp. 173-187.

4 Hadamard, J., "Sur Les Transformations Planes," C. R. Acad. Sci., Paris, vol. 142, 1906, pp. 71-64.

5 Hadamard, J., "Sur Les Transformations Ponctuelles," Bull. Soc. Math. France, vol. 34, 1906, pp. 77-84.

6 Hadamard, J., "Sur Les Correspondences," Oeuvres, pp. 383-384.

7 Hunt, L. R. and Su, R., "Global Transformation of Nonlinear Systems," submitted to IEEE Trans. on Autom. Contr.

8 Hunt, L. R. and Su, R., "Linear Equivalents of Time Varying Linear Systems," 1981 International Symposium on Mathematical Theory of Networks and Systems, to appear.

9 Hunt, L. R. and Su, R., "The Poincare Lemma and Transformations of Nonlinear Systems," 1981 International Symposium on Mathematical theory of Networks and Systems, to appear.

10 Hunt, L. R. and Su, R., "Multi-Input Nonlinear Systems," in progress.

11 Jakubczyk, B. and Respondek, W., "On Linearization of Control Systems," preprint.

12 Kou, S. R., Elliott, D. L. and Tarn, T. J., "Observability of Nonlinear Systems," Information and Control, Vol. 22, 1973, pp. 89-99.

13 Krener, A. J., "On the Equivalence of Control Systems and the Linearization of Nonlinear Systems," SIAM J. Control, Vol. 11, 1973, pp. 670-676.

14 Meyer, G. and Cicolani, L., "A Formal Structure for Advanced Automatic Flight Control Systems," NASA TN D-7940, 1975.

15 Meyer, G. and Cicolani, L., "Applications of Nonlinear System Inverses to Automatic Flight Control Design-System Concepts and Flight Evaluations," AGARDograph on Theory and Applications of Optimal Control in Aerospace Systems, P. Kant, ed., 1980.

16 Palais, R. S., "Natural Operations on Differential Forms," Trans. Amer. Math. Soc., Vol. 92, 1959, pp. 125-141.

17 Sandberg, I. W., "Global Implicit Function Theorems," IEEE Trans. on Circuits and Systems, Vol. 28, 1981, pp. 145-149.

18 Su, R., "On the Linear Equivalents of Nonlinear Systems," submitted to Systems and Control Letters.

19 Wu, F. F. and Desoer, C. A., "Global Inverse Function Theorem," IEEE Trans. on Circuit Theory, Vol. 19, 1972, pp. 199-201.

LOCAL TRANSFORMATIONS FOR MULTI-INPUT NONLINEAR SYSTEMS

L. R. HUNT AND R. SU

AMES RESEARCH CENTER, NASA
MOFFETT FIELD, CALIFORNIA

Local Transformations for Multi-Input Nonlinear Systems

L.R. HUNT* AND R. SU†

Ames Research Center, NASA
Moffett Field, California

ABSTRACT

In this paper Brockett's feedback invariants of nonlinear systems are generalized to a larger class of transformations. In terms of these invariants, we also extend our recent results on linear equivalents of nonlinear systems to the multiple-input case.

*Research supported by Ames Research Center under the IPA program and the Joint Services Electronics Program at Texas Tech U. under Contract N00014-76-C-1136.

†Research Associate of National Research Council at Ames Research Center.

PRECEDING PAGE BLANK-NOT FILMED

INTRODUCTION

By means of state space coordinate changes and feedback, a controllable linear system can be transformed into decoupled series of integrators, that is

$$\dot{x}_1 = x_2,\ \dot{x}_2 = x_3,\ \ldots,\ \dot{x}_{k_1} = u_1,$$

$$\dot{x}_{k_1+1} = x_{k_1+2},\ \dot{x}_{k_1+2} = x_{k_1+3},\ \ldots,$$

$$\dot{x}_{k_1+k_2} = u_2,$$

$$\dot{x}_{k_1+\ldots+k_{m-1}+1} = x_{k_1+\ldots+k_{m-1}+2},\ \ldots,$$

$$\dot{x}_{k_1+\ldots+k_m} = u_m.$$

This particular form is sometimes called the Brunovsky canonical form. It is also well known that the orders of these series of integrators are invariant under the transformations (1, 2); they are usually called the Kronecker indices.

In the literature, there have been some efforts made to generalize these linear results to nonlinear systems. Meyer and Cicolani (3) showed that the class of block-triangular systems can be transformed into the Brunovsky canonical form. Brockett (4) discovered an important class of invariants (discussed later) and also gave a necessary and sufficient condition for a single-input nonlinear system to be transformed into a single series of integrators. By enlarging the set of transformations, Su (5) obtained a still larger equivalence class. Later in (6) the authors extended the results to global transformations and also gave a way of constructing them.

The goal of this paper is to generalize the previous results to obtain a characterization of the class of linear equivalents of multiple-input nonlinear systems. Recently the authors were informed by Professor E. Sontag about similar results obtained by Jakubczyk and Respondek (7); we shall briefly discuss those results in the final section.

$\mathcal{G}$-EQUIVALENCE

We consider nonlinear systems in R^n of the form

$$\dot{x} = F(x, u_1, \ldots, u_m) \qquad (1)$$

where F defines a $\mathcal{C}^\infty$ vector field and $F(0) = 0$.

A *$\mathcal{G}$-transformation* is defined to be a $\mathcal{C}^\infty$ deffeomorphism which preserves the origin. A system Σ_1 is said to be *$\mathcal{G}$-related* to another system Σ_2 on a neighborhood $U \subset R^{n+m}$ if there is a $\mathcal{G}$-transformation $T = (T_1, T_2, \ldots, T_{n+m})$ such that for any state and control trajectory $(x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t))$ in U of Σ_1, the image $T(x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t))$ corresponds to a state and control trajectory of the system Σ_2, with $T_{n+1}, \ldots, T_{n+m}$ being the controls of Σ_2. By arguments similar to those in (5) we have the following results.

<u>Proposition 1</u>

A system $\dot{x} = F(x, u_1, \ldots, u_m)$ is $\mathcal{G}$-related to a system $\dot{y} = G(y, v_1, \ldots, v_m)$ if and only if there is a transformation $T = (T_1, \ldots, T_n, T_{n+1}, T_{n+m})$ such that

1. $T_1, \ldots, T_n$ are independent of the controls $u_1, \ldots, u_m$

2. The Jacobian

$$\begin{bmatrix} \frac{\partial T_{n+1}}{\partial u_1} & \cdots & \frac{\partial T_{n+1}}{\partial u_m} \\ \cdots & & \\ \frac{\partial T_{n+m}}{\partial u_1} & \cdots & \frac{\partial T_{n+m}}{\partial u_m} \end{bmatrix}$$

is nonsingular near the origin of R^{n+m} and

3. T satisfies the system of partial differential equations

$$\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} F_j = G_i \circ T , \qquad i = 1, \ldots, n$$

where $F = (F_1, \ldots, F_n)$, $G = (G_1, \ldots, G_n)$, and $G_i \circ T$ denotes the composition.

Proposition 2

The $\mathcal{F}$-relation is an equivalence relation.

With this equivalence relation we now examine the equivalence classes that contain the controllable linear systems. It is clear that a nonlinear system that is $\mathcal{F}$-equivalent to a controllable linear system is also $\mathcal{F}$-equivalent to the Brunovsky canonical form of the linear system. More precisely, for such a nonlinear system

$$\dot{x}_i = F_i(x_1, \ldots, x_n, u_1, \ldots, u_m) , \qquad i = 1, \ldots, n$$

it is associated with a transformation $T = (T_1, \ldots, T_n, T_{n+1}, \ldots, T_{n+m})$ and a set of m positive integers $(k_1, \ldots, k_m)$ such that by letting

$$a_p = \sum_{\ell=0}^{p-1} k_\ell$$

with $k_0 = 0$, we have

$$\left.\begin{aligned} &\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} F_j(x, u_1, \ldots, u_m) = T_{i+1} , \\ &\qquad a_{p-1} + 1 \le i \le a_p - 1 \end{aligned}\right\} \qquad (2)$$

and

$$\sum_{j=1}^{n} \frac{\partial T_{a_p}}{\partial x_j} F_j(x, u_1, \ldots, u_m) = T_{n+p} , \qquad (3)$$

where $p = 1, \ldots, m$. Observe that for each fixed state $x \in R^n$, Eq. (2) is a system of $n - m$ linear equations with constant coefficients satisfied by $F_1, \ldots, F_n$. As an extension of the result in (5), we have the following necessary condition.

Theorem 1

If a system $\dot{x}_i = F_i(x, u_1, \ldots, u_m)$, $i = 1, \ldots, n$ is $\mathcal{F}$-equivalent to a controllable linear system, then $F = (F_1, \ldots, F_n)$ must take the form

$$F_i = f_i(x) + \sum_{j=1}^{m} g_{ij}(x)\phi_j(x,u)$$

for all $i = 1, \ldots, n$, where $\phi_j(0,0) = 0$ and the Jacobian $\partial(\phi_1, \ldots, \phi_m)/\partial(u_1, \ldots, u_m)$ is locally nonsingular.

Since the replacement of the scalar functions $\phi_1, \ldots, \phi_m$ by a set of new controls $v_1, \ldots, v_m$ is a legitimate $\mathcal{F}$-transformation, we shall discuss a [illegible] system only in terms of the $m + 1$ vector fields $f, g_1, \ldots, g_m$.

$\mathcal{F}$-INVARIANTS

In this section we show that the invariants discovered by Brockett (4) are also invariants under our $\mathcal{F}$-transformations. First we need the following observation.

Observation 1

If a system $\Sigma: \dot{x} = F(x, u_1, \ldots, u_m)$ is $\mathcal{F}$-equivalent to a system of the form

$$\dot{x} = f(x) = \sum_{i=1}^{m} g_i(x)\phi_i(x, u_1, \ldots, u_m) ,$$

then Σ must take the similar form as

$$\dot{y} = f(y) = \sum_{i=1}^{m} g_i(y)\phi_i(y, v_1, \ldots, v_m) .$$

This can be verified by simple computations which we leave to the reader. With this fact we can now generalize Brockett's results.

Let $L_{i,j}$ be the linear span of all the vector fields ψ that are Lie derivatives of $f, g_1, \ldots, g_m$ with the total degree of ψ with respect to f being less than or equal to i and the total degree of ψ with respect to the g's being less than or equal to j. For example

$$L_{2,1} = \text{linear span } \{f, g_1, \ldots, g_m, (ad^1 f, g_1), \ldots, (ad^1 f, g_m), (ad^2 f, g_1), \ldots, (ad^2 f, g_m)\} .$$

Here the *Lie derivatives* are defined inductively as

$$(ad^0 f, g) = g, \quad (ad^1 f, g) = \frac{\partial f}{\partial x} g - \frac{\partial g}{\partial x} f ,$$

and

$$(ad^i f, g) = (ad^1 f, (ad^{i-1} f, g)) ,$$

where

$$\frac{\partial f}{\partial x} , \frac{\partial g}{\partial x}$$

are the Jacobian matrices.

Let $S_{i,j}$ be the subspace $\{p \mid p = h(0), h \in L_{i,j}\}$ of R^n.

Theorem 2

The dimensions of $S_{i,j}$ are $\mathcal{F}$-invariant.

The proof is omitted, but we remark that essentially Brockett's proof in (4) works in our case.

Now we examine some important dimensions of the $S_{i,j}$ which we need in characterizing the equivalence class of interest to us.

Given a system $(f, g_1, \ldots, g_m)$ in R^n we construct a matrix of vector fields

B_0	g_1	g_2		g_m
B_1	$(ad^1 f, g_1)$	$(ad^1 f, g_2)$		$(ad^1 f, g_m)$
$\vdots$			$\cdots$	
B_{n-1}	$(ad^{n-1} f, g_1)$	$(ad^{n-1} f, g_2)$		$(ad^{n-1} f, g_m)$

Let B_i denote the ith row of the matrix. In terms of $L_{i,1}$ we have $L_{0,1} = \text{span}\{B_0\}$, $L_{1,1} = \text{span}\{B_0, B_1\}$, etc. We assume in the rest of the paper that the dimension of $L_{i,1}$ is constant on some neighborhood and define the following indices.

1. Indices $\alpha = (\alpha_0, \ldots, \alpha_{n-1})$
   $\alpha_i \triangleq \dim S_{i,1}$ for $0 \le i \le n-1$
2. Indices $\beta = (\beta_0, \ldots, \beta_{n-1})$
   $\beta_0 \triangleq \alpha_0$, and $\beta_i \triangleq \alpha_i - \alpha_{i-1}$ for $1 \le i \le n-1$
3. Indices $\gamma = (\gamma_0, \ldots, \gamma_{n-1})$
   $\gamma_i \triangleq \beta_i - \beta_{i+1}$ for $0 \le i \le n-1$ with $\beta_n = 0$
4. Indices $k = (k_1, \ldots, k_m)$
   $k_i \triangleq$ the number of β_j's with $\beta_j \ge i$

Clearly, by way of construction, the indices $(\alpha_0, \alpha_1, \ldots, \alpha_{n-1})$ are an increasing sequence. By computation, it also can be shown that $(\beta_0, \beta_1, \ldots, \beta_{n-1})$ is always a decreasing sequence, and, thus, $\gamma_i \ge 0$ for all i. In this paper we are most interested in the case in which $\alpha_0 = m$ and $\alpha_{n-1} = n$. If this is the case, one has

$$\sum_{i=0}^{n-1} \gamma_i = m ,$$

and

$$\sum_{i=1}^{m} k_i = n .$$

MAIN RESULTS

We continue the development begun in the second section ($\mathcal{F}$-equivalence). Suppose a system $(f, g_1, \ldots, g_m)$ is $\mathcal{F}$-equivalent to a controllable linear system associated with a set of (Kronecker) indices $(k_1, \ldots, k_m)$ where $k_1, \ldots, k_m$ are positive integers and

$$\sum_{i=1}^{m} k_i = n .$$

Equations (2) and (3) then become

$$\left.\begin{aligned} &\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} (f_j + g_{1j}u_1 + \ldots + g_{mj}u_m) = T_{i+1} , \\ &\qquad a_{p-1} + 1 \le i \le a_p - 1 \end{aligned}\right\} \quad (4)$$

and

$$\sum_{j=1}^{n} \frac{\partial T_{a_p}}{\partial x_j} (f_j + g_{1j}u_1 + \ldots + g_{mj}u_m) = t_{n+p} , \quad (5)$$

where $1 \le p \le m$ and

$$c_p = \sum_{i=0}^{p-1} k_i$$

with $c_0 = 0$. Since the right-hand sides of Eqs. (4) are [illegible] and of the controls, we have

$$\left.\begin{aligned} &\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} g_{kj} = 0 , \quad 1 \le k \le m, \\ &a_{p-1} + 1 \le i \le a_p - 1, \ 1 \le p \le m \end{aligned}\right\} \quad (6)$$

$$\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} f_j = T_{i+1} . \quad (7)$$

The nonsingularity of the transformation T in turn implies that the matrix

$$\begin{bmatrix} \sum_{j=1}^{n} \frac{\partial T_{a_1}}{\partial x_j} g_{1j} & \cdots & \sum_{j=1}^{n} \frac{\partial T_{a_1}}{\partial x_j} g_{mj} \\ \sum_{j=1}^{n} \frac{\partial T_{a_m}}{\partial x_j} g_{1j} & \cdots & \sum_{j=1}^{n} \frac{\partial T_{a_m}}{\partial x_j} g_{mj} \end{bmatrix} \quad (8)$$

is nonsingular.

Recognizing that the summation

$$\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} g_{kj}$$

can be expressed as a duality product of *forms* and vector fields, Eqs. (6), (7), and (8) are equivalently rewritten as

$$\left.\begin{aligned} &\langle dT_i, g_k \rangle = 0 , \quad 1 \le k \le m, \\ &a_{p-1} + 1 \le i \le a_p - 1, \ 1 \le p \le m \end{aligned}\right\} \quad (9)$$

$$\langle dT_i, f \rangle = T_{i+1} \quad (10)$$

and

$$\begin{bmatrix} \langle dT_{a_1}, g_1 \rangle & \cdots & \langle dT_{a_1}, g_m \rangle \\ \langle dT_{a_m}, g_1 \rangle & \cdots & \langle dT_{a_m}, g_m \rangle \end{bmatrix} \quad (11)$$

is nonsingular, where dT_i is the *form*

$$\left(\frac{\partial T_i}{\partial x_1}, \ldots, \frac{\partial T_i}{\partial x_n} \right) ,$$

and the *duality product* $\langle dT_i, f \rangle$ is

$$\sum_{j=1}^{n} \frac{\partial T_i}{\partial x_j} f_j .$$

Using the well-known formula

$$\langle d(dT, f), g \rangle = d(dT, g) - \langle dT, (ad^1 f, g) \rangle , \quad (12)$$

the preceding discussion results in the following theorem.

Theorem 3

A system $(f, g_1, \ldots, g_m)$ is $\mathcal{F}$-equivalent to a controllable linear system if and only if there is a set of positive integers $k_1, k_2, \ldots, k_m$ with

$$\sum_{i=1}^{m} k_i = n ,$$

and m scalar functions

$$T_{a_1+1}, T_{a_2+1}, \ldots, T_{a_m+1}$$

such that

1. The forms $dT_{a_1+1}, \ldots, dT_{a_m+1}$ are linearly independent,

2. $\left\langle dT_{a_p+1}, (ad^i f, g_j)\right\rangle = 0, 1 \le j \le m,$

$0 \le i \le a_{p+1} - 2,$

3. The matrix

$$\begin{bmatrix} \langle dT_{a_1+1}, (ad^{a_1-1} f, g_1)\rangle & \cdots & \langle dT_{a_1+1}, (ad^{a_1-1} f, g_m)\rangle \\ \cdots & & \\ \langle dT_{a_m+1}, (ad^{a_{m+1}-1} f, g_1)\rangle & \cdots & \langle dT_{a_m+1}, (ad^{a_{m+1}-1} f, g_m)\rangle \end{bmatrix}$$

is nonsingular, where

$$a_p = \sum_{i=0}^{p-1} k_i , \quad 1 \le p \le m, \; k_0 = 0 .$$

Therefore, the problem of the existence of a transformation T becomes the existence of m positive integers and m linearly independent exact one-forms which satisfy the above conditions. Once these m one-forms are obtained, the rest of the transformation can be constructed by Eqs. (10). We remark that in the proof of the above theorem one has to show that a transformation so constructed in indeed nonsingular. For details the reader is referred to our paper (8).

Next we come to another main theorem which gives conditions on the characteristics of a system (including the invariant indices discussed in the preceding section — $\mathcal{F}$-invariants) for being $\mathcal{F}$-equivalent to a linear system.

Theorem 4

A system $(f, g_1, \ldots, g_m)$ is $\mathcal{F}$-equivalent to a controllable linear system with m controls if and only if

1. $\alpha_0 = m, \alpha_{n-1} = n$

2. For each i such that $\gamma_i \ne 0$, the set of vector fields

$$\{g_1, \ldots, g_m, \ldots, (ad^{i-1} f, g_1), \ldots, (ad^{i-1} f, g_m)\}$$

is involutive.

Now suppose $i_1 - i_2 \ge 2$, and $\omega_1, \ldots, \omega_{\gamma_{i_1}}$ are needed one-forms that vanish on the integral manifold associated with i_1, namely, the integral manifold of $L_{i_1-1,1}$. One step of Lie derivative gives us forms $L_f(\omega_1), \ldots, L_f(\omega_{i_1})$ which, by definition of $\omega_1, \ldots, \omega_{\gamma_{i_1}}$, have zero duality products with the vector fields in $L_{i_1-2,1}$. From the Frobenius theorem, $L_{i_1-2,1}$ is involutive. Repeating the argument we claim the following result.

Theorem 5

If a system $(f, g_1, \ldots, g_m)$ satisfies the conditions in Theorem 4, namely,

1. $\alpha_0 = m, \alpha_{n-1} = n$

2. For each i such that $\gamma_i \ne 0$, $L_{i_1-1,1}$ is involutive

then $L_{j,1}$ is involutive for every $0 \le j \le i_1 - 1$.

CONCLUDING REMARKS

In this paper we have obtained a complete characterization of these nonlinear systems which are $\mathcal{F}$-equivalent to a controllable linear system. These nonlinear systems are grouped into difference equivalence classes indexed by $\mathcal{F}$-invariant indices.

The conditions stated in Theorem 4 are weaker than those in the paper by Jakubczyk and Respondek (7). Some of their conditions are redundant, as indicated by our Theorem 5. Also their proof is an existence proof, whereas ours is constructive in nature. Details and explicit construction of the desired transformations are omitted here. They will appear in our paper (8).

ACKNOWLEDGMENT

The authors would like to acknowledge a brief discussion of this problem with Professor Brockett and the courtesy of Professor Sontag in providing a copy of the paper by Jakubczyk and Respondek.

REFERENCES

1 Brunovsky, P., "A Classification of Linear Controllable Systems," Kybernetika (Praha), Vol. 3, 1970.

2 Kalman, R. E., "Kronecker Invariants and Feedback," in Ordinary Differential Equations, L. Weiss, ed., Academic Press, 1972.

3 Meyer, G., and Cicolani, L., "Application of Nonlinear System Inverses to Automatic Flight Control Design — System Concepts and Flight Evaluations," Theory and Applications of Optimal Control in Aerospace Systems, P. Kant, ed., AGARDograph, 1980.

4 Brockett, R. W., "Feedback Invariants for Nonlinear Systems," IFAC Congress, Helsinki, 1978.

5 Su, R., "On the Linear Equivalents of Nonlinear Systems," submitted to Systems and Control Letters.

6 Hunt, L. R., and Su, R., "Global Mappings of Nonlinear Systems," submitted to 1981 Joint Automatic Control Conference.

7 Jakubczyk, B., and Respondek, W., "On Linearization of Control Systems," preprint.

8 Hunt, L. R., and Su, R., "Multi-input Nonlinear Systems," submitted to SIAM J. on Control and Optimization.

CONTROLLABILITY OF NONLINEAR HYPERSURFACE SYSTEMS

L. R. HUNT

LECTURES IN APPLIED MATHEMATICS
VOLUME 18, 1980

PRECEDING PAGE BLANK-NOT FILMED

Lectures in Applied Mathematics
Volume 18, 1980

CONTROLLABILITY OF NONLINEAR HYPERSURFACE SYSTEMS

L. R. Hunt*

ABSTRACT. Consider the nonlinear system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), x(o) = x_o \in M$$

where M is a connected real-analytic n-dimensional manifold, $f, g_1, \ldots, g_{n-1}$ are real-analytic vector fields on M, and $u_1, \ldots, u_{n-1}$ are real-valued controls. We are interested in characterizing the largest open subset U of M, if any, which is reachable from x_o and which we call the region of reachability of our system from x_o. If the Lie algebra L_A generated by $f, g_1, \ldots, g_{n-1}$ and successive Lie brackets has vector space dimension n at x_o, and if $f, g_1, \ldots, g_{n-1}$ are linearly independent at some point in M, we find the region of reachability from x_o. Suppose U is the smallest open subset of M with $x_o \in \bar{U}$ so that ∂U contains the integral manifolds of the Lie algebra L'_A generated by $g_1, \ldots, g_{n-1}$ that intersect it and f assigns vectors on ∂U which point in the direction

*Research supported in part by the National Science Foundation under NSF Grant MCS76-05267-A01 and by the Joint Services Electronics Program under ONR Contract 76-C-1136.

© American Mathematical Society 1980

PRECEDING PAGE BLANK-NOT FI

of $\bar{U}$. Then U is the region of reachability from x_0 for our system. Much of the work is involved in proving a similar result in the more general $\mathscr{C}^\infty$ case under the stronger assumption that $f, g_1, \ldots, g_{n-1}$ are linearly independent on the connected $\mathscr{C}^\infty$ n-dimensional manifold M.

1. INTRODUCTION. Let M be a connected real-analytic n-dimensional manifold, $f, g_1, \ldots, g_{n-1}$ be $\mathscr{C}^\infty$ vector fields on M, and $u_1, \ldots, u_{n-1}$ be real-valued controls. The system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), \; x(o) = x_0 \in M ,$$

is called a hypersurface system since the number of controls is one less than the dimension of the manifold M. We assume that the vector space dimension of the Lie algebra L_A generated by $f, g_1, \ldots, g_{n-1}$ and successive Lie brackets is n at x_0 (i.e. this Lie algebra spans the tangent space to M at x_0) and that $f, g_1, \ldots, g_{n-1}$ are linear independent at some point of M. Results due to Sussmann and Jurdjevic [18] and Krener [15] show that we can reach a nonempty open subset of M from x_0. Let U be the largest open subset which is reachable for our system from x_0. We characterize U by proving that U is the smallest open subset of M with $x_0 \in \bar{U}$ (the closure of U in M) satisfying i) ∂U contains the integral manifolds which intersect it of the Lie algebra L'_A generated by $g_1, \ldots, g_{n-1}$ and successive Lie brackets and ii) the vector field f points in the direction of $\bar{U}$ on ∂U. This set U is called the region of reachability from x_0, and if $U = M$, the system is controllable from x_0.

The real-analytic theory depends on results proved for a system

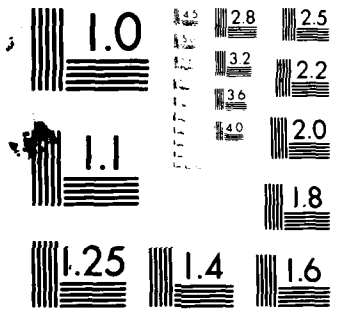
$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), \; x(o) = x_0 \in M$$

where $f, g_1, \ldots, g_{n-1}$ are $\mathscr{C}^\infty$ linearly independent vector fields

AD-A112 113 TEXAS TECH UNIV LUBBOCK INST FOR ELECTRONIC SCIENCE F/G 9/5
ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICES ELECTRONICS --ETC(U)
DEC 81 R SAEKS, L R HUNT, J MURRAY, J WALKUP N00014-76-C-1136

UNCLASSIFIED NL

2 of 3



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

on a connected $\mathscr{C}^\infty$ n-dimensional manifold M. For this case we show that the region of reachability U is the smallest open subset of M with $x_0 \in U$ satisfying U is an (n-1)-dimensional integral manifold of $g_1,\ldots,g_{n-1}$ and f assigns vectors on ∂U in the direction of U. We view the control problem much as one would the famous Holmgren's Uniqueness Theorem (see [4]) of partial differential equations. A solution of the partial differential equation is unique up to the characteristics. A similar situation occurs in the study of uniqueness of analytic continuation for the CR-functions on a $\mathscr{C}^\infty$ real hypersurface in $\mathbb{C}^n$, $n > 1$ [11]. In this instance one gets uniqueness up to the characteristics of the tangential CR-equations, which are the integral manifolds of a subbundle of the tangent bundle to the hypersurface of codimension 1. If $f,g_1,\ldots,g_{n-1}$ in our system are linearly independent on M, then $g_1,\ldots,g_{n-1}$ give us a subbundle of the tangent bundle of codimension 1. Thus the only possible way to have a set which is not reachable from x_0 is to have it disconnected in some fashion from the reachable set by an integral manifold of $g_1,\ldots,g_{n-1}$. Of course such integral manifolds may not even exist in which case we expect controllability of the system from x_0. Michael Freeman has results giving conditions at a point for there to be an integral manifold of a collection of real-analytic vector fields through that point.

A nice expository paper containing the problems considered in this article has been written by Brockett [1]. Related results can be found in the work of Krener [15], Lobry [16], [17], Sussmann and Jurdjevic [18], and the author [12], [13]. Theorems concerning the problem of local controllability along a reference trajectory are due to Hermes [7], [8], [9]. Results dealing with control theory for linear systems are in [14]. This paper is arranged in the following way. In section 2 we give definitions and a relevant example. Section 3 contains a local theory concerning the boundary of the region of reachability U of our $\mathscr{C}^\infty$ system under the assumption that the boundary is $\mathscr{C}^1$ near one of its points. In section 4 we state a theorem from [11]

concerning a subbundle of the tangent bundle to M and allowing us to remove the $\mathscr{C}^1$ restriction. Then we prove our main result for the $\mathscr{C}^\infty$ case and give several applications. Section 5 contains our result for the real-analytic system on a real-analytic manifold.

2. DEFINITIONS. We shall use the classical Frobenius Theorem and Chow's Theorem [2]. For a statement of these results and their applications to control theory we refer the reader to [1].

Of interest to us is the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t)g_i(x(t)),\ x(o) = x_0 \in M \qquad (2.1)$$

where M is a connected $\mathscr{C}^\infty$ n-dimensional manifold, $f, g_1, \ldots, g_{n-1}$ are $\mathscr{C}^\infty$ vector fields on M, and $u_1, \ldots, u_{n-1}$ are controls.

Let $T(M)$ be the tangent bundle of M with $T_x(M)$ the tangent space for $x \in M$. Recall that if X is a vector field on M (i.e. X is a section of $T(M)$) then α is an *integral curve* of X if α is a $\mathscr{C}^\infty$ mapping from a closed interval $I \subset \mathbb{R}$ into M such that $\frac{d\alpha(t)}{dt} = X(\alpha(t))$ for all $t \in I$.

DEFINITION 2.1 [18]. *If D is a subset of $T(M)$, then an* <u>*integral curve of D*</u> *is a mapping α from a real interval $[t,t']$ into M such that there exist $t = t_0 < t_1 < \ldots < t_k = t'$ and vector fields $X_1, \ldots, X_k$ in D with the restriction of α to $[t_{i-1}, t_i]$ being an integral curve of X_i, for each $i = 1,2,\ldots,k$.*

DEFINITION 2.2. *Let D be a subset of $T(M)$ and let $x_0 \in M$. A point $x \in M$ is* <u>*D-reachable from x_0*</u> *if there is an integral curve α of D and some $T \geq 0$ in the interval for α such that $\alpha(0) = x_0$ and $\alpha(T) = x$. A subset A of M is* <u>*D-reachable from x_0*</u> *if every point $x \in A$ is reachable from x_0.*

Since the D we consider is the subset of $T(M)$ given by the vector fields of a system of the form (2.1) we drop the D from D-reachable. We shall make assumptions of $f, g_1, \ldots, g_{n-1}$

that assure us that we can reach an open subset of M from x_0.

DEFINITION 2.3. *The largest open subset* U *of* M *which is reachable from* x_0 *is called the* <u>*region of reachability from* x_0</u>. *If* $U = M$, *we say that the system is* <u>*controllable from* x_0</u>.

DEFINITION 2.4. *Let* O *be an open subset of* M *and let* $x \in \partial O$ *such that* ∂O *is a* $\mathscr{C}^1$ *manifold near* x. *Then* f <u>*points in the direction of* O *(or towards* O*) at* x</u> *if* $f(x)$ *is not tangent to* ∂O *at* x *and if there exists an open neighborhood* W *of* x *in* M *such that the vector assigned by* f *at* x, *projected into* M *(by the exponential map), and intersected with* $W-\{x\}$ *is contained in* O. *If* O *is a* $\mathscr{C}^1$ *manifold and if the above is true for every* $x \in \partial O$, *then* f <u>*points in the direction of* O *on* ∂O</u>.

DEFINITION 2.5. *Let* O *be an open set in* M *and let* $x \in \partial O$. *Then* f <u>*points in the direction of* $\bar{O}$ *(or towards* $\bar{O}$*) at* x</u> *if there is an open neighborhood* W *of* x *in* M *such that the integral curve of* f *starting at* x *and intersected with* W *is contained in* $\bar{O}$. *If this is true for all* $x \in \partial O$, *then* f <u>*points in the direction of* $\bar{O}$ *on* ∂O</u>.

If f and g are $\mathscr{C}^\infty$ vector fields defined on M, we define the *Lie bracket* of f and g by $[f,g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g$. A set of $\mathscr{C}^\infty$ vector fields $\{f_1,\ldots,f_r\}$ is called *involutive* if there exist $\mathscr{C}^\infty$ functions $\gamma_{ijk}(x)$ on M such that $[f_i,f_j](x) = \sum_{k=1}^{r} \gamma_{ijk}(x) f_k(x)$ for all $i,j,1 \le 1$, $j \le r$, $i \ne j$.

We introduce one example from [1] which may help us understand the problems involved in trying to determine the reachable set of a system.

EXAMPLE Consider the system

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + u(t) \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= f(x(t)) + u(t)g(x(t))$$

where M is the set $\mathbb{R}^2 - \{(o,o)\}$. Notice that f and g are linearly independent on M. Brockett [1] states that if $x_o = (x_1^o, x_2^o)$ is in the positive quadrant, then the region of reachability U from x_o is contained in this quadrant. The integral curve of g through a point on $x_1 = o$, $x_2 > o$ is the line (o, x_2) with $x_2 > o$. Moreover, the integral curve of g through a point on $x_2 = o$, $x_1 > o$ is the line (x_1, o) with $x_1 > o$. These together form the boundary of the first quadrant in M. The vector field f assigns vectors to this boundary which point toward the first quadrant. Thus there is no hope of a solution of the system starting in this quadrant to leave it.

We could give more examples at this time, but they would all hint at the same conclusion. If $f, g_1, \ldots, g_{n-1}$ are linearly independent in the system (2.1), the important items to check appear to be the integral manifolds of $g_1, \ldots, g_{n-1}$, if any exist, and the direction of the vector field f on these integral manifolds. We next examine these conditions for regions of reachability with $\mathscr{C}^1$ boundaries.

3. $\mathscr{C}^1$ BOUNDARIES. Suppose we consider the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), \quad x(o) = x_o \in M, \qquad (3.1)$$

where M is a connected $\mathscr{C}^\infty$ n-dimensional manifold, $f, g_1, \ldots, g_{n-1}$ are $\mathscr{C}^\infty$ linearly independent vector fields on M, and $u_1, \ldots, u_{n-1}$ are controls. It is easy to see that we can reach an open subset of M which contains x_o in its closure.

Let U be the region of reachability of the hypersurface system given in (3.1). Let x be an element of the boundary of U, and assume ∂U is $\mathscr{C}^1$ is some open neighborhood W of x in M. As just mentioned we need to consider the directions of f on $W \cap \partial U$ and the possibility of having an integral manifold of $g_1, \ldots, g_{n-1}$ through x. Recall that a differentiable submanifold S of M is an integral manifold of $g_1, \ldots, g_{n-1}$ if $T_y S$ is the space spanned by $g_1, \ldots, g_{n-1}$ at y for each $y \in S$.

HYPERSURFACE SYSTEMS

For a more thorough discussion of integral manifolds we must consider the Lie bracket which we defined earlier. If g_i and g_j are different vector fields on M then

$$[g_i, g_j] = \frac{\partial g_i}{\partial x} g_j - \frac{\partial g_j}{\partial x} g_i .$$

This may or may not give us a "new" direction in which to move (see [1]), depending on whether the collection $\{g_i, g_j\}$ is involutive or not. Let L'_A be the smallest Lie algebra generated by taking successive Lie brackets of the $g_1, \ldots, g_{n-1}$ given in equation (3.1). If we get a vector space of the same dimension at each point of M, then L'_A is a vector subbundle of the tangent bundle to M.

The following definition is essential to our work. Let S_1 and S_2 be $\mathscr{C}^1$ submanifolds of M of dimensions k and $n-k$ respectively.

DEFINITION 3.1. *The manifolds* S_1 *and* S_2 *<u>intersect transversally</u> at a point* $y \in S_1 \cap S_2$ *if and only if* $T_y(S_1) \oplus T_y(S_2) = T_y(M)$. *Here* $\oplus$ *denotes the direct sum.*

We now prove a result under the assumption that locally our open set has a $\mathscr{C}^1$ boundary.

THEOREM 3.2. *Let* O *be an open set in* M *which is reachable from* x_0 *for system* (3.1), *and let* x *be an arbitrary point in* ∂O. *Suppose there is an open neighborhood* W *of* x *in* M *such that* $W \cap \partial O$ *is a* $\mathscr{C}^1$ *real* $(n-1)$*-dimensional submanifold of* M. *If any one of the following conditions holds, then* O *is not the region of the reachability from* x_0;

i) *the fiber dimension of* L'_A *at* x *is* n,

ii) *the integral curve of some* g_i, $1 \le i \le n-1$ *is transversal to* ∂O *at* x,

iii) f *assigns at* x *a vector pointing in the direction of the complement of* O.

<u>Proof</u>. If i) holds then the fiber dimension of L'_A at all points in some open neighborhood of x in M must be n (since n is maximal). Thus ∂O cannot be an integrable manifold of

$g_1,\ldots,g_{n-1}$ near x by the Frobenius Theorem, and there exist a point $y \in \partial O$ arbitrarily close to x and a g_i, $1 \le i \le n-1$, such that the integral curve of g_i is transversal to ∂O at y. Hence, i) reduces to ii).

Next we assume that ii) is true. If the integral curve of g_1, chosen arbitrarily from the set $g_1,\ldots,g_{n-1}$ and renumbered if necessary, is transversal to ∂O at x, then it is transversal to ∂U in $W \cap \partial O$, W being an open neighborhood of x in M (this W may be smaller set than our original W). Following the integral curves of g_1 that start in $W \cap O$, a reachable set from x_0, and continuing past $W \cap \partial O$, we have that O cannot be the region of reachability from x_0. We have used the fact that we may as well assume we can move along the integral curve of any g_i since u_i is unbounded.

If iii) holds at x, then it holds for all points in $W \cap \partial O$, and the argument given in ii) with g_1 replaced by f implies the desired result. ¤

It is interesting to note that condition i) does not depend on $W \cap \partial O$ being a $\mathscr{C}^1$ manifold.

We seek a minimum number of necessary conditions that an open set $U \subset M$ be the region of reachability from x_0.

THEOREM 3.3. *Let* U *be the region of reachability from* x_0 *of the system* (3.1). *Suppose* ∂U *is a* $\mathscr{C}^1$ *manifold for an open neighborhood* W *of* $x \in \partial U$ *in* M. *Then* $W \cap \partial U$ *is an integral manifold of* $g_1,\ldots,g_{n-1}$, *and the vector field* f *assigns to* $W \cap \partial U$ *vectors pointing in the direction of* U.

<u>Proof</u>. It follows from part ii) of the preceding theorem that $W \cap \partial U$ is an integral manifold of $g_1,\ldots,g_{n-1}$. Hence $W \cap \partial U$ is actually a $\mathscr{C}^\infty$ submanifold of M. Since $f, g_1,\ldots,g_{n-1}$ form a linearly independent set on M and $W \cap \partial U$ in an integral manifold of $g_1,\ldots,g_{n-1}$, part iii) implies the statement concerning f. ¤

We shall prove in the next section that the hypothesis ∂U is $\mathscr{C}^1$ near x is superfluous.

4. THE RESULT FOR $\mathscr{C}^\infty$ MANIFOLDS. The following theorem was proved in [11] for use in the uniqueness of analytic continuation problem for CR-distributions on CR-hypersurfaces in $\mathbb{C}^n$, $n > 1$. The statement concerning a $\mathscr{C}^2$ boundary can be relaxed to $\mathscr{C}^1$, or we can simply replace $\mathscr{C}^1$ by $\mathscr{C}^2$ everywhere in the preceding section.

THEOREM 4.1. *Let* M *be a* $\mathscr{C}^\infty$ *manifold of dimension* n, *and let* H *be a subbundle of the tangent bundle of* M *with fiber (vector space) dimension* $n-1$. *Suppose* $U \subset M$ *is an open set with the property that if* $O \subset U$ *is an open set having a* $\mathscr{C}^2$ *boundary, then for each* $x \in \partial O \cap \partial U$ *we have* $T_x(\partial O) = H_x$ *(the fiber of* H *at* x*). Then for each point* $x \in \partial U$, *there is a neighborhood* V *of* x, *a real-valued function* $h \in \mathscr{C}^\infty(V)$ *with nonzero differential for all points in* V, *and a closed nowhere dense set* $E \subset \mathbb{R}$ *such that*

(1) $\partial U \cap V = \{x \in V | h(x) \in E\}$,

(2) *for each* $\ell \in E$, $S_\ell = \{x \in V | h(x) = \ell\}$ *is an integral manifold of* H; *i.e. the boundary of* U *is foliated by integral manifolds of* H.

We now restate Theorem 3.3 under more general conditions.

THEOREM 4.2. *Let* U *be the region of reachability from* x_0 *of the system* (3.1). *Then* ∂U *is a* $\mathscr{C}^\infty$ *integral manifold of* $g_1,\ldots,g_{n-1}$ *and* f *assigns vectors on* ∂U *which point in the direction of* U.

<u>Proof</u>. Let H be the subbundle of $T(M)$ spanned by $g_1,\ldots,g_{n-1}$. If O is an open subset of U with a $\mathscr{C}^2$ boundary, then an application of Theorem 3.2 and Theorem 4.1 gives us the stated conclusion. ∎

We have the following important corollary, the proof of which is obvious.

COROLLARY 4.3. *Suppose* M *contains no integral manifolds of* $g_1,\ldots,g_{n-1}$ *for which both of the following statements hold:*

a) *The closure of the integral manifold in* M *is foliated by integral manifolds.*

b) *The vectors assigned by* f *on this integral manifold always point in the same direction relative to the integral manifold (i.e. if this manifold divides* M *into two components, the vectors must point toward the same component).*

Then the system (3.1) *is controllable from any* $x_0 \in M$.

Let Λ^d denote Hausdorff measure (see [3]) in dimension d on M. Suppose L is the set of points on which the Lie algebra L'_A has dimension n-1. Then L is a closed set in M, and the Frobenius Theorem implies that L contains the integral manifolds of $g_1,\ldots,g_{n-1}$, if any exist. For such an integral manifold we must have $\Lambda^{n-1}(L) > 0$, and we have proved our next result.

THEOREM 4.4. *If* $\Lambda^{n-1}(L) = 0$ *then the system* (3.1) *is controllable from any* $x_0 \in M$.

Notice that if M is of dimension 2, we always have integral curves of g for the system $\dot{x}(t) = f(x(t)) + u(t)g(x(t)), x(0) = x_0$. Thus Theorem 4.4 does not apply in this case.

We state two theorems from [1] and indicate in a rather superficial way the relation of these theorems to this present work. We restrict our attention to dimension 2 and to a hypersurface system.

THEOREM 4.5. *Suppose* f *and* g *are vector fields on a connected* $\mathscr{C}^\infty$ *real 2-dimensional manifold* M. *Suppose that* {f,g} *meet the conditions of Chow's Theorem for* $\mathscr{C}^\infty$ *vector fields, and suppose that for each initial condition* x_0 *the solution of* $\dot{x}(t) = f(x(t))$ *is periodic with a least period* $T(x_0)$. *Then the reachable set from* x_0 *of the system* $\dot{x}(t) = f(x(t)) + u(t)g(x(t))$ *is the set given by Chow's Theorem.*

We start at $x_0 \in M$ and take the integral curve of g through x_0. Suppose this curve divides M into two connected components M^+ and M^-. If the solution of $\dot{x}(t) = f(x(t))$ starts at x_0 in the direction of M^+, then since the solution

is periodic, there is some point on the integral curve of g through x_0 at which the vector of f is in the M^- direction. Of course, this is in keeping with Theorem 4.2.

Hirschorn proved a very nice generalization of the following result, which we state in dimension 2.

THEOREM 4.6 [10]. *Consider the system*

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)), \quad x(o) = x_0 ,$$

for a $\mathscr{C}^\infty$ *real 2-dimensional manifold* M. *Suppose* $[f,g] = hg$ *on* M, *where* h *is a* $\mathscr{C}^\infty$ *function on* M. *Then the reachable set from* x_0 *is obtained by taking the integral curve* α *of* f *through* x_0 *(in the positive time sense) and then all integral curves of* g *intersecting* α.

This 2-dimensional version can be seen in light of the following result found in [6]. The one-parameter group of transformations generated by f permutes the integral curves of g with a change of parametrization if $[f,g] = hg$ for some $\mathscr{C}^\infty$ function h on M. Interpreted freely, once an integral curve of f passes through an integral curve of g it can never return. This seems to be in agreement with Theorem 4.2.

An obvious question to ask is if the necessary conditions of Theorem 4.2 are also sufficient.

THEOREM 4.7. *Let* $x_0 \in M$ *and suppose* U *is the smallest open subset of* M *with* $x_0 \in \bar{U}$ *satisfying* ∂U *is an integral manifold of* $g_1,\ldots,g_{n-1}$ *and* f *assigns vectors to* ∂U *in the direction of* U. *Then* U *is the region of reachability from* x_0 *for the system* (3.1).

In the statement of this theorem, we add the assumption that if $U \neq M$, every open neighborhood of any point $p \in \partial U$ contains points from U and the complement of $\bar{U}$. G. Stefani and A. Baccioti have pointed out that the correct conclusion to the theorem as stated above is $U \subset$ region of reachability $\subset$ interior of $\bar{U}$. The author wishes to thank Professors Stefani and Baccioti for their comments.

Proof. We know that we can reach an open set and by the theory developed in this paper we have that we can reach U. The important fact to remember is that to leave U we must break through ∂U near some point $x \in \partial U$. In the system $\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t)g_i(x(t))$ at the point x we can move in the directions $f, g_1, \ldots, g_{n-1}$, $-g_1, \ldots, -g_{n+1}$ and $f + \sum_{i=1}^{n-1} u_i(t)g_i$ for the appropriate finite u_i's. Since ∂U is an integral manifold of $g_1, \ldots, g_{n-1}$, Lie brackets like $[g_i, g_j]$ with $i \neq j$ will give us no "new" directions in which to move from x. Also, since $f, g_1, \ldots, g_{n-1}$ span $T(M)$, the brackets $[f, g_i]$, $i = 1, \ldots, n-1$ will yield vector fields which are linear combinations of $f, g_1, \ldots, g_{n-1}$ (the same is also true for successive Lie brackets). The only linear combinations here which can be used at x are those already indicated by the system. $\square$

The proof of Theorem 4.7 is applicable only for hypersurface systems (or certain general systems that behave like hypersurface systems). We prove results like those in this paper for general systems of the form

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t)g_i(x(t))$$

in [12] and [13]. Such a system with $m < n-1$ is more difficult to handle than a hypersurface system, and we make stronger assumptions in order to prove theorems concerning controllability.

5. REAL-ANALYTIC MANIFOLDS. We examine the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t)g_i(x(t)), \quad x(o) = x_o \in M \qquad (5.1)$$

where M is a connected real-analytic n-dimensional manifold $f, g_1, \ldots, g_{n-1}$ are real-analytic complete vector fields on M, and $u_1, \ldots, u_{n-1}$ are controls. We take the Lie aglebras L_A and L'_A as defined previously. If the vector space dimension of L_A at x_o is less than n, we apply the real analytic

version of Chow's Theorem and refer the reader to [13]. We prove the following result, which is an improvement of Theorem 4.7 for the real-analytic case.

THEOREM 5.1. *Assume the vector space dimension of* L_A *at* x_0 *is* n *and that* $f, g_1, \ldots, g_{n-1}$ *are linearly independent at some point of* M. *Let* U *be the smallest open subset of* M *with* $x_0 \in \bar{U}$ *satisfying* ∂U *contains the integral manifolds of* L'_A *which intersect it (and which are given by Chow's Theorem) and* f *points in the direction of* $\bar{U}$ *on* ∂U. *Then* U *is the region of reachability from* x_0 *for the system* (5.1).

<u>Proof</u>. Because M, $f, g_1, \ldots, g_{n-1}$ are real analytic, $f_1, g_1, \ldots, g_{n-1}$ are linearly independent at some point of M, and M is connected, the set of points P in M at which $f, g_1, \ldots, g_{n-1}$ are linearly dependent is nowhere dense in M. Of course this set contains the points where the dimension of L_A is less than n. Since the dimension of L_A at x_0 is n, this is true for an open neighborhood of x_0 in M. From Krener's [15] proof we can reach an open set O in M which is arbitrarily close to x_0. By remarks made earlier in this proof, we may as well assume that $f, g_1, \ldots, g_{n-1}$ are linearly independent on O. Let O' be the largest connected component of M containing O on which $f, g_1, \ldots, g_{n-1}$ are linearly independent. Choose a point $x_1 \in O$ and apply Theorem 4.7 with x replaced by x_1 and M by O' to get the region of reachability U' of the system from x_1 in O'.

We examine the boundary of U' in M, which has two components $\partial U' \cap \tilde{P}$ and $\partial U' \cap \tilde{P}$, where $\tilde{P}$ denotes the complement of P in M. If $\partial U' \cap \tilde{P}$ is a nonempty set, Theorem 4.7 implies that $\partial U' \cap \tilde{P}$ is a $(n-1)$ dimensional integral manifold of $g_1, \ldots, g_{n-1}$ and f points towards U' on $\partial U' \cap \tilde{P}$.

We assume that $\partial U' \cap P$ is nonempty and that $\partial U' \cap P$ is a real-analtyic $(n-1)$ dimensional manifold in an open neighborhood of a point $x \in \partial U' \cap P$. If the integral curve of some $g_i, 1 \le i \le n-1$, is transversal to $\partial U' \cap P$ at x or if f assigns at x a vector which does not point towards $\bar{U}$;

arguments like those given in the proof of Theorem 3.2 show that we can reach an open set in another connected component of M like O', and we simply start all over there.

Thus we assume that the $(n-1)$-dimensional manifold parts of $\partial U' \cap P$ contain the integral manifolds of L'_A which intersect them and f points towards $\tilde{U}'$ on these parts. Let $x \in \partial U' \cap P$ and suppose the unique integral manifold N of L'_A through x contains the integral curves of f that start at every point of N. In this case we have the vector space dimensions of L_A and L'_A agree on N and if we reach any point of N, then we can reach all points of N (this is the important part of the real-analytic theory in [13]). Then N will not be in U if ∂N contains the integral manifolds of L'_A that intersect it and f points in the direction of $\tilde{N}$ on ∂N. Of course this means that if $\tilde{U}$ and ∂N intersect, then the points in this intersection are in ∂U and satisfy the required conditions on the integral manifolds of L'_A and on the direction of f. An easy example of such a manifold N is a common equilibrium point of $f, g_1, \ldots, g_{n-1}$, which is certainly not reachable.

Hence we have that $\partial U' \cap P$ contains the integral manifolds of L'_A (given by Chow's Theorem) which intersect it and f points in the direction of $\tilde{U}'$ on $\partial U' \cap P$. Therefore $\partial U'$ must consist of the integral manifolds of L'_A intersecting it and f points towards $\tilde{U}'$ on $\partial U'$. A repeat of the proof of Theorem 4.3 found in [12] shows that we cannot reach an open subset of M in $\tilde{U}'$ from x_1.

Recall there is an open neighborhood of x_0 in M on which the dimension of L_A at x_0 is n. Arbitrarily close to any point in this neighborhood that can be reached from x_0 is an open set which is reachable from x_0. Thus we can reach an open V of M which contains x_0 in its closure, and we may assume that ∂V contains the integral manifolds of L'_A intersecting it and f points towards V on ∂V. We need to show that $V = U$, and we know that $x_0 \in \tilde{V} \cap \tilde{U}$.

If $x_0 \in V \cap U$, then an open subset of U can be reached.

By the theory developed in this paper U is reachable and no larger open subset containing U of M is reachable. Also V can be reached and we cannot leave V once we get to it. Hence we must have $V = U$. If $x_0 \in \partial V \cap U$, we reach an open subset of M arbitrarily close to x_0 (this is how we found V) which must be in U. The same proof just mentioned implies $V = U$.

Suppose that $x_0 \in \partial U$. Then the integral manifold of L'_A through x_0 must remain in ∂U and the integral curve of f starting at x_0 (and moving in positive time) moves in $\bar{U}$. If the curve for f reaches U, then we know we can reach an open subset of U, which will also be an open subset of V by the way in which V was formed. Above arguments show that we have $V = U$.

Assume that the integral curve of f starting at x_0 stays in ∂U. Taking the unique integral manifold of L'_A through each point of this integral curve keeps us in ∂U. Let N' be the set defined as the union of these integral manifolds of L'_A. At each point of N' we can start an integral curve of f, and we suppose that all such curves remain in ∂U. We can continue to repeat this process and we assume that we cannot leave ∂U. Perhaps Lie brackets like $[f,g]$ and higher order brackets will allow us to get out of ∂U. Helgason [5] interprets the Lie bracket $[f,g]$ at a point x as the tangent vector to a curve segment (starting at x) and moving in the g direction, the f direction, the $-g$ direction, and then the $-f$ direction all for t units of time. Thus Lie brackets will not help us reach an open subset of M. Since we know that we can reach the open set V, an integral curve of f must take us into U. Once we have this, we know that $V = U$. $\square$

REFERENCES

1. Brockett, R. W., *Nonlinear systems and differential geometry*, Proc. IEEE 64 (1976), 61-72.

2. Chow, W. L., *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann. 117 (1939), 98-105.

3. Federer, H., *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

4. Garabedian, P. R., *Partial Differential Equations*, John Wiley and Sons, New York, 1964.

5. Helgason, S., *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.

6. Hermann, R., *Differential Geometry and the Calculus of Variations*, Academic Press, New York, 1968.

7. Hermes, H., *On necessary and sufficient conditions for local controllability along a reference trajectory*, Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, Eds., Dordrecht, Holland: Reidel Publishing Co., 1973.

8. Hermes, H., *Local controllability and sufficient conditions in singular problems*, J. Differential Equations 20 (1976), 213-232.

9. Hermes, H., *Local controllability and sufficient conditions in singular problems, II*, SIAM J. Control 14 (1976), 1049-1062.

10. Hirschorn, R., *Topological semigroups, sets of generators, and controllability*, Duke Math J. 40 (1973), 937-947.

11. Hunt, L. R., J. C. Polking, and M. J. Strauss, *Unique continuation for solutions to the induced Cauchy-Riemann equations*, J. Differential Equations 23 (1977), 436-447.

12. Hunt, L. R., *Controllabiltiy of general nonlinear systems*, Math. Systems Theory, 12 (1979), 361-370.

13. Hunt, L. R., *Control theory for nonlinear systems*, 4th International Symposium on the Mathematical Theory of Networks and Systems, 3 (1979), 339-343.

14. Kalman, R. E., P. L. Falb, and M. A. Arbib, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York, 1969.

15. Krener, A. J., *A generalization of Chow's Theorem and the bang-bang theorem to nonlinear control problems*, SIAM J. Control 12 (1974), 43-52.

16. Lobry, C., *Contrôlabilité des systèmes non linéaires*, SIAM J. Control 8 (1970), 573-605.

17. Lobry, C., *Quelques aspects qualitatifs de la theories de la commande*, L'Universite Scientificue et Medical de Grenoble, pour obtenir le titre de Docteur es Sciences Mathematiques, May 19, 1972.

18. Sussmann, H., and V. Judjevic, *Controllability of nonlinear systems*, J. Differential Equations 12 (1972), 95-116.

DEPARTMENT OF MATHEMATICS
TEXAS TECH UNIVERSITY
LUBBOCK, TEXAS 79409

ABSTRACT OF

MULTI-INPUT NONLINEAR SYSTEMS

L. R. HUNT AND RENJENG SU

MULTI-INPUT NONLINEAR SYSTÉMS

L. R. Hunt and Renjeng Su

Abstract

Consider the nonlinear system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)) ,$$

where $f, g_1, \ldots, g_m$ are C^∞ vector fields on some neighborhood of the origin in $\mathbb{R}^n$ and $f(0) = 0$. We present necessary and sufficient conditions for this system to be transformed to a controllable linear system. Our results are constructive and depend upon the solutions of overdetermined systems of partial differential equations.

PRECEDING PAGE BLANK-NOT FI

ABSTRACT OF

GLOBAL TRANSFORMATIONS OF NONLINEAR SYSTEMS

BY

L. R. HUNT AND RENJENG SU

PRECEDING PAGE BLANK-NOT FILMED

Global Transformations of Nonlinear Systems

L. R. HUNT AND RENJENG SU

Abstract — Recent results have established necessary and sufficient conditions for a nonlinear system of the form

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t))$$

with $f(0) = 0$, to be locally equivalent in a neighborhood of the origin in $\mathbb{R}^n$ to a controllable linear system. We combine these results with several versions of the global inverse function theorem to prove sufficient conditions for the transformation of a nonlinear system to a linear system. In doing so we introduce a technique for constructing a transformation under the assumptions that $\{g,[f,g], \ldots, (ad^{n-1}f,g)\}$ span an n-dimensional space and that $\{g,[f,g], \ldots, (ad^{n-2}f,g)\}$ is an involutive set.

PRECEDING PAGE BLANK-NOT FI

ABSTRACT OF

CONTROL OF NONLINEAR TIME-VARYING SYSTEMS

BY

L. R. HUNT AND RENJENG SU

PRECEDING PAGE BLANK-NOT FILMED

CONTROL OF NONLINEAR TIME-VARYING SYSTEMS

L. R. Hunt and Renjeng Su

Abstract

Consider the time-varying nonlinear system of the form

$$\dot{x}(t) = f(x,t) + \sum_{i=1}^{m} u_i(t) g_i(x,t),$$

with $f, g_1, \ldots, g_m$ being C^∞ vector fields on $\mathbb{R}^{n+1}$. We give necessary and sufficient conditions for this system to be transformable to a time-invariant controllable linear system. In order to control the nonlinear system, we map to the linear system, choose a desired control there, and return to the nonlinear system by the inverse of the transformation.

PRECEDING PAGE BLANK-NOT F

ABSTRACT OF

TRANSFORMATION OF NONHOMOGENEOUS NONLINEAR SYSTEMS

BY

R. SU, G. MEYER, AND L. R. HUNT

PRECEDING PAGE BLANK-NOT FI

TRANSFORMATION OF NONHOMOGENEOUS NONLINEAR SYSTEMS

R. Su, G. Meyer, and L. R. Hunt

Abstract

The problem of when a nonlinear system can be transformed to a linear system is treated here. Previous results are further generalized.

PRECEDING PAGE BLANK-NOT FI

ABSTRACT OF

SUFFICIENT CONDITIONS FOR CONTROLLABILITY

BY

L. R. HUNT

PRECEDING PAGE BLANK-NOT FILMED

ABSTRACT

Sufficient Conditions for Controllability

L. R. Hunt

The problem is to find sufficient conditions for the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M$$

to be controllable. Here M is a connected $\mathcal{C}^\infty$ n-dimensional manifold, $f, g_1, \ldots, g_m$ are complete $\mathcal{C}^\infty$ vector fields on M, and $u_1, \ldots, u_m$ are real-valued controls. If $m = n-1$, M, $f, g_1, \ldots, g_{n-1}$ are real-analytic, M is simply connected, and $g_1, \ldots, g_{n-1}$ are linearly independent on M, then necessary and sufficient conditions are known. For the case of our $\mathcal{C}^\infty$ system with general m, we assume that the space spanned by the Lie algebra L_A generated by $f, g_1, \ldots, g_m$ and successive Lie brackets has constant dimension p on M and the algebra L_A' generated by $g_1, \ldots, g_m$ and successive Lie brackets has constant dimension $p' \leq p$ on M. If $p' = p$, Chow's Theorem implies controllability for a p-dimensional submanifold of M containing x_0. If $p' < p$, sufficient conditions are found involving the computation of certain Lie brackets at points where the vector field f is tangent to the integral manifolds of L_A'. Here we assume that every integral manifold of L_A' contains such a point.

PRECEDING PAGE BLANK-NOT

ABSTRACT OF

N-DIMENSIONAL CONTROLLABILITY WITH (N-1) CONTROLS

BY

L. R. HUNT

PRECEDING PAGE BLANK-NOT FILMED

ABSTRACT

n-Dimensional Controllability with (n-1) Controls

L.R. Hunt

Let M be a connected real-analytic n-dimensional manifold, $f, g_1, \ldots, g_{n-1}$ be complete real-analytic vector fields on M which are linearly independent at some point of M, and $u_1, \ldots, u_{n-1}$ be real-valued controls. Consider the controllability of the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M.$$

Necessary and sufficent conditions are given so that this system is controllable on any simply connected domain D contained in M on which $g_1, \ldots, g_{n-1}$ are linearly independent. These conditions depend on the computation of Lie brackets at those points where $f, g_1, \ldots, g_{n-1}$ are linearly dependent.

PRECEDING PAGE BLANK-NOT FI

NONLINEAR FAULT ANALYSIS

R. SAEKS

PRECEDING PAGE BLANK-NOT FILMED

Texas Tech University | Institute for Electronic Science

Joint Services Electronics Program | Research Unit: 3

1. Title of Investigation: Nonlinear Fault Analysis
2. Senior Investigator: Richard Saeks Telephone: (806)-742-3528
3. JSEP Funds: $25,875
4. Other Funds:
5. Total Number of Professionals: PI's 2 (3 months) RA's _____
6. Summary:

A decade ago the author initiated a research program directed at the formulation of an algorithm for *fault diagnosis in analog circuits and systems* which was capable of running *efficiently* in the dual mode "*depot/field*" *environment* associated with most DOD maintenance systems. Specifically, it was desired to formulate an algorithm which:

i). is applicable to both *linear and nonlinear systems* modeled in either the *time or frequency domain*,

ii). can be used to locate *multiple hard or soft faults*,

iii). and is capable of locating failures in "*replaceable modules*" such as an *IC chip*, *PC board*, or *subsystem* rather than dis-discrete components;

all of this being achieved with a minimum number of *test points* and at acceptable *computational cost*. Although a number of algorithms which achieve these goals in the linear case have been proposed by the author and others, little progress had been made in the nonlinear case until the past year. During the past year, however, we have found a long sought mechanism for incorporating a *failure bound* into a *simulation-after-test algorithm* thereby combining the

PRECEDING PAGE BLANK-NOT FILMED

best attributes of the classical simulation-before-test and simulation-after-test algorithms into a single package. Indeed, the success of our simulated experiments with the new algorithm has been phenomenal; and, as such, we believe that *our new algorithm represents an essentially complete solution to the analog fault diagnosis problem.*

7. Publications and Activities:

A. Refereed Journal Articles

1. Saeks, R., and R.-w. Liu, "Fault Diagnosis in Electronic Circuits," Jour. of the Soc. of Instr. and Cont. Engrgs. Vol. 20, pp. 20-22, (1981, a prelimary version of this paper also appeared in the IEEE CHMT Society Newsletter, Vol. 3, No. 3, 1980).

2. Wu, C.-c., and R. Saeks, "A Data Base for Symbolic Network Analysis," IEE Proc. Part G, (to appear).

3. Saeks, R., Sangiovanni-Vincentelli, A., and V. Vishvanathan, "Diagnosibility of Nonlinear Circuits and Systems - Part II *Dynamical Systems*, IEEE Trans. on Computers/Circuits and Systems, (to appear).

4. Wu, C.-c., Nakajima, K., Wey, C.-L., and R. Saeks, "Analog Fault Diagnosis with Failure Bounds," IEEE Trans. on Circuits and Systems, (to appear).

B. Conference Papers and Abstracts

1. Wu, C.-c., Sangiovanni-Vincentelli, A., and R. Saeks, "A Differential - Interpolative Approach to Analog Fault Simulation," Proc. of the IEEE Inter. Symp. on Circuits and Systems, Chicago, April 1981, pp. 266-269.

2. Saeks, R., "Criteria for Analog Fault Diagnosis," Proc. of the European Conf. on Circuit Theory and Design, The Hague, Aug. 1981, pp. 75-78.

3. Wu, C.-c., Nakajima, K., Wey, C.-L., and R. Saeks, "Analog Fault Diagnosis with Failure Bounds," Proc. of the 24th Midwest Symp. on Circuits and Systems, Albuquerque, June 1981, pp. 515-520.

C. Dissertations and Theses

1. Wu, C.-c., Ph.D. Dissertation, Texas Tech Univ., 1981.

2. Wey, C.-L., Ph.D. Dissertation, Texas Tech Univ., (in preparation).

3. Brandon., D., Ph.D. Dissertation, Texas Tech Univ., (in preparation).

D. Conferences and Symposia

1. Nakajima, K., Wu, C.-c., Wey, C.-L., and R. Saeks, 24th Midwest Symp. on Circuits and Systems, Albuquerque, June, 1981.

2. Saeks, R., IEEE Design for Testability Workshop, Vail, April 1981.

3. Saeks, R., IEEE Inter. Symp. on Circuits and Systems, Chicago, April 1981.

4. Saeks, R., ONR Workshop on Analog Fault Diagnosis, Notre Dame, May 1981, (Co-Chairman).

5. Saeks, R., European Conf. on Circuit Theory and Design, The Hague, Aug. 1981.

FAULT DIAGNOSIS IN ELECTRONIC CIRCUITS

R. SAEKS

AND

R. LIU

PRECEDING PAGE BLANK-NOT FILMED

Fault Diagnosis in Electronic Circuits

R. Saeks* and R. Liu**

During the past quarter century the engineering community has been witness to tremendous strides in the art of electronics design. The graphical algorithms of the previous generation have given way to the modern CAD package, the breadboard has been subsumed by the simulator. Indeed, even the universal building block has become a reality. To the contrary electronics maintenance has changed little since the day of the vacuum tube, remaining the responsibility of the experienced technician with scope and multimeter. As such, our ability to design a complex electronic circuit is quickly out-distancing our ability to maintain it. In turn, the price reductions which have accompanied modern electronics technology have been paralleled by increasing maintenance and operations costs. Indeed, many industries are finding that the life cycle maintenance costs for their electronic equipment now exceed their original capitol investment.

Given the above, it is quickly becoming apparent that the electronics maintenance process, like the design process, must be automated. Unfortunately, the 50 years of progress in circuit theory, on which our electronics design automation has been predicated, does not exist in the maintenance area. As such, the past decade has witnessed the inauguration of a basic research program to lay the foundations for a theory of electronics maintenance and a parallel effort to develop operational electronic maintenance codes.

Thus far the greatest success has been achieved in the digital electronics area, wherein the finite state nature of the UUT (unit under test) may be exploited[3]. Typically, one assumes that all failures manifest themselves in the form of component outputs which are either "stuck-at-one" or "stuck-at-zero" and/or shorts and opens[5]. Under such an assumption a theory for digital system maintenance has been developed and practical fault diagnosis algorithms are in the formative stages of development. Typically, one hypothesizes some limit on the number of simultaneous faults and then simulates the responses of the UUT to a family of test vectors for each allowed combination of faults. The actual responses of the UUT are then compared with the simulated responses to locate the failure. Although lacking in asthetic appeal the above approach, termed *fault simulation*, is ideally suited for the maintenance environment, wherein, the actual simulation process need only be done once at the factory or a maintenance depot with the simulated response data being distributed via magnetic tape to the various field locations where the actual test actual test is conducted. As such, with the aid of some sophisticated software engineering, this apparently "brute force" approach to the fault diagnosis problem has slowly evolved into a workable concept[4]. Indeed, at the present time a number of automatic test program generators which classify faults, choose test vectors, and carry out the appropriate simulation (often in a parallel processing mode), are commercially available and, as such, the automated maintenance of digital electronic circuits is becoming a reality[4].

Unfortunately, the above described success in the digital world has not been paralleled by progress in the analog world. Indeed, test engineers complain that while 80% of the boards are digital, 80% of their headaches are analog and hybrid. This difficulty arises from a number of characteristics

* Texas Tech University
** University of Notre Dame

PRECEDING PAGE BLANK-NOT FILMED

of the analog problem which are not encountered in digital circuits. Indeed, in an analog circuit:

(i) there is a continuum of possible failures,
(ii) a component may be "in tolerance" but not nominal,
(iii) complex feedback structures are encountered,
(iv) simulation is slow and costly.
(v) post-fault component characteristics may not be known,
(vi) and a fault in one component may induce an apparent fault in another

Items (i) and (ii) imply that an extremely large number of simulations will be required for analog testing. Items (iii) and (iv) suggest that these simulations will be far more expensive than similar digital simulations. Finally, items (v) and (vi) indicate that the simulation of a post-fault circuit by itself may not be a tractable problem. As such, it is by no means clear that the kind of "brute force" fault simulation algorithm associated with the digital problem will be applicable to the analog or hybrid case.

As an alternative to fault simulation, a number of academic researchers have proposed a variety of "post test" fault diagnosis algorithms, wherein, an "equation solving like" algorithm is used to locate the faulty component given the test data from UUT[2),8)]. Although these algorithms are, in some sense, "smarter" than the simulation algorithms, most of the required computing must be done in the field after the UUT has been tested. Moreover, these computational requirements must be replicated each time a unit fails. As such, the success of such "post test" algorithms is contingent on reducing their computational requirements to a bare minimum. Although no system is yet operational, with the aid of the powerful linear circuit theory developed over the past half century, a computationally efficient solution to the fault diagnosis problem for linear analog circuits appears to be within reach[1),2)]. Unfortunately, no such light exists at the end of the nonlinear tunnel, wherein progress appears to be limited by a "computational complexity/test point" bound.

Not suprisingly, the computational cost of an analog fault diagnosis algorithm is an inverse function of the number of test points at which measurements of the UUT may be made. Indeed, if one lets n be a measure of UUT complexity (which may loosely be taken to be the total number of terminals for all of the circuit components), then if one has access to $O(n)$(注1) test points the fault diagnosis problem can be resolved using linear algorithms[7),10)]. Moreover, by combining such algorithms with the above mentioned linear algorithms, acceptable computational efficiency can be obtained with $O(m)$ test points where m is a measure of the complexity of the "nonlinear subsystem" of the UUT[6),7)]. Although such algorithms can be effective on the typical academic example a "real world" PC (printed circuit) board does not have terminal space for the 20 or 30 test points which are required even for a routine board made up of discrete components and/or SSI (Small Scale Integration). Although the problem can be partially alleviated by making internal measurements with the aid of a "bed-of-nails" tester it has been our experience that such testers cause as many failures as they locate while their applicability to two-sided, multilayer, and coated boards is severely limited. As such, we would like to limit the number of test points to the terminal space available at the edge of a PC board. On the other hand, the UUT complexity, n, increases with the area of the board. As such, the number of test points required by an analog fault diagnosis algorithm should increase at a rate of no greater than $O(n^{1/2})$. A further study of the possible tradeoff between test points and computational cost appears in references 11) and 12).

Unfortunately, all computationally acceptable "post test" algorithms which have thus far been proposed have test point requirements which grow linearly with UUT complexity (assuming that m grows linearly with n). As such, many researchers are looking at the classical fault simulation algorithms with renewed vigor. Indeed, these algorithms have minimal on-line computational costs, while the number of test points employed, can easily be

(注1) $f(n)=O(n)$ means f increases in the order of n; more precisely, $|f(n)|\leq c|n|$ for some $c>0$.

kept below $0(n^{2/1})$. The difficulty lies with the required number of simulations and the development of decision algorithms which will allow us to "interpolate" between simulated data points.

Thus, while the state-of-the-art in digital diagnosis is fast maturing, a serious investigation of analog fault diagnosis problems is only just beginning. Indeed, a satisfactory fault diagnosis code for linear analog circuits has yet to be demonstrated while the nonlinear problem has yet to progress beyond the basic research stage.

References

1) Chen, H. S. M. and R. Saeks: A Search Algorithm for the Solution of the Multifrequency Fault Diagnosis Equations, IEEE Trans. on Circuits and Systems, CAS-26, 589/594 (1979)
2) Duhamel, P. and J. C. Rault: Automatic Test Generation Techniques for Analog Circuits—A Review, IEEE Trans. on Circuits and Systems, 25, 411/439 (1979)
3) Friedman, A. D. and Memon, P. R.: Fault Detection in Digital Circuits, New York, Prentice Hall (1971)
4) Greenbaum, J. R.: Computer-Aided Fault Analysis —Today, Tomorrow, or Never, in Rational Fault Analysis (ed. R. Saeks, and S. R. Liberty), New York, Marcel Dekkar, 96/111 (1977)
5) Hayes, J. P.: Modeling Faults in Digital Circuits, in Rational Fault Analysis, (ed. R. Saeks, and S. R. Liberty), New York, Marcel Dekker, 78/95 (1977)
6) Hsieh, M.: Ph. D. Dissertation, Texas Tech Univ. (1980)
7) Ngo, Q.-D.: M. S. Thesis, Texas Tech Univ. (1980)
8) Plice, W. A.: Automatic Generation of Fault Isolation Tests for Analog Circuit Boards-A Survey, Presnted at ATEX EAST 78, Boston, Sept. (1978)
9) Saeks, R., Singh, S. P. and R. W. Liu: Fault Isolation via Components Simulation, IEEE Trans. on Circuit Theory, CT-19, 634/640 (1972)
10) Trick, T. N., Mayeda, W. and A. A., Sakla: Calculation of Parameter Values from Node Measurements, IEEE Trans. on Circuits and Systems, CAS-26, 466/474 (1979)
11) R. W. Liu and V. Visvanathan: Sequentially Linear Fault Diagnosis: Part I—Theory, IEEF Trans. on Circuits and Systems, 490/496, July (1979)
12) V. Visvanathan and R. W. Liu: Sequentially Linear Diagnosis: Part II—The Design of Diagnosable Systems, IEEE Trans. on Circuits and Systems, 558/564, July (1979)

CRITERIA FOR ANALOG FAULT DIAGNOSIS

R. SAEKS

PROCEEDINGS OF THE 1931 EUROPEAN CONFERENCE
ON CIRCUIT THEORY AND DESIGN

THE HAGUE, THE NETHERLANDS, 25-23 AUGUST, 1931

PRECEDING PAGE BLANK-NOT FILMED

CRITERIA FOR ANALOG FAULT DIAGNOSIS*

R. Saeks
Department of Electrical Engineering
Texas Tech University
Lubbock, TX 79409 USA

INTRODUCTION

After a half century of neglect by the electronics community the past decade has witnessed an expanding effort in the analog fault diagnosis area. Indeed, the ever increasing complexity of electronic circuits combined with the decreasing availability of trained maintenance technicians has pushed *computer-aided testing* (CAT) to the forefront of electronics research. Unfortunately, the tremendous strides which have been made in digital test technology have not been paralleled by equal progress in the analog area. As such, even though "80% of the boards are digital 80% of the problems are analog".

The lack of progress in analog CAT vis-a-vis digital CAT may be attributed to four factors:

i). the cost of analog circuit simulation,

ii). the continuous nature of analog failure phenomena,

iii). tolerances on the "good" components in an analog circuit,

iv). and the lack of viable models for the components in a faulty circuit.

Moreover, these difficulties have been exagerated by the economics of the maintenance environment which limits the degree to which many of the classical tools of analog circuit design can be used in a CAT package.

The purpose of the present paper is to describe a set of criteria which we believe a practical analog CAT algorithm should achieve and to indicate the degree to which they are met by the various algorithms which have thus far been proposed.[1] These criteria include computational requirements, numbers of test points and test vectors employed, robustness to tolerance effects, availability of models, and the degree to which the algorithm is amenable to parallel processing. Although many specific algorithms have been proposed they may naturally be classified into three categories:

i). simulation-before-test,

ii). simulation-after-test with a single test vector,

iii). and simulation-after-test with multiple test vectors.

Each of these three approaches to the analog CAT problem is compared against our criteria, and, interestingly, each approach fails to meet at least one of the proposed criteria.

CRITERIA

A. Computational Requirements: Unlike a CAD algorithm which is used only in the initial design of a circuit or system, a CAT algorithm lives in an operational environment and thus must be used repeatedly each time a system fails. As such, a viable measure for the computational cost of a CAT algorithm must distinguish between on-line computation which is done in the field and must be repeated for each unit under test (UUT) and off-line computation which is independent of the unit under test and thus need only be done once at the factory or a maintenance depot. Indeed, the distinction between on-line and off-line computation is further exaggerated by the high cost of computing and the dearth of trained personnel in a field maintenance environment vis-a-vis that is available at a maintenance depot. Thus in a CAT algorithm a *great priority must be placed on reducing the on-line computational requirements* even at the cost of significantly increasing the off-line computation. As such, an algorithm which is viable in a design environment might not be acceptable in a maintenance environment and vice-versa. Indeed, in a CAT algorithm one would be happy to accept the cost of generating a complex data base in an off-line environment to achieve a reduction in on-line computational requirements.

B. Test Points: Historically, analog circuits have been tested with the aid of a "bed of nails" tester which allows one to make use of test data which is not accessible via the input and output terminals of the circuit board. Unfortunately, modern circuit boards are often multilayered and/or coated, thereby limiting the applicability of the "bed of nails" concept. As such, a modern CAT algorithm must be designed to work with the test data which is available at the externally accessible terminals of a printed circuit board. In practice, this proves to be a dominating factor in the design of a CAT package, which precludes the use of some of the more attractive algorithms with test point requirements which grow linearly with circuit complexity. In fact, circuit complexity is proportional to the area of a printed circuit board (if not a power thereof) while the number of accessible test points is proportional to the edge length of the board. As such, in a practical CAT package it is reasonable to require that the *number of test points grow with the square root of circuit complexity* (or less).

C. Robustness: Unlike a digital system wherein a device is either good or bad in an analog circuit a device is either "in-tolerance" or "out-of-tolerance" and, as such, an analog CAT algorithm must be able to cope with the effects of *components which are in-tolerance but not nominal*. Although, at the time of this writing, there is insufficient experimental data to determine the import of robustness in an analog algorithm it is, at minimum, a factor of which one must be cognizent and may, in fact, prove to be a dominating factor in the design of a

* This research supported in part by the Joint Services Electronic Program at Texas Tech University under ONR Contract 76-C-1136.

PRECEDING PAGE BLANK-NOT FILMED

viable CAT package.

D. Models: Since most CAT algorithms presuppose some form of circuit simulation in their operation and design of such an algorithm must consider the type and availability of circuit models which are required and/or available. In particular, does the algorithm use nominal circuit models or *faulted circuit models?* Indeed, even if nominal circuit models are used do they operate in their normal range? Finally, one must consider whether or not the algorithm is capable of dealing with "fuzzy" components which do not admit viable simulation models.

E. Module vs. Parameter Testing: Most analog fault diagnosis algorithms can be catagorized as either module oriented or parameter oriented. In the former case the algorithm tests the input-output performance of the individual modules or subsystems which make up the UUT while in the latter case the algorithm estimates a set of parameter values which determine the performance of a given circuit component. Although one can often formulate a circuit model for a given module thereby permitting one to use a parameter oriented algorithm to test modules, such a process may unnecessarily complicate the test procedure. As such, *a module oriented CAT algorithm is preferred* over a parameter oriented algorithm if it can be formulated without compromising other factors.

F. In-Situe Testing: Although secondary to the above considerations the ideal CAT algorithm should allow for *in-situe testing*. Since one cannot control the input signals applied to the UUT in-situe such an algorithm must work with an arbitrary input signal rather than a fixed set of test vectors.

G. Parallel Processing: Since the CAT problem is inherently a large scale systems problem it is essential to exploit whatever computational power is available to reduce both on-line and off-line computational requirements. In particular, digital CAT algorithms often use some degree of parallel processing in their implementation. Given the additional computational problems associated with an analog CAT algorithm *the degree to which an algorithm can be implemented in parallel becomes a significant factor* in determining its viability and should therefore be included among our criteria for an analog CAT package.

In the above paragraphs we have described seven aspects of the CAT problem which must be considered in judging an analog CAT algorithm. Although we would ideally like to formulate an algorithm with minimal computational requirements a *moderate amount of off-line computation* is acceptable since the off-line computation need only be done once and is carried out in a depot environment where good computational facilities and high level personnel are available. On the other hand since the *on-line computation* associated with a CAT algorithm is replicated for each UUT and carried out in a field environment it must be kept to a minimum. Likewise the test point requirements for an analog CAT algorithm must be kept to a minimum. Although the requirement that the number of test points used by a CAT algorithm grow with the square root of circuit complexity is open to debate it is indicative of a fundamental limitation to the effect that the number of test points should grow at less than a linear rate with circuit complexity. Concerning the remaining criteria we want an algorithm that is robust though the significance of this requirement is not fully understood at this time. Similarly, the availability of circuit models to implement an algorithm must be considered. Finally, but secondary to the above requirements, it would be desirable to have a module oriented algorithm which is amenable to in-situe testing and parallel processing. These criteria are summarized in table 1, along with a set of goals which one would wish to achieve in an "ideal analog fault diagnosis algorithm".

CAT ALGORITHMS

A. Simulation-Before-Test: Although it is essentially a brute force search algorithm, simulation-before-test is well suited to the depot/field computational environment of the CAT problem and, as such, it predominates in most state-of-the-art digital CAT packages.[3] On the other hand its weaknesses become more pronounced in the analog problem wherein it has yet to be successfully implemented. Basically, a simulation-before-test algorithm is a search algorithm in which one simulates the expected test data which would result from various hypothesized failures in an off-line environment. Then when the actual test data is obtained in the field it is compared with the simulated results to determine the failure. Needless to say the technique requires immense amounts of off-line computer time to generate the required data base but is extremely efficient on-line, wherein one need only compare the test results with the simulated data base.

Unfortunately, the cost of an analog simulation is much greater that that of a digital simulation. Moreover, one requires a much larger data base in the analog problem than in the digital problem to cope with the continuous nature of the analog failure phenomena and the robustness problem. As such, there is considerable doubt about the applicability of the simulation-before-test concept in an analog CAT package.

Vis-a-vis our criteria for analog fault diagnosis simulation-before-test requires extremely large amounts of off-line computer time but only a minimum of on-line computer time. Additionally, the test point requirements for the algorithm are minimal. On the other hand the technique has no inherent robustness and uses faulted simulation models for all components. With regard to the secondary factors the algorithm is module oriented and amenable to parallel processing but not in-situe testing. These considerations are summarized in Table 1.

B. Simulation-After-Test with a Single Test Vector: Rather than using a search algorithm for fault diagnosis one can attempt to model the analog faul diagnosis problem as a nonlinear equation in which one solves for the internal variables or component parameters in terms of the test data. Although this may, at first, seem to totally bypass the repetitive simulation-before-test algorithm, a careful analysis will reveal that each iteration o the required numerical equation solver amounts to simulation of the UUT. In this case, however, the particular simulations which one carries out are based on known test data rather than a-priori fault hypotheses. As such, the simulations are

done on-line after the test data has been obtained and the technique is thus termed simulation-after-test.[2]

In the case where only a single test vector is employed the resultant fault diagnosis equations are "almost linear" and may be solved with the aid of a single (off-line) sparse matrix inversion.[4,5] The test point requirements for the algorithm, however, grow linearly with circuit complexity. Interestingly, this class of algorithms has been discovered independently by a number of authors over the years, most of whom though that they had found the "ideal algorithm" until they fully appreciated the significance of the test point requirement which severely limits its applicability. From the point of view of our other criteria, however, the algorithm is, indeed, "ideal". Off-line computational requirements are moderate while on-line computational requirements are minimal. Moreover, the algorithm is inherently robust and requires no simulation models of any kind, it tests modules, and it is amenable to in-situe testing. Finally, the computational requirements associated with the algorithm are sufficiently low so as to render the parallel processing question moot.

C. Simulation-After-Test with Multiple Test Vectors: One approach to reducing the test point requirements of the simulation-after-test algorithm is to use multiple test vectors to increase the number of equations obtained from a given set of test points, thereby rendering the fault diagnosis equation soluable with a restricted number of test points. The most common form of the multiple test vector algorithm is the multifrequency algorithm used in linear fault diagnosis, though the concept extends to the nonlinear case via the use of multiple test vectors of any type.[1,2]

The reduced test point requirement obtained via the use of multiple test vectors is, however, achieved at the cost of greatly increasing the complexity of the resultant fault diagnosis equations. Indeed, the "almost linear" equations of the single test vector algorithm are replaced by an extremely complex set of nonlinear equations (even for linear systems) in the multiple test vector algorithm. Although these equations can be made trackable in the linear case they appear to be totally untrackable in the nonlinear case and, as such, most of the advantages of the simulation-after-test concept are lost when multiple test vectors are employed.

With regard to our criteria the multiple test vector algorithms require large amounts of on-line computer time though relatively little off-line computer time is required. In its most obvious form the technique is robust, though this robustness is compromised by most of the "tricks" which have been proposed to make the multiple test vector fault diagnosis equations trackable. Faulted simulation models are required and the algorithm is inherently parameter oriented. Finally, it is not suited to either in-situe testing or parallel implementation.

CONCLUSIONS

The above concepts are summarized in Table 1, wherein the various criteria, by which an analog CAT algorithm should be measured are tabulated, the goals for an ideal algorithm are described, and the degree to which the various algorithms achieve these goals is indicated. From the table it is apparent that none of the algorithms is fully acceptable. Indeed, even if one neglects the secondary considerations regarding modules vs. parameters, in-situe testing, and parallel processing all three approaches fail to meet one or more of the primary criteria (indicated by capital letters in the table). As such, the proper approach to the solution of the analog CAT problem remains an open question.

REFERENCES

1. Duhamel, P., and J.C. Rault, "Automatict Test Generation for Analog Circuits and Systems: A Review", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 411-440, (1979).

2. Plice, W.A., "Automatic Generation of Fault Isolation Tests for Analog Circuit Boards, A Survey", Presented at ATEX East '78, Boston, pp. 26-28, Sept. 1978.

3. Saeks, R., and S.R. Liberty, Rational Fault Analysis, New York, Marcel Dekker, 1977.

4. Saeks, R., Singh, S.P., and R.-w. Liu, "Fault Isolation via Components Simulation", IEEE Trans. on Circuit Theory, Vol. CT-19, pp. 634-640, (1972).

5. Trick, T.N. Mayeda, W., and A. Sakla, "Calculation of Parameter Values from Node Voltage Measurements", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 466-474, (1979).

Criteria	Goal	Simulation-Before Test	Simulation-After-Test with a Single Test Vector	Simulation-After-Test with Multiple Test Vectors
Off-line Comp.	Moderate	VERY HIGH	Moderate	Moderate
On-line Comp.	Minimal	Minimal	Minimal	HIGH
Test Points	Less than linear growth	Less than linear growth	LINEAR GROWTH	Less than linear growth
Robustness	Yes	NO	Yes	Yes (sometimes)
Models	Nominal	FAULTED	None	FAULTED
Modules/Param.	Modules	Modules	Modules	Parameters
In-situe Test	Yes	No	Yes	No
Parallel Proc.	Yes	Yes	—	No

Table 1: Performance of Analog Fault Diagnosis Algorithms. Unacceptable performance factors are indicated by capital letters.

ANALOG FAULT DIAGNOSIS WITH FAILURE BOUNDS

C.-C. WU, K. NAKAJIMA,

C.-L. WEY AND R. SAEKS

24TH MIDWEST SYMPOSIUM ON CIRCUITS AND SYSTEMS

UNIVERSITY OF NEW MEXICO, ALBUQUERQUE, NM

JUNE 29-30, 1931

ANALOG FAULT DIAGNOSIS WITH FAILURE BOUNDS*

C.-c. Wu, K. Nakajima, C.-L. Wey, and R. Saeks
Department of Electrical Engineering
Texas Tech University
Lubbock, Texas 79409

Abstract

A simulation-after-test algorithm for the analog fault diagnosis problem is proposed in which a bound on the maximum number of simultaneous failures is used to minimize the number of test points required. The resultant algorithm is applicable to both linear and nonlinear systems and can be used to isolate a fault up to an arbitrarily specified "replaceable module".

I. INTRODUCTION

Conceptually, analog fault diagnosis algorithms can be subdivided into three classes;[3] simulation-before-test, simulation-after-test with a single test vector, and simulation-after-test with multiple test vectors. The former is commonly employed in digital testing and is characterized by minimal on-line computational requirements. Unfortunately, the high cost of analog circuit simulation coupled with the large number of potential fault modes which must be simulated in an analog circuit limits the applicability of simulation-before-test algorithms in an analog test environment. As an alternative to simulation-before-test, a number of researchers have proposed simulation-after test algorithms, in which the internal system variables or component parameters are computed from the test data via a "nonlinear equation solver - like" algorithm. Indeed, in the case where sufficiently many test points are available only a single test vector is required and the fault diagnosis problem reduces to the solution of a linear equation.[8,9] Except for the large number of test points required this approach is ideally suited to the analog fault diagnosis problem and, as such, a considerable research effort has been directed towards the problem of reducing its test point requirements.[3] One such approach uses multiple test vectors to increase the number of equations obtained from a given set of test points. Unfortunately, this is achieved at the cost of greatly complicating the set of simultaneous equations which must be solved and, as such, the applicability of the approach is limited.

The purpose of the present paper is to describe an alternative simulation-after-test algorithm in which a bound on the maximum number of simultaneous failures is used to reduce the test point requirements while still retaining the computational simplicity inherent in a single test vector algorithm. Indeed, even though a circuit may contain several hundred components it is reasonable to assume that at most two or three have failed simultaneously. As such, rather than solving a set of simultaneous equations in n-space the solution to our fault diagnosis problem actually lies in a two or three dimensional submanifold which should yield a

* This research supported in part by the Joint Services Electronic Program of Texas Tech University under ONR Contract 76-C-1136.

PRECEDING PAGE BLANK-NOT FI

commensurate reduction in test point requirements. Unfortunately, even though we may assume that at most two or three components have failed we do not know which two or three, and as such, some type of search is still required. Fortunately, with the aid of an appropriate decision algorithm the required search can be implemented quite simply.

Consider the circuit or system which is illustrated in figure 1.

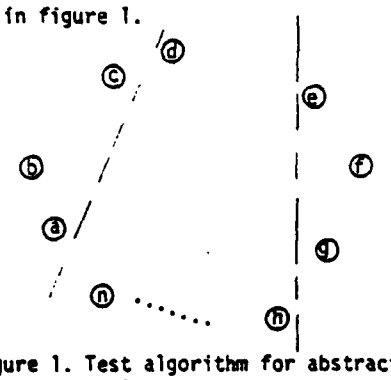


Figure 1. Test algorithm for abstract circuit or system.

Here, the individual circuit components or subsystems are denoted by circles indexed from a to n. These components are subdivided into two groups, at each step of the test algorithm, as indicated by the dashed lines in figure 1. At each step we assume that one group; say, d through n; is composed of good components and we use the known characteristics of these components together with the test data to determine whether or not the remaining components; a,b and c in this case; are good. Of course, if components d through n are actually good then the resultant test results for components a, b and c will be reliable. On the other hand, if any one of the components d through n is faulty the test data on a, b and c will be unreliable. As such, we repeat the process at the next step of the test algorithm with a different subdivision of components. For instance, we may assume that a through d and h through n are good and use their characteristics to test components e, f and g. Finally, after a number of such repetitions the test results obtained at various steps are analyzed to determine the faulty components.

Of course, the number of components which may be tested at one step is dependent on the number of test points available while the number of steps required is determined by the number of components which may be tested at any one step and the bound on the maximum number of simultaneous failures. As such, the procedure yields a natural set of tradeoffs between the numbers of test points, simultaneous failures and steps required by the algorithm. Indeed, since the computational cost associated with each step of the algorithm is essentially the cost of a single system simulation the latter parameter is a natural measure of the computational cost.

In the following section we describe the simulation model used to test one set of components under the assumption that the remaining components are good. In section three two decision algorithms for analyzing the resultant test data are described. Indeed the required theory is reminiscent, though not identical to, the t-diagnosibility theory developed for digital testing over the past decade.[4,6] Finally, section four is devoted to a number of examples including linear circuits with 12 and 22 components which were run on a desktop calculator and a 16 bit mini, respectively.

II. THE SIMULATION MODEL

Although our test algorithm can be formulated in terms of any of the standard system models for the purpose of this exposition we will assume a *component connection model* for the circuit or system test.[2] In the nonlinear case the *unit under test* is represented by a set of decoupled state models characterizing its components and/or subsystems together with an algebraic connection equation as follows.

$$\dot{x}_i = f_i(x_i, a_i) \quad ; \; x_i(0) = 0, i=1,2, \ldots ,n \qquad (2.1)$$
$$b_i = g_i(x_i, a_i)$$

and

$$a = L_{11}b + L_{21}u \qquad (2.2)$$

$$y = L_{21}b + L_{22}u \qquad (2.3)$$

Here, $a = \mathrm{col}(a_i)$ is the column vector composed of

the *component input variables*, $b = \mathrm{col}(b_i)$ is the column vector composed of *component output variables*, u is the vector of *external test inputs* applied to the system and y is the vector of *system responses measured at the various test points*. Although the component connection model is not universal it is quite general and subsumes most of the classical topological connection models commonly used in circuit and system theory.[2] Moreover, its inherently decoupled nature is ideally suited to the test problem wherein we desire to distinguish between the characteristics of the individual system components. Although these components may be taken to be elementary RLC components and/or discrete semiconductor devices, in practice the "components" are taken to be the "replaceable modules" within the circuit or system, under test; say, an IC or a "throw-away" circuit board.

At each step of the test algorithm we subdivide the "components into two groups denoted by "1" and "2" with the components in group "1" assumed to be good and used together with the known values of u and y to compute the component input and output variables, a_i and b_i, for the components in group "2". Although computationally we prefer to work with the decoupled component equations for notational brevity we combine the equations for the components in each group into a single equation

$$\dot{x}^1 = f^1(x^1,a^1) \quad ; \; x^1(0) = 0 \qquad (2.4)$$
$$b^1 = g^1(x^1,a^1)$$

and

$$\dot{x}^2 = f^2(x^2,a^2) \quad ; \; x^2(0) = 0 \qquad (2.5)$$
$$b^2 = g^2(x^2,a^2)$$

Here, x^1, a^1 and b^1 are the vectors of group "1" component state variables, inputs and outputs; and similarly for x^2, a^2 and b^2. To retain notational compatibility with 2.4 and 2.5 we reorder and partition the connection equations of 2.2 and 2.3 to be conformable with 2.4 and 2.5 as follows

$$a^1 = L_{11}^{11}b^1 + L_{11}^{12}b^2 + L_{12}^{1}u \qquad (2.6)$$

$$a^2 = L_{11}^{21}b^1 + L_{11}^{22}b^2 + L_{12}^{2}u \qquad (2.7)$$

$$y = L_{21}^{1}b^1 + L_{21}^{2}b^2 + L_{22}u \qquad (2.8)$$

Given equations 2.4 through 2.8 our goal is to compute the goup "2" component variables, a^2 and b^2, given the *test input*, u, the *measured test responses*, y, and an *assumption to the effect that the group "1" components are not faulty*. To this end we assume that L_{21}^2 admits a left inverse, $[L_{21}^2]^{-L}$, which, in turn, determines the allowable component subdivisions. Under this assumption one may then formulate a component connection model for a *"pseudo circuit"* composed of the group "1" components with external input vector $u^p = \mathrm{col}(u\ ,\ y)$ and external output vector $y^p = \mathrm{col}(a^2,\ b^2)$ in the form

$$\dot{x}^1 = f^1(x^1,a^1) \quad ; \; x^1(0) = 0 \qquad (2.9)$$
$$b^1 = g^1(x^1,a^1)$$

$$a^1 = K_{11}b^1 + K_{21}u^p \qquad (2.10)$$

$$y^p = K_{21}b^1 + K_{22}u^p \qquad (2.11)$$

Since, in our test problem u and y are known, the above equations can be solved via any standard analysis code to compute $y^p = (a^2,b^2)$. Now, under our assumption that the group "1" components are not faulty $y^p = (a^2,b^2)$ represents the inputs and outputs which actually appeared at the terminals of the group "2" components during the test. As such, we may determine which of the group "2" components are faulty by solving equation 2.5 with input a^2 and checking to determine whether or not the resultant output coincides with b^2. Of course, since our assumption to the effect that the group "1" components are not faulty may not be valid *the results of this test are not reliable*. As such, we repeat the process a number of times with different choices for the subdivision of the components into group "1" and group "2". Here, the only constraint on the choice of subdivisions is the requirement that $[L_{21}^2]^{-L}$ exist while the number of combinations employed is limited only by

the cost of the required simulations. The results of the several steps in the test algorithm are then analyzed via the techniques described in the following section to determine those components which are actually faulty.

III. DECISION ALGORITHMS

Since the results of the test described in the preceding section are dependent on our assumption that the group "1" components are not faulty they are not immediately applicable. Following the philosophy initiated by Preparata, Metze, and Chein[6] in their study of self testing computer networks, however, if one assumes a bound on the maximum number of faulty components it is possible to determine the actual fault(s) from an analysis of the test results obtained at various steps in the algorithm. To this end we will give a complete analysis of the theory required to locate a single fault together with an heuristic which is applicable to the multiple fault case.

Let us assume that at most one circuit component is faulty and that the test results obtained from a given step of the algorithm indicate that all group "2" components are good as indicated in the following table, where the binary anotation to the left of the group "2" components indicates those which were found to be good(0) and bad (1) at this step of the test algorithm

"2" \ "1"	a	b	c	...	k
0	x				
0	y				
⋮	⋮				
0	z				

In this case we claim that the group "2" components are, in fact, good. Indeed, if a group two component were actually faulty then our test results are incorrect, which could only happen if one of the group "1" components was faulty. As such, the system would have two faulty components contradicting our assumption to the effect that at most one component is faulty.

Now, consider the case where the results from a given step of the test algorithm indicate that exactly one group "2" component is faulty; say, x.

"2" \ "1"	a	b	c	....	k
1	x				
0	y				
⋮	⋮				
0	z				

In this case the same argument we used above will guarantee that the components which test good; say, y through z; are, in fact, good. On the other hand we have no information about x. It may be faulty, or alternatively, the test result may be due to a faulty group "1" component.

Finally, consider the case where two or more group "2" components test bad in a given step as indicated in the following table.

"2" \ "1"	a	b	c	...	k
1	x				
1	y				
⋮	⋮				
0	z				

Since, under our assumption of a single failure, it is impossible for two or more group "2" components to be faulty, this test result implies that at least one of the group "1" components is bad. On the other hand since we have assumed that there is at most one faulty component the faulty group "1" component is the only faulty component and, as such, the group "2" components are all good.

Consistent with the above, at each step of the test algorithm, either all or all but one group "1" components are found to be good. As such, if we choose our subdivisions so that the components which are found to be good at one step of the algorithm are included in group "1" in all succeeding steps we eventually will arrove at a group "1", all of whose components are known to be good. As such, the test results obtained at that step will be reliable thereby allowing us to accurately determine the faulty components in group "2".

Unlike the single fault case, at the time of this writing, we do not yet have an exact decision algorithm for the multiple fault case. Following Liu, however, the problem can be greatly simplified if one adopts an "analog heuristic" to the effect that two independent analog failures will never cancel.[5]

Recall from our discussion of the single fault case that whenever a test result indicates that a component is good then it is, in fact, good. Although this is not rigorously true in the multiple failure case it is true under the assumption of our heuristic. For instance, consider the test results indicated in the following table in which x is found to be good.

"2"	"1"	a	b	c	...	k
0	x					
1	y					
⋮	⋮					
0	z					

Now, if x is actually faulty there must be a faulty group "1" component whose effect is to cancel the error in x as observed during this step of the test algorithm. This is, however, forbidden by our heuristic and, as such, we conclude that x is actually good.

Interestingly, our heuristic can be carried a step further than indicated above since, under our heuristic, a bad group "1" component would normally yield erroneous test results. An exception would, however, occur if some of the group "1" components are totally decoupled from some of the group "2" components. As such, if prior to our test we generate a coupling table (by simulation or a sensitivity analysis) which indicates whether or not a faulty group "1" component will effect test results on a group "2" component, our heuristic may be used to verify that certain group "1" components are good whenever a good group "2" component is located. Consider, for example, the following table in which a "1" in the i-j position indicates that the test results for component i

"2"	"1"	a	b	c	...	k
0	x	1	0	1		1
1	y	1	1	0		0
⋮	⋮	⋮	⋮	⋮		⋮
0	z	0	1	1		0

are effected by component j while a "0" in the i-j position indicates that component j does not effect the test results for component i. Now, since component z has been found to be good in this test our heuristic implies that b and c are also good. Similarly, since component x is good so are a,c, and k. Thus, with a single test we have verified that x,z, a,b,c and k are all good.

IV. EXAMPLES

To obtain examples the above techniques were applied to the 12 and 22 component linear amplifier circuits shown in figure 2 using simulated test data for various numbers of simultaneous failures, choices of test point locations, and both decision algorithms. All analysis for the 12 component circuit was done on an HP 9825 desktop calculator while the 22 component examples were run on a TI 990/20 minicomputer. The results of some 150 simulations of the algorithms are tabulated in table 1. where the number of test points, simultaneous faults, and the decision algorithm employed are indicated. The results of the various simulations are indicated by the ambiguity of the resultant diagnosis. For instance, in our simulation of the 12 component circuit with 3 test points, one failure and the exact algorithm 12 runs were made (one with each component fauly). On 10 occasions the fault was located exactly while the fault was located up to an ambiguity set composed of two components on 2 occasions. Finally, we note that the 5th run of the 12 component circuit indicated by an asterick in the table represents a simulation in which the good components were set at +/-2% off of nominal to test the robustness of the algorithm.

VI. REFERENCES

1. Amin, T., unpublished notes, Bell Laboratories,

1980.

2. DeCarlo, R.A., and R.Saeks, Interconnected Dynamical Systems, New York, Marcell Dekker, (to appear).

3. Duhamal, P., and J.C. Rault, "Automatic Test Generation Techniques for Analog Circuits and Systems: A Review", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 411-440, (1979).

4. Hakimi, S.L. "Fault Analysis in Digital Systems - A Graph Theoretic Approach", in Rational Fault Analysis, New York, Marcal Dekker, 1977, pp. 1-12.

5. Liu, R.-w., unpublished notes, Univ. of Notre Dame, 1980.

6. Preparata, F.P., Metze, G., and R.T. Chein, "On the Connection Assignment Problem of Diagnosible Systems", IEEE Trans. on Electronic Computers, Vol. EC-16, pp. 448-454, (1967).

7. Saeks, R., "Criteria for Analog Fault Diagnosis", Proc. of the 1981 European Conf. on Circuit Theory and Design, The Hague, Aug. 1981, (to appear).

8. Saeks, R., Singh, S.P., and R.-w. Liu "Fault Isolation via Components Simulation", IEEE Trans. on Circuit Theory, Vol. CT-19, pp. 634-640, (1972).

9. Trick, T.N., Mayeda, W., and A. Sakla, "Calculation of Parameter Values from Node Voltage Measurements", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 466-474, (1979).

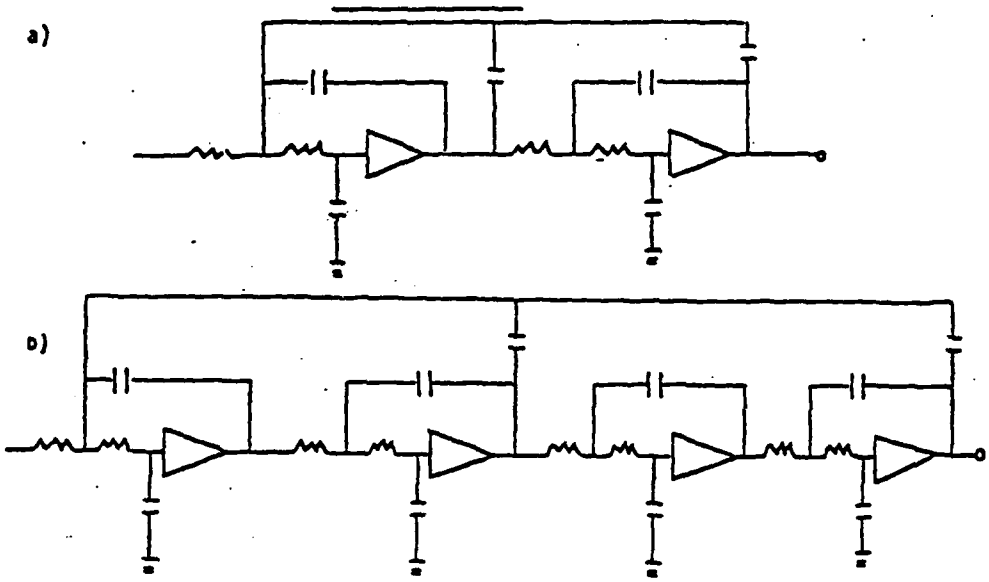


Figure 2. a) 12 component amplifier and b) 22 component amplifier. All stages of the amplifier circuits have nominal op-amp gains of 1.6, nominal resistance of 10K ohms, and nominal capacitance values of .001uf while the feedback capicators have nominal values of 100pf.

Circuit/Computer	#Test Points	#Faults	Dec. Alg	Ambiguity set 1	2	4	6
12 component circuit simulated on an HP 9825 desktop calculator	4	1	Exact	12			
	4	2	Heuristic	12			
	3	1	Exact	10	2		
	3	1	Heuristic	12			
	3	1	Exact	10*	2*		
22 component circuit simulated on a TI 990/20 minicomputer	8	1	Exact	22			
	6	1	Exact	18		4	
	5	1	Exact	16			6
	5	1	Heuristic	22			

Table 1. Simulated test data. * indicates a simulated test in which the good components were taken to be +/- 2 off of nominal.

A DIFFERENTIAL-INTERPOLATIVE APPROACH TO ANALOG FAULT SIMULATION

C.-C. Wu

A. Sangiovanni-Vincentelli

AND

R. Saeks

1981 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS PROCEEDINGS

Radisson Chicago Hotel, Chicago, Ill., April 27-29, 1981

A DIFFERENTIAL-INTERPOLATIVE APPROACH TO ANALOG FAULT SIMULATION

C.-c. Wu**, A. Sangiovanni-Vencentelli*, and R. Saeks**

I. INTRODUCTION

After a half century of neglect by the circuits and systems community the past decade has witnessed the emergence of a research effort in the analog circuit maintenance area. The various algorithms which have been thus far proposed for the analog fault diagnosis problem may naturally be subdivided into two classes termed "simulation-before-test" and "simulation-after-test". The former are commonly used in digital system test algorithms and require an automatic test program generator (ATPG) which simulates the responses of "all possible" failures. This is typically done at a maintenance depot with the simulated responses being recorded and shipped to the field where the response of the unit under test (UUT) is compared with the simulated responses to determine the failure. The major advantage of simulation-before-test is that it is ideally matched to the depot/field maintenance environment with the largest part of the computation done only once. As such, the technique is ideally suited for digital testing where the binary nature of the problem keeps the number of failures to be simulated within bounds and eliminates tolerance problems. Unfortunately, in the analog problem we must cope with a continuum of possible failures and simultaneously deal with good components which are in tolerance but not nominal. As such, a tremendous number of simulations are required by a simulation-before-test algorithm, while some type of decision algorithm is required to cope with the tolerance effects.

Unlike simulation-before-test, simulation-after-test uses an "equation solver-like" algorithm to compute the parameters of the UUT components in the field. Since most such algorithms require iterative evaluation of the equation to be solved, the UUT is effectively simulated at each iteration, though the simulation is based on actual test data rather than hypothesized failure data. The simulation process is, thus, carried out after testing the UUT and hence the choice of terminology. The advantage to such an approach is that the faulty component parameters are computed explicitly, thereby, eliminating the ambiguity caused by the use of discrete simulation-before-test data and tolerance effects. Although relatively few simulations are required for each UUT, they must be carried out in the field rather than the depot and they must be repeated for each UUT.

The purpose of the present paper is to describe a research effort directed at alleviating some of the difficulties in developing a simulation-before-test algorithm for analog fault diagnosis. The underlying philosophy and motivation for our formulation is discussed in section 2, along with a derivation of the required differential-interpolative fault diagnosis formula. Finally, section 3 is devoted to a number of illustrative examples of the approach. These include both linear and nonlinear examples formulated in the frequency and time domains, respectively.

II. A DIFFERENTIAL-INTERPOLATIVE ALGORITHM

Although any practical fault diagnosis algorithm must be able to handle systems with a hundred or more components, from an intuitive point of view our algorithm is best illustrated in the two component cases where the parameter space can be displayed graphically. Say, we are dealing with an RC circuit for which the parameter space is illustrated in figure 1.

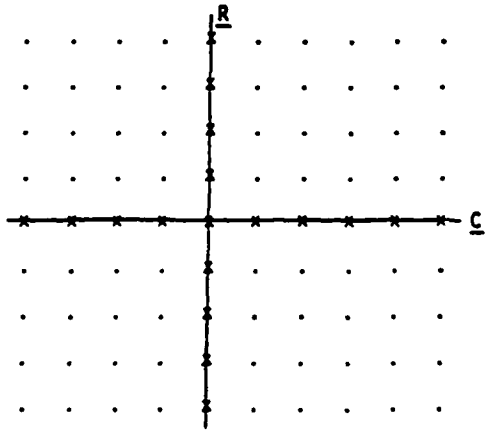


Figure 1: Parameter space for RC circuit.

Here, $\underline{R}$ and $\underline{C}$ represent normalized parameter values, wherein, the nominal parameter values are transformed to the origin. In the most general simulation-before-test algorithms one assumes that the faulty parameter values may lie anywhere in the $\underline{R}$-$\underline{C}$ plane and therefore carries out simulations along a two

* Dept. of Elec. Engrg. and Comp. Science, Univ. of California at Berkeley, Berkeley, CA. 90024.
** Dept. of Elec. Engrg., Texas Tech Univ., Lubbock, TX 79409. This research supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

PRECEDING PAGE BLANK-NOT FI

dimensional discrete array spread over the entire plane.

Fortunately, in a "real world" testing environment one can assume that only a "limited number of components" fail simultaneously. In our two component example we may therefore assume that either R or C has failed with the other remaining nominal, in which case the circuit need only be simulated at a discrete set of points along the coordinate axes in the R-C plane denoted by x's in figure 1. As such, the number of simulations required is significantly decreased. Indeed, this is one of the major advantages of the simulation-before-test concept as compared to simulation-after-test algorithms which typically fail to exploit a "limited number of failures" assumption.

While the above described approach has been used with considerable success in digital system testing, wherein, the axes are binary and no tolerance problems are encountered, it is not well suited to the analog test problem. First, an analog failure may occur anywhere along the axis and hence some type of approximation scheme is required to interpolate between the discrete simulations. Secondly, a "good" component is assumed to be in-tolerance, though it may not be nominal. As such, in an analog environment the "limited number of failures" assumption implies that the solution to our fault diagnosis problem lies near, but not necessarily on, the coordinate axes as indicated by the shaded regions in figure 2a.

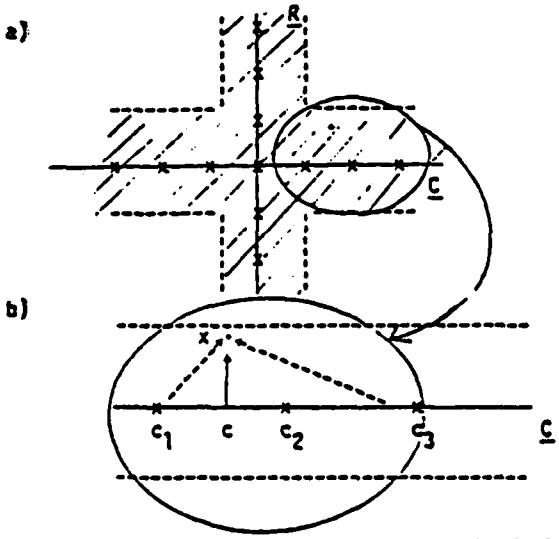


Figure 2: a) Solution space under a single failure assumption.
b) Illustration of the differential-interpolative diagnosis algorithm.

While we might choose to simply fill the shaded region with additional simulations, the cost of such a process may prove to be excessive. Rather, we exploit the fact that the deviations of good component parameters from nominal are small and use a 1st order Taylor series approximation to approximate the deviation. We note that such an approach cannot be used to locate the faulty parameter values which may be far from nominal; indeed, it is often infinite or zero; though it can be used to cope with the tolerance effects.

Our differential-interpolative approach thus uses a classical minimum distance algorithm to locate the general region of the faulty parameter values indicated by the circle in figure 2a (which is magnified in figure 2b). Now, it is assumed that the simulated values of the system responses; f_1, f_2, and f_3; corresponding to the points; c_1, c_2, and c_3; are available along with the associated inverse sensitivity matrices; J_1^{-1}, J_2^{-1}, and J_3^{-1}. We then interpolate these data points to approximate the system responses and the associated inverse sensitivity matrice along the axis by functions $f(c)$ and $J(c)^{-1}$. Although any interpolation can be employed we have had our best results using a bilinear interpolation for f (which gives exact results in the linear case) and a second order polynomic interpolation of J^{-1}. Now, if x denotes the faulty parameter vector and m denotes the measured system responses then a 1st order Taylor series approximation combined with our interpolation will yield the (approximate)equality

$$m = f(c) + J(c)[x - c] \qquad 1.$$

for those values of c near x. Equivalently,

$$[x - c] = J(c)^{-1}[m - f(c)] \qquad 2.$$

Interestingly, by invoking the Projection theorem one can reduce the above vector equation to a scalar equation and simultaneously eliminate the requirement for storing the inverse sensitivity matrices. Indeed, the vector [x-c] will be perpendicular to the axis at the point c which makes the closest approach to the fault. As such, if e_c denotes the unit vector in the direction of the axis then

$$0 = e_c^t[x - c] = e_c^t J(c)^{-1}[m - f(c)] \qquad 3.$$

which can be solved for the faulty parameter value, c. Note, our goal is to solve for c, not x, since we are interested in locating the faulty parameter value in the presence of the tolerance problem, but we really do not care to compute the deviations from nominal in the good parameters.

To summarize, if rather than simply storing the simulated circuit responses, f_i, we also store the vectors $e_c^t J_i^{-1}$ then the tolerance effects associated with the good components can be completely removed from our fault diagnosis algorithm at least up to the approximation error induced by the interpolation process and Taylor series expansion. Since most good circuit simulation codes include a package for generating sensitivity matrices at little additional cost over and above that involved in simply simulating the circuit responses, the approach can be implemented with only a minimal increase in simulation costs. As such, the major expense associated with the approach lies with the storage requirements (for

the f_i and $e_c^t J_i^{-1}$ vectors) which are approximately double that of a classical fault simulation algorithm.

Although the above derivation has been illustrated in the two dimensional case with a single faulty parameter it can be readily extended to a general setting, say with several hundred components and three or four simultaneous faults. If one assumes p simultaneous faults then p inner products are required to apply the Projection theorem yielding p equations and p unknowns to be solved for the faulty parameter values. Otherwise the formulation for the general case is identical to the single fault case described above.

III. EXAMPLES

In this section, two examples are given, one of them for linear systems and one for the nonlinear case. All of these examples were simulated on an HP9825A programmable calculator, and yielded fairly good results.

Our first example is a second order low pass filter. The filter contains five components, K, R_1, R_2, C_1, and C_2, while, the circuit diagram is shown in figure 3-1

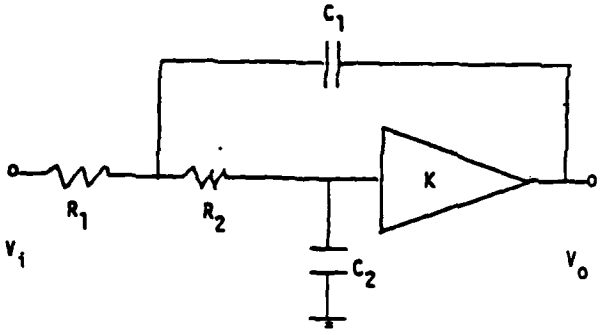


Figure 3-1.

The transfer function for this circuit is given by

$$f(\bar{r},s) = \frac{K}{S^2C_1C_2R_1R_2 + s[R_2C_2 + R_1C_2 + R_1C_1(1-K)]+1} \quad (3\text{-}1)$$

The partial derivatives of the transfer function with respect to each parameter take the form

$$\frac{\delta f}{\delta f}(\bar{r},s) = \frac{D + SC_1R_1K}{D^2} \quad (3\text{-}2)$$

$$\frac{\delta f}{\delta R_1}(\bar{r},s) = \frac{-K[S^2C_1C_2R_2 + SC_2 + SC_1(1-K)]}{D^2} \quad (3\text{-}3)$$

$$\frac{\delta f}{\delta R_2}(\bar{r},s) = \frac{-K[S\ C_1C_2R_1 + SC + SC_2]}{D^2} \quad (3\text{-}4)$$

$$\frac{\delta f}{\delta C_1}(\bar{r},s) = \frac{-K[S^2C_1R_2R_1 + SR_1(1-K)]}{D_2} \quad (3\text{-}5)$$

$$\frac{\delta f}{\delta C_2}(\bar{r},s) = \frac{-K[S^2C_1R_1R_2 + SR_2 + SR_1]}{D^2} \quad (3\text{-}6)$$

where $D = S^2C_1C_2R_1R_2 + S[R_2C_2 + R_1C_2 + R_1C_1(1-K)] + 1$

Since we have five parameters in the transfer function, five distinct test frequencies are required to provide sufficient information for diagnosis.

The fault diagnosis results are listed in table 3-1. Here, the nominal values of K, R_1, R_2, C_1, C_2 are 1.6, 1KΩ, 1KΩ, 0.16 F and 0.16μF respectively and the faulty parameter is underlined in the table

Table 3-1

	1	2	3	4	5	6
K	0.6	1.62	1.62	1.58	1.62	4
R_1	1090	2500	1090	1070	1050	1050
R_2	930	1040	10	930	930	1045
C_1	0.163u	0.161u	0.157u	0.25u	0.157u	0.162u
C_2	0.162u	0.162u	0.162u	0.157u	0.25u	0.162u
Result	K 0.591	R_1 2492	R_2 19.4	C_1 0.239u	C_2 0.238u	K 4.01

In the first simulation, K is the faulty component with a value of 0.6, while the other four components are 5% or so off their nominal values, the simulation result shows that K failed, and locates it at K=0.591. The same remarks apply to the other five simulations.

Although the technique generally yields satisfactory results occasional errors occur when the good components are too far out of tolerance. For instance, the following parameter values K=1.62, R_1=1070, R_2=910, C_1=0.5 , and C_2=0.172u led to an erroneous result. The simulation shows that C_2 has failed with the value of 0.179u. However, the faulty component, in this simulation, is actually C_1. If we sketch a two dimensional representation of the C_1, C_2 plane the difficulty becomes clear. Figure 3-2 shows that C_2 is too far away from its own nominal value, and thus instead of locating the error at α as we expect the simulation resut locates the failure at β, with the differential term still pointing toward the actual failure denoted by x.

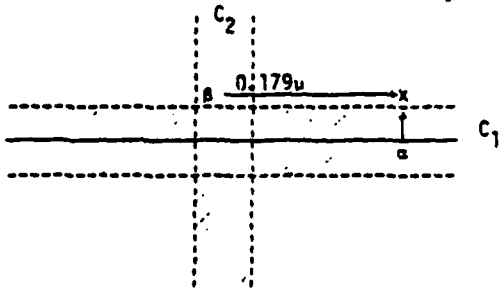


Figure 3-2

Our nonlinear example is composed of a diode loaded by a shunt RC circuit as illustrated in Figure 3-3. The diode is modeled by the characteristic function

$$I = I_o(e^{V_o/V_t} - 1) \qquad (3\text{-}7)$$

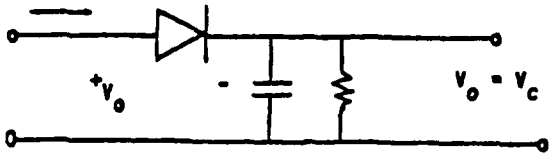


Figure 3-3

Now, instead of working with frequency domain transfer functions, we work in the time domain. A state equation for this circuit is given by

$$V_c = \frac{1}{C} I_o(e^{(V_i - V_c)/V_t} - 1) - \frac{V_c}{RC} \cdot \qquad (3\text{-}8)$$

The goal is to integrate this differential equation so as to build a $V_c(\bar{r})$ vector and $f(\bar{r})$ vectors as in the previous examples.

Numerical techniques can be used to compute $V_c(t)$ at any instant t. In this example, the $V_c(\bar{r})$ vector was elevated by applying the fourth Runge-Kutta method. Note that since there are four independent parameters, R, C, I_o, and V_T; in equation (3-8) $V_v(\bar{r}, t)$ should be evaluated at four different time instants to build a $V_c(\bar{r})$ vector.

The simulation results are summerized in table 3-2. The nominal values of I_o, V_T, R, C are 0.2, 0.1, 1K, and 0.25 respectively.

Table 3-2

	1	2	3	4	5
I_o	0.035	0.21	0.2	0.22	0.002
V_T	0.11	5.3	0.11	0.11	0.09
R	920	930	6250	1070	1090
C	0.23	0.27	0.255	2.6	0.23
Result	I_o 0.0312	C 5.449	R 6001	C 2.578	I_o 0.003

IV. CONCLUSIONS

For the simulation-before-test-approach to fault diagnosis, we gain from the fact that most computation can be done by off-line computation, thus greatly reducing the repetitive on-line computation associated with many fault diagnosis algorithms. From a practical point of view, the economics of such an approach are extremely attractive. Unfortunately, the simulation-before-test approach is subject to a certain degree of ambiguity introduced by good components which are in-tolerance but not nominal.

In this paper, we have proposed a simulation-before-test algorithm for analog fault diagnosis, in which a differential-interpolative technique is used to eliminate the ambiguity caused by tolerance effects. Our approach has been tested with satisfactory results in both the linear and nonlinear cases. In fact, for the linear case, the approach provides an exact interpolation for f(c) on the axes, and thus reduces the amount of simulation-before-test data required on each axis. Although this is not true for nonlinear case, the diagnosis results are still very attractive. Of course, occasional errors may occur when the good components are too far out of tolerance. This phenomena is, however, expected and well understood. Indeed, the difficulty occurs only when the 1st order Taylor series approximation is too good. Because this phenomena will rarely occur in the real world, we believe that it may be neglected in a practical algorithm.

ABSTRACT OF

A DATA BASE FOR SYMBOLIC NETWORK ANALYSIS

C.-C HU AND R. SAEKS

183

Abstract

Historically, symbolic network analysis has been motivated by the problems of circuit design and, as such, the emphasis has been placed on quickly and efficiently obtaining a symbolic transfer function from a given set of circuit specifications. In an operational or maintenance environment, however, one is typically given a prescribed nominal circuit and desires determine the effect of various (possibly large) perturbations thereon. This is the case in a power system where one is given a fixed network and desires to determine the effect of proposed modifications thereto. Alternatively, in the problem of analog circuit fault diagnosis one desires to simulate the effect of a number of alternative failures to compare the simulated data with the observed failure data.

In such an operational or maintenance environment numerous perturbations of the nominal circuit are studied and, as such, significant computational efficiencies can be obtained if one first generates a data base in terms of the nominal circuit parameters and then extracts the appropriate symbolic transfer function from the data base each time a different symbolic transfer is required. Of course the benefit to be achieved via such an approach is dependent on the size of the data base and the ease with which a symbolic transfer function may be retrieved therefrom.

The obvious manner in which to generate such a data base is to simply pre-compute the coefficients of all required symbolic transfer functions

PRECEDING PAGE BLANK-NOT FILMED

and store them in the data base. Retreival from such a data base is, of course, immediate but the data base may become overly large. Indeed, the number of transfer functions which must be stored is $O(\underline{k}^{\underline{p}})$ where $\underline{k}$ is the total number of potentially variable circuit parameters and $\underline{p}$ is the maximum number of circuit parameters which may vary simultaneously. An alternative approach is to store the nominal transfer function information and then use Householder's formula to compute the required symbolic transfer functions. In such a data base we need only store $O(\underline{n}^2)$ transfer functions where $\underline{n}$ is the total number of component output terminals but retreival requires $O(\underline{n}^3+p^3)$ multiplications where p is the actual number of circuit parameters which vary simultaneously. Since, in practice, $\underline{n} >> p$ the retreival process requires approximately $O(\underline{n}^3)$ multiplications and is dominated by the large dimensional matrix multiplication required by Householder's formula rather than the low dimensional inverse.

In the present paper we will formulate an alternative data base for the symbolic transfer functions which also requires $O(\underline{n}^2)$ entries, but for which retreival requires only $O(p^3)$ multiplications. Since p is typically small this is tantamount to immediate retreival.

ABSTRACT OF

DIAGNOSABILITY OF NONLINEAR CIRCUITS AND SYSTEMS

RICHARD SAEKS,
ALBERTO SANGIOVANNI-VINCENTELLI
AND V. VISVANATHAN

Abstract

A theory for the diagnosability of nonlinear dynamical systems is developed. It is based on an input-output model of the system in a Hilbert space setting. A necessary and sufficient condition for the local diagnosability of the system, which is a rank test on a matrix, is derived. A simple sufficient condition is also derived. It is shown that, for locally diagnosable systems, there exist a finite number of test inputs that are sufficient to diagnose the system. Illustrative examples are presented.

PRECEDING PAGE BLANK-NOT FILMED

ABSTRACT OF

ANALOG FAULT DIAGNOSIS WITH FAILURE BOUNDS

C.-C. WU, K. NAKAJIMA,
C.-L. WEY AND R. SAEKS

PRECEDING PAGE BLANK-NOT FI

Abstract

A simulation-after-test algorithm for the analog fault diagnosis problem is proposed in which a bound on the maximum number of simultaneous failures is used to minimize the number of test points required. The resultant algorithm is applicable to both linear and nonlinear systems with multiple hard or soft faults and can be used to isolate failure up to an arbitrarily specified "replaceable chip or subsystem".

PRECEDING PAGE BLANK-NOT FILMED

MULTIDIMENSIONAL SYSTEM THEORY

J. MURRAY

PRECEDING PAGE BLANK-NOT FILMED

Texas Tech University | Institute for Electronic Science

Joint Services Electronics Program | Research Unit: 4

1. Title of Investigation: Multidimensional System Theory
2. Senior Investigator: J. Murray Telephone: (806) 742-3506
3. JSEP Funds: Current $25,875
4. Other Funds: Current
5. Total Number of Professionals: PI's 1 (1 mo.) RA's 1 (1/2 time
6. Summary:

Although most research in the image processing area is motivated by the computational problems associated with the actual image processing algorithms, progress in the area has also been limited by the cost of designing an efficient 2-D signal processing algorithms. Indeed, in many image processing applications simple non-recursive algorithms are used in lieu of far superior recursive algorithms because of the prohibitive design costs associated with the recursive algorithms. As such, this work unit is directed at the problem of developing efficient design techniques for stable 2-D digital signal processors. In this endeavor we have developed, and reported upon, a 2-dimensional design algorithm based on a spatially-invariant symmetric half-plane recursive model and are in the process of developing an algorithm which uses a 2-D frequency domain model for a periodically varying system originally introduced by Jury and Mullin. The latter model completely eliminates the analytical difficulties classically associated with 2-dimensional design but this is achieved at the price of working with high dimensional matrices. As such, the computational cost of the design process based thereon is prohibitive. We, however, believe that the structure of the matrices which arise in this

PRECEDING PAGE BLANK-NOT FILMED

model can be exploited to formulate a class of efficient design algorithms and are presently investigating this possibility.

7. Publications and Activities:

A. Refereed Journal Articles

1. Murray, J., "Lumped-Distributed Networks and Differential Delay Systems," in Algebraic and Geometric Methods in Linear System Theory, Providence, AMS, 1980.

2. Murray, J., "A Design Method for 2-Dimensional Recursive Digital Filters," IEEE Trans. on Acoustics, Speech, and Signal Processing, (to appear), February 1982.

B. Conference Papers and Abstracts

1. Murray, J., "A Time-Varying Approach to Two-Dimensional Digital Filtering," Proc. of the 24th Midwest Symp. on Circuits and Systems, Albuquerque, July 1981, pp. 351-355.

2. Murray, J., "The Design of 2-D Filters as 1-D Time-Varying Systems," to be presented at the 1982 IEEE International Conference on Acoustics, Speech and Signal Processing.

C. Theses: Chen, S-H, Ph.D. Dissertation (in preparation).

D. Conferences and Symposia:

1. Murray, J., 1981 International C onference on Acoustics, Speech and Signal Processing, Atlanta, GA., March 1981.

2. Murray, J., 24th Midwest Symposium on Circuits and Systems, Albuquerque, NM, June 1981.

3. Murray, J., 2nd ASSP Workshop on Two-Dimensional Signal Processing, New Paltz, NY, Oct. 1981.

E. Lectures:

"Stability Problems in Filter Design," 2nd ASSP Workshop on Two-Dimensional Signal Processing, New Paltz, Oct 1981.

A TIME-VARYING APPROACH TO TWO-DIMENSIONAL DIGITAL FILTERING

JOHN MURRAY

PROCEEDINGS OF THE 24TH MIDWEST SYMPOSIUM ON CIRCUITS AND SYSTEMS

ALBUQUERQUE, NM, PP. 351-355, JULY 1981

199

"A TIME-VARYING APPROACH TO TWO-DIMENSIONAL DIGITAL FILTERING" *

John Murray
Department of Electrical Engineering
Texas Tech University
Lubbock, TX 79409 USA

Abstract

A new approach to two-dimensional digital filtering is presented. This approach is based on a one-dimensional periodically time-varying model which accurately reflects the scanning process inherent in most recursive multidimensional signal processing. Time-varying models are in general intractable; however, periodically time-varying discrete-time models such as occur in the present case are essentially equivalent to multi-input, multi-output, one-dimensional time-invariant systems. They therefore permit the application of classical techniques to design and analysis problems. Two further advantages of the approach are the fact that it bypasses the problem of boundary conditions, and that allowing time-variation gives a degree of design flexibility not available in the shift-invariant case. Some possible design methods using these time-varying ideas are presented.

1. INTRODUCTION

The general field of two-dimensional data processing has been the subject of extensive investigation during the past several years. The simplest situation, and that which has received most attention in theoretical work, is the shift-invariant case, where the processing operations are assumed to commute with translations in both directions. Much of the effort in this direction has been devoted to the analysis and design of two-dimensional digital filters. In studying this work, one quickly becomes aware of a basic dichotomy between finite impulse response (FIR) filters, which are usually implemented nonrecursively, and infinite impulse response (IIR) filters, which are usually implemented recursively. The theory and design of FIR filtering are well understood and, although there are problems in some areas (such as uniform approximation) and difficulties with the data-rates and amount of computation required (which are inevitable in two-dimensional work) the use of FIR filters is almost routine by now in two-dimensional data processing. The design and use of IIR filters, on the other hand, very often presents serious difficulties. In what follows, we will point out some theoretical problems associated with shift-invariant IIR filtering, and show how the theory of time-varying systems offers a (theoretical) resolution of these problems.

2. IIR FILTERING AND CAUSALITY

2.1. Two-Dimensional Causality

In one-dimensional data processing, the dimension in question is usually time, and

* This research supported in part by the Joint Services Electronic Program of Texas Tech University under ONR Contract 75-C-1136.

PRECEDING PAGE BLANK-NOT FILMED

so there is a natural concept of causality. In two dimensions, the situation is more complicated. In many cases, the dimensions are both spatial, and so there is no intrinsic causality. Even in the case where one dimension is time and the other is space (e.g., transducer arrays) one will often have the data already recorded, and so causality will not be significant. However, in this latter case, the appropriate causality is a two-dimensional "symmetric half-plane" causality - but it should be emphasized that this causality is not intrinsic to the data; it is, rather, a property of the processing used. At this point, a distinction must be made between FIR and IIR filters.

2.2. One-Dimensional Causality.

As was mentioned in the previous paragraph, there may or may not be a notion of causality inherent in two-dimensional data. However the data must be processed in time, and this introduces a concept of causality which (at least with classical processing) is one-dimensional. How compatible this is with two-dimensional causality in the data will depend on the specific problem. In the context of nonrecursive processing, the order in which the data are processed is irrelevant, and so no real conflict can arise; but the very nature of recursive processing, in which the output at a point depends on the output at previous points, demands that an order be specified. This order is almost always taken to be a "scanning" order, in which a line is processed from left to right, followed by the next line below it, etc.. We will assume that this is the order in which processing occurs for the remainder of this paper.

There is, however, an inconsistency between this ordering and the assumption of shift-invariance. While not serious in practice, it does indicate that the theory may run into difficulties. This inconsistency arises from the fact that the above ordering requires that the horizontal lines be finite, while the use of shift-invariance requires in principle that they be infinite in at least one direction. A first approximation to resolving this difficulty is given by the McClellan transform,[1] which essentially concatenates the lines into one long line, and applies one-dimensional linear shift-invariant theory to the resulting signal; this however has the disadvantage in principle of treating the edges of the image in the same way as points in the interior. (In practice, of course, one applies suitable boundary conditions). It is clear that the model which most accurately reflects the actual processing being done in this situation is a one-dimensional periodically time-varying model.

3. PERIODICALLY VARYING DISCRETE-TIME SYSTEMS

Time-varying systems in general are extremely difficult to work with; it is therefore fortunate that one of the few classes of time-varying systems for which a complete closed-form theory exists is the class of periodically varying discrete-time systems.[2,3,4] There are several versions of this theory, but they all describe scalar-input scalar-output systems whose period of variation is n sampling periods by nxn time-invariant transfer matrices. We proceed to outline one version.

Input and output sequences are represented by the transform:

$$(u_0, u_1, \ldots\ ) \longmapsto U(Z)$$

where

$$U(Z) = \begin{bmatrix} u_0 + u_n Z^n + u_{2n} Z^{2n} \ldots \\ u_1 Z + u_{n+1} Z^{n+1} + \ldots \\ u_2 Z^2 + u_{n+2} Z^{n+2} + \ldots \\ \vdots \\ u_{n-1} Z^{n-1} + u_{2n-1} Z^{2n-1} + \ldots \end{bmatrix}$$

so that the transform of a scalar signal is an n-dimensional column vector. The input-output relationship of a scalar system of the type under consideration is then given by

$$Y(Z) = P(Z)U(Z)$$

where P(Z) is an nxn matrix of rational functions of the form

$$p_{ij}(Z) = Z^k q_{ij}(Z^n)$$

where

$$k = \begin{cases} i-j & i \geq j \\ n+i-j & i < j \end{cases}$$

For example, a memoryless time-varying gain a(m) is given by the diagonal matrix

$$\begin{bmatrix} a(o) & & & \\ & a(1) & & 0 \\ & & \ddots & \\ & 0 & & a(n) \end{bmatrix}$$

while the unit delay is given by the matrix

$$\begin{bmatrix} 0 & 0 \ldots\ldots & Z \\ Z & 0. & . \\ 0 & Z \ddots & . \\ \vdots & \vdots \ddots & \vdots \\ 0 & 0 \ldots\ldots & Z \; 0 \end{bmatrix}.$$

(It follows that in this representation, time-invariant systems are described by circulant matrices).

It can be shown that by using the above transformation, systems which vary with a period equal to n sampling periods can be treated precisely as if they were n-input n-output time-invariant systems; in particular, such a system is stable if and only if the matrix P(Z) is invertible for all Z with $|Z| \leq 1$.

APPLICATION TO TWO-DIMENSIONAL PROCESSING

In applying the theory outlined in the previous section to two-dimensional data processing, we immediately encounter an immense difficulty; namely, n must be taken to be the number of points in a line of data, and so is usually very large. This implies that nxn matrices of rational functions will be totally intractable. Before discussing this problem, however, we will outline how the two-dimensional situation relates to the periodically time-varying approach. From now on, n will denote the number of points in a horizontal line.

As an example, we consider a mask of the form

$$\begin{matrix} a_{11} & a_{10} & a_{1,-1} \\ a_{01} & a_{00} & \end{matrix}$$

where the (0,0) subscript indicates the point at which processing is occurring. Specifically, if the above were the input mask of a purely nonrecursive filter, the output would be the convolution of the above array with the input array.

The two-dimensional Z-transform of the above array is given by

$$p(Z_1,Z_2) = a_{00}+a_{01}Z_2 \; (a_{11} Z_2+a_{10}+a_{1,-1}Z_2^{-1})Z_1.$$

If one uses the McClellan transform, that is, concatenates the rows and regards the filter as a one-dimensional shift-invariant filter, the resulting one-dimensional Z-transform is

$$p(Z) = a_{00}+a_{01}Z + (a_{11}Z+a_{10}+a_{1,-1}Z^{-1})Z^n. \quad (1)$$

To put this in clearer perspective, we will write it in terms of our time-varying model: the resulting transfer operator is

$$P(Z)=\begin{bmatrix} a_{00}+a_{10}Z^n & ,a_{1,-1}Z^{n-1}, \ldots 0, & a_{01}Z+a_{11}Z^{n+1} \\ a_{01}Z+a_{11}Z^{n+1}, & \ddots & 0 \\ 0 & & \\ \vdots & & \\ a_{1,-1}Z^{n-1}, & & a_{00}+a_{10}Z^n \end{bmatrix} \quad (2)$$

-a circulant matrix, as expected.

In practice, of course one does not implement this as a strictly shift-invariant one-dimensional filter; one puts in appropriate boundary conditions when a boundary is crossed. It is easy to see that, if zero boundary conditions are assumed, the transfer operator which describes the actual processing performed is

$$P(Z) = \begin{bmatrix} a_{00}+a_{10}Z^n & , a_{1,-1}Z^n & ,\dots & 0, & 0 \\ a_{01}Z+a_{11}Z^{n+1}, & & & & \\ 0 & & & & \\ \vdots & & & & \\ 0 & & & & a_{00}+a_{10}Z^n \end{bmatrix} \qquad (3)$$

-a Toeplitz, rather than a circulant, matrix, since the processing is now (slightly) time-varying.

While this may seem like a trivial modification, it can in some situations have serious theoretical effects. For instance if one specializes to a quarter-plane filter by setting $a_{1,-1} = 0$, the time-invariant stability condition is (from (1))

$$a_{00}+a_{01}Z+a_{10}Z^n+a_{11}Z^{n+1} \neq 0 \quad \text{for } |Z| \leq 1. \qquad (4)$$

However, if one uses the time-varying version (3) in this situation, the resulting matrix is a lower triangular Toeplitz matrix whose diagonal element is

$$a_{00}+a_{10}Z^n$$

and so the stability condition in this case is

$$a_{00}a_{10}Z^n \neq 0 \quad \text{for } |Z| \leq 1 \qquad (5)$$

which is substantially different from (4).

(The problem here is that if (4) is not satisfied and (5) is, the matrix (3) will be invertible for all $|Z| \leq 1$, but the inverse matrix may contain extermely large elements, and so while the system is stable in principle, it will be unstable in practice.)

5. POSSIBLE DESIGN APPROACHES

As was pointed out previously, a major problem blocking any realistic application of the above approach is the size of the matrices involved. This problem is not completely hopeless, however, for the following reasons. Firstly, the matrices which occur are banded with bandwidth equal to the horizontal width of the filter array. Thus for reasonably-sized filters, the matrices are both sparse and structured. Secondly, one does not normally want to design filters which are wildly shift-varying; it is usually desireable to have a filter which is approximately shift-invariant far from the boundaries. In this connection the change from circulant to Toeplitz which occurred in the previous section (as a result of inserting boundary conditions) comes to mind, and suggests that "displacement rank" ideas [5] may have some relevance here.

Another possibility is to replace the shift in the horizontal direction with some other "basic dymanical operator" and to design in terms of this operator.

The most obvious choice is a descrete approximation to the derivative, which has the following advantages:

1) It is local, and so can be calculated efficiently.
2) At the boundaries, it can be calculated as a one-sided derivative, and so avoids the problem of boundary values.

The simplest operator which is a discrete approximation to the derivative and enjoys both of these advantages is the matrix

$$T = \begin{bmatrix} -1, & 1, & 0, & 0 \dots\dots & 0 \\ \frac{a_2-1}{2}, & -a_2, & \frac{a_2+1}{2}, & 0 \dots\dots & 0 \\ 0, & \frac{a_3-1}{2}, & -a_3, & \frac{a_3+1}{2}, & 0 \\ \vdots & & & & \vdots \\ 0 & & & -1, & 1 \end{bmatrix} \qquad (6)$$

where $\alpha_2, \alpha_3, \ldots$ are arbitrary but fixed.

The algebra generated by T, the constants, and the vertical shift is then commutative, and so offers the possibility of a tractable theory. The major remaining question concerns the choice of the constants α_2, $\alpha_3 \ldots$ While this is currently under investigation, and no definitive results have been established, the following points may be noted.

1) If one designs a filter in terms of the derivative in the horizontal direction and the shift in the vertical direction one can then replace the derivative by the discretized version T.

2) Since the spectrum of the derivative is the entire imaginary axis, while the spectrum of T can be adjusted by varying the α_k, this approach yields the possibility of choosing the approximation so that the discretized filter is stable even when the original designed filter is unstable, for example by concentrating the spectrum of T on a few points. To be specific, if the denominator of the original filter is $\sum_{j=0}^{n} f_j(D)Z^j$, where D is the derivative in the horizontal direction, and Z the shift in the vertical direction, the stability condition for the discretized filter will be

$$\sum_{j=0}^{n} f_j(\lambda_i)Z^j \neq 0 \ , \ \forall |Z| \leq 1 \qquad (7)$$

for all eigenvalues λ_i of T. Thus the smaller the set of eigenvalues, the less stringent the stability conditions.

3) For practical stability, if the eigenvalue λ_i is associated with a large Jordan block, the theoretical condition (7) must be replaced by

$$\sum_{j=0}^{n} f_j(w)Z^j \neq 0, \ \forall |Z| \leq 1,$$

for all w such that $|w - \lambda_i| \leq 1$. Thus using large Jordan blocks can actually make the stability conditions more, rather than less, stringent. We note in passing that T always contains a two-dimensional Jordan block with zero eigenvalue.

4) The previous two paragraphs taken together imply that the minimal polynomial should be of low order. This has the further benefit of ensuring that the support of the filter in the horizontal direction is of limited width; achieving this is a major problem in the shift-invariant design approach.

6. CONCLUSIONS

We have pointed out some theoretical inconsistencies between the assumption of two-dimensional shift-invariance and the usual scanning model employed in the processing of two-dimensional data. As a remedy for this inconsistency, a one-dimensional periodically time-varying model has been proposed; this also has the advantage of not requiring boundary conditions. Finally, two possible design approaches have been mentioned.

REFERENCES

1. J.H. McClellan, "The Design of 2-D Digital Filters by Transformation", in Proc. 7th Annual Princeton Conf. on Information Sciences and Systems, 1973, pp. 247-251.
2. E.I. Jury and F.J. Mullin, "The Analysis of Sampled-Data Control Systems, with a Periodically Time-Varying Sampling Rate", IRE Trans. Auto. Control, AC-4(1959) 15-21.
3. J.H. Davis, "Stability Conditions Derived from Spectral Theory: Discrete Systems with Periodic Feedback", SIAM J. Cont. 10 (1972) 1-13.
4. R.A. Meyer and C.S. Burrus, "A Unified Analysis of Multirate and Periodically Time-Varying Digital Filters", IEEE Trans. Circuits and Systems, CAS 22 (3) (1975), 162-168.
5. T. Kailath, A. Viera and M. Morf, "Inverses of Toeplitz Operators, Innovations, and Orthogonal Polynomials", SIAM Rev. 20 (1978) 106-119.

LUMPED-DISTRIBUTED NETWORKS AND DIFFERENTIAL-DELAY SYSTEMS

JOHN MURRAY

ALGEBRAIC AND GEOMETRIC METHODS IN LINEAR SYSTEM THEORY
PROVIDENCE, AMS, 1980

PRECEDING PAGE BLANK-NOT FI

Lectures in Applied Mathematics
Volume 18, 1980

LUMPED-DISTRIBUTED NETWORKS AND DIFFERENTIAL-DELAY SYSTEMS

J. Murray

1. INTRODUCTION. In this paper we will consider some known properties of differential-delay systems and their relationship to the lumped-distributed networks studied by classical circuit theorists. The two theories are fundamentally the same, but the emphasis is different; in particular, the first question asked in the system-theoretic approach tends to be about stability, while the circuit-theorists' primary interests have had to do with passivity. A further major difference is that the system-theorists are concerned with the (infinite-dimensional) state spaces associated with these systems, while the circuit-theorists tend to ignore the state space, and concentrate entirely on input-output properties.

Actually, the similarities and differences between these two fields (and others) have been treated recently (and excellently) by Kamen [1]. The present paper gives a different viewpoint, however, being an analytic approach in contrast to the algebraic approach in [1]. Further, it considers only input-output properties of systems, and may be considered as a study of the simplest case of the convolution algebra approach in [2,3]. It is hoped that the following discussion will give some intuition for the last-mentioned approach, and in particular for the relationship between it and the classical circuit-theorists' use of several complex variables to model lumped-distributed networks.

2. DIFFERENTIAL-DELAY SYSTEMS: ALGEBRAIC ASPECTS. The subject matter of this section has been extensively treated in many

© American Mathematical Society 1980

PRECEDING PAGE BLANK-NOT FILMED

places; we give a quick summary simply to fix ideas. We will work with the simplest possible case of differential-delay systems, namely, the case where all the delays are integral multiples of one fundamental delay; time-units are assumed to be normalized so that this fundamental delay is of length one. The input-output mapping for such a system (assuming BIBO stability) is given by convolution with an expression of the form

$$H(t) = F(t) + \sum_{i=-\infty}^{\infty} K_i \, \delta(t-n) \tag{1}$$

where $F(t)$ is a matrix-valued function in $L_1(-\infty,\infty)$ and the K_i are matrices with $\sum_{i=-\infty}^{\infty} \|K_i\| < \infty$; this is the simplest case of the algebras studied by Callier and Desoer [2,3].

For conceptual purposes, we have included non-causal systems in the above; in the real-world case of causal systems, we have

$$F(t) = 0, \quad t < 0$$

and

$$K_i = 0, \quad i < 0 .$$

For a system composed of a finite number of differentiators and integer delays, the Fourier transform of a transfer operator of the above type is well known to be of the form

$$R(s,e^{-s}) \qquad (s = iw)$$

where R is a rational matrix function of two complex variables. Since the functions s and e^{-s} are algebraically independent, R is unique. Also, since the transform of a composition (convolution) is the product of the transforms, it follows that an algebra of input-output operators of the type (1) arising from finite systems is isomorphic to a subalgebra of the field of rational functions in two variables (over $\mathbb{R}$ or $\mathbb{C}$, as appropriate). The realization problem for differential-delay systems consists of identifying this subalgebra, and has been treated in several places (e.g. [4]). Since it is not our purpose here to treat either the algebraic aspects of these systems or the

problems of their realization, we merely repeat that the algebraic treatment rests on the fact that there is an algebra isomorphism between the appropriate set of input-output operators and an algebra of rational functions in two variables. Thus one may say that from the algebraic point of view, systems of the type (1) can be treated as two-dimensional.

3. DIFFERENTIAL-DELAY SYSTEMS: ANALYTIC ASPECTS. Much of the power of transform methods in electrical engineering arises not simply from the algebraic isomorphism between a convolution algebra and a function algebra which these transforms define, but from the more "analytic" properties of the isomorphism; e.g., the relationship between pole-location and stability, or between passivity and the positive-real property. Both of these relationships will be discussed below as they apply to the present class of operators. The most obvious analytic property of these operators is that they have a norm defined by

$$\|H\| = \int_{-\infty}^{\infty} |F(t)|dt + \sum_{i=-\infty}^{\infty} |K_i| \quad . \tag{2}$$

For convenience, we will restrict ourselves to scalar-input, scalar-output systems from here on; this case contains all the essential features which we wish to discuss. With this assumption, it is easy to check that the operators of the form (1), with the norm (2), form a commutative Banach algebra which we will denote by B. It is therefore natural to try to compute the Gelfand spectrum of this algebra, and see if it can contribute to the understanding of these operators. This can be done in various ways, but one of the most natural is to begin with the spectrum of $L_1(-\infty,\infty)$, and investigate its behaviour under the transformation discussed in the previous section, which changed the original one-dimensional problem into a two-dimensional problem. This transformation is given by the mapping

$$\tilde{f}: \mathbb{C} \to \mathbb{C}^2$$

defined by

$$\tilde{f}(s) = (s, e^{-s}) \quad .$$

In order to avoid working with points at infinity, it is convenient to take a bilinear transform of the first coordinate, and work instead with the mapping

$$f: \mathbb{C} \to \mathbb{C}^2$$

defined by

$$f(s) = \left(\frac{1-s}{1+s}, e^{-s}\right) .$$

This has the advantage that f maps the right half-plane into the unit bidisk, U^2, and maps the imaginary axis (which is the spectrum of $L_1(-\infty,\infty)$) into the distinguished boundary T^2, of U^2. (We will use the notation

$$U = \{Z \in \mathbb{C} \mid |Z| < 1\}$$

$$\bar{U} = \{Z \in \mathbb{C} \mid |Z| \leq 1\}$$

$$T = \{Z \in \mathbb{C} \mid |Z| = 1\}$$

and

$$U^2 = U \times U, \text{ etc. . .)}$$

Since B has an identity, its spectrum is compact, and it is natural to conjecture that this spectrum, $\sigma(B)$, is the closure of the image of the imaginary axis. This is in fact the case [5]. The image is defined by

$$\left(\frac{1-iw}{1+iw}, e^{-iw}\right) \qquad w \in \mathbb{R}$$

Representing the torus as a square with its opposite edges identified in the usual way, we can draw an approximation to $\sigma(B)$ as in Fig. 1. It consists of the circle $\theta_1 = \pi$ together with a line which is asymptotic to this circle.

As mentioned in the introduction, one of the most important questions in system theory is stability. In virtually every situation in which it arises, input-output stability is equivalent to some operator having a bounded, causal inverse. It is for this reason that transforms are useful; invertibility of an

element in a Banach algebra is equivalent to invertibility of its (Gelfand) transform at every point in the spectrum. However, the spectrum in Fig. 1 is that of $\mathcal{B}$ and so nonvanishing on this set implies only the existence of an inverse--not necessarily a causal inverse. There are two equivalent ways of deciding whether or not a causal inverse exists; the first ("Hurwitz") approach is to find the spectrum, S, of the causal subalgebra of $\mathcal{B}$, and check for nonvanishing on this set. (The second ("Nyquist") approach will be discussed in the next section.)

Exactly as in the case of the spectrum of $\mathcal{B}$, one is led to conjecture that S (the causal spectrum) is the closure in $\bar{U}^2$ of the image of the half-plane (the spectrum of $L_1[0,\infty)$) under f. Again, this is the case. While we can no longer draw a picture of this spectrum, we can get a good idea of what kind of object it is. Its intersection with U^2 is the image of a one-(complex) dimensional manifold under a proper holomorphic map, and so is a two-dimensional analytic subset of U^2. The intersection of S with the boundary of U^2 consists of two parts; the spectrum of $\mathcal{B}$ described in the previous section, and the disk

$$\{(1,Z_2) \,\big|\, |Z_2| < 1\} .$$

The upshot of all this is that the spectrum is a very small subset of $\bar{U}^2$. While nonvanishing of a function on a one-dimensional analytic subset of U^2 together with nonvanishing on T^2 can imply nonvanishing on all of $\bar{U}^2$ [6,7], $\sigma(\mathcal{B})$ is much too small a subset of T^2 for any such conclusion to be possible in this case. Thus we are led to the conclusion that stability of a system is equivalent to the nonvanishing of a two-variable rational function $R(Z_1,Z_2)$ on the fairly complicated one-dimensional subset, S, of $\mathbb{C}^2$; or equivalently, to the transcendental function

$$R\left(\frac{1-s}{1+s}\,,\ e^{-s}\right)$$

being bounded away from zero in the half-plane. In either case, from the analytic point of view, one has a strictly

one-dimensional problem. The two-variable approach appears merely as a device which gives one a convenient way of calculating the spectrum of the appropriate convolution algebra.

4. DIFFERENTIAL-DELAY SYSTEMS: A TOPOLOGICAL ASPECT. We digress from the main purpose of the paper in this section in order to discuss the "Nyquist" approach to stability mentioned above. The essence of this approach is that instead of looking for nonvanishing of the transform on a spectrum bigger than $\sigma(B)$ one looks for conditions (in addition to nonvanishing) on the behavior of the transform on $\sigma(B)$ itself. The classic case of this is, of course, the Nyquist criterion itself, where one demands that the transform in question does not vanish on the imaginary axis, and in addition that the image of the imaginary axis under the transform does not encircle 0. In other words, one associates an index with the operator in question and, assuming that the operator is invertible, demands that the index be zero for causal invertibility. (We are assuming here that the original operator is itself a bounded, causal operator). In the case of the Nyquist criterion itself, the index is an integer, but one can not expect this to be true in general. The most one can expect is that the index will take its values in some (possibly partially ordered) group. Since it is known (for fairly general convolution algebras over $\mathbb{R}$) that a causal invertible element a has a causal inverse if and only if a is in the connected component of the identity in the group of invertibles in the algebra (see [8]), the appropriate "Index Group" here is the quotient group: Invertibles/component of the identity. (For an arbitrary Banach algebra, this group is actually known as the abstract index group of the algebra [9]).

This would be of little use were it not for the fact that the structure of this group is known for commutative Banach algebras; for such an algebra, the abstract index group is given by

$$\check{H}^1(\text{spectrum}, \mathbb{Z})$$

the first Čech cohomology group of the spectrum of the algebra.

simpler proof using the cascade loading formula is given in [13], there appears to be little point in making the above discussion rigorous. The important point is that whereas stability imposes restrictions on the behavior of the two-variable transform only on thin subsets of T^2 and U^2, passive synthesis imposes constraints over the entirety of both sets. From any point of view, the passive synthesis problem is two-dimensional.

6. STABILITY FOR VARIABLE DELAYS. It is clear from the previous section that if one considers the stability of a system for *all* lengths of delay time, one will have a two-dimensional problem. Various results can be derived using this approach [14]. As an example we have:

PROPOSITION: *Suppose a system* R *is composed of a finite number of differentiators and delays of equal length* α, *so that its two-variable transfer function is rational:*

$$R(Z_1,Z_2) = \frac{P(Z_1,Z_2)}{Q(Z_1,Z_2)}$$

Assume that R *has no indeterminacies on* T^2. *If the following conditions are satisfied:*

i) *There exists a number* M *such that* R *is stable for all* $\alpha > M$

ii) $Q(1,Z_2) \neq 0$, $|Z_2| < 1$

iii) R *is stable for* $\alpha = 0$,

then R *is stable for all* $\alpha > 0$.

<u>Proof</u>.

$$\bigcup_{\alpha>M} \sigma(B_\alpha) \cup \{(1,e^{i\theta}) | 0 < \theta < \pi\} = T^2$$

so that the hypotheses imply that $Q(Z_1,Z_2)$ has no zeros on T^2. Condition iii) implies that $Q(Z_1,1) \neq 0$ for $|Z_1| < 1$, and this together with condition ii) implies that $Q(Z_1,Z_2) \neq 0$ for $|Z_1| < 1$, $|Z_2| < 1$ [6,7]. It follows immediately that R is stable for all $\alpha > 0$.

Condition iii) is actually unnecessary here; it can be eliminated by a slightly more sophisticated argument.

with integer coefficients. It is intuitively clear (and not difficult to prove) that

$$\check{H}^1(\sigma(\mathcal{B}), \mathbb{Z}) \simeq \mathbb{Z}^2 .$$

Thus in the present situation, stability requires that two distinct integer indices be zero, rather than one as in the classical Nyquist criterion. This is not surprising--intuitively one might expect to have one index for each "independent" kind of delay element. While one can get some feeling for these indices by examining the generators of $\check{H}^1(\sigma(\mathcal{B}), \mathbb{Z})$, it must be admitted that this is by no means the best way of actually calculating the indices for any given element of $\mathcal{B}$. In fact, the problem of calculating these indices for elements of algebras considerably more general than $\mathcal{B}$ has already been solved [10,11] in a much more straight-forward manner than that discussed above.

However, the above discussion does bring out a number of points. The major one is that the stability of lumped-distributed systems is associated with the algebraic topology of the spectrum of the appropriate Banach algebra. Further, the important topological entity is the first Čech cohomology group--and we note that in the case of $\sigma(\mathcal{B})$ this is not the same as, e.g., the first singular cohomology group (in contrast with the classical case). Thirdly, one can tell simply by looking at the spectrum what kinds of conditions are needed for stability--in this case, that two integers vanish. Finally, since $\check{H}^1(\sigma(\mathcal{B}), \mathbb{Z}) \simeq \mathbb{Z}^2$, one can say in some vague sense that, from the topological or index point of view, the stability problem in $\mathcal{B}$ is two-dimensional.

5. LUMPED-DISTRIBUTED CIRCUITS: PASSIVE SYNTHESIS. As was mentioned in the introduction, the treatment of circuits consisting of lumped elements and commensurable transmission lines is formally similar to that of differential-delay systems. There are two major differences however; in the first place, the functions involved are input parameters (impedance, admittance, or scattering) rather than transfer functions (we will again confine

our attention to the scalar case): secondly, the really interesting problem is passive synthesis.

The purpose of this section is to give some feeling for the fact that, in contrast with the stability problem, the problem of passive synthesis is genuinely two-dimensional.

To this end, suppose that we are given a two-variable rational function $R(Z_1,Z_2)$ and that we wish to synthesize a passive one-port whose scattering parameter is

$$p(s) = R(\frac{1-s}{1+s}, e^{-s}) .$$

Exactly as in the classical case, it is necessary that p be bounded real on the right half-plane or, equivalently, that R be bounded real on the set S discussed in section 3. However, if a circuit devoid of sources is constructed from lumped components and unit delay lines, it remains passive when delay lines *of any nonnegative length replace the unit delays.* It follows that the function

$$p_\alpha(s) = R(\frac{1-s}{1+s}, e^{-\alpha s})$$

is bounded real on the right half-plane for all $\alpha > 0$ or equivalently (in an obvious notation) that R is bounded real on the set S_α, $\alpha > 0$. While it is *not* true that $U_\alpha S_\alpha$ *like this* fills out the bidisk ($U_\alpha S_\alpha$ is three-dimensional, while the bidisk is four-dimensional), it *is* true that as α varies over the nonnegative real numbers, the set $\sigma(B_\alpha)$ drawn in Fig. 1 moves in such a way as to sweep out all of T^2, with the exception of the set (labeled C in Fig. 1):

$$C = \{(\theta_1,\theta_2) | \theta_1 = 0, \theta_2 \neq 0\} .$$

Considering the known results about the stability of two-variable systems in terms of their behavior on T^2 [6,7], and of their behavior on three-dimensional subsets of U^2 [12] it is then quite plausible that a necessary condition for passive synthesis is that $R(Z_1,Z_2)$ be bounded real on all of U^2. This in fact is well known to be the case, and can be proved by considerations along the above lines. However, since a much

7. SYSTEMS WITH IRRATIONALLY RELATED DELAYS. We have been concerned throughout with systems whose delays are all integer multiples of one fundamental delay. If there are n independent delays over the rationals, then one gets, in the usual way, rational functions in $n+1$ variables. We remark in passing that if one has two independent delays without any differentiations (lumped elements) one will again get rational functions in two variables. However, the theory here is very different from the lumped-distributed case discussed previously, since now the spectrum is all of T^2, and the causal spectrum is all of U^2. For this reason, the stability theory of pure-delay systems with several incommensurable delays is much simpler than that of differential-delay systems.

In the general case, we will merely indicate what the spectrum looks like. For systems involving n independent delays and differentiations, $\sigma(\mathcal{B})$ consists of a line together with an n-dimensional torus, the line being asymptotic to a dense line (i.e., one which winds around the torus at an irrational angle) on the torus.

The set S (the causal spectrum) consists of $\sigma(\mathcal{B})$ as described above together with one-dimensional analytic set in U^{n+1}, and an n-dimensional polydisc (whose distinguished boundary is the n-dimensional torus mentioned above) contained in the boundary of U^{n+1}. Again, the spectrum of the convolution algebra, $\sigma(\mathcal{B})$, can be found by examining the image of the imaginary axis under the mapping which transforms the one-dimensional problem into an (n+1)-dimensional problem; that is, the several-variable approach may be regarded as a vehicle for computing the spectrum.

Finally, the "stability index group" is isomorphic to $\mathbb{Z}^{n+1}$; a more detailed analysis shows that when one allows arbitrarily many delays of arbitrary lengths, this group is isomorphic to $\mathbb{Z} \times \mathbb{R}$ (see [10,11]).

8. CONCLUSIONS. Our main purpose here has been to give a discussion of the radical differences which can occur between algebras of systems which are formally similar. These differences may arise from the problems under consideration, as in the difference between the problems of stability and passive synthesis; or from differences in the classes of systems themselves, as in the stability problem for systems composed of differentiators and integral multiples of one delay compared to the same problem for systems involving two irrationally related delays.

A further objective has been to demonstrate the utility of the Gelfand spectrum in connection with these problems. In the classical cases of either lumped, continuous-time systems or discrete-time systems the Gelfand spectrum is a circle, (or a disk in the causal case), and the usefulness of the representation of elements of the system as functions from this set into $\mathbb{C}$ is well known. In mixed differential-delay systems, the spectra are considerably more complicated objects, and the intuition to be gained by studying the action of individual systems on these spectra is not quite so transparent. Nonetheless, it is hoped that the above discussion has shown that even in the nonclassical cases the spectrum is significant, and that knowledge of the spectrum of the algebras involved can give a deeper insight into the behavior of the systems.

REFERENCES

[1] Kamen, E. W., *A note on the representation and realization of lumped-distributed networks, delay-differential systems, and 2-D systems*, IEEE Trans. Circ. and Syst. To appear.

[2] Callier, F. M., and C. A. Desoer, *An algebra of transfer functions for distributed linear time-invariant systems*, IEEE Trans. Circ. and Syst., CAS-25 (1978), 651-662.

[3] _____, *Simplifications and clarifications on the paper "An algebra of transfer functions for distributed linear time-invariant systems,"* IEEE Trans. Circ. and Syst. To appear.

[4] Kamen, E. W., *On an algebraic theory of systems defined by convolution operators*, Math. Syst. Theory, 9 (1975), 57-74.

[5] Gelfand, I., D. Raikov, and G. Shilov, *Commutative Normed Rings*, New York, Chelsea, 1964.

[6] Rudin, W., *Function Theory in Polydiscs.* New York, Benjamin, 1969.

[7] Decarlo, R. A., J. Murray, and R. Saeks, *Multivariable Nyquist theory,* Int. J. Control, 25 (1977), 657-675.

[8] Taylor, J. L., *Measure algebras,* CBMS Regional Conference Series, No. 16, Providence, Rhode Island, 1973, AMS.

[9] Douglas, R. G., *Banach Algebra Techniques in Operator Theory.* New York, Academic Press, 1972.

[10] Callier, F. M., and C. A. Desoer, *A graphical test for checking the stability of a linear time-invariant feedback system,* IEEE Trans. Autom. Contr. AC-17 (1972), 773-780.

[11] Davis, J. H., *Encirclement conditions for stability and instability of feedback systems with delays,* Int. J. Control, 15 (1972), 793-799.

[12] Stoll, W., *Holomorphic functions of finite order in several complex variables,* CBMS Regional Conference Series, No. 21, Providence, Rhode Island, 1974, AMS.

[13] Youla, D. C., *The synthesis of networks containing lumped and distributed elements,* in Proc. Symp. Generalized Networks, Vol. XVI, P.I.B. 1966, pp. 289-343.

[14] Bose, N. K., K. Zaki, and R. W. Newcomb, *A multivariable bounded-reality criterion,* J. Franklin Inst., 297 (6). 1974, pp. 479-484.

DEPARTMENT OF ELECTRICAL ENGINEERING
TEXAS TECH UNIVERSITY
LUBBOCK, TEXAS 79409

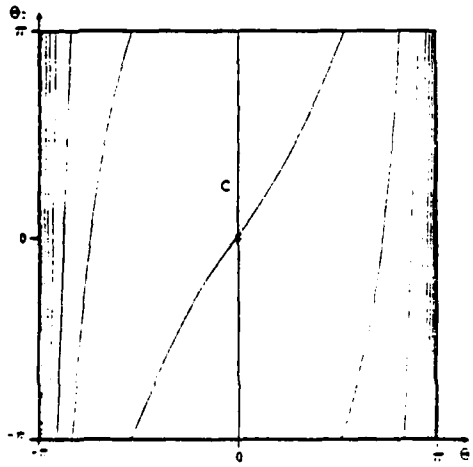


Figure 1.

ABSTRACT OF

A DESIGN METHOD FOR 2-D RECURSIVE DIGITAL FILTERS

JOHN MURRAY

Abstract

A method is described for the design of two-dimensional half-plane recursive digital filters, in the form of a cascade connection of filters which are of second order in the (principal) direction of recursion, and of arbitrarily high order in the other direction. The filters thus derived are shown to be automatically stable, but yield poor responses in the vicinity of very wide or very narrow bandwidths. Some techniques for tackling these difficulties are discussed, and the results of applying these design procedures are shown.

PRECEDING PAGE BLANK-NOT FILMED

ABSTRACT OF

THE DESIGN OF 2-D FILTERS AS 1-D TIME-VARYING SYSTEMS

JOHN MURRAY

PRECEDING PAGE BLANK-NOT FI

THE DESIGN OF 2-D FILTERS AS 1-D TIME-VARYING SYSTEMS

Abstract

In a previous paper a theoretical approach to two-dimensional recursive digital filtering was proposed. This approach was based on the fact that two-dimensional recursive filtering with the normal "scanning" recursion is in reality a one (time) - dimensional periodically time-varying discrete-time operation. Time-varying systems are normally intractable; periodically time-varying discrete-time systems, however, are among the few classes of time-varying systems for which a complete, closed-form theory exists. This theory transforms single-input single-output time-varying systems of period N (where we assume that the sampling period is 1) into a subclass of N-input, N-output time-invariant systems. There is a possibility of using classical multi-input multi-output time-invariant system theory to design 2-D filters in this setting.

There are two problems which arise. Firstly, N is usually very large, and so direct manipulations are impossible. Secondly, the time-varying systems correspond to only a subset of multi-input multi-output systems, and the classical design methods do not necessarily yield a design in this subset.

This paper exhibits a design procedure which takes care of both of these problems, and shows examples of filters designed by using this procedure.

PRECEDING PAGE BLANK-NOT FILMED

DETECTION AND ESTIMATION IN IMAGERY

J. WALKUP

PRECEDING PAGE BLANK-NOT FILMED

Texas Tech University | Institute for Electronic Science

Joint Services Electronics Program | Research Unit: 5

1. Title of Investigation: Detection and Estimation in Imagery
2. Senior Investigators: J.F. Walkup Telephone: (806)-742-3500
3. JSEP Funds: Current $25,875
4. Other Funds: Current
5. Total Number of Professionals: PI's 2 (1 mo.) RA's 1 (1/2 time)
6. Summary

Although one might, upon a cursory investigation, conclude that the image processing problem was simply a two dimensional generalization of the standard 1-D signal processing problem this is not the case since both the *signal and noise phenomena encountered in the image processing problem* tend to be either *nonlinear and/or space variant* (the spatial analog of time-varying). Indeed, both *photo-electric shot noise* and *film grain noise* are highly signal dependent in nature while the *edge effects* in a finite image introduce space-variant effects. Although a theory for coping with these nonlinear and space-variant effects has been developed in a 1-D setting, in the 2-D image processing problem these techniques have proven to be *computationally prohibitive*. As such, the present work unit is directed towards the problem of developing *computationally viable algorithms for the digital image processing problem*. These include *sub-optimal algorithms* for estimation in signal-dependent noise, analytic techniques for *reducing the effective dimensionality* of an image and techniques for *exploiting the signal dependent* nature of the *noise phenomena*.

7. Publications and Activities:

A. Refereed Journal Articles

PRECEDING PAGE BLANK-NOT FILMED

1. Froehlich, G.K., Walkup, J.F., and T.F. Krile, "Estimation in Signal-Dependent Film-Grain Noise", Applied Optics, Vol. 20, pp. pp. 3619-3626, (1981).

B. Conference Papers and Abstracts

1. Froehlich, G.K., Walkup, J.F., and T.F. Krile, "Some Effects of Signal-Dependent Noise on Estimation Structures", presented at the Annual Meeting of the Optical Soc. of Amer., Oct. 1980, (an abstract for this presentation appeared in the Jour. of the OSA, Vol. 20, p. 613).

C. Theses

1. Kasturi, R., Ph.D. Dissertation, (in preparation).

D. Conferences and Symposia

1. Walkup, J.F. Froehlich, and T.F. Krile, Annual Meeting of the Optical Soc. of Amer., Oct. 1980.

E. Lecture

1. Walkup, J.F., Univ. of Washington, June 1981.

ESTIMATION IN SIGNAL-DEPENDENT FILM-GRAIN NOISE

GARY K. FROEHLICH

JOHN F. WALKUP

AND

THOMAS F. KRILE

Estimation in signal-dependent film-grain noise

Gary K. Froehlich, John F. Walkup, and Thomas F. Krile

Optimal estimators are derived for a signal-dependent film-grain noise model, and the effect of signal-dependence on the estimators's structures is investigated. Due to the mathematical complexity of these optimal estimators, various suboptimal estimators are proposed. Computer simulations are then presented which compare the optimal and suboptimal estimators with regard to mean square estimation error, sensitivity to signal-dependence, and robustness with respect to the *a priori* signal probability density function.

I. Introduction

A number of physical noise processes are inherently signal-dependent. These include photoelectronic shot noise,[1] magnetic tape recording noise,[2] and of course photographic film-grain noise.[1,3,4] In an earlier paper,[4] a very general model was presented which embodies all the just mentioned processes and more. That model is presented again in

$$r = s + kf(s)n_1 + n_2, \tag{1}$$

where r is the noisy measurement; s is the underlying signal to be estimated, which is generally characterized by its probability density function $p(s)$; $f(s)$ is a zero-memory (spatial) function of the signal; n_1 and n_2 are signal-independent random noise processes; and k is a scalar constant, which, when set equal to zero, yields the signal-independent additive noise model. It is further assumed that n_1, n_2, and s are mutually statistically independent.

It has already been shown that estimators which ignore signal-dependence of noise pay a penalty in terms of mean square estimation error (MSEE) and that inclusion of the signal-dependence, while increasing the complexity of estimators, results in potentially superior performances.[4,5] In this paper, we intend to compare various estimators with regard to ease of implementation, MSEE, sensitivity to signal-dependence, and robustness with respect to the *a priori* probability density $p(s)$. This is achieved through a comparison of the estimators's structures and by using computer simulations.

To simplify the comparisons as much as possible, various assumptions about the model of Eq. (1) are necessary. The signal-independent noise terms, n_1 and n_2, are assumed to be zero-mean normal random variables with variances σ_1^2 and σ_2^2, respectively. Notationally, this is represented by

$$n_i \sim N(0,\sigma_i^2), \quad i = 1,2. \tag{2}$$

Furthermore, a specific function for $f(s)$ must be chosen. The function $f(s) = s^p$ is of particular interest, as it represents photographic film-grain noise when s represents photographic density, p is between 0.2 and 0.7,[3,4] and k is a scanning constant. The exponent p is taken to be 0.5 in the remainder of this paper. Thus

$$f(s) = \sqrt{s}. \tag{3}$$

The final assumption is with regard to the *a priori* probability density function (pdf) of the signal, i.e., $p(s)$. Several cases are treated in Refs. 4 and 5. These include the Gaussian, Rayleigh, uniform, discrete-uniform, and folded-normal[6] pdf's, which were chosen for their tractability or their positivity constraints.[7] For purposes of comparison and brevity, only the Gaussian, Rayleigh, and uniform cases will be presented here. With the above assumptions, the model becomes

$$r = s + k\sqrt{s}n_1 + n_2, \tag{4}$$

and the conditional measurement r, given s, is distributed normally with mean s and variance

$$v(s) = k^2\sigma_1^2 s + \sigma_2^2, \tag{5}$$

or, in the notation of Eq. (2),

$$p(r|s) \sim N[s, v(s)]. \tag{6}$$

When this work was done all authors were with Texas Tech University, Department of Electrical Engineering, P.O. Box 4439, Lubbock, Texas 79409; Gary Froehlich is now with Sandia National Laboratories, Albuquerque, New Mexico 87185.

PRECEDING PAGE BLANK-NOT FI

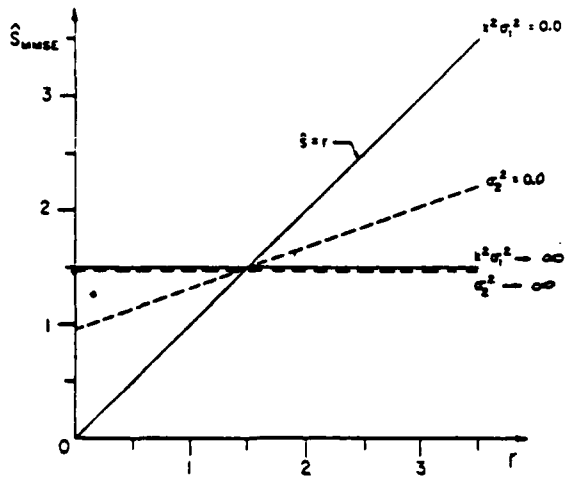


Fig. 1. MMSE estimator structure for Gaussian $p(s)$: solid line ($\sigma_2^2 = 0.3$, $k^2\sigma_1^2$ varies); dashed line ($k^2\sigma_1^2 = 0.3$, σ_2^2 varies); $p(s) \sim N(1.5, 0.25)$.

II. Optimal Estimation—Minimum Mean Square Error

Optimal estimators based on the measurement model of Eq. (4) are now presented. The first optimality criterion considered is minimization of MSEE, or equivalently, minimization of the Bayes's risk for a quadratic cost function.[8,9] Computationally, this is merely the conditional mean of the *a posteriori* probability density function $p(s|r)$. Thus the minimum mean square error (MMSE) estimate is given by

$$\hat{s}_{MMSE} = \int_{-\infty}^{\infty} s p(s|r) ds. \tag{7}$$

Unfortunately, the *a posteriori* density $p(s|r)$ is very difficult to compute. Bayes's rule can be applied, however, to rewrite Eq. (7) as

$$\hat{s}_{MMSE} = \frac{\int_{-\infty}^{\infty} s p(r|s) p(s) ds}{\int_{-\infty}^{\infty} p(r|s) p(s) ds}, \tag{8}$$

where $p(r|s)$ is given by Eq. (6), and some particular form is assumed for $p(s)$. As an example of the complexity of this estimator, consider the computation of Eq. (8) when $p(s)$ is Gaussian, i.e.,

$$p(s) \sim N(\mu_s, \sigma_s^2). \tag{9}$$

In this case $p(r|s)p(s)$ is given by

$$p(r|s)p(s) = \frac{1}{2\pi\sigma_s\sqrt{v(s)}} \exp\left\{-\tfrac{1}{2}\left[\frac{(r-s)^2}{v(s)} + \frac{(s-\mu_s)^2}{\sigma_s^2}\right]\right\}, \tag{10}$$

where $v(s)$ is given in Eq. (5). Similar complexity results when $p(s)$ is assumed to be either Rayleigh distributed or uniformly distributed.

To arrive at some general conclusions about the MMSE estimators, the structure of the estimators is presented. The estimate $\hat{s}_{MMSE}$ is plotted vs the measurement r in Figs. 1–3. In Fig. 1, $p(s)$ is assumed to be Gaussian, in Fig. 2 it is assumed Rayleigh, and in Fig. 3 it is assumed uniform. In each figure, the solid lines represent cases in which σ_2^2, the variance of the signal-independent noise term, is fixed while $k^2\sigma_1^2$, which is proportional to the variance of the signal-dependent noise term, is allowed to vary. The dashed lines represent cases in which $k^2\sigma_1^2$ is fixed and σ_2^2 is allowed to vary. The actual numerical values of all the parameters are realistic, and they serve to restrict the photographic density variables to a practical range of zero to three.[10] Also, inspection of Eq. (4) shows that when $k = 0$, the model reduces to the classical signal-independent additive noise model $r = s + n_1$.

Several general conclusions are implied by examination of Figs. 1–3. In all three cases, for example, the estimators exhibit greater sensitivity to the signal-dependent noise term than to the signal-independent

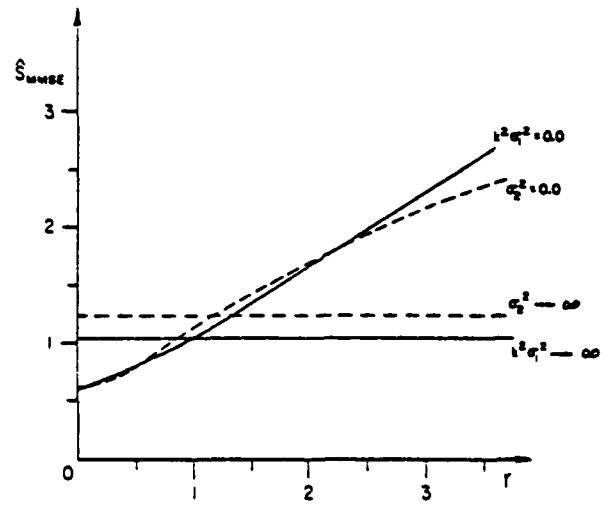


Fig. 2. MMSE estimator structure for Rayleigh $p(s)$: solid line ($\sigma_2^2 = 0.3$, $k^2\sigma_1^2$ varies); dashed line ($k^2\sigma_1^2 = 0.3$, σ_2^2 varies); $p(s) \sim$ Rayleigh with mean = 1.25, $\sigma^2 = 0.43$.

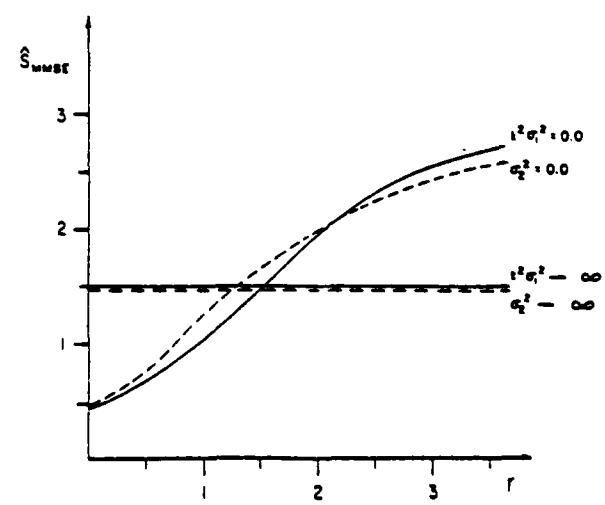


Fig. 3. MMSE estimator structure for uniform $p(s)$: Solid line ($\sigma_2^2 = 0.3$, $k^2\sigma_1^2$ varies); dashed line ($k^2\sigma_1^2 = 0.3$, σ_2^2 varies); $p(s) \sim$ uniform with mean = 1.5, $\sigma^2 = 0.75$.

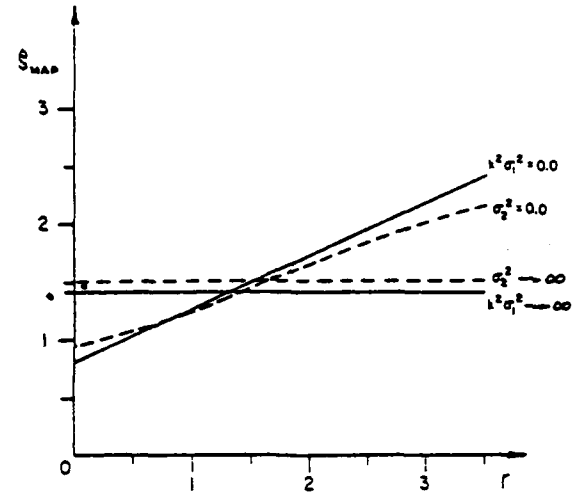


Fig. 4. MAP estimator structure for Gaussian $p(s)$: solid line ($\sigma_2^2 = 0.3$, $k^2\sigma_1^2$ varies); dashed line ($k^2\sigma_1^2 = 0.3$, σ_2^2 varies); $p(s) \sim N(1.5, 0.25)$.

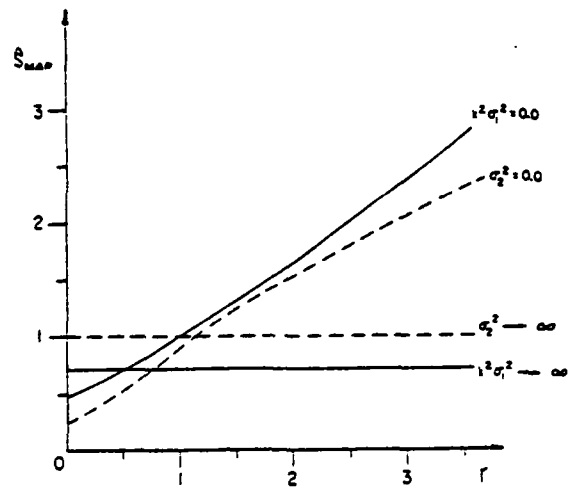


Fig. 5. MAP estimator structure for Rayleigh $p(s)$: solid line ($\sigma_2^2 = 0.3$, $k^2\sigma_1^2$ varies); dashed line ($k^2\sigma_1^2 = 0.3$, σ_2^2 varies); $p(s) \sim$ Rayleigh with mean = 1.25, $\sigma^2 = 0.43$.

one. Also, in all three cases, if either noise term increases without bound, the estimate becomes the (assumed known) mean of s. Furthermore, note that the more $p(s)$ deviates from normality, the more nonlinear the estimator structure becomes. This implies that estimator performance will suffer if the assumed form of $p(s)$ is wrong. Ideally we would like to find a robust estimator, i.e., an estimator for which the structure is invariant to changes in $p(s)$.

One final note about the MMSE estimator is in order. The complexity of Eq. (8) would seem to discourage the use of this estimator on a point-by-point basis over an image. However, if the parameters $k, \sigma_1^2, \sigma_2^2$, and the moments of s do not change for the class of images under study, the numerical integration of Eq. (8) can be done off-line for the entire practical range of measurements r. The only on-line operation then is reduced to a table look-up procedure to match the precalculated estimate with the corresponding measurement.

III. Optimal Estimation—Maximum *a posteriori*

As an alternative to minimization of mean square error, we now consider an estimator which minimizes the Bayes's risk[8] for a uniform cost function. This turns out to be the conditional mode of the *a posteriori* probability density function $p(s|r)$. Since the mode is merely the peak value of the pdf, the estimator is often referred to as the maximum *a posteriori* (MAP) estimator.[8,9] The MAP estimate is treated in detail in Refs. 4 and 5. For the model of Eq. (4), the MAP estimate is the solution of a polynomial. For $p(s)$ normally distributed, the polynomial is cubic. For Rayleigh $p(s)$, the MAP equation is of degree four, and for uniform $p(s)$, the polynomial is quadratic. The latter case, with uniform $p(s)$, is equivalent to maximum likelihood (ML) estimation.[4,5,8,9] The ML estimate is used when no prior information about the statistics of s is assumed or known. The uniform distribution for $p(s)$ corresponds to this worst case.

As with the MMSE estimator, the MAP estimator structures are presented as plots of the MAP estimates vs the measurements r. Figure 4 is the MAP estimator structure for Gaussian $p(s)$, Fig. 5 is the MAP estimator structure for Rayleigh $p(s)$, and Fig. 6 is the ML estimate. Once again, the solid lines represent the cases wherein σ_2^2 is fixed and $k^2\sigma_1^2$ is allowed to vary, and the dashed lines correspond to the case where $k^2\sigma_1^2$ is fixed and σ_2^2 is allowed to vary.

The overall trends in these three figures are very similar to those indicated in Figs. 1–3. Again, the estimators become increasingly nonlinear as $p(s)$ departs

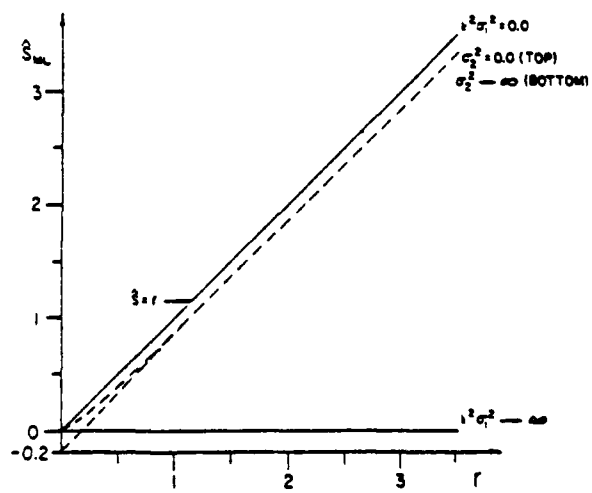


Fig. 6. ML estimator structure: solid line ($\sigma_2^2 = 0.3$, $k^2\sigma_1^2$ varies); dashed line ($k^2\sigma_1^2 = 0.3$, σ_2^2 varies).

from normality. Also the sensitivity to the signal-dependent noise term is much greater than the sensitivity to the signal-independent noise term for all three cases. However, the value of the estimate when $k^2\sigma_1^2$ increases without bound is always lower than the estimate when σ_2^2 increases without bound. This is quite different from the MMSE estimators which converged to the same value regardless of which noise term became unbounded.

Another characteristic shared by both the MAP and MMSE estimators is computational complexity. It is generally undesirable to have to solve a polynomial at every point of measurement (pixel), just as it was undesirable to integrate numerically at each point. However, if the parameters k, σ_1^2, σ_2^2, and the statistics of s do not change appreciably for the class of images under study, the problem is again reduced to one of an off-line table generation followed by on-line table look-ups.

IV. Suboptimal Estimation

The major limitations of the optimal estimators just presented were their general computational complexity and their sensitivity to choice of $p(s)$. It might prove acceptable to sacrifice some theoretical performance for ease of implementation and some measure of robustness. Toward this end, suboptimal estimators are next investigated.

A. Weighted Spatial Averaging

As a first step, the basic sample mean will be exploited. The sample mean is certainly easy to implement, and it is also known to be the most robust estimator of the true mean.[11-14] The estimation procedure then is to replace each measurement with the average of that measurement and its neighbors. In a 2-D sampled image, for example, this might correspond to a pixel and its eight nearest neighbor pixels. Defining the sample mean at the point j as $\bar{r}_j$,

$$\bar{r}_j = \sum_{i=1}^{n} r_{ij}, \tag{11}$$

the weighted spatial average (WSA) estimate is then

$$\hat{s}_{\mathrm{WSA}j} = \bar{r}_j. \tag{12}$$

Unfortunately, this estimator has one rather severe limitation: it is not robust with respect to the spatial noise power spectrum. This occurs because the WSA algorithm is in effect a finite-window spatial low-pass filtering operation. As such, it does not affect low-frequency noise, and it destroys high-frequency signal information.

B. Modified Signal-Independent MAP

In an attempt to combine the desirable features of the WSA algorithm and the sophistication of a slightly more complex estimator, the MAP estimator designed for signal-independent additive noise and Gaussian signal statistics was modified. The modification allows the statistics to vary spatially—a consequence of signal-dependence. The signal-independent MAP estimator is given by[4,5]

$$\hat{s}_{\mathrm{MAP}} = \frac{\sigma_s^2}{\sigma_2^2 + \sigma_s^2} r + \frac{\sigma_2^2}{\sigma_2^2 + \sigma_s^2} \mu_s, \tag{13}$$

where it was assumed that $s \sim N(\mu_s, \sigma_s^2)$. This requires *a priori* knowledge of μ_s and σ_s^2, which remain fixed in the estimator of Eq. (13).

It can be easily shown that the expected value of $\bar{r}_j$ is μ_s, i.e.,

$$E[\bar{r}_j] = \mu_s. \tag{14}$$

Thus we can eliminate the need for *a priori* knowledge of μ_s and simultaneously allow the statistics to vary spatially by replacing μ_s in Eq. (13) with an estimate of μ_s, namely, $\bar{r}_j$. The modified signal-independent MAP (MSIMAP) estimator then becomes

$$\hat{s}_{\mathrm{MAP}j} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_2^2} r_j + \frac{\sigma_2^2}{\sigma_s^2 + \sigma_2^2} \bar{r}_j, \tag{15}$$

where $\bar{r}_j$ was given by Eq. (11). Because the MSIMAP estimator of Eq. (15) adapts spatially over the image, it should out perform the unmodified MAP estimator of Eq. (13) when both are applied to the signal-dependent noise model of Eq. (4).

A potential limitation of the MSIMAP algorithm is the requirement that σ_s^2 be known *a priori*. There may be situations where this is not unreasonable, such as when the energy of the signal is known; however, there may well be situations when it is quite unreasonable to require *a priori* knowledge of any of the signal statistics. An additional modification is required to eliminate this defect.

C. James-Stein

Just as the signal mean is adaptively estimated in the MSIMAP algorithm, the signal variance σ_s^2 can also be adaptively estimated within the overall estimator structure. Note that Eq. (15) can be rewritten as

$$\hat{s}_j = Qr_j + (1 - Q)\bar{r}_j, \tag{16}$$

where

$$Q \triangleq \sigma_s^2/(\sigma_s^2 + \sigma_2^2). \tag{17}$$

Equation (16) can in turn be rewritten as

$$\hat{s}_j = \bar{r}_j + Q(r_j - \bar{r}_j). \tag{18}$$

Now the signal-independent additive noise model for which the signal-independent MAP estimator was designed has r distributed normally with mean μ_s and variance σ_r^2 given by

$$\sigma_r^2 = \sigma_s^2 + \sigma_2^2. \tag{19}$$

Thus

$$Q = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_2^2} = \frac{\sigma_s^2}{\sigma_r^2} = 1 - \frac{\sigma_2^2}{\sigma_r^2}. \tag{20}$$

Rewriting Q in this fashion eliminates the explicit requirement to know σ_s^2. It was assumed at the outset that σ_2^2 was known, so the problem is reduced to one of estimating σ_r^2. Since r is the measured quantity, the obvious intuitive estimate for σ_r^2 is the sample variance, employing the same nearest-neighbor approach used to compute the sample mean. Defining the sample variance at location j as v_j, it is given by

Table I. Sample MSEE for Gaussian $p(s)$ with $\mu_s = 1.5$ and $\sigma_s^2 = 0.25$

Case	$k^2\sigma_1^2 = 0$, $\sigma_2^2 = 0.03$	$k^2\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.01$	$k^2\sigma_1^2 = 0.03$, $\sigma_2^2 = 0$
Noise	0.027	0.014	0.012
MMSE	0.031	0.011	0.011
MAP	0.024	0.014	0.012
ML	0.027	0.014	0.012
WSA	0.163	0.164	0.162
MSIMAP	0.026	0.014	0.012
JS	0.026	0.015	0.012

Table II. Sample MSEE for Rayleigh $p(s)$ with $\mu_s = 1.25$ and $\sigma_s^2 = 0.43$

Case	$k^2\sigma_1^2 = 0$, $\sigma_2^2 = 0.03$	$k^2\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.01$	$k^2\sigma_1^2 = 0.03$, $\sigma_2^2 = 0$
Noise	0.026	0.012	0.011
MMSE	0.024	0.012	0.009
MAP	0.025	0.012	0.010
ML	0.026	0.012	0.011
WSA	0.297	0.290	0.293
MSIMAP	0.025	0.012	0.011
JS	0.025	0.012	0.011

Table III. Sample MSEE for Discrete-Uniform $p(s)$ with Mean = 0.25 and Var. = 0.02

Case	$k^2\sigma_1^2 = 0$, $\sigma_2^2 = 0.03$	$k^2\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.01$	$k^2\sigma_1^2 = 0.03$, $\sigma_2^2 = 0$
Noise	0.038	0.011	0.002
MMSE	0.009	0.007	0.004
MAP	0.015	0.009	0.001
ML	0.038	0.011	0.002
WSA	0.015	0.005	0.001
MSIMAP	0.022	0.009	0.002
JS	0.027	0.007	0.002

Table IV. Sample MSEE for Discrete-Uniform $p(s)$ with Mean = 1.5 and Var. = 0.75

Case	$k^2\sigma_1^2 = 0$, $\sigma_2^2 = 0.03$	$k^2\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.01$	$k^2\sigma_1^2 = 0.03$, $\sigma_2^2 = 0$
Noise	0.026	0.014	0.012
MMSE	0.021	0.014	0.014
MAP	0.023	0.014	0.012
ML	0.026	0.014	0.012
WSA	0.021	0.015	0.015
MSIMAP	0.025	0.013	0.012
JS	0.016	0.009	0.012

$$v_j = \frac{1}{n-1}\sum_{i=1}^{n}(r_{ij} - \bar{r}_j)^2, \tag{21}$$

where $\bar{r}_j$ was defined by Eq. (11).

With σ_r^2 adaptively estimated by v_j, the variable Q now varies with location and thus is subscripted as Q_j, where

$$Q_j = 1 - \frac{\sigma_2^2}{v_j}. \tag{22}$$

The final estimator resulting from these manipulations is given by

$$\hat{s}_j = \bar{r}_j + Q_j(r_j - \bar{r}_j). \tag{23}$$

As it turns out, this is an empirical Bayesian estimator which estimates the mean of a multivariate normal distribution and exhibits uniformly lower MSEE than the sample mean. This estimator is known as the James-Stein (JS) estimator.[15,16] It can be shown[16] that the MSEE is further lowered by restricting Q_j to be nonnegative. With this final modification, the (JS) estimator is given by

$$\hat{s}_j = \bar{r}_j + Q_j^+(r_j - \bar{r}_j), \tag{24}$$

where $a^+ \triangleq \max(0,a)$, Q_j is given by Eq. (22), and $\bar{r}_j$ is given by Eq. (11).

An estimator which is in some sense intermediate between the MSIMAP estimator [Eq. (15)] and the JS estimator [Eq. (24)] can be obtained by using the sample variance v_j of Eq. (21) to estimate the signal variance σ_s^2 rather than the measurement variance σ_r^2. This estimator is discussed in Ref. 5, where it is called the MSIMAP2 estimator. It was also discussed by Lee,[17] who showed some experimental results with noise-degraded images.

V. Simulations

To compare the optimal and suboptimal estimators with each other, it was necessary to perform several computer simulations. Only a few results are presented in this paper. For a very extensive tabulation of results

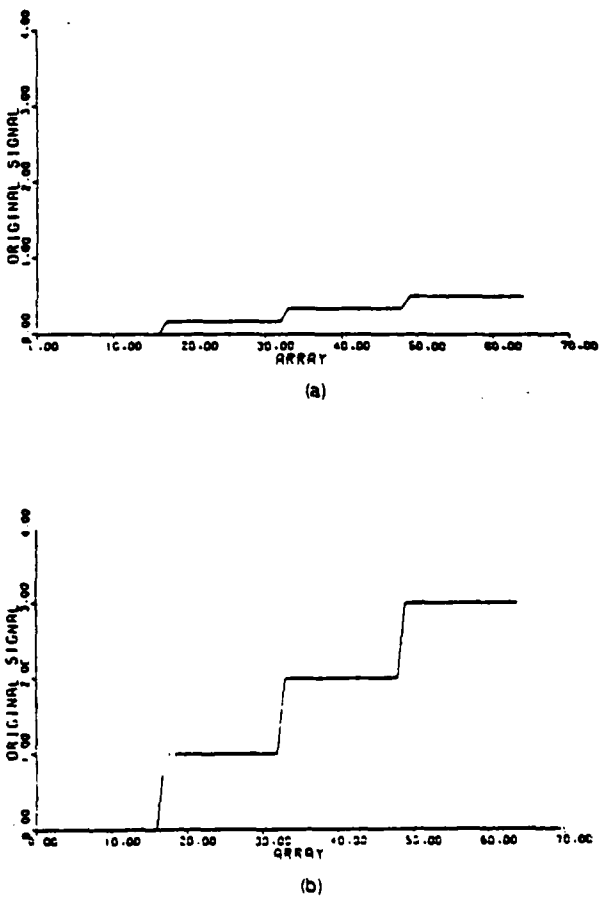


Fig. 7. Uncorrupted signals with discrete-uniform statistics: (a) low-contrast low-signal-mean case; (b) high-contrast case.

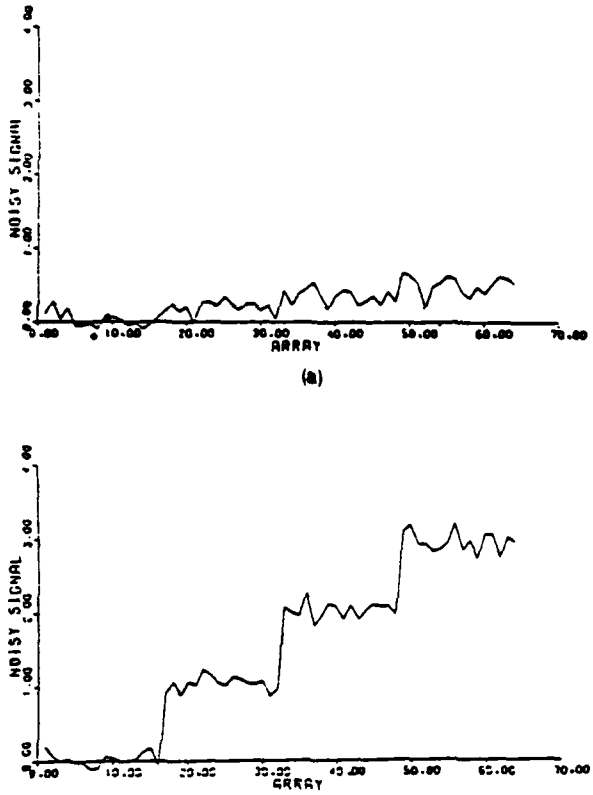


Fig. 8. Waveforms of Fig. 7 corrupted by equal parts signal-independent and signal-dependent noise ($\sigma_1^2 = \sigma_2^2 = 0.01$).

see Ref. 5. The estimators treated here are as follows: (1) the MMSE estimator with $p(s)$ Gaussian; (2) the MAP estimator with $p(s)$ Gaussian; (3) the ML estimator; (4) the WSA estimator; (5) the MSIMAP estimator; and (6) the JS estimator.

These six estimators were applied to measurements generated by the noise model of Eq. (4). Here three noise regimes are considered. They are (1) a signal-independent noise only case ($k^2\sigma_1^2 = 0, \sigma_2^2 = 0.03$). (2) a signal-dependent noise only case ($\sigma_2^2 = 0, k^2\sigma_1^2 = 0.03$), and (3) a case with equal parts signal-dependent and signal-independent noise ($\sigma_2^2 = k^2\sigma_1^2 = 0.01$). The values were chosen to insure reasonable SNRs for purposes of visual comparison. Each Monte Carlo simulation employed 256 sample measurements, and 200 simulations were run for each case.

The sample MSEE for each of the six estimators, as well as the squared deviation of the noisy measurement from the true signal (labeled simply noise), is tabulated under columns corresponding to the three noise mixtures described above. There are four such tables presented in this paper. Table I is for Gaussian $p(s)$. To get some feeling about robustness, $p(s)$ is next treated (Table II) as a Rayleigh pdf. To allow still further deviation from normality, $p(s)$ is next taken to be a discrete-uniform density. Here, however, two sets of parameters are considered. The first corresponds to a low-contrast low-signal-mean case (Table III), while the second corresponds to a high-contrast case (Table IV).

Looking at Table I, two things are immediately apparent. First, the MMSE estimator does indeed exhibit minimum sample MSEE, except in the case where signal-independent noise dominates. Second, the WSA estimator performs very poorly. This latter condition occurs because of the low-pass filtering nature of the algorithm. The signal s in this case is a sample function from a pseudorandom Gaussian process. Consequently, a low-pass filtered version deviates more from the true signal than the noisy measurement itself does. Note also that the other estimates perform more or less equally well.

Turning next to Table II, where s is a Rayleigh-distributed random variable, note that the MMSE estimator again has a slight edge over the other estimators in most cases. As before, the WSA estimator performs very badly, again due to the nature of s and the effect of low-pass filtering. Also the remaining estimators again perform about equally.

Table III shows quite different behavior on the part of the WSA estimator. This is primarily due once again to the nature of the signal. The particular sample function used as a signal for this case is shown in Fig. 7(a) and is seen to have very little high-frequency content. Thus a low-pass filtering operation is ideal for

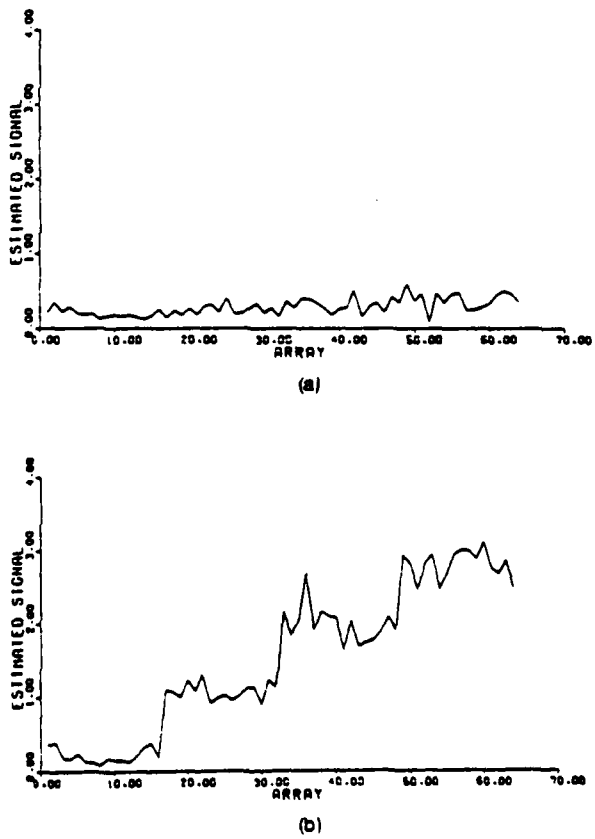


Fig. 9. MMSE estimates for the noisy measurements of Fig. 8.

eliminating noise with a large high-spatial frequency component. The other estimators performed relatively well in most cases here also.

The last case considered is the high-contrast case, with s shown in Fig. 7(b), and the results of the simulations given in Table IV. Due to the large step size in this signal, there is a sizable high-spatial frequency component. Because of this, the superior performance exhibited by the WSA process in the similar case above is lost here. The various optimal estimators perform rather poorly as well due to drastic deviation from the normality assumption. It is the spatially adaptive estimators which performed well in this case, most notably the JS estimator.

To aid interpretation of the data presented in Tables I–IV, consider the pictorial representations. Figure 8 illustrates the signals of Fig. 7 after corruption by equal parts signal-dependent and signal-independent noise. These corrupted waveforms are then used as noisy measurements, and three of the estimators are applied. The MMSE estimate is illustrated in Fig. 9, the WSA estimate in Fig. 10 (where the low-pass filtering effect is evident), and the JS estimate is shown in Fig. 11.

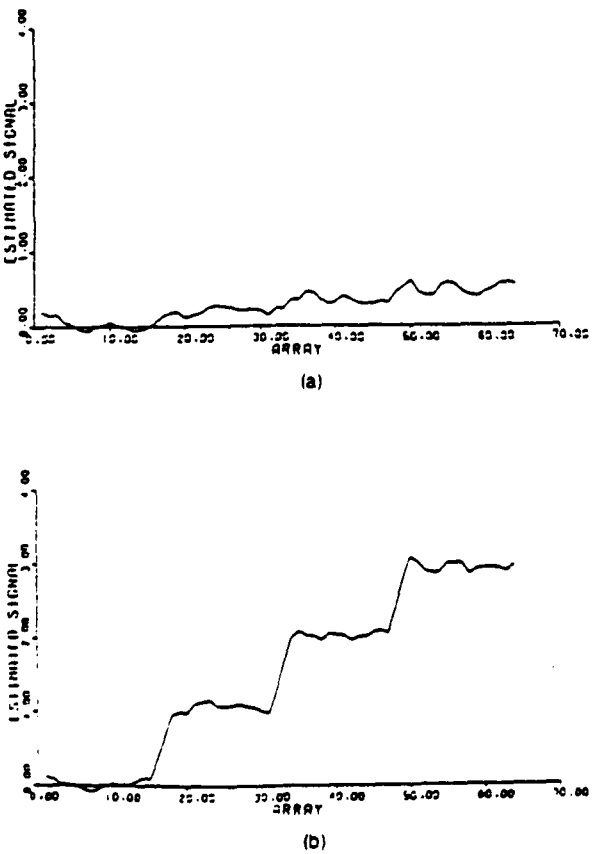


Fig. 10. WSA estimates for the noisy measurements of Fig. 8.

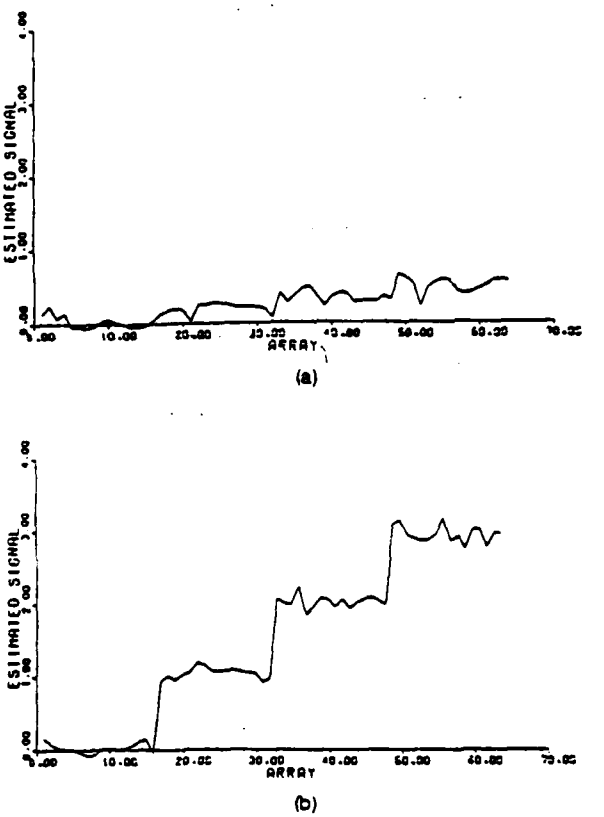


Fig. 11. James-Stein estimates for the noisy measurements of Fig. 8.

VI. Conclusions

Overall, then, using the MSEE as a performance measure, several observations may be made. If the original signal has a great deal of high-frequency content, as does an image with a lot of fine detail, the WSA estimator should be avoided. The MMSE estimator is the most desirable when the signal statistics do not deviate too far from normality, with the JS estimator becoming the better choice in cases of large deviations from the normality assumption. On the other hand, if the signal is known (or suspected) to have little high-frequency content, the superior estimator is the WSA algorithm, especially when substantial deviations from the Gaussian assumption are also encountered.

This work was supported by the Joint Services Electronics Program at Texas Tech U. under ONR contract N0014-76-C-1136. The assistance of Gus Oliver with the computer simulations, the typing of the manuscript by Heidi Hanssen, and the assistance of Rangachar Kasturi with the figures are gratefully acknowledged.

References

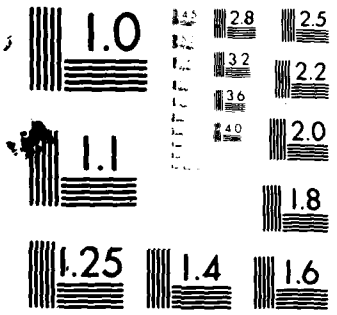
1. J. F. Walkup and R. C. Choens, Opt. Eng. 13, 258 (1974).
2. J. C. Mallinson, Proc. *IEEE* 64, 196 (1976).
3. F. Naderi and A. A. Sawchuk, Appl. Opt. 17, 1228 (1978).
4. G. K. Froehlich, J. F. Walkup, and R. B. Asher, J. Opt. Soc. Am 68, 1665 (1978).
5. G. K. Froehlich, "Estimation in Signal-Dependent Noise," Ph.D. Dissertation, Department of Electrical Engineering, Texas Tech U., Lubbock (Dec. 1980).
6. Suppose that X is a normally distributed random variable. Define the random variable Y by $Y = |X|$. Now Y is said to be a folded-normal random variable since the portion of $p_X(x)$ for $x < 0$ is folded back onto the portion of $x \geq 0$ to arrive at $p_Y(y)$.
7. Many, if indeed not all, of these densities are idealizations, as such behavior is rarely evident in real images, although it may well occur in nonimage type signals. B. R. Hunt and T. M. Cannon have demonstrated [IEEE Trans. Syst. Man. Cybern. SMC-6, 876 (1976)] that images can be modeled as consisting of intensity fluctuations about a nonstationary ensemble mean, and that these fluctuations about that mean are shown experimentally to be highly Gaussian. Unfortunately, the signal-dependent nature of our model introduces the nonstationarity of the mean into the variance term as well.
8. H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part 1* (Wiley, New York, 1968).
9. A. P. Sage and J. L. Melsa, *Estimation Theory With Applications to Communications and Control* (McGraw-Hill, New York, 1971).
10. P. L. Bachman, "Silver Halide Photography," in *Handbook of Optical Holography*, H. J. Caulfield, Ed. (Academic, New York, 1979)
11. P. J. Huber, Ann. Math. Stat. 35, 73 (1964).
12. P. J. Huber, Ann. Math. Stat. 43, 1041 (1972).
13. F. R. Hampel, Ann. Math. Stat. 42, 1887 (1971).
14. P. Papantoni-Kazakos, IEEE Trans. Inf. Theory IT-23, 223 (1977).
15. B. Efron and C. Morris, Sci. Am. 119 (May 1977).
16. B. Efron and C. Morris, J. Am. Stat. Assoc. 70, 311 (1975).
17. J. S. Lee, IEEE Trans. Patt. Anal. Mach. Intel. *PAMI-2*, 165 (1980).

AD-A112 113 TEXAS TECH UNIV LUBBOCK INST FOR ELECTRONIC SCIENCE F/G 9/5
ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICES ELECTRONICS --ETC(U)
DEC 81 R SAEKS, L R HUNT, J MURRAY, J WALKUP N00014-76-C-1136

UNCLASSIFIED NL

3 OF 3
AD
A112113

END
DATE
FILMED
4 82
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

POINTING AND TRACKING

T.G. NEWMAN

Texas Tech University | Institute for Electronic Science

Joint Services Electronics Program | Research Unit: 6

1. Title of Investigation: Pointing and Tracking

2. Senior Investigator: Thomas G. Newman Telephone: (806)-742-2571

3. JSEP Funds: Current $25,875

4. Other Funds: Current _____

5. Total Number of Professionals: PI's 1 (1 mo.) RA's 1 (1/2 time)

6. Summary:

The present work unit represents the continuation of an on-going research program in which modern *group theoretic* methods are used to obtain a solution to the *pointing and tracking problem*. Initially, we developed an algorithm applicable to imagery in the plane whose motion is characterized by a four parameter Lie group (two translations, rotation, and magnification). At the present time we are in the process of *implementing the* resultant *pointing and tracking algorithm* (hopefully in real-time) on our video image processing system and simultaneously extending the theory to a true three dimensional model.

7. Publications and Activities

A. Conference Papers and Abstracts

1. Newman, T.G., and L. Zlobec, "Adaptive Pattern Matching Using Control Theory on Lie Groups", Proc. of the Inter. Symp. on the Mathematics of Networks and Systems, Santa Monica, Aug. 1981, pp. 206-210.

B. Theses

1. Demus, A., M.S. Report, (in preparation).

2. D. Tarrel, M.S. Thesis, (in preparation).

3. C. Hsia, M.S. Report, (in preparation)

PRECEDING PAGE BLANK-NOT FILMED

C. Conferences and Symposia

1. Newman, T.G., Inter. Symp. on the Mathematics of Networks and Systems, Santa Monica, Aug. 1981.

ADAPTIVE PATTERN MATCHING USING CONTROL THEORY ON LIE GROUPS

THOMAS G. NEWMAN

AND

LEOPOLD ZLOBEC

ADAPTIVE PATTERN MATCHING USING CONTROL THEORY ON LIE GROUPS*

Thomas G. Newman and Leopold Zlobec
Texas Tech University
Lubbock, Texas

Abstract

A method is given for matching a subpattern of a two-dimensional image against a stored prototype, where the latter is defined on a window whose position and shape is determined by the action of a Lie group of transformations. The method involves the construction of a path in the control group along which the matching error decreases to a local minimum.

1. INTRODUCTION

A problem of classical interest in pattern recognition is that of determining the presence or absence of a particular subpattern or subpattern class. In the analysis of two-dimensional imagery this can take the form of detection of corners and edges or the location of a specific silhouette. More particularly, we may be interested in obtaining an exact match of a specific portion of the image to a subimage, often a prototype, which may appear in an arbitrary manner, varying in size, location and orientation. This is the problem which is herein addressed.

A related question was considered by Dirilten and Newman [3] where it was shown how two planar images could be matched under arbitrary affine transformation of the plane, if a match were at all possible. In addition to affine transformations, an allowance was also made for dilation of intensity scale such as that which results from under or over exposure of film within latitude limits. The results cited, however, are of little use in matching subpatterns, since the algorithms are highly sensitive to the background context. Nevertheless, the utility of a group theoretic approach to pattern matching was clearly demonstrated.

In the following we present a method for performing a local search for an imbedded subpattern of a two-dimensional image. The

*This research was supported by the Army Research Office, Contract DAAG29-80-C-0087 and by the Office of Naval Research, Contract N0014-76-C-1136.

PRECEDING PAGE BLANK-NOT FILMED

method is one involving adaptive control of a retina which seeks the desired subpattern by evolving along a curve in the space of parameters in a direction which assures improvement in the goodness of fit.

2. BACKGROUND

Let G be a Lie group of transformation on an analytic manifold M. Suppose G has dimension n while M has dimension m. Let x and y denote the coordinates of elements f and g in G, respectively, in a patch containing the identity element e of G. Also, let p denote coordinates of an element u of M in some patch in M. We may then express the coordinates z of the product h = fg and the coordinates q of the element v = gu, relative to suitable patches, by means of analytic functions

$$z = J(x,y) \tag{2.1}$$

$$q = K(y,p) \tag{2.2}$$

K and J are vector-valued, having values in n-dimensional space R^n or C^n and m-dimensional space R^m or C^m. Hereafter we shall assume that these underlying spaces are real. We denote the ith component of J by J_i and the jth component of K by K_j.

In order to define the Lie algebra of G we first introduce real-valued maps on G by

$$P_{ij}(x) = \frac{\partial J_i}{\partial y_j}(x,y)\,|_{y=e}, \tag{2.3}$$

where i and j each range from 1 to n. The cross-section P_{*j}, which consists of the P_{ij} as i ranges from 1 to n, and j is fixed, may be thought of as a vector field in R^n. Such a vector field attaches to a point x the vector $P_{*j}(x)$. As such, P_{*1}, $P_{*2},\ldots,P_{*n}$ form a basis for the tangent space at the point x [1,2]. The infinitesimal transformations of G may now be defined by

$$X_j = \sum_{i=1}^{n} P_{ij}(x)\frac{\partial}{\partial x_i}, \tag{2.4}$$

for j = 1,2,...,n.

The differential operators so defined are to be considered as linear operators on the space of analytic functions on G, or, more generally, on the space of differentiable functions on G. The Lie algebra of G is simply the n-dimensional vector space consisting of all linear combinations of these operators, and will be denoted by L(G) [2].

The Lie algebra of G may also be defined in terms of its actions on the manifold M. Analogous to (2.3) we define

$$Q_{\alpha j}(p) = \frac{\partial K_\alpha}{\partial y_j}(y,p)\,|_{y=e} \tag{2.5}$$

for $\alpha = 1,2,\ldots,m$ and $j = 1,2,\ldots,n$. Finally, as in (2.4) above we set

$$X_j' = \sum_{\alpha=1}^{m} Q_{\alpha j}\frac{\partial}{\partial p_\alpha}. \tag{2.6}$$

The operators X_1', X_2', ..., X_n' apply to functions defined on M and span a Lie algebra isomorphic to L(G).

The following result from [4] will be used later, and is stated for reference:

<u>Theorem 2.1</u>. Let $f: M \to R$ be differentialbe and define $F: G \times M \to R$, in terms of coordinates, by

$$F(x,p) = f(K(x,p)). \tag{2.7}$$

Then for each j = 1,2,...,n we have

$$X_jF = X_j'F. \tag{2.8}$$

Let us consider a curve $t \to g(t)$ in G satisfying g(0) = e. In terms of a coordinate patch at e, g(t) may be described by a curve x(t) in R^n satisfying x(0) = 0. We shall consider the case in which x(t) is given as the solution of an evolution equation of the form

$$\dot{x}(t) = \sum_{i=1}^{n} \lambda_i(t)P_{*i}(x(t)),\ x(0) = 0, \tag{2.9}$$

where $P_{*1},\ldots P_{*n}$ are cross-sections of the array of functions given by (2.3), and $\lambda_1(t),\ldots,\lambda_n(t)$ are suitable control functions.

Now let p denote the coordinates of a point u in some coordinate patch. For a differentiable map $f: M \to R$ we may define $H: R \times M \to R$ by setting

$$H(t,p) = f(g(t)u). \qquad (2.10)$$

We recognize that $H(t,p) = F(x(t),p)$ where F is the extension of f to $G \times M$ as in Theorem 2.1 above. From the point of view of application, if we regard $f: M \to R$ as an image, then $H(t,p)$ represents the moving image obtained by translation due to the curve $g(t)$. Also from [4], we have

Theorem 2.2. In the context above,

$$\frac{\partial H}{\partial t} = \sum_{i=1}^{n} \lambda_i(t) X_i' H. \qquad (2.11)$$

3. THE CONTROL MODEL

By an image we mean a map $f: M \to R$, where the value $f(p)$ at a point $p \in M$ represents the gray value at the picture element at p. In practice, values are observed on a subset W M, which we regard as a window which may be translated by the action of G on M. Thus, upon translation by an element $x \in G$, the value observed at $p \in W$ is given by $F(x,p) = f(K(x,p))$, as in (2.7) above.

We consider a given prototype sub-image V defined on the window W, $V: W \to R$. The problem then is to determine $x \in G$ such that $F(x,p) = V(p)$ for all $p \in W$, or determine that no such x exists. As a matter of practice, we seek $x \in G$ which minimizes the objective function

$$\Psi(x) = \frac{1}{2} \int_W (F(x,p) - V(p))^2 dp, \qquad (3.1)$$

where dp represents a volume element and the integral is over the window W, which is assumed to be of bounded volume.

In general, for any two functions f_1, f_2: $W \to M$ we define

$$\langle f_1, f_2 \rangle = \int_W f_1 f_2 dp \quad \text{and}$$

$$||f_1|| = \langle f_1, f_1 \rangle^{1/2} .$$

Thus, $\Psi(x) = ||F - V||^2/2$, where x is regarded as a parameter.

The following is a well-known property of the Lie group G [2]:

Lemma 1. In order that the differential $d\ (x) = 0$ at a point $x \in G$, it is necessary and sufficient that each $X_i\Psi(x) = 0$ where $X_1, X_2, \ldots, X_n$ are the generators of L(G) given by (2.4).

By direct calculation, we obtain $X_i\Psi(x) = \int_W (F(x,p) - V(p)) X_i F(x,p) dp$. In practice, this expression is difficult to compute numerically, due to the presence of the term $X_i F$, which cannot be computed directly from observed data. However, by Theorem (2.1) we have $X_i F = X_i' F$, and the latter can be calculated from a single value of x.

Suppose now that a curve in G is given by coordinates $x(t)$ obtained as a solution of Equation (2.9). We seek to find $\lambda(t) = (\lambda_1(t), \ldots, \lambda_n(t))$ so that $\psi(t) = \Psi(x(t))$ decreases to a minimum value. Defining $H(t,p) = F(x(t),p)$ we obtain,

$$\dot{\psi}(t) = \int_W (H(t,p) - V(p)) \frac{\partial H}{\partial t}(t,p) dp \qquad (3.2)$$

which, by application of Theorem (2.2), becomes

$$\begin{aligned} \dot{\psi}(t) &= \sum_{i=1}^{n} \lambda_i(t) \int_W (H(z,p) - V(p)) X_i' H(t,p) dp \\ &= \sum_{i=1}^{n} \lambda_i(t) \langle H - V, X_i' H \rangle \end{aligned} \qquad (3.3)$$

Upon observing that $\langle H - V, X_i' H \rangle = \langle F - V, X_i F \rangle = X_i \Psi$ at $x = x(t)$, we deduce:

Theorem 3.1. If $\lambda_i(t)$ is chosen so that $\operatorname{sgn} \lambda_i(t) = -\operatorname{sgn} \langle H - V, X_i' H \rangle$, we have $\dot{\psi}(t) \le 0$ for all t, with equality at $t = t_0$ if and only if $d\Psi = 0$ at $x = x(t_0)$.

Among the class of bounded controls, $|\lambda_i - (t)| \le 1$, we see that the rate of decrease of (t) is maximized by the choice

$$\lambda_i(t) = -\operatorname{sgn} \langle H - V, X_i' H \rangle, \qquad (3.4)$$

for $i = 1,2,\ldots,n$. Of course, other strategies can be formulated, including steepest descent, and some methods using unbounded controls. By proceeding along trajectories defined by the solution of (2.9) with $\lambda(t)$ given by (3.4), we approach a critical point of Ψ (i.e. $\dot{d} = 0$). Since maxima and saddle points are unstable under perturbation, in practice this extreme point will always be a minimum.

4. SIMULATION RESULTS

The results discussed in the previous section have been implemented by a discrete algorithm and tested on simulated data [5]. A digitized two-dimensional image was first generated in the form of a large two-dimensional array, and the prototype was generated in a 20 × 20 window array.

The image space was assumed to be subject to translation, magnification and rotation, giving rise to a four parameter Lie group of transformations in the plane, R^2.

A number of cases were considered, including some involving multiple (false) targets and others in which the prototype was absent from the image being searched. In some cases the image was contaminated by 5% random noise. In all cases the search was started with overlap between the prototype target and the image target.

The differential equation (2.9) was solved by means of a Runge-Kutta fourth order method, with a dynamic step size, which was increased as necessary to accelerate convergence and decreased as necessary to maintain stability. Integration was replaced by summation, although we conjecture that convergence could have been accelerated by the use of a trapezoid rule.

Generally, search times ranged from 30 to 50 steps, with the longer search times prevailing for the more difficult cases.

In all cases, the final results were quite reasonable, even in those cases where the prototype was absent. In the latter cases, the search terminated with a "best" match, with a commensurately large final error.

As an example, Figure 1 shows that starting position for a noisy image containing two objects. The prototype is indicated by the central silhouette, while the true target is shifted upward, slightly to the right and is reduced in size. A false target overlaps the lower right corner of the prototype.

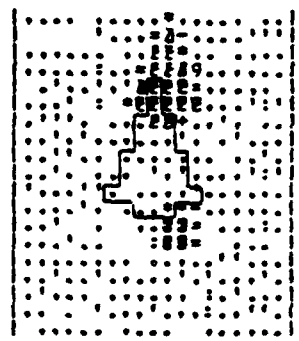


Fig. 1. Initial Window Position.

The termination conditions are shown in Figure 2, where the true target was located after 49 steps. All parameters were correct with the exception of magnification, which was about 5% too large. Smaller values of magnification, however, increase the error due to the presence of the false object, which is barely touching the bottom edge of the window in Figure 2.

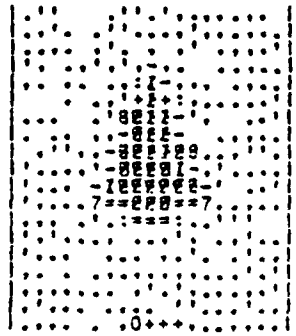


Fig. 2. Terminal Window Position.

BIBLIOGRAPHY

[1] Auslander, L., Differential Geometry, Harper and Row, New York, 1967.

[2] Cohn, P. M., Lie Groups, Cambridge University Press, London, 1957.

[3] Dirilten, H. and T. G. Newman, Pattern Matching under Affine Transformations, IEEE Trans. Comp., Vol. C-24, No. 12, 1975, pp. 1191-1201.

[4] Newman, T. G. and D. A. Demus, Lie Theoretic Methods in Video Tracking, Proceedings of the MICOM Workshop on Imaging Trackers and Autonomous Acquisition Applications, Redstone Arsenal, Nov. 1979.

[5] Zlobec, L., Pattern Matching by Means of Adaptive Control, Masters Report, Texas Tech University, 1980.

DIRECTOR'S DISCRETIONARY FUND

R. SAEKS

PRECEDING PAGE BLANK-NOT FILMED

Texas Tech University | Institute for Electronic Science

Joint Services Electronics Program | Research Unit: 7

1. Title of Investigation: Director's Discretionary Fund

2. Senior Investigator: R. Saeks Telephone: (806) 742-3528

3. JSEP Funds: Current $20,081

4. Other Funds: Current_____

5. Total Number of Professionals: To be determined

6. Summary:

During the past year the director's discretionary fund has been used to fund several preliminary investigations into the VLSI design problem, to initiate a new investigation directed at the application of parallel processing techniques in system theory, and to follow up on a previous JSEP work unit on large scale systems.

7. Publications and Activity:

A. Refereed Journal Publications

1. Karmokolias, C., Portnoy, W.M., and R. Saeks, "Optimal Selection of IC Fabrication Parameters", Circuit Theory and its Applications, Vol. 9, pp. 211-215, (1981).

2. Green, B., Iyer, A., Saeks, R., and K.-S. Chao, "Continuation Algorithms for the Eigenvalue Problem", Circuits, Systems, and Signal Processing, (to appear).

B. Theses

1. Tai, C.-T., M.S. Thesis, (in preparation).

PRECEDING PAGE BLANK-NOT FILMED

OPTIMAL SELECTION OF IC FABRICATION PARAMETERS

C. KARMOKOLIAS

W. M. PORTNOY

AND

R.E. SAEKS

PRECEDING PAGE BLANK-NOT FI

OPTIMAL SELECTION OF IC FABRICATION PARAMETERS

C. KARMOKOLIAS, W. M. PORTNOY AND R. E. SAEKS
Department of Electrical Engineering, Texas Tech University, Lubbock, Texas 79409. U.S.A.

SUMMARY

A procedure is described in which the output characteristics of an integrated circuit are optimized with respect to a set of variable fabrication parameters. A simple *RC* coupled audio amplifier is used as an example. The gain–bandwidth product is obtained as a function of oxidation and diffusion times and temperatures, and the optimization is performed by way of a line search using these variables as the parameters of the optimization. The values established for the process parameters are consistent with those employed for conventional fabrication, and desired changes in performance can be obtained, in general, by a straightforward readjustment of the values of the process variables. Although limited by certain assumptions and a relatively primitive circuit, the results demonstrate the validity of the procedure.

1. INTRODUCTION

The manufacture of an integrated circuit can be considered as a three-stage procedure. Initially, performance requirements are provided or established by the circuit application. These requirements suggest an interconnection of passive and active elements whose values and geometry are determined during the second stage. Finally, a suitable process is chosen to fabricate the circuit design of the previous stages. Every fabrication process generating the necessary impurity profiles is controlled by a number of independent variables which must be assigned appropriate values. The specification of these values is an integral part of the design and constitutes an implicit relation between the performance requirements and the fabrication process. In practice, the process is usually known in advance, but the independent variables must still be specified.

A situation which often occurs is one where a circuit is desired whose output characteristics are close and yet not identical to those of a generically related circuit for which the entire three-stage design has been completed. What is often done in these cases is to introduce appropriate changes in the second stage of the procedure and then redesign the third. Because of the number of iterations involved, this practice turns out to be quite laborious and expensive. In addition, more often than not, the new design calls for a new geometry, and the creation of new masks adds significantly to labour, cost and delays.

The technique described here introduces optimal changes at the fabrication stage, several steps beyond the circuit design stage. A given process model is incorporated into a given circuit model and the predicted output characteristics are optimized with respect to a specified set of variable fabrication parameters. The predicted output characteristics can be quite accurate if accurate circuit and process models are employed. There is certainly an abundance of reliable circuit models; very detailed process models are available also, (see, for example, Reference 1). The optimization is performed by way of an index which measures the difference between desired and predicted performance over a physically prescribed range of the fabrication variables. For all practical purposes, the fabrication of the new circuit is obtained without any delay and is essentially cost-free. Additional constraints can, of course, be imposed to resolve questions of realizability, sensitivity, thermal variation, and so on.

The technique is particularly useful for those integrated circuits which are designed on a modular basis. In these cases, a circuit consisting of several modules is desired, where most or all of the modules may have been previously designed independently. The designs are frequently incompatible; the technique would then provide a unified fabrication design using the existing geometry of the given modules.

PRECEDING PAGE BLANK-NOT FILMED

The following work applies the technique to the fabrication of an audio amplifier which might itself form the modular basis for an operational amplifier array. Although simplified circuit and process models are used, the results indicate the feasibility of the method, which is the objective of this paper.

2. CIRCUIT MODEL

The amplifier circuit, which incorporates an *n–p–n* transistor, is illustrated in Figure 1a. The specified output function is the voltage gain, $|A(\omega)|$, which depends on the circuit variables shown in Figure 1b; these, excepting the emitter bypass and coupling capacitors, depend on impurity concentration profiles established by the process. Because the emitter bypass capacitance is usually quite large, it was convenient to make it a fixed external capacitor. The values of C_1 and C_2 are very large and do not enter the calculations.

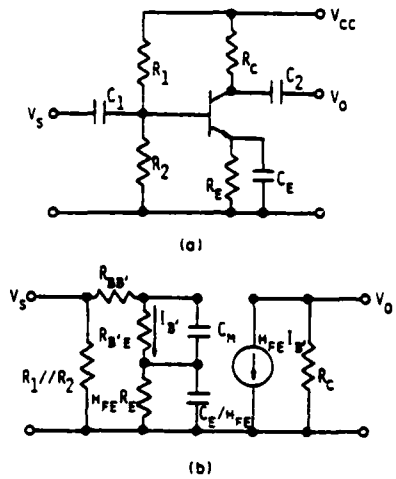


Figure 1. (a) Audio amplifier circuit; (b) equivalent circuit of the audio amplifier

The resistances $r_{bb'}$ and $r_{b'e}$ are obtained from the fixed dimensions of the base and emitter regions and the base sheet resistance. Distributed resistances R_1, R_2, $R_{e'}$ and R_c are formed during the base diffusion, and also depend on the base sheet resistance; their dimensions are fixed. The Miller capacitance, C_M, includes the depletion capacitances of the reverse-biased base-collector and forward-biased base-emitter junctions (charge storage in the base is neglected).

The average impurity concentrations in the emitter and base regions (and the fixed uniform collector concentration) determine the junction capacitances. These averages, and the base sheet resistance (which includes a constant hole mobility term), require the emitter and base impurity profiles and the junction depths in their calculation; the latter are obtained by equating emitter and base, and base and collector, concentrations. The current gain, h_{fe}, is calculated from the average concentrations, the distance between the junctions, and the minority carrier lifetimes and diffusion constants. Although these latter are complicated functions of impurity concentration,[2] approximate expressions can be obtained.

3. PROCESS MODEL

Each term appearing in the expression for the gain is implicitly related, through the impurity concentration profiles, to the parameters of the process. The concentration profiles are the solutions of the diffusion equation,

$$\frac{\partial N(x,t)}{\partial t} = D(T)\frac{\partial^2 N(x,t)}{\partial x^2}$$

Table I. Optimal fabrication parameters for specified gain in the bandwidth 20 Hz to 20 kHz. Gain is in dB, time (t) is in seconds, and temperature (T) is in °C

Parameter	Gain 20	25	30	35	40
t_b	3,300	7,140	7,200	4,690	6,210
T_b	1,100	1,130	1,170	1,200	1,200
t_1	1,200	1,200	1,230	1,230	1,260
T_1	1,190	1,190	1,190	1,180	1,180
t_e	2,400	2,400	2,370	2,310	2,370
T_e	1,150	1,150	1,150	1,150	1,150
t_2	1,230	930	930	870	870
T_2	1,100	1,100	1,090	1,090	1,090

Table II. Optimal fabrication parameters for specified gain in the bandwidth 10 kHz to 40 kHz. Gain is in dB, time (t) is in seconds, and temperature (T) is in °C

Parameter	Gain 30	35	40
t_b	5,800	6,330	4,140
T_b	1,160	1,200	1,200
t_1	1,150	1,260	1,150
T_1	1,180	1,190	1,180
t_e	2,370	2,340	2,370
T_e	1,150	1,160	1,150
t_2	880	630	940
T_2	1,090	1,090	1,090

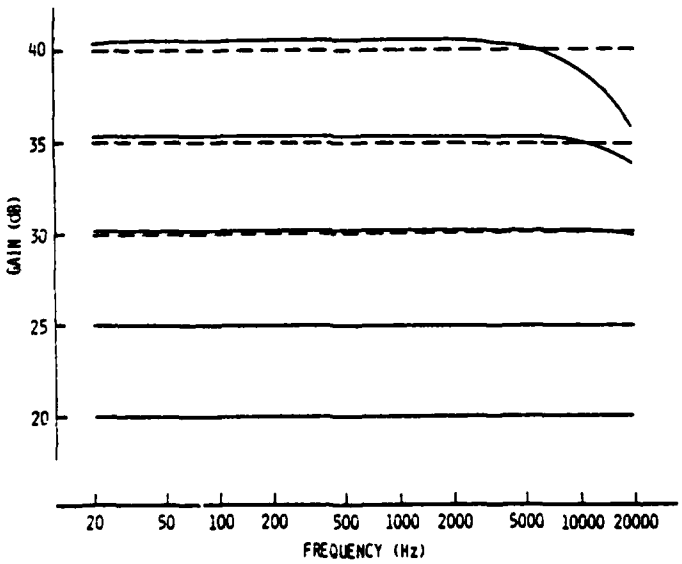


Figure 2. Specified (dashed lines) and optimally obtained (solid lines) gain *versus* frequency in the range 20 Hz to 20 kHz

assuming one-dimensional diffusion and a diffusion coefficient which depends on temperature only. The process steps establish the boundary conditions on the solutions. A typical process sequence was chosen in this work, consisting of a base diffusion blocking oxidation; a p-type base deposition followed by a base drive-in diffusion; an emitter diffusion blocking oxidation; an n-type emitter diffusion; and a pre-metalization oxidation. The bias resistors are derived from the base diffusion. Photolithography and etching were assumed to take place wherever required; metalization is not considered to affect the impurity profiles. The circuit variables then depend on the times and temperatures of the process sequence. These times and temperatures are the independent parameters of the optimization; quantities which do not depend on them are considered constant. Eight significant process parameters were identified; these are

t_b, the base diffusion time;
T_b, the base diffusion temperature;
t_1, the emitter diffusion blocking oxide growth time;
T_1, the emitter diffusion blocking oxide growth temperature;
t_e, the emitter diffusion time;
T_e, the emitter diffusion temperature;
t_2, the pre-metalization oxidation time; and
T_2, the pre-metalization oxidation temperature.

These parameters were optimized assuming a fixed set of mask dimensions to obtain the desired frequency response, $|A(\omega)|$.

4. OPTIMIZATION

The gain, $|A(\omega)|$, was calculated for an inital point, that is, for an 8-tuple of the independent parameters, t_b, T_b, t_1, T_1, t_e, T_e, t_2, T_2. $|A(\omega)|$ was then compared with the desired gain, A_0, by way of the integral

$$J = \int_{\omega_1}^{\omega_2} [A_0 - |A(\omega)|]\, d\omega$$

where $\omega_2 - \omega_1$ is the desired bandwidth. J is minimized using a simple line search, that is, by varying each independent parameter while holding the others constant. In order to ensure reasonable values for the optimal process variables, these were constrained; the maximum and minimum values permitted for temperature were 1,200°C and 900°C, respectively, and for time, 7200 s (2 h) and 600 s (10 min), respectively.

The optimization was performed first for several gains in the bandwidth 20 Hz to 20 kHz, and again for the higher gains in the higher frequency range, 10 kHz to 40 kHz. The integral was evaluated using the trapezoidal rule. 1,000 Hz increments were used throughout the calculation for the higher bandwidth; for the lower, 1,000 Hz increments were used above 100 Hz, and 20 Hz increments below 100 Hz. Temperature and time were varied in 2°C and 30 s increments, respectively.

5. RESULTS AND DISCUSSION

The results of the optimization, obtained with two to three iterations, are tabulated in Tables I and II. The values which were obtained for the elements of the equivalent circuit were quite reasonable in terms of a conventional circuit design; however, these values are not particularly important, because they are only the results of the optimization, and do not participate in it. Even atypical values would not be significant, as long as they could be obtained within the constraints set on the process times and temperatures.

Figures 2 and 3 illustrate the behaviour of the optimal gains *versus* frequency in the two frequency ranges. Changes which match moderate frequency requirements are obtained easily by an optimal readjustment of the values of the fabrication parameters. As requirements become more severe, as at the higher frequencies of Figure 3, deviation of circuit gain from desired values increases, finally becoming so large that a redesign at circuit level is unavoidable.

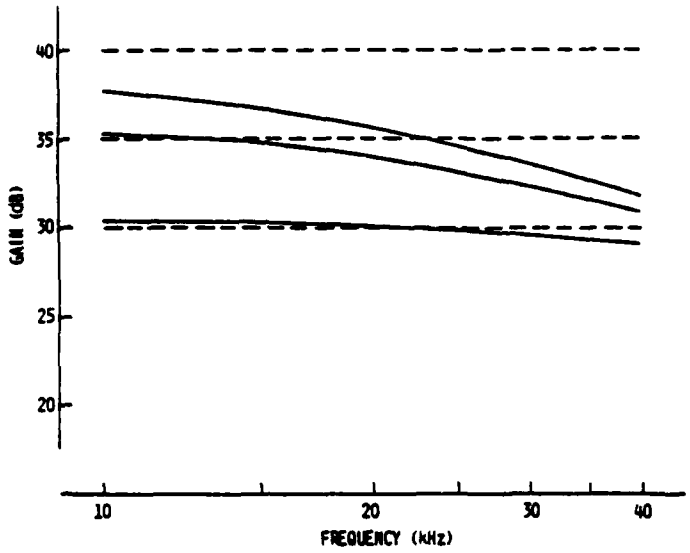


Figure 3. Specified (dashed lines) and optimally obtained (solid lines) gain *versus* frequency in the bandwidth 10 kHz to 40 kHz

There are two major limitations on the accuracy of the calculation. First, the process model suffers from certain deficiencies because of several assumptions which were made. For example, the emitter blocking oxide must be grown thick enough to be effective. For the tabulated values of t_1 and T_1, the oxide thickness is around 2,000 Å,[3] which is not thick enough to block the required 40 min phosphorous emitter diffusion.[3] However, a thickness constraint can be introduced, although it was not done here, to assure appropriate oxide thickness. Also, the line search which was used to obtain the optimum does not guarantee a global minimum. It is important to emphasize again that these results are not intended to establish a practical process, but to demonstrate the use and feasibility of the technique.

ACKNOWLEDGEMENT

This research was supported in part by the Joint Services Electronics Program at Texas Tech University, under ONR Contract 76–C–1136.

REFERENCES

1. D. A. Antoniadis and R. W. Dutton, 'Models for computer simulation of complete IC processes', *IEEE Trans. Electron Devices*, **ED-26**, 490–500 (1979).
2. S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, 1969.
3. S. K. Ghandi, *The Theory and Practice of Microelectronics*, Wiley, New York, 1968.

ABSTRACT OF

CONTINUATION ALGORITHMS FOR THE EIGENVALUE PROBLEM

B. GREEN, A. IYER, R. SAEKS AND K.-S. CHAO

PRECEDING PAGE BLANK-NOT FILMED

Continuation Algorithms for the Eigenvalue Problem*

B. Green, A. Iyer, R. Saeks and K.-S. Chao
Department of Electrical Engineering
Texas Tech University
Lubbock, Texas 79409

Abstract

Three algorithms for the solution of the eigenvalue problem for a continuous parameterized family of sparse matrices are presented; a continuous LU (or LR) algorithm, a continuous QR algorithm, and a continuous Hessenberg algorithm. Each of the three algorithms may be implemented recursively and the sparsity of the given matrices is preserved throughout the numerical process.

PRECEDING PAGE BLANK-NOT FILMED

*This research supported by the Joint Services Electronics Program at Texas Tech under ONR Contract 76-C-1176.

Grants and Contracts Administered by JSEP Personnel

A. Funded

Hunt, L.R., NASA Grant in support of Professor Hunt's leave of absence at NASA/AMES, $52,090, 1 yr.

Krile, T.F., NSF Grant SER-800-1394, "Fiber Optic Experiments for Undergraduate Engineers", $31,533, 2 yrs.

Krile, T.F., SORF Grant, "Fiber Optic Experiments for Undergraduate Engineers", $11,896, 2 yrs.

Krile, T.F., E. Systems Corp. Contract, "Digital and Optical Signal Processing and Detection", $48,973, 2 yrs.

Murray, J., AFOSR Grant 80-0156, "The Application of Crossed Products to the Stability and Design of Time-Varying Systems", $47,365, 2 yrs.

Saeks, R., ONR Contract 76-V-1136, "Joint Services Electronics Program", $1,206,500, 6 1/2 yers. (with JSEP Staff).

Saeks, R., NSF Grant ENG-78-24414, "Frequency Domain-Like Methods for the Analysis and Design of Nonlinear and Time-Varying Systems", $79,240, 3 yrs.

Walkup, J.F., AFOSR Grants 75-2855 and 79-0076, "Space-Variant Optical Systems", $570,954, 7 1/4 yrs.

Walkup, J.F., SPIE Grant, "Optical Engineering Education", $3,500, 2 yrs.

Walkup, J.F., ARO Contract DAAG 29-80-0110, "Workshop on Future Directions for Optical Information Processing", $11,867, 1 yr.

Total Annual Funding: $491,754

B. Proposed

Nakajima, K., Proposal to ONR, "Scheduling and Parallel Computation for Sparse Matrices", $106,152, 3 yrs.

Nakajima, K., Proposal to NSF, "A Study of t-Fault Diagnosibility in Analog and Digital Self-Testing Systems", $30,800, 1 yr.

Saeks, R., Proposal to AFOSR, "Feedback Systems and the Simultaneous Design Problem", $35,743, 1 yr.

Saeks, R., Proposal to ONR "Softwar Development Program for Analog Fault Diagnosis", $241,880, 2 yrs.

PRECEDING PAGE BLANK-NOT FILMED

GRANTS AND CONTRACTS IN ELECTRICAL ENGINEERING*

A. Systems

Saeks, R., ONR Contract, "Joint Services Electronics Program," $220,000.

Saeks, R., NSF Grant, "Frequency Domain-Like Methods for the Analysis and Design of Time-Varying and Nonlinear Systems," $25,740.

Saeks, R., State of Texas Matching, "Frequency Domain-Like Methods for the Analysis and Design of Time-Varying and Nonlinear Systems," $10,000.

Walkup, J.F., AFOSR Grant, "Space Variant Optical Systems," $103,420.

Murray, J., AFOSR Grant, "The Application of Crossed Products to the Stability and Design of Time-Varying Systems," $25,274, 1 yr.

Chao, K.-S., NSF Grant "Continuation Methods in Nonlinear Network Analysis," $22,376.

Gustafson, D., and T. Krile, E-Systems Corp. Contract, "Digital and Optical Signal Processing Data," $50,000.

Total Annual Funding in Systems: $456,810 .

B. Electro Physics

Trost, T., NASA Grant, "Lightning Sensors and Data Interpretation," $63,300.

Hagler, M.O., NSF Grant, "Investigation of RF Plasma Heating in Toroidal Geometry," $55,000.

Hagler, M.O., State of Texas Matching, "Investigation of RF Plasma Heating in Toroidal Geometry," $3,270.

Portnoy, W.M., NRL Contract "Reliability Study of Refractory Gate Gallium Arsenide MESFETS," $49,033.

Portnoy, W.M., AFOSR Grant, "Investigation of the Physics of Failure in Semiconductor Resulting from Electrical Transients," $49,523.

Total Annual Funding in Systems: $230,126 .

C. Pulsed Power Research

Kristiansen, M., AFOSR Grant, "Pulsed Power Research Colloquim," $34,000.

* Funding shown represents the annual funding for most recent year.

Kristiansen, M., AFOSR Grant, "Coordinated Research Program in Pulsed Power Physics", $639,170.

Kristiansen, M., ARO Contract, "Coordinated Research Program in Pulsed Power Physics", $112,577.

Kunhardt, E., NATO/NSWC/AFWL Contract, "Nato Advance Study Institute", $60,000.

Kunhardt, E., LASL Contract, "Hot Electron Phenomena in Semiconductor", $10,000.

Kunhardt, E., State of Texas Matching, "Hot Electron Distribution Function", $5,000.

Kunhardt, E., NSWC Contract, "Breakdowns at High Overvoltage", $116,332.

Schoenback, K., ARO Contract, "Workshop on Diffuse Discharge Modelling", $12,650.

Kunhardt, E., OnR Contract, "Non-Stationary Ionization Phenomena in Gases" $303,000.

Portnoy, W.M., RADC Contract, "Fast Transient Turn-on in Thristor Switch", $60,000.

Total Annual Funding in Pulsed Power Research: $1,352,729.

D. Power Systems

Craig, J.P., Texas Power and Light Co., "Power System Studies", $8,000.

Reichert, J.D., DOE Contract, "Crosbyton Solar Power Project", $500,000.

Total Annual Funding in Power Systems: $508,000.

E. Other

Seacat, R.H., State of Texas Matching, "Research and Development in Electrical Engineering", $19,557.

Walkup, J.F., SPIE Contract, "Optical Engineering Education", $2,000.

Krile, T., NSF Grant, "Fiber Optics Experiments for Undergraduate Engineers", $10,717.

Total Annual Funding in Other Activities: $32,274.

F. Sources of Funding in Electrical Engineering

Air Force	$ 921,387
Army	125,227
Navy	688,365
**DOE	510,000
NASA	63,300
State of Texas	37,827
NSF	113,833
Industry	60,000
NATO	60,000
Total Annual Funding in Electrical Engineering	$2,579,939

**This includes $500,000 of approximately $1,000,000 of funding for the Crosbyton Solar Project.

Grants and Contracts in Mathematics

Anderson, R., and Wayne Ford, State of Texas, "Mathematical Methodology for Evaluating Simulations of Flow in Porous Media" 1½ yrs. $10,000.

Barnard, R., NSF Grant, "Some Extremal Problems in Complex Function Theory", 2½ yrs., $8,023.

Emerson, W., State of Texas, " Models for Plasmid Replication of Partition in Biology", 1 yr. $500.

Ford, W., and R. Anderson, DOE, "Mathematical Methodology for Evaluating Simulations of Flow in Porous Media", 2 yrs. $70,683.

Lutzer, D., NSF Grant, "Abstract Spaces, Function Spaces & Ordered Spaces", 3 yrs. $8,128.

Nelson, P., State of Texas, "Computational & Mathematical Aspects of Radiation Transport", 2½ yrs. $5,000.

Nelson, P., NSF Grant, "Computational & Mathematical Aspects of Radiation Transport", 2½ yrs. $34,162.

Strauss, M., NSF Grant, "Uniqueness and Norm Convexity in the Cauchy Problem", 1½ yrs. $4,416.

Sources of Funding in Mathematics

DOE	$ 70,683
NSF	54,729
State of Texas	15,500
Total Annual Funding in Mathematics	$ 140,912

Publications for JSEP Personnel*

A. Refereed Journal Articles

Froehlich, G.K., Walkup, J.F., and T.F. Krile, "Estimation in Signal-Dependent Film-Grain Noise", Applied Optics, Vol. 20, pp. 3619-3626, (1981, JSEP).

Hagler, M.O., Marks, R.J., II, Kral, E.L., Walkup, J.F., and T.F. Krile, "Scanning Technique for Coherent Processors", Appl. Optics, Vol. 19, pp. 4253-4257 (1980, AFOSR).

Hunt, L.R., "Controllability of Nonlinear Hypersurface Systems", in Algebraic and Geometric Methods of Linear System Theory (eds. C.I. Byrnes and C.F. Martin), Providence, AMS, pp. 209-224, (1980, JSEP)

Kasturi, R., Krile, T.F., and J.F. Walkup, "Multiplex Holography for Space-Variant Processing: A Transfer Function Sampling Approach", Appl. Optics, Vol. 20, pp. 881-886, (1981, AFOSR).

Murray, J., "Lumped-Distributed Networks and Differential Delay Systems", in Algebraic and Geometric Methods in Linear System Theory, Providence, AMS, (1980, JSEP).

Saeks, R., and J. Murray, "Feedback System Design: The Tracking and Disturbance Rejection Problems", IEEE Trans. on Auto. Cont., Vol. AC-26, pp. 203-217, (1981, JSEP).

Saeks, R., and R.-w. Liu, "Fault Diagnosis in Electronic Circuits", Jour. of the Soc. of Instr. and Cont. Engrgs. Vol. 20, pp. 20-22, (1981, a preliminary version of this paper also appeared in the IEEE CHMT Society Newsletter, Vol. 3, No. 3, 1980, JSEP).

B. Conference Papers and Abstracts

Carson, R.F., Walkup, J.F., and T.F. Krile, "Tristimulus-Based Approach to Incoherent Optical Processing", J. Opt. Soc. Am., 1594A. (1981). Paper presented at the 1981 Annual Meeting, Optical Society of America, Kissimmee, FL, Oct. 1981. (Abstract in J. Opt. Soc. of Am., Vol. 70, p. 1594A, 1981, AFOSR).

Froehlich, G.K., Walkup, J.F., and T.F. Krile, "Some Effects of Signal-Dependent Noise on Estimation Structures", presented at the Annual Meeting of the Optical Soc. of Amer., Oct. 1980, (abstract in the Jour. of OSA, Vol. 20, p. 613, JSEP).

Hunt, L.R., Meyer, G., and R. Su, "Transformations of Nonhomogeneous Nonlinear Systems", Proc. of the 19th Allerton Conf. on Communications, Control and Computing, Oct 1981, (to appear, JSEP/NASA).

*Includes all publications by JSEP personnel with source of support.

PRECEDING PAGE BLANK-NOT FI

Hunt, L.R., and R. Su, "Local Transformations for Multi-input Nonlinear Systems", Proc. of the Joint Auto. Cont. Conf., Charlottesville, June 1981, paper FA3B, (JSEP/NASA).

Hunt, L.R. and R. Su, "Global Mappings of Nonlinear Systems", Proc. of the Joint Auto Conf., Charlottesville, June 1981, paper FA3C, (JSEP/NASA).

Hunt, L.R., and R. Su, "Linear Equivalents of Nonlinear Time-Varying Systems", Proc. of the Inter. Symp on the Mathematics of Networks and Systems, Santa Monica, Aug. 1981, pp. 119-123, (JSEP/NASA).

Hunt, L.R., and R. Su, "Poincare Lemma and Transformations of Nonlinear Systems", Proc. of the Inter. Symp. on the Mathematics of Networks and Systems", Santa Monica, Aug. 1981, pp. 111-118, (JSEP/NASA).

Hunt, L.R., and R. Su, "Transforming Nonlinear Systems", Proc. of the 24th Midwest Symp. on Circuits and Systems, Albuquerque, June 1981, pp. 341-345, (JSEP/NASA).

Jones, B.H., Walkup, J.F., and T.F. Drile, "Hybrid Laser Plotter for Optical", J. Opt. Soc. Am., 1595A (1981). Paper presented at the 1981 Annual Mtg., Optical Society of America, Rissiminee, FL. Oct. 1981., (Abstract in J. Opt. Soc. of Am., Vol. 70, p. 1595A, 1981, AFOSR).

Karmokolias, C., and R. Saeks, "Suboptimal Control with Optimal Quadratic Regulators", Proc. of the Conf. on Information Sciences Systems, Johns Hopkins Univ., April 1981, pp. 53-58, (JSEP).

Karmokolias, C., and R. Saeks, "A Fractional Representation Approach to Adaptive Control", Proc. of the IEEE Conf. on Decision and Control, Albuquerque, NM, Dec. 1980, pp. 272-273, (JSEP).

Kasturi, R., Krile, T.F., and J.W. Walkup, "Space-Variant Processing Techniques - a Comparison". Paper presented at the 1980 Annual Meeting, Optical Society of America, Chicago, IL, Oct. 1980. (Abstract in J. Opt. Soc. of Am. Vol-70, P. 1590A, 1980, JSEP).

Murray, J., "A Time-Varying Approach to Two-Dimensional Digital Filtering", Proc. of the 24th Midwest Symp. on Circuits and Systems, Albuquerque, July 1981, pp. 351-355, (AFOSR).

Murray, J., "The Design of 2-D Filters as 1-D Time-Varying Systems", to be presented at the 1982 IEEE International Conference on Acoustics, Speech and Signal Processing, (JSEP).

Murray, J., and R. Saeks, "Simultaneous Design of Control Systems", Proc. of the EEE Conf. on Decision and Control, San Diego, Dec. 1981, (to appear, JSEP).

Saeks, R., "Criteria for Analog Fault Diagnosis" Proc. of the European Conf. on Circuit Theory and Design, The Hague, Aug. 1981, pp. 75-78, (JSEP).

Saeks, R., and J. Murray, "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem", Proc. of the IEEE Inter. Symp. on Circuits and Systems, Chicago, April 1981, pp. 463-464, (JSEP).

Walkup, J.F., "Space-Variant Coherent Optical Processing", invited paper presented at the Second SPSE Symposium on Optical Data Display, Processing, and Storage, Las Vegas, NV, March, 1981, (AFOSR).

Wu, C.-c., Nakajima, K., Wey, C.-L., and R. Saeks, "Analog Fault Diagnosis with Failure Bounds", Proc. of the 24th Midwest Symp. on Circuits and Systems, Albuquerque, June 1981, pp. 515-520, (JSEP).

Wu, C.-c., Sangiovanni-Vincentelli, A., and R. Saeks, "A Differential - Interpolative Approach to Analog Fault Simulation", Proc. of the IEEE Inter. Symp. on Circuits and Systems, Chicago, April 1981, pp. 266-269, (JSEP).

C. Preprints

Hunt, L.R., "N-dimensional Controllability with n-1 Controls", IEEE Trans. on Auto. Cont., (to appear, JSEP).

Hunt, L.R., "Sufficient Conditions for Controllability", IEEE Trans on Circuits and Systems, (to appear, JSEP).

Jones, M.I., Walkup, J.F., and M.O. Hagler, "Multiplex Hologram Representations of Space-Variant Optical Systems Using Ground Glass Encoded Reference Beams", (in press, Appl. Optics, AFOSR).

Kral, E.L., Walkup, J.F., and M.O. Hagler, "Correlation Properties of Random Phase Diffusers for Multiplex Holography", (in press, Appl. Optics, AFOSR).

Murray, J., "A Design Method for 2-Dimensional Recursive Digital Filters", IEEE Trans. on Acoustics, Speech, and Signal Processing, (to appear, February 1982, JSEP).

Saeks, R., Murray, J., Chua, O., Karmokolias, C., and A. Iyer, "Feedback System Design: The Single Variate Case", Circuits, Systems, and Signal Processing, (to appear, JSEP).

Saeks, R., Sangiovanni Vincentelli, A., and V. Vishvanathan, "Diagnosibility of Nonlinear Circuits and Systems - Part II Dynamical Systems, IEEE Trans. on Computers/Circuits and Systems, (to appear, JSEP).

Wu, C.-c., Nakajima, K., Wey, C.-L., and R. Saeks, "Analog Fault Diagnosis with Failure Bounds", IEEE Trans. on Circuits and Systems. (to appear, JSEP).

Wu, C.-c., and R. Saeks, "A Data Base for Symbolic Network Analysis", IEE Proc. Part G, (to appear, JSEP).

DAT
FILM
4-