AD-A077 652 TEXAS TECH UNIV LUBBOCK INST FOR ELECTRONICS SCIENCE F/G 9/4
ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICE ELECTRONICS P--ETC(U)
OCT 79 R SAEKS , K S CHAO , J WALKUP N00014-76-C-1136

UNCLASSIFIED NL

1 OF 2
ADA 077652

LEVEL III

AD A077652

AD A059977

ANNUAL REVIEW OF RESEARCH

under the

JOINT SERVICES ELECTRONICS PROGRAM

October 1979

Institute for Electronics Science

TEXAS TECH UNIVERSITY
Lubbock, Texas 79409

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

(6) ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICE ELECTRONICS PROGRAM

5. TYPE OF REPORT & PERIOD COVERED

(9) INTERIM (reptg)

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

(16) R. SAEKS, K.S. CHAO, J. WALKUP, T. NEWMAN
L.R. HUNT, G.J. MURRAY

8. CONTRACT OR GRANT NUMBER(s)

(15) N0014-76-C-1136

9. PERFORMING ORGANIZATION NAME AND ADDRESS

TEXAS TECH UNIVERSITY
INSTITUTE FOR ELECTRONIC SCIENCE
LUBBOCK, TEXAS

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

122105

11. CONTROLLING OFFICE NAME AND ADDRESS

(11) 1 Oct 79

12. REPORT DATE

10/1/79

13. NUMBER OF PAGES

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

OFFICE OF NAVAL RESEARCH
800 N. QUINCY AVENUE
ARLINGTON, VA.

(12) 165

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE-DISTRIBUTION UNLIMITED

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

18. SUPPLEMENTARY NOTES

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

FAULT ANALYSIS
CONTROL AND ESTIMATION
NONLINEAR CONTROL
MULTIDIMENSIONAL SYSTEM THEORY
OPTICAL NOISE
PATTERN RECORNITION
LARGE-SCALE SYSTEMS

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

THE FOLLOWING REPORT REPRESENTS A REVIEW OF RESEARCH CONDUCTED UNDER THE AUSPICES OF THE JOINT SERVICES ELECTRONICS PROGRAM AT THE INSTITUTE FOR ELECTRONIC SCIENCE AT TEXAS TECH UNIVERSITY. SPECIFIC TOPICS COVERED INCLUDE FAULT ANALYSIS, CONTROL AND ESTIMATION, NONLINEAR CONTROL, MULTIDIMENSIONAL SYSTEM THEORY, OPTICAL NOISE, AND PATTERN RECOGNITION.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

410 354

REVIEW OF RESEARCH

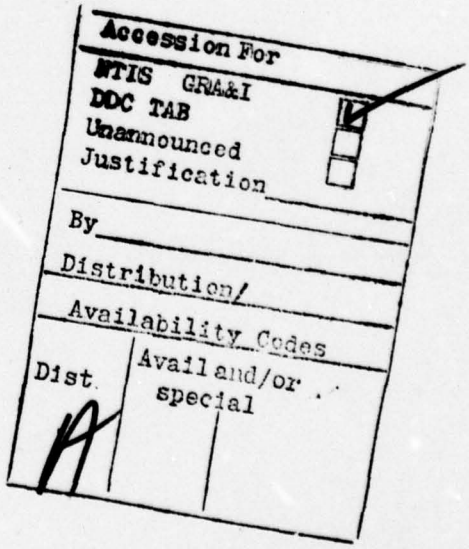
under the

JOINT SERVICES ELECTRONICS PROGRAM

at the

INSTITUTE FOR ELECTRONIC SCIENCE

TEXAS TECH UNIVERSITY



October 1979

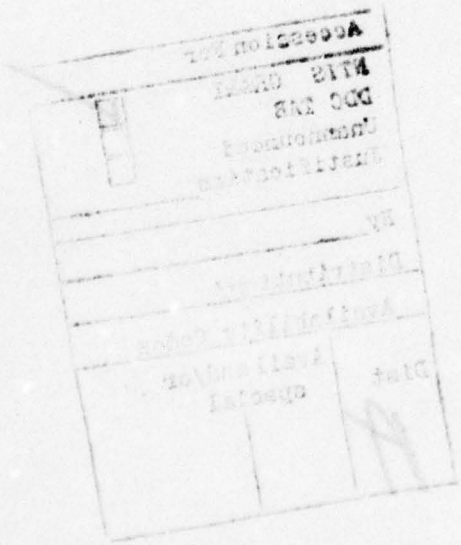
Lubbock, Texas 79409

AD A10 354

Preface

This report represents the third 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 Quadratic Optimization Problems, Nonlinear Control, Nonlinear Fault Analysis, the Qualitative Analysis of Large-Scale Systems, Multidimensional System Theory, Optical Noise, and Pattern Recognition.

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 a list of all grants and contracts administered by JSEP personnel and/or the department of Electrical Engineering and a list of all publications prepared by JSEP personnel.



Contents:

Significant Accomplishments Report

A. Feedback System Design

The feedback system design problem may naturally be subdivided into two tasks:

i. satisfaction of design constraints and

ii. optimization of system performance.

The first and foremost design constraint is stability, though system specifications may also call for an asymptotic tracking and/or disturbance rejection constraint. During the past year; working with a team of investigators from the University of Notre Dame, the University of California, and Texas Tech; we have formulated a new algebraic fractional representation approach to the feedback system design problem. Unlike classical design theories wherein a single solution to the given design problem is formulated, the key to our approach is a parameterization of the set of compensators which achieve the design constraints. As such, the design constraints of task i. are satisfied and the stage is simultaneously set for the optimization problem of task ii.

The design theory is formulated in a very general algebraic setting and is therefore applicable to any class of linear systems; distributed, multivariable, time-varying, multidimensional, etc. Moreover, we believe that the techniques developed can potentially form the basis of an entire family of design techniques for adaptive and robust control system. The initial part of this research is described in a paper which will appear in a forthcoming issue of the IEEE Transactions on Automatic Control while a second paper is in preparation. Furthermore, we have initiated work on the robust and adaptive control problems for which an M.S. thesis is presently in preparation.

B. Pointing and Tracking

During the summer of 1979 Professor T.G. Newman, while working on his JSEP project at the White Sands Missile Range, developed an entirely new approach to the problem of identifying multiple moving targets in a scene. The target motion is assumed to be modeled by the elements of a Lie group (of translations, rotations, magnifications, etc.) and the key to the theory is the formulation of an equation of motion in the coordinate system of the Lie group rather than in Euclidian coordinates. In this coordinate system every point of a rigid body is moving at exactly the same speed. As such, if one numerically computes a velocity profile from photographic data, the resultant profile will be piecewise constant with distinct levels corresponding to distinct objects.

Although formulated in a Lie group the required equation of motion can be represented by a nonlinear partial differential equation in Euclidian space, thereby, permitting the theory to be implemented via standard numerical techniques. Indeed, Newman has experimentally implemented the theory using actual photographic tracking data taken at White Sands. In particular, he successfully applied the algorithm to an extremely noisy sequence of photographs of an aircraft moving in front of a mountain which had been the subject of several previous unsuccessful attempts at analysis.

C. Eigenvalue Computation

In large-scale system theory, one often encounters the problem of computing the eigenvalues for a continuously parameterized family of large sparse matrices. As an alternative to the classical approach of discretizing the parameter and using a standard eigenvalue code at each parameter value we have developed a new continuations algorithm for eigenvalue computation. Basically, one

formulates a nonlinear ordinary differential equation whose trajectories represent the eigenvalue loci of the given family of matrices. One then computes the eigenvalues for an initial matrix, using a standard algorithm, and uses these eigenvalues as initial conditions for the differential equation in a numerical integration scheme to compute the eigenvalues of the remaining matrices in the family.

The key to our continuations algorithm is the formulation of the required differential equation, so as to minimize the computational effort required to implement the numerical integration scheme. To achieve this goal one must exploit the sparseness of the given family of matrices and simultaneously minimize the number of matrix inversions employed. With these points in mind we have formulated and tested three alternative continuations algorithms for the solution of the eigenvalue problem; one based on the LU algorithm, one based on the QR algorithm, and one which employs a Hessenberg form.

Texas Tech University | Institute for Electronic Science

Joint Services Electronics Program | Research Unit: 1

1. Title of Investigation: Quadratic Optimization Problems
2. Senior Investigator: Richard Saeks Telephone: (806) 742-3528
3. JSEP Funds: $23,500
4. Other Funds:
5. Total Number of Professionals: PI's 1 (1 mo.) RA's 1 (1/2 time)
6. Summary:

The goal of the work unit is the development of techniques for the design of modern robust, adaptive, and decentralized control systems. To this end a powerful quadratic optimization theory previously developed by the author will be employed along with a new feedback system design theory developed by the senior investigator and several colleagues under the present work unit. By combining these techniques we have developed a new quadratic optimal control theory for feedback systems with unstable plants. Moreover, a modified version of this theory has been developed in which one includes an additional term in the performance measure to reduce the sensitivity of the system to plant perturbations. As such, by controlling the weight of this term, one may obtain a tradeoff between system performance and robustness. In another direction we have developed a new approach to suboptimal control theory which is capable of handling systems with non-quadratic performance measures, decentralized systems, and nonlinear systems.

The major result obtained during the past year has been the formulation of a new feedback system design theory in which we give an explicit parameterization of the class of all possible compensators which stabilize a given feedback system. Moreover, the resultant feedback system gains are linear in the design

parameter. As such, by working with this parameterization, we characterize all compensators which achieve the stability constraint while simultaneously simplifying the process of choosing a compensator within this class. A paper describing this work has been accepted for publication in the IEEE Transactions on Automatic Control. At the present time we are in the process of extending this theory to obtain a similar characterization of the compensators which stabilize a given plant and simultaneously cause it to track and/or reject prescribed inputs.

In parallel with the above described work we have developed a new approach to sub-optimal control theory which is applicable to non-quadratic, nonlinear, and decentralized control problems. Basically, we approximate the given problem by a linear quadratic problem and compute the classical linear regulator for this problem, relative to a specified set of weighting matrices. This regulator is then used in the actual system with its performance measure being minimized over the choice of weighting matrices used to construct the linear quadratic regulator. A reprint of a conference paper in this area is included in the present report. This describes the general technique and its application to the design of an aircraft landing system.

A final aspect of the work is in the decentralized control area and is represented by a reprint of a paper recently published in the IEEE Transactions on Automatic Control. This paper proves a surprising theorem to the effect that decentralized control is just as powerful as decentralized control from the point of view of pole placement (but not optimization) in an interconnected dynamical system.

7. Publications and Activities:

A. Refereed Journal Articles

1. Saeks, R., "On the Decentralized Control of Interconnected Dynamical Systems", IEEE Trans. on Auto. Control, Vol, AC-24, pp. 269-271, (1979).

2. Desoer, C.A., Liu, R.-W., Murray, J., and R. Saeks, "Feedback System Design: The Fractional Representation Approach to Analysis and Synthesis", IEEE Trans. on Auto. Cont., (to appear).

B. Conference Papers and Abstracts

1. Karmokilias, C., and R. Saeks, "Optimal Selection of Weighting Matrices in Kalman Regulators", Proc. of the 21st Midwest Symp. on Circuits and Systems, Iowa State Univ., Ames, Ia., Aug. 1978, pp. 71-72.

C. Preprints

1. Karmokolias, C., and R. Saeks, "Suboptimal Control with Optimal Quadratic Regulators", submitted for publication.

2. Karmokolias, C., and R. Saeks, "Suboptimal Design of an Aircraft Landing System", submitted for publication.

D. Theses

1. Karmokolias, C., "Suboptimal Control with Optimal Quadratic Regulators", Ph.D. Dissertation, Texas Tech Univ., 1979.

2. Chua, O., M.S. Thesis, Texas Tech Univ., (in preparation).

E. Conferences and Symposia

1. Saeks, R., 21st Midwest Symp. on Circuits and Systems, Iowa State Univ., Aug. 1978.

2. Karmokolias, C., 21st Midwest Symp. on Circuits and Systems, Iowa State Univ., Aug. 1978.

3. Saeks, R., IEEE Decision and Control Conf., San Diego, Jan. 1979.

F. Lectures

1. Saeks, R., "Feedback System Design", Elec. Engrg. Colloquim, Univ. of Texas at El Paso, March 1979.

2. Saeks, R., "Feedback System Design", Texas Systems Workshop, Southern Methodist Univ., April 1979.

8. Reprint of "On the Decentralized Control of Interconnected Dynamic Systems", by R. Saeks from the IEEE Transactions on Automatic Control, Vol. AC-24, pp. 269-271, (1979).

On the Decentralized Control of Interconnected Dynamical Systems

R. SAEKS, FELLOW, IEEE

Abstract—It is shown that the fixed modes of an interconnected dynamical system under decentralized control are precisely the uncontrollable and unobservable states of the individual system compoonents. As such, the system can be stabilized by decentralized controllers if and only if its individual system components can be stabilized. Moreover, these conditions are shown to be equivalent to the conditions for stabilizing the system using a global controller.

Introduction

Given a linear system with partitioned inputs and outputs

$$\dot{X} = FX + \sum_{i=1}^{n} B^i u_i$$
$$y_i = C^i X \qquad i = 1, 2, \cdots, n \tag{1}$$

it is desired to design a family of dynamic decentralized controllers

$$\begin{aligned} \dot{Z}_i &= S_i Z_i + R_i y_i \\ u_i &= Q_i Z_i + K_i y_i \end{aligned} \qquad i = 1, 2, \cdots, n \tag{2}$$

which place the poles (eigenvalues) of the resultant feedback system in prescribed locations. In its most general form the solution to this problem was given by Wang and Davison [1]. Their solution is formulated in terms of the (diagonally) fixed modes of the system

$$\theta_d(F, B, C) = \cap \lambda(F + BK_dC). \tag{3}$$

Here, B and C are the matrices $B = \text{row}(B^i)$ and $C = \text{col}(C^i)$, respectively, $\lambda(M)$ denotes the set of eigenvalues of the matrix M, and the intersection is take over the set of block diagonal (complex[1]) matrices K_d

Manuscript received February 23, 1978; revised October 16, 1978. Paper recommended by D. Siljak, Chairman of the Large Scale Systems, Differential Games Committee. This work was supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

The author is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

[1]Precisely the same theory can be formulated for systems characterized by real matrices, although in that case the arguments are complicated by the fact that one must work with pairs of complex conjugate eigenvalues to preserve reality.

whose partition is comformable with the partitions of B and C. Using this concept of fixed modes, Wang and Davison [1] and Corfmat and Morse [2] showed that the eigenvalues of the system can be placed in a prespecified open region of the complex plane using the dynamic decentralized controllers of (2) *if and only if* θ_d *lies in that region*. More precisely, they showed that θ_d represents the set of eigenvalues of F which cannot be moved by any family of decentralized dynamic controllers, while all remaining eigenvalues of F can be arbitrarily placed by an appropriate choice of decentralized dynamic controllers [2].

If one observes that $F+BK_dC$ is just the state matrix for the given system with the static decentralized feedback matrix K_d, the above result can be interpreted as a characterization of the eigenvalue placement properties of the system under dynamic decentralized control in terms of its eigenvalue placement properties under static decentralized control. Indeed, the theory states that those eigenvalues which can be moved at all by static controllers can be arbitrarily placed by dynamic controllers, whereas those eigenvalues which are fixed under all static controllers are also fixed by dynamic controllers [1], [2].

Since the partitioning in (1) is arbitrary, the above-described theorem can be applied to the classical case wherein the given system has only a single input and output, in which case the fixed modes of the system are given by

$$\theta(F,B,C)=\cap\lambda(F+BKC) \tag{4}$$

where the intersection is now taken over arbitrary matrices K which are conformable with B and C. Of course, in this special case $\theta(F,B,C)$ reduces to the usual set of eigenvalues which are either uncontrollable or unobservable [4]. Moreover,

$$\theta(F,B,C)\subset\theta_d(F,B,C) \tag{5}$$

since the intersection used to define θ is taken over a larger set of matrices than that used to define θ_d. Equation (5) formalizes the intuitively obvious fact that a system which is "stabilizable" by a family of decentralized controllers is also "stabilizable" by a global (centralized) controller.

The purpose of the present paper is to show that (5) holds with equality in the case where F, B, and C represent the dynamics of an interconnected dynamical system [4] in which y_i and u_i denote the local inputs and outputs associated with a given system component. As such, in that special case the eigenvalues of the system can be placed in prespecified locations by decentralized dynamic controllers whenever they can be placed in the same locations by a global dynamic controller. Although the class of interconnected dynamical systems is considerably smaller than the class of decentralized systems studied by Wang and Davison *et al.*, the design of the local controllers for the components of an interconnected dynamical system is the "physical problem" which usually motivates the study of the general decentralized control problem. As such, we believe that the above result is significant.

The class of interconnected dynamical systems which we consider is characterized schematically in Fig. 1 and mathematically by the set of equations

$$\begin{aligned}\dot{X}_i&=A_iX_i+B_ia_i\\ y_i&=C_iX_i\\ a_i&=\sum_{j=1}^{n}L^{ij}y_j+u_i\end{aligned}\qquad i=1,2,\cdots,n. \tag{6}$$

Here the first two equations represent the dynamics of the ith component, whereas the third equation defines the interconnection structure in which the input to the ith component is taken to be a linear combination of the outputs of the various components (including the ith) and an external control. The fact that the control inputs u_i are not multiplied by compensator matrices implies that the local controllers, given by (2), have full access to the inputs of the individual system components, and similarly, that the output of the individual system components is fully accessible to the controllers.

For notational brevity (6) may be restated in block matrix form as

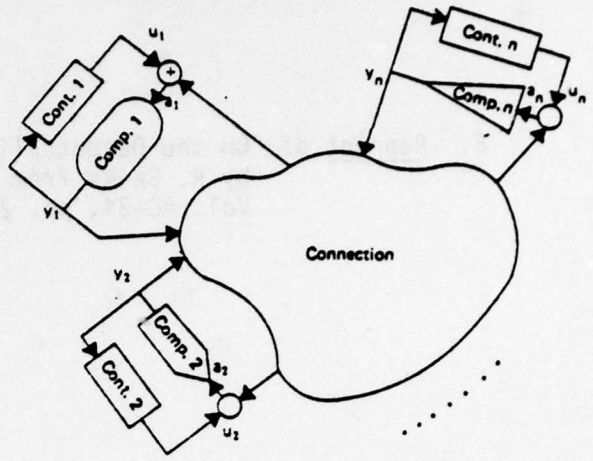


Fig. 1. Interconnected dynamical system with local controllers.

$$\begin{aligned}\dot{X}&=AX+Ba\\ y&=CX\\ a&=Ly+u\end{aligned} \tag{7}$$

where $X=\mathrm{col}(X_i)$, $a=\mathrm{col}(a_i)$, $y=\mathrm{col}(y_i)$, $u=\mathrm{col}(u_i)$, $A=\mathrm{diag}(A_i)$, $B=\mathrm{diag}(B_i)$, $C=\mathrm{diag}(C_i)$, and $L=\mathrm{mat}(L^{ij})$. Combining these into a single equation for the overall composite system, we obtain the composite state model

$$\begin{aligned}\dot{X}&=FX+Bu\\ y&=CX\end{aligned} \tag{8}$$

where

$$F=A+BLC. \tag{9}$$

Moreover, upon observing that

$$Bu=\sum_{i=1}^{n}B^iu_i \tag{10}$$

and

$$y_i=C^iX \qquad i=1,2,\cdots,n \tag{11}$$

where $B^i=\mathrm{col}(0,0,\cdots,0,B_i,0,\cdots,0)$ and $C^i=\mathrm{row}(0,0,\cdots,0,C_i,0,\cdots,0)$, we see that (8) naturally decomposes into the form of the decentralized control problem of (1). In (8), however, the B^i and C_i matrices take on a special form, whereas they are arbitrary in (1). Indeed, it is this special form which yields the desired equality in (5). Intuitively, this implies that the ith local controller may drive only the state of the ith system component, although that state may, in turn, drive the remainder of the system through the connection equations. Similarly, the ith local controller may observe only the state of the ith component, with the remaining components being observed only indirectly through the state of the ith component.

Main Theorem

Using the above notation, our main theorem may be stated as follows.

Theorem: For the system of (8)

$$\theta_d(F,B,C)=\theta(F,B,C)=\theta(A,B,C)=\bigcup_i\theta(A_i,B_i,C_i). \tag{12}$$

Proof: To show that $\phi(F,B,C)=\phi(A,B,C)$ we simply observe that

$$\lambda(F+BKC)=\lambda(A+B(L+K)C)=\lambda(A+BK'C) \tag{13}$$

where $K'=L+K$. As such, the same set of matrices are spanned if one takes the intersection of the $\lambda(F+BKC)$ over K or the intersection of the $\lambda(A+BK'C)$ over K', and hence $\theta(F,B,C)=\theta(A,B,C)$. Moreover, since A, B, and C are block diagonal, $\theta(A,B,C)$ is just the union of the fixed

modes associated with each block. Given (5) to prove the validity of the first equality of (12), it suffices to show that $\theta_d(F,B,C) \subset \theta(F,B,C)$. For this purpose, we desire to show that if λ is not in $\theta(F,B,C)$ then it is not in $\theta_d(F,B,C)$. Initially, we assume that A, B, and C are partitioned as 2 by 2 matrices, the general case following therefrom by induction. If λ is not in $\theta(F,B,C)$, then there exists a K (dependent on λ) such that $\det(\lambda I-(F+BKC)) \neq 0$ and we desire to construct a block diagonal K_d, also dependent on λ, such that $\det(\lambda I-(F+BK_dC)) \neq 0$. To this end we write out the matrix $F+BKC$ in partitioned form and expand its determinant via the formula for the determinant of a 2 by 2 partitioned matrix [3], obtaining

$$
\begin{aligned}
0 \neq \det(\lambda I-(F+BKC)) &= \det(\lambda I-(A+B(L+K)C)) \\
&= \det\left[\begin{array}{c|c} \lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1 & -B_1L^{12}C_2-B_1K^{12}C_2 \\ \hline -B_2L^{21}C_1-B_2K^{21}C_1 & \lambda I-A_2-B_2L^{22}C_2-B_2K^{22}C_2 \end{array}\right] \\
&= \det(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1) \\
&\quad \cdot \det\left[\lambda I-A_2-B_2L^{22}C_2-B_2K^{22}C_2(B_2L^{21}C_1+B_2K^{21}C_1)\right. \\
&\quad \left.\cdot(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}(B_1L^{12}C_2+B_1K^{12}C_2)\right] \\
&= \det(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1)\det\left[\lambda I-A_2-B_2L^{22}C_2\right. \\
&\quad \left.-B_2\underline{K}^{22}C_2-B_2L^{21}C_1(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1L^{12}C_2\right]
\end{aligned} \tag{14}
$$

where

$$
\begin{aligned}
\underline{K}^{22} = K^{22} &+ K^{21}C_1(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1L^{12} \\
&+ L^{21}C_1(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1K^{12} \\
&+ K^{21}C_1(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1K^{12}
\end{aligned} \tag{15}
$$

Here

$$
L = \left[\begin{array}{c|c} L^{11} & L^{12} \\ \hline L^{21} & L^{22} \end{array}\right] \tag{16}
$$

and

$$
K = \left[\begin{array}{c|c} K^{11} & K^{12} \\ \hline K^{21} & K^{22} \end{array}\right] \tag{17}
$$

are partitioned to be conformable with A, B, and C. Now, if we define K_d via

$$
K_d = \left[\begin{array}{c|c} K^{11} & 0 \\ \hline 0 & \underline{K}^{22} \end{array}\right] \tag{18}
$$

and compute $\det(\lambda I-(F+BK_dC))$, we obtain

$$
\begin{aligned}
\det(\lambda I-(F+BK_dC)) &= \det(\lambda I-(A+B(L+K_d)C)) \\
&= \det\left[\begin{array}{c|c} \lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1 & -B_1L^{12}C_2 \\ \hline -B_2L^{21}C_1 & \lambda I-A_2-B_2L^{22}C_2-B_2\underline{K}^{22}C_2 \end{array}\right] \\
&= \det(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1) \\
&\quad \cdot \det\left[\lambda I-A_2-B_2L^{22}C_2-B_2\underline{K}^{22}C_2\right. \\
&\quad \left.-B_2L^{21}C_1(\lambda I-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1L^{12}C_2\right] \\
&= \det(\lambda I-(F+BKC)) \neq 0.
\end{aligned} \tag{19}
$$

Thus, there is at least one block diagonal K_d matrix such that λ is not an eigenvalue of $F+BK_dC$ showing that λ is not in $\theta_d(F,B,C)$. Note that, in general, K_d is complex even for a real K if λ is complex. To obtain a real K_d one would have to work with complex conjugate pairs of eigenvalues rather than single eigenvalues. Since this would further complicate an already complex derivation, we will not attempt to construct a real K_d here. Also note that we have assumed that the upper left-hand corner of the matrix of (14) is nonsingular.

To extend the above argument from 2 by 2 partitioned matrices to n by n partitioned matrices, we repeat the above construction $n-1$ times as follows. Given an n by n matrix

$$
K = \left[\begin{array}{c|c|c|c|c} K^{11} & K^{12} & K^{13} & \cdots & K^{1n} \\ \hline K^{21} & K^{22} & K^{23} & & K^{2n} \\ \hline K^{31} & K^{32} & K^{33} & & K^{3n} \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline K^{n1} & K^{n2} & K^{n3} & \cdots & K^{nn} \end{array}\right] \tag{20}
$$

such that $\det(\lambda I-F+BKC) \neq 0$ it is partitioned into a 2 by 2 matrix as shown by the double line whence the above argument is employed to formulate a matrix

$$
\underline{K} = \left[\begin{array}{c|c|c|c|c} K^{11} & 0 & 0 & \cdots & 0 \\ \hline 0 & \underline{K}^{22} & \underline{K}^{23} & & \underline{K}^{2n} \\ \hline 0 & \underline{K}^{32} & \underline{K}^{33} & & \underline{K}^{3n} \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline 0 & \underline{K}^{n2} & \underline{K}^{n3} & \cdots & \underline{K}^{nn} \end{array}\right] \tag{21}
$$

such that $\det(\lambda I-(F+B\underline{K}C)) \neq 0$. This matrix is then repartitioned into a new 2 by 2 matrix as shown by the double line in (22) and the process is repeated. Since the 1-1 entry in the partitioned matrix is not affected by the process, this results in a new matrix of the form

$$
\underline{K} = \left[\begin{array}{c|c|c|c|c} K^{11} & 0 & 0 & \cdots & 0 \\ \hline 0 & \underline{K}^{22} & 0 & & 0 \\ \hline 0 & 0 & \underline{K}^{33} & & \underline{K}^{3n} \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline 0 & 0 & \underline{K}^{n3} & \cdots & \underline{K}^{nn} \end{array}\right] \tag{22}
$$

such that $(\lambda I-(F+B\underline{K}C) \neq 0$. Repeating the process $n-1$ times eventually results in a block diagonal matrix K_d such that $\det(\lambda I-(F+BK_dC) \neq 0$, showing that if λ is not in $\theta(F,B,C)$, then it is also not in $\theta_d(F,B,C)$, thereby verifying that $\theta_d(F,B,C)$ is contained in $\theta(F,B,C)$ and completing the proof of the theorem.

Given the theorem, for the class of interconnected dynamical systems, local control is just as good a global control from the point of view of pole placement. From the point of view of optimal control, however, global control will, in general, still be superior since it gives one a greater range of options [4].

References

[1] S. H. Wang and E. J. Davison, "On the stabilization of decentralized control," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 473–478, 1973.

[2] J. P. Corfmat and A. S. Morse, "Decentral control of linear multivariable systems," *Automatica*, vol. 12, pp. 479–495, 1976.

[3] F. R. Gantmacher, *The Theory of Matrices*, 2 vols. New York: Chelsea, 1959.

[4] R. Saeks and R. A. DeCarlo, *Interconnected Dynamical Systems*. New York: Marcel Dekker, to be published.

9. Reprint of "Optimal Selection of Weighting Matrices in Kalman Regulators", by C. Karmokolias and R. Saeks from the Proceedings of the 21st Midwest Symposium on Circuits and Systems, Iowa State Univ., August 1979, pp. 72-76.

Abstract

This paper suggests an approach to optimally selecting the weighting matrices in a Kalman regulator, by minimizing an arbitrarily defined performance index.

1. INTRODUCTION AND PROBLEM STATEMENT

The solution of a common class of control problems involves minimizing a functional J, termed the "performance index" over a class of functions which are the allowable inputs to the system. In general, J represents a compromise between the system's performance and the energy content of the system's inputs. More precisely, the problem may be formulated as follows

$$\min_{u(t)} \quad J = \int_{t_o}^{t_f} \{x'(t)\, Q\, x(t) + u'(t)\, R\, u(t)\}\, dt \qquad (1.1)$$

subject to

$$\dot{x}(t) = F(t)\, x(t) + G(t)\, u(t) \qquad (1.2)$$

$$x(t_o) = x_o. \qquad (1.3)$$

The Q and R matrices in Equation (1.1) are called the "state weighting matrix" and the "input weighting matrix", respectively. Both terms in the integrand of Equation (1.1) can be thought of as cost functions. The $x'(t)\, Qx(t)$ term represents the energy expenses needed to achieve such a trajectory. Very often the off-diagonal entries in the Q and R matrices are assumed to be zero. Then, the two matrices also represent measures of relative importance among the states and the various inputs of the system.[1]

Thus, the designer must select the two weighting matrices Q and R. Frequently, the selection is made by trial and error.[1] An approximate procedure is suggested in (2). In this technique, an entry in Q is arbitrarily selected and the remaining entries in Q are R are obtained by assuming:

(1) The maximum contributions to J by the $x'(t)\, Qx(t)$ must occur simultaneously in time.

(2) The total contribution of the $x'(t)\, Qx(t)$ term must equal the total contribution of the $u'(t)\, Ru(t)$ term.

Obviously, the first assumption is not always valid, whereas, the second is indeed quite arbitrary. In some cases, the Q and R matrices are obtained by solving the Inverse Control Problem where a linear control law is assumed for the feedback and then the weighting matrices are obtained from Equation (1.1) subject to this condition.[3,4]

In any case, the selection of weighting matrices, although a matter of experience and ingenuity, is generally suggested by factors external to the system. In most cases, the plan S, which is to be controlled, is a subsystem of a general system Σ and so it is factors in Σ which dictate the performance specifications of S. Thus, Q and R

could be selected by examining the effect that the performance of S has upon Σ.

Hence, consider a linear dynamic system S which is a subsystem of a system Σ, as shown in Figure 1-1. Assume that S is controlled by a Kalman regulator

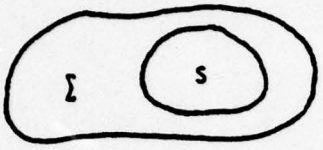


Figure 1-1. A General System Σ

where the Q and R matrices are to be specified. Assume that a functional on Σ can be defined as

$$J_\Sigma = \int_{t_o}^{t_f} h(x(t), u(t), t)\, dt \tag{1.4}$$

where $h(\cdot,\cdot,\cdot)$ is some given function, possibly nonlinear. Then the problem may in general be formulated as follows:

$$\underset{Q,R}{\text{Min}}\ J_\Sigma = \int_{t_o}^{t_f} h(x^*(t), u^*(t), t)\, dt \tag{1.5}$$

subject to

$$\underset{u(t)}{\text{Min}}\ J_S = \int_{t_o}^{t_f} \{x'(t)\, Qx(t) + u'(t)\, R\, u(t)\}\, dt \tag{1.6}$$

subject to

$$\overset{o}{x}(t) = F(t)\, x(t) + G(t)\, u(t) \tag{1.7}$$

$$x(t_o) = x_o \tag{1.8}$$

where $u^*(t)$ is the optimal solution of Equation (1.6) and $x^*(t)$ is obtained from Equation (1.7) and Equation (1.8) by setting $u(t) = u^*(t)$. To ensure a unique solution of Equation (1.5), it is assumed that $h(\cdot,\cdot,\cdot)$ is convex in Q and R and that one of the entries of R is arbirarily set to 1.

The solution of Equation (1.6) is given by[3]

$$u^*(t) = R^{-1}G'(t)P(t)x(t) \tag{1.9}$$

where P(t) is the solution of the matrix Riccati Equation

$$\frac{dP(t)}{dt} = -F'(t)P(t) - P(t)F(t) + P(t)G(t)R^{-1}G'(t)P(t) - Q \tag{1.10}$$

$$P(t_f) = 0 \tag{1.11}$$

Then substituting Equation (1.9) into Equation (1.7) $x^*(t)$ is given by[3]

$$x^*(t) = \phi_x(t,t_o)x_o \tag{1.12}$$

where $\phi_x(t,t_o)$ is the solution of

$$\overset{o}{\phi}_x(t,t_o) = [F(t) - G(t)R^{-1}G'(t)P(t)]\phi_x(t,t_o) \tag{1.13}$$

$$\phi_x(t,t) = I. \tag{1.14}$$

Since, in general, $h(x^*(t), u^*(t),t)$ is not given, it appears on first sight as if one has traded the arbitrariness in selecting Q and R for the arbitrariness in selecting the performance index J_Σ. Although this is in fact true, the non-linearity of $h(x^*(t), u^*(t), t)$ makes the selection of J_Σ much easier than the selection of Q and R. In fact, concepts from utility theory and decision theory, are readily applicable thus facilitating the selection of J_Σ.[4]

On the other hand, the problem of Equation (1.5) is to be solved only once over extended periods of time. Thus, though $h(x^*(t), u^*(t), t)$ is in general a complicated functional, the computational costs are not prohibitive.

2. SCALAR EXAMPLES

It is assumed that the system dynamics are given as

$$\overset{o}{x}(t) = -x(t) + u(t) \tag{2.1}$$

$$x(t_o) = x_o \tag{2.2}$$

and the S performance index is given as

$$J_S = \int_0^1 (qx^2(t) + u^2(t))\, dt. \tag{2.3}$$

The Riccati equation is

$$\frac{d}{dt}P(q,t) = -2P(q,t) + P^2(q,t) - 9 \tag{2.4}$$

$$P(q,1) = 0 \tag{2.5}$$

Following a method described in (3), an analytic solution for Equation (2.4) is obtained by solving the augmented differential equation

$$\frac{d}{dt}\,\Theta(t,T) = \begin{bmatrix} -1 & -1 \\ -q & 1 \end{bmatrix}\Theta(t,T) \qquad (2.6)$$

$$\Theta(t,t) = I \qquad (2.7)$$

The eigenvalues of Equation (2.6) are

$$\lambda_{1,2} = \pm \sqrt{q+1}\,. \qquad (2.8)$$

The eigenvectors associated with the eigenvalues of Equation (2.8) are

$$p_1 = \begin{bmatrix} 1 \\ -(1+\sqrt{q+1}) \end{bmatrix} \quad p_2 = \begin{bmatrix} 1 \\ -1+\sqrt{q+1} \end{bmatrix}. \qquad (2.9)$$

Then

$$p = \begin{bmatrix} 1 & 1 \\ -(1+\sqrt{q+1}) & (-1+\sqrt{q+1}) \end{bmatrix} \qquad (2.10)$$

and

$$p^{-1} = \frac{1}{2\sqrt{q+1}} \begin{bmatrix} (-1+\sqrt{q+1}) & -1 \\ (1+\sqrt{q+1}) & 1 \end{bmatrix} \qquad (2.11)$$

Thus, a fundamental matrix for Equation (2.6) is

$$\Theta(t) = pe^{tJ}p^{-1} \qquad (2.12)$$

where

$$e^{tJ} \triangleq \begin{bmatrix} e^{t\sqrt{q+1}} & 0 \\ 0 & e^{-t\sqrt{q+1}} \end{bmatrix} \qquad (2.13)$$

Thus, a transition matrix for Equation (2.6) is

$$\Psi(t,T) = \Theta(t)\,\Theta(t)^{-1} \qquad (2.14)$$

But since

$$\Theta^{-1}(T) = (p^{-1})^{-1}\,e^{-TJ}p^{-1} = pe^{-TJ}p^{-1} \qquad (2.15)$$

Equation (2.14) becomes

$$\Psi(t,T) = pe^{(t-T)J}p^{-1} \qquad (2.16)$$

Hence, since $\Theta(T,T) = I$, the solution of Equation (2.6) is $\Theta(t,T) = \Psi(t,T)$. (2.17)

Then, the solution of Equation (2.4) and Equation (2.5) is

$$P(q,t) = \Theta_{21}(t,T)\cdot\Theta_{11}(t,T)^{-1} \qquad (2.18)$$

Substituting and performing the indicated calculations,

$$P(q,t) = \frac{q[-e^{(t-1)\sqrt{q+1}} + e^{-(t-1)\sqrt{q+1}}]}{(-1+\sqrt{q+1})e^{(t-1)\sqrt{q+1}} + (1+\sqrt{q+1})e^{-(t-1)\sqrt{q+1}}} \qquad (2.19)$$

Hence, by (3), the optimal input is

$$u^*(t) = -P(q,t)x(t) \qquad (2.20)$$

Then, substituting Equation (2.20) into Equation (2.1),

$$\frac{d}{dt}\,x^*(t) = -(1+P(q,t)x^*(t) \qquad (2.21)$$

Integrating Equation (2.21) and using Equation (2.2),

$$x^*(t) = x_o \cdot \exp\left[-t - \int_o^t P(q,\sigma)\,d\sigma\right] \qquad (2.22)$$

It is now assumed that the Σ performance index is of the form

Case 1 $$J_\Sigma = \int_o^1 \{[1 - x^*(t)]^2 + [u^*(t)]^2\}dt \qquad (2.23)$$

Thus, substituting Equation (2.20) and Equation (2.22) into Equation (2.23)

$$J_\Sigma = \int_o^1 \{[1 - x_o e^{-(t+\int_o^t P(q,\sigma)d\sigma)}]^2 + [-P(q,t)x_o e^{-(t+\int_o^t P(q,\sigma)d\sigma)}]^2\}dt \qquad (2.24)$$

where P(q,t) is given by Equation (2.19). The q minimizing Equation (2.24) is calculated for several values of x_o, using numerical integration techniques.

At this point a brief discussion of the problem may be useful. Refer to Figure 2-1. As indicated the performance required from the system is indeed quite unrealistic. But this should be expected since J_Σ does not necessarily consider the physical limitations of the system. Furthermore, given any initial condition x_o, the optimal selection of q is not obvious.

In Figure 2-2, the q minimizing Equation (2.24) is plotted versus the initial condition x_o. As seen as $x_o \to \pm\infty$, $q_{opt} \to 1$. This is reasonable since

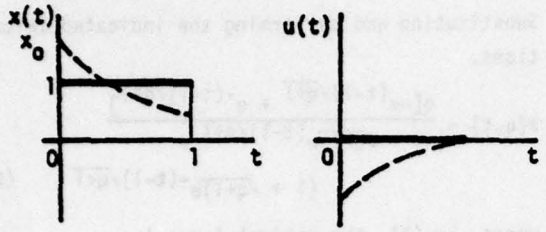


Figure 2-1. Required Versus Actual Performance. Solid Lines are the Required Performance Whereas Dotted Lines are the Actual.

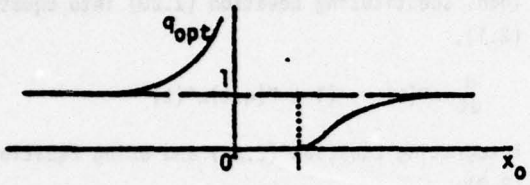


Figure 2-2. Optimal q Versus Initial Condition x_0.

for $|x_0|$ being very large,

$$(x(t)-1)^2 \simeq x(t)^2 \tag{2.25}$$

and hence, equal importance is placed upon the state and the input terms of J_Σ, Equation (2.23). As x_0 goes to zero from the positive side, the state approximates its requirement better and better and hence the regulator is "instructed" to emphasize the input cost by decreasing q. For x_0 between 0 and 1, q_{opt} is equal to zero. This also was to be expected since for positive initial value the response of the first order system is always bounded above by its natural response. Hence, the regulator is "instructed" to totally "ignore" the state in an attempt to force the system to achieve the state upper bound. When $x_0 = 0$, by Equation (2.20) and Equation (2.22), $x^*(t) = u^*(t) = 0$. Thus, by Equation (2.3), q_{opt} is indeterminate. Once x_0 becomes slightly negative, the regulator is "instructed" to drive x(t) to zero as soon as possible since x(t) is now adding to the error. As x_0 becomes more and more negative, the error in the input becomes significant also, and thus, as stated earlier, q_{opt} approaches 1. Thus, it was seen that even in this case where one normally would not use a Kalman regulator, the results agree with intuition.

Case 2 $$J_\Sigma = \int_0^1 ([1 - t - x^*(t)]^2 + [u^*(t)]^2)dt \tag{2.26}$$

A Kalman regulator is a more "natural" controller for this case. Yet the comments made for the previous case are also applicable here. Figure 2-3 shows required versus actual performance and Figure 2-4 shows the q minimizing Equation (2.26) versus the initial condition x_0.

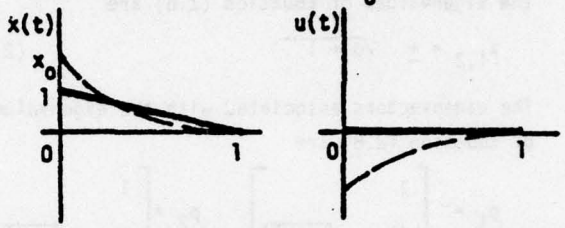


Figure 2-3. Required Versus Actual Performance. Solid Lines are the Required Performance Whereas Dotted Lines are the Actual.

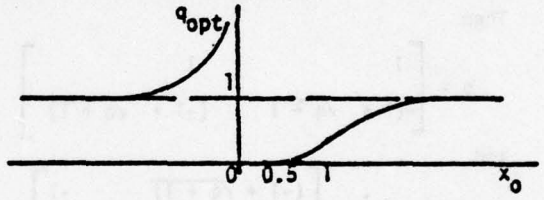


Figure 2-4. Optimal q Versus Initial Condition x_0.

3. APPLICATIONS

The proposed approach could be applied in several real world situations. One such case is the landing of an aircraft, where a Kalman regulator could be used to drive the aircraft from some initial state x_0 to a zero state. Generally speaking, one is interested in a safe, comfortable and economical landing. Towards achieving these goals, certain restrictions have been or could be imposed on the states and the inputs of the system, either by law or regulation or as a matter of policy or merely due to the physical limitations of the aircraft. These restrictions are then used to determine the overall system performance of J_Σ.

Another case is the control of a large system consisting of a number of interconnected subsystems. Several schemes have been proposed[5,6,7] where a Kalman regulator is used to control each subsystem

independently from the others. The main differences in these schemes arise in how should the couplings be handled. More often than not, the assumption of weak couplings is made. In the numerical approach suggested here, no assumption is made about the nature of the couplings and thus the method is equally applicable to systems with either strong or weak couplings. The J_Σ index could even be assumed to be quadratic in this case. So, assuming that the Q and R matrices for the overall system are given, one seeks the weighting matrices for each of the subsystems.

To efficiently perform the desired minimization, an efficient algorithm for the calculation of the solution of the Riccati equation is desired for the various entries in the Q and R matrices.[8] Though not presented here, an algorithm has been derived where the solution P is calculated by first obtaining the value of P for the initial point (q_0, r_0, t_0) and then solving two linear, time-invariant differential equations to obtain the value of P at some other point (q,r,t).

In fact, it was the efficiency of this calculation that dictated the choice of the present approach over the alternative offered by the Inverse Control Problem, since in the latter case, the optimization had to be carried over a time varying matrix K(t), rather than a time independent matrix (Q + R), a task considerably more difficult.

REFERENCES

1. D.E. Kirk, "Optimal Control Theory, An Introduction," Englewood Cliffs, New Jersey: Prentice Hall Inc., 1970.
2. C.W. Merriam, "Optimization Theory and the Design of Feedback Control System, "New York: McGraw-Hill, 1964.
3. B.D.O. Anderson, J.B. Moore, "Linear Optimal Control," Englewood Cliffs, New Jersey: Prentice Hall Inc., 1971.
4. M.F. Rubinstein, "Patterns of Problem Solving," Englewood Cliffs, New Jersey: Prentice Hall Inc., 1975.
5. D.D. Siljak, M.K. Sundareshan, "A Multilevel Optimization of Large-Scale Dynamic Systems," IEEE Trans. Automat. Contr., pp. 79-84, Fev. 1976.
6. F.N. Bailey, F.C. Wang, "Decentralized Control Strategies for Linear Systems," Proc. Sixth Asilomar Conf., Circuits and Systems, Nov. 1972.
7. L. Isaksen, H.J. Payne, "Suboptimal Control of Linear Systems by Augmentation with Application to Freeway Traffic Regulation," IEEE Trans. Automat. Contr., Vol. AC-18, No. 3, pp. 210-219, June 1973.
8. J. Medanic, M. Andjelic, "Convex Approximation of the Solution of the Matrix Riccati Equation," IEEE Trans. Automat. Contr., pp. 234-238, April, 1975.

10. Abstract of "Feedback Systems Design: The Factional Representation Approach to Analysis and Synthesis" by Desoer, C.A., Liu, R.-W., Murray,J.J. and R.Saeks, to appear in the IEEE Transactions on Automatic Control.

The problem of designing a feedback system with prescribed properties is attacked via a fractional representation approach to feedback system analysis, and synthesis. To this end we let H denote a ring of operators with the prescribed properties and model a given plant as the ratio of two operators in H. This, in turn, leads to a simplified test to determine whether or not a veedback system in which that plant is embedded has the prescribed properties and a complete characterization of those compensators which will "place" the feedback system in H. The theory is formulated axiomatically to permit its application in a wide variety of system design problems and is extremely elementary in nature, requiring no more than addition, multiplication, subtraction, and inversion for its derivation even in the most general settings.

11. Abstract of "Suboptimal Control with Optimal Quadratic Regulators" by C. Karmokolias and R. Saeks.

The purpose of the present paper is to describe an approach to the control system design problem, wherein, one designs a quadratic regulator for an approximation of the given 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 advantage is that the system performance is suboptimal.

12. Abstract of "Suboptimal Design of an Aircraft Landing Systems" by C. Karmokolias and R. Saeks.

The design of an aircraft landing system is carried out via a suboptimal algorithm. In particular, a specific optimal controller is obtained by restricting the design to the class of quadratic regulators for an unconstrained linear approximation to the given system. A suboptimal controller is then chosen from within that class by optimizing over the choice of weighting matrices. The resultant control system compares well with previous designs and was obtained without undue computational effort.

Texas Tech University — Institute for Electronic Science

Joint Services Electronics Program — Research Unit: 2

1. Title of Investigation: Nonlinear Control
2. Senior Investigator: L. R. Hunt
3. JSEP Funds: $23,500
4. Other Funds:
5. Total Number of Professionals: PI's 2 (3 mo.) RA's
6. Summary:

The goal of the work unit is the development of a control theory for a large class of nonlinear system via differential geometric techniques. To this end we have formulated a global controllability theory for the differential equation

$$\dot{x} = f(x) + \sum_{i=1}^{k} u_i g_i(x) \qquad 1.$$

in which the reachable sets from any initial state are characterized and we are presently investigating the possible extension of this theory to the observability and stabilizability problems.

Our main activity during the past year has been the formulation of the aforementioned controllability theory. This is based on the additive nature of equation 1. and exploits the fact that a state trajectory must move with the flow of the homogeneous equation

$$\dot{x} = f(x) \qquad 2.$$

or along the integral curves of the $g_i(x)$. A natural candidate for the boundary of a reachable set is therefore an integral curve since the controls cause the

trajectory to move along the integral curve rather than across it. As such, the only way a state trajectory can cross an integral curve is with the flow of the homogeneous equation. Although some rather powerful differential geometry is required to formalize these ideas they can be put together into a rigorous characterization of the reachable set in state space from a given initial condition. In particular, equation 1. is globally controllable if, and only if, no integral curve on which the flow of the homogeneous equation is unidirectional separates the state space.

These ideas have been described in six publications, three of which have appeared and are reprinted in this report, while we are presently pursuing the extension of the approach to the observability and stabilizability problems.

7. Publications and Activities:

A. Refereed Journal Articles

1. Hunt, L.R., "Controllability of General Nonlinear Systems", Math. Sys. Theory, Vol. 12, pp. 361-370, (1979).

B. Conference Papers and Abstracts

1. Hunt, L.R., "Control Theory for Nonlinear Systems", Proc. of the 12th Asilomar Conf. on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1979. pp. 339-343.

2. Hunt, L.R., "Controllability of Nonlinear Systems", Proc. of the 1979 Inter. Symp. on the Mathematics of Networks and Systems, T.H., Delft, July 1979, pp. 466-467.

C. Preprints

1. Hunt, L.R., "Global Controllability in Two Dimensions", submitted for publication.

2. Hunt, L.R., "Controllability of Nonlinear Hypersurface Systems", submitted for publication.

3. Hunt, L.R., "Controllability and Stability", submitted for publication.

D. Conferences and Symposia

1. Hunt, L.R., 1979 Inter. Symp. on the Mathematics of Networks and Systems, T.H. Delft, July 1979.

2. Hunt, L.R., NATO Workshop on Algebra and Algebraic Geometry in Linear System Theory, Harvard Univ., June 1979.

3. Hunt, L.R., Texas Systems Workshop, Southern Methodist Univ., April 1979.

4. Hunt, L.R., 12th Asilomar Conf. on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1978

8. Reprint of "Controllability of General Nonlinear Systems" by L.R. Hunt from Mathematical System Theory, Vol. 12, pp. 361-370, (1979).

Controllability of General Nonlinear Systems

L. R. Hunt*

Department of Mathematics, Texas Tech University, Lubbock, Texas 79409

Abstract. Consider the nonlinear system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \qquad x(0) = x_0 \in M$$

where M is a C^∞ real n-dimensional manifold, $f, g_1, \ldots, g_m$ are C^∞ vector fields on M, and $u_1, \ldots, u_m$ are real-valued controls. If $m = n-1$ and $f, g_1, \ldots, g_m$ are linearly independent, then the system is called a hypersurface system, and necessary and sufficient conditions for controllability are known. For a general m, $1 \leq m \leq n-1$, and arbitrary C^∞ vector fields, $f, g_1, \ldots, g_m$, assume that the Lie algebra generated by $f, g_1, \ldots, g_m$ and by taking successive Lie brackets of these vector fields is a vector bundle with constant fiber (vector space) dimension p on M. By Chow's Theorem there exists a maximal C^∞ real p-dimensional submanifold S of M containing x_0 with the generated bundle as its tangent bundle. It is known that the reachable set from x_0 must contain an open set in S. The largest open subset U of S which is reachable from x_0 is called the region of reachability from x_0. If O is an open subset of S which is reachable from x_0, we find necessary conditions and sufficient conditions on the boundary of O in S so that $O = U$. Best results are obtained when it is assumed that the Lie algebra generated by $g_1, \ldots, g_m$ and their Lie brackets is a vector bundle on M.

---

*Research supported in part by the National Science Foundation under NSF Grant MCS 76-05267-A01 and by the Joint Services Electronics Program under ONR Contract 76-C-1136.

AMS (MOS) SUBJECT CLASSIFICATION: 93C10, 93C15.

KEY WORDS AND PHRASES: Nonlinear systems, controllability, reachable set, vector bundle.

0025/5661/79/0012-0361$02.00

I. Introduction

Let $f, g_1,\ldots,g_m$ be C^∞ vector fields on a connected C^∞ real n-dimensional manifold M. If $u_1,\ldots,u_m$ are real-valued controls, then we examine the controllability of the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t)g_i(x(t)), \qquad x(0) = x_0 \in M.$$

The system is called a hypersurface system if $m = n-1$ and if $f, g_1,\ldots,g_{n-1}$ are linearly independent on M, and necessary and sufficient conditions for controllability are given in [5]. For a general m, $1 \leqslant m \leqslant n-1$, we know that the reachable set of this system contains an open subset of the real p-dimensional submanifold S of M given to us by Chow's Theorem. Here we assume that the Lie algebra generated by $f, g_1,\ldots,g_m$ and by taking successive Lie brackets of $f, g_1,\ldots,g_m$ is a vector bundle with constant fiber (vector space) dimension p on M and that $f, g_1,\ldots,g_m$ are complete vector fields on M (which may not be linearly independent on M).

Let U be the largest open subset of S which is reachable from x_0. We call such a set the region of reachability from x_0. If $U \neq S$, we find necessary conditions on the boundary of $U(\partial U)$ in S so that U is the region of reachability from x_0. If O is an open subset of S containing x_0 which is reachable from x_0, we give sufficient conditions on ∂O in S so that $O = U$ for certain systems.

The hypersurface case is the nice case in which complete answers can be derived. In section 2 of this paper we discuss the results from [5] for hypersurface systems. Section 3 contains necessary conditions that an open subset of S be the region of reachability from x_0 for a general system. In section 4 we discuss sufficient conditions on the boundary of an open set in S so that this open set is the region of reachability.

II. Hypersurface Systems

Even though the results of this section are restricted to hypersurface systems (with linearly independent vector fields $f, g_1,\ldots,g_{n-1}$)

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t)g_i(x(t)), \qquad x(0) = x_0 \in M \tag{2.1}$$

our definitions (with the exception of Definitions 2.3 and 2.4) apply to general systems

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t)g_i(x(t)), \qquad x(0) = x_0 \in M. \tag{2.2}$$

Let $T(M)$ denote the tangent bundle of M with fiber $T_x(M)$ for $x \in M$. If X is a vector field on M (i.e. X is a global section of $T(M)$) then α is an *integral*

curve of X if α is a C^∞ mapping from a closed interval $I \subset \mathbf{R}$ into M such that

$$\frac{d\alpha(t)}{dt} = X(\alpha(t)) \qquad \text{for all} \qquad t \in I.$$

Definition 2.1 [8]. If D is a subset of $T(M)$, then an *integral curve of* D 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 *D-reachable from* x_0 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 *D-reachable from* x_0 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 in (2.1) or (2.2) we drop the D from D-reachable. For hypersurface systems as (2.1) we know that there is an open set in M which is reachable from x_0 and which contains x_0 in its closure.

Definition 2.3. The largest open subset U of M which is reachable from x_0 for system (2.1) is called the *region of reachability from* x_0. If $U = M$, we say that the system is *controllable from* x_0.

Definition 2.4. Let O be an open subset of M and let $x \in \partial O$. Then f *points in the direction of* O (or *towards* O) *at* x 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 this is true for every $x \in \partial O$, then f *points in the direction of* O *on* ∂O.

If f points in the direction of O at $x \in \partial O$ and if ∂O is C^1 near x, then there is some open neighborhood W of x in M so that the integral curve of f (moving in positive time) starting at x and intersected with $W - \{x\}$ is contained in O.

The following result from [5] gives necessary and sufficient conditions on the boundary of an open set of M for this set to be the region of reachability of a hypersurface system. Recall that a C^∞ submanifold B of M is an integral manifold of the linearly independent vector fields $Y_1,\ldots,Y_k$ on M if $T_y(B)$ is the space spanned by $Y_1,\ldots,Y_k$ at y for each point $y \in B$.

Theorem 2.5 [5]. *Suppose $x_0 \in M$ and O is an open subset of M which contains x_0 in its closure and which is reachable from x_0. Then O is the region of reachability U from x_0 of the system (2.1) if and only if ∂O is an integral manifold of $g_1,\ldots,g_{n-1}$ and f assigns vectors on ∂O which point in the direction of O. In fact, the smallest open subset U of M with $x_0 \in \overline{U}$ (the closure of U) satisfying ∂U is an integral manifold of $g_1,\ldots,g_{n-1}$ and f points in the direction of U on ∂U is the region of reachability from x_0.*

As a corollary we proved in [5] that if there is no integral manifold of $g_1,\ldots,g_{n-1}$ which disconnects M in some sense, then the system (2.1) is controllable from any $x_0 \in M$. Also if such integral manifold exists, but the vector field f

assigns vectors on each which always point in both directions relative to the manifold (i.e. if an integral manifold N divides M into two components, then f assigns vectors in the direction of one component at some points of N and toward the second component at other points of N), then the system (2.1) is controllable from any $x_0 \in M$.

The hypersurface case is ideal in the sense that we obtain clear cut solutions to problems concerning reachability and controllability. For this reason we use it as a model for our system (2.2), even though we don't get as satisfying results in general due to the inherent nature of the system. The difference seems to be that ∂U is an integral manifold of $g_1,\ldots,g_{n-1}$ in the hypersurface case, which implies that ∂U is a C^∞ manifold. This is not true in general for nonhypersurface systems.

III. Necessary Conditions

We need to state several definitions and results before considering the general problem. Again, let M be a connected C^∞ real n-dimensional manifold. If $f_1,\ldots,f_r$ are C^∞ vector fields on M, we define the *Lie bracket* of f_i and f_j, $1 \leq i, j \leq r, i \neq j$, by

$$[f_i,f_j] = \frac{\partial f_j}{\partial x} f_i - \frac{\partial f_i}{\partial x} f_j.$$

The set of vector fields $\{f_1,\ldots,f_r\}$ is called *involutive* if there exist C^∞ functions $\gamma_{ijk}(x)$ on M such that

$$[f_i,f_j](x) = \sum_{k=1}^{r} \gamma_{ijk}(x) f_k(x).$$

We state the following version of the Frobenius Theorem from [1].

Theorem 3.1. *Let $f_1(x),\ldots,f_r(x)$ be an involutive collection of C^∞ linearly independent vector fields on M. Given any point $x_0 \in M$ there exists a unique maximal C^∞ submanifold S of M containing x_0 such that $T_x(S)$ is the space spanned by $f_1(x),\ldots,f_r(x)$ for every $x \in S$: i.e. S is the unique integral manifold of $f_1(x),\ldots,f_r(x)$ through x_0 and contained in M.*

It is well known that under the conditions of the Frobenius Theorem, M is a C^∞ $(n-r)$-parameter family of C^∞ r-dimensional manifolds.

Next we consider the possibility that the set of vector fields $f_1(x),\ldots,f_r(x)$ is not involutive. Suppose $f_1(x),\ldots,f_r(x)$ are complete (i.e. the integral curves of each f_i are defined for all $-\infty < t < \infty$). Given f, for each t $(\exp tf)$ defines a map of M into itself which is produced by the flow on M defined by the differential equation $\dot{x} = f(x(t))$, and $(\exp tf)x_0$ denotes the solution starting at x_0 and moving in the positive time sense. Repeated exponentials like $(\exp tf_2)(\exp tf_1)x_0$ mean that we start at x_0, move along the integral curve of f_1 for positive t units of time, and then along the integral curve of f_2 for positive t units of time. We denote the smallest subgroup of the diffeomorphisms of M with itself which contains $\exp tf$ for all f in $\{f_1,\ldots,f_r\}$ by $\{\exp\{f_i\}\}_G$.

Let L_A be the Lie algebra of vector fields generated by $f_1,\ldots,f_r$ and by taking successive Lie brackets of these vector fields. We assume that this is a vector bundle with vector space dimension p on M. Defining $\{\exp\{f_i\}_{L_A}\}_G$, $\{\exp\{f_i\}_G\}x_0$ and $\{\exp\{f_i\}_{L_A}\}x_0$ in the obvious manner, we state the following version of Chow's Theorem (see [1]).

Theorem 3.2 [2]. *Given any point $x_0 \in M$, there exists a unique maximal C^∞ real p-dimensional submanifold $S \subset M$ containing x_0 such that $\{exp\{f_i\}_G\}x_0 = \{exp\{f_i\}_{L_A}\}_G x_0 = S$. This S is the unique submanifold of M through x_0 having L_A as its tangent bundle.*

Thus, under the hypotheses of Chow's Theorem, M is a C^∞ $(n-p)$-parameter family of C^∞ p-dimensional manifolds.

We now return to our system (2.2) and let L_A be the Lie algebra of vector fields on M generated by $f, g_1,\ldots,g_m$ and by taking successive Lie brackets. If we could control the drift term f from (2.2), then the reachable set from $x_0 \in M$ is equal to S, assuming that the hypotheses of Chow's Theorem are satisfied for the vector fields $f, g_1,\ldots,g_m$. We generalize Definition 2.3 as follows.

Definition 3.3. Let $f, g_1,\ldots,g_m$ satisfy the hypotheses of Chow's Theorem with the dimension of L_A being p. If S is the C^∞ p-dimensional submanifold of M through x_0 of Chow's Theorem and if f is treated as a drift term without control, then the largest open subset U of S which is reachable from x_0 for system (2.2) is called the *region of reachability from x_0*. If $U = S$, we say that the system is *S-controllable from x_0*.

An obvious question is that given a system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \qquad x(0) = x_0 \in M,$$

satisfying Chow's Theorem, is there a nonempty open subset of S which is reachable from x_0 when f is the drift term? This question is answered affirmatively by the proof of Theorem 3.1 in [8]. Thus Definition 3.3 is not vacuous. We remark that the work in [8] is for real analytic vector fields and real analytic manifolds. However, the advantage of real analytic over C^∞ is that the Jacobian matrix of a real analytic map that has maximal rank at some point must have maximal rank at almost all points in the appropriate sense (see [8]). In [7] Krener gives an interesting proof of the existence of a reachable open set in S.

Unlike the hypersurface case, we do not require that $f, g_1,\ldots,g_m$ be linearly independent on M. Unless otherwise specified, for the remainder of this article we assume L_A is a vector bundle with vector space dimension p, that $f, g_1,\ldots,g_m$ are complete vector fields, and that S is the manifold through x_0 given by Chow's Theorem. Also, since we can use unbounded (both positive and negative) controls, we may as well assume it is possible to move along the integral curves of $g_1,\ldots,g_m$.

In the following definition A and B are C^1 submanifolds of M of dimensions k and $n-k$ respectively.

Definition 3.4. The manifolds A and B *intersect transversally* at a point $x \in A \cap B$ if and only if $T_x(A) \oplus T_x(B) = T_x(M)$, where $\oplus$ denotes the direct sum.

We need a sequel for Definition 2.4.

Definition 3.5. Let O be an open set in S and let $x \in \partial O$ (taken relative to S). Then f *points in the direction of* O (or *towards* O) *at* x if there exists an open neighborhood W of x in M such that the vector assigned by f at x, projected into S (by the exponential map), and intersected with $W - \{x\}$ is contained in O. If this is true for every $x \in \partial O$, then f *points in the direction of* O *on* ∂O.

If f points in the direction of O at $x \in \partial O$ and if ∂O is C^1 near x, then there is some open neighborhood W of x in M so that the integral curve of f (moving in positive time) starting at x and intersected with $W - \{x\}$ is contained in O.

Definition 3.5'. Let O be an open set in S and let $x \in \partial O$. Then f *points in the direction of* $\overline{O}$ (the closure of O in S) *at* x if and only if there exists 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 $\overline{O}$. If this is true for every $x \in \partial O$, then f *points in the direction of* $\overline{O}$ *on* ∂O.

Our first result concerning necessary conditions parallels Theorem 3.2 found in [5]. Let L_A' be the Lie algebra generated by $g_1, \ldots, g_m$ and their Lie brackets, and let $\{L_A'\}_x$ be the restriction of L_A' to the point x.

Theorem 3.6. *Let O be an open set in S which is reachable from x_0 for system (2.2), 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 C^1 real $(p-1)$-dimensional submanifold of S. If any one of the following conditions holds, then O is not the region of reachability from x_0:*

i) *the dimension of $\{L_A'\}_x$ as a vector space is p.*
ii) *the integral curve of some g_i, $1 \leq i \leq m$, is transversal to ∂O at x.*
iii) *f assigns at x a vector which does not point in the direction of $\overline{O}$.*

Proof. The neighborhood W of x in M will be made smaller whenever necessary.

If i) holds at x then it holds for all points in $W \cap \partial O$ since p is maximal by the assumption on L_A and $L_A' \subset L_A$. Suppose that each g_i, $1 \leq i \leq m$, assigns only tangent vectors to $W \cap \partial O$; i.e. none of the g_i's is transversal at any point of $W \cap \partial O$. Then the Lie algebra generated by $g_1, \ldots, g_m$ and successive Lie brackets is contained in the tangent bundle to $W \cap \partial O$, a contradiction since this bundle is $(p-1)$-dimensional. Thus there is a point $y \in W \cap \partial O$ arbitrarily close to x with the integral curve of some g_i, $1 \leq i \leq m$, transversal to ∂O at y, and i) reduces to ii).

Suppose that condition ii) holds at x. If the integral curve of g_1, chosen arbitrarily from $g_1, \ldots, g_m$ and renumbered if necessary, is transversal to ∂O at x, then it is transversal to ∂O in $W \cap \partial O$. Following the integral curves of g_1 in S that begin at points in O which are sufficiently close to x, and continuing past $W \cap \partial O$, we have that $O \subset S$ is not the region of reachability from x_0. This is true since O is reachable from x_0.

If iii) is true at x, then the arguments given in ii) with f replacing g_1 implies that O is not the largest reachable region from x_0. □

We state a result similar to the necessary part of Theorem 2.5 under the local assumption of a C^1 boundary.

Theorem 3.7. *Let U be the region of reachability from x_0 of the system* (2.2). *Suppose ∂U is a C^1 manifold for an open neighborhood W of $x \in \partial U$ in M. Then $W \cap \partial U$ contains the integral curves of $g_1, \ldots, g_m$ in W which intersect ∂U, and f assigns vectors to $W \cap \partial U$ which point in the direction of $\overline{U}$. Moreover, there is an open dense set in $W \cap \partial U$ for which the vectors from f point in the direction of U.*

Proof. It is obvious from Theorem 3.6 that we need only prove the last statement of Theorem 3.7. Suppose there is an open set V in $W \cap \partial U$ for which the vector field f is contained in the tangent bundle to V. Then the bundle generated by $f, g_1, \ldots, g_m$ and their Lie brackets on V is contained in the tangent bundle to the $(p-1)$-dimensional manifold V. This contradicts our assumption that the vector space dimension of L_A is p. □

In addition to our hypothesis that the dimension of L_A is p, suppose that we also let the dimension of L_A' be constant on M. Given any point $x \in M$, Chow's Theorem gives us a C^∞ maximal integral manifold through x with L_A' as its tangent bundle. In Theorem 3.7 we would have that ∂U must contain these integral manifolds if $x \in \partial U$.

The following theorem from [6] will allow us to reduce somewhat the C^1 assumption of Theorem 3.7. The statement concerning a C^2 boundary can be relaxed to C^1, or we can simply replace C^1 in our preceding results by C^2.

Theorem 3.8 [6]. *Let M be a C^∞ manifold of dimension n, and let H be a subbundle of the tangent bundle to M with 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 C^2 boundary then for each $x \in \partial O \cap \partial U$ we have $T_x(\partial O) = H_x$ (the vector space of H at x). Then for each point $x \in \partial U$, there is a neighborhood V of x, a real valued-function $h \in C^\infty(V)$ with nonzero differential for all points in V, and a closed nowhere dense set $E \subset \mathbb{R}$ such that*

1) $\partial U \cap V = \{x \in V \mid h(x) \in E\}$,

2) *for each $l \in E$, $S_l = \{x \in V \mid h(x) = l\}$ is an integral manifold of H, i.e. the boundary of U is foliated by integral manifolds of H.*

Under the restriction that L_A' is a bundle we need no differentiability restrictions as in Theorem 3.7.

Theorem 3.9. *Let U be the region of reachability from x_0 of the system* (2.2). *If the vector space dimension of L_A' is the constant $p' < p$ on M, then ∂U contains the C^∞ p'-dimensional integral manifolds (or more generally, the foliation of such manifolds) of the bundle L_A' that intersect ∂U. Also the vector field f always points in the direction of $\overline{U}$ on ∂U.*

Proof. Let x be any point in ∂U. There is an open neighborhood W of x in M which consists of a C^∞ $(n-p')$-parameter family of p'-dimensional integral manifolds of L_A'. Since $L_A' \subset L_A$, $W \cap S$ consists of a C^∞ $(p-p')$-parameter family of p'-dimensional integral manifolds of L_A'. Take an arbitrary C^∞ 1-parameter family of the p'-dimensional integral manifolds the union of which

contains x in its interior. This 1-parameter family forms a C^∞ $(p'+1)$-dimensional manifold L containing x, and L_A' is a p'-dimensional subbundle of its tangent bundle. If L (or an open neighborhood of x in L) is contained in ∂U, there is nothing to prove. If $\partial U \cap L$ contains an open set in L and x is a boundary point of this open set, then we simply choose another L so that this does not occur. Otherwise, we apply Theorem 3.8 with U replaced by $U \cap L$, ∂U by $\partial U \cap L$, M by L, and H by L_A', and Theorem 3.7 (see the remarks after the proof of Theorem 3.7) to complete the proof of the first conclusion.

Suppose $x \in \partial U$ and f does not point towards $\overline{U}$ at x. Moving along the integral curve of f starting at x we reach a point in $\overline{U}$, the complement of $\overline{U}$ in S. Starting at all points in ∂U for an open neighborhood V of x in ∂U, by continuous dependence on initial data and uniqueness of integral curves, we find that we can reach an open set in $\overline{U}$ from V. It remains to be shown that we can reach an open set in $\overline{U}$ from U, a reachable set. Now all integral curves of f which pass through V fill up an open set in S containing x in its closure. Since the intersection of this open set with U contains an open set with V in its boundary, we can reach an open subset of V along integral curves of f from U. Hence U is not the region of reachability from x_0, a contradiction. □

Suppose there exist no sets in S like ∂U of Theorem 3.9 which disconnect S in the appropriate sense (see Corollary 4.3 in [5]). Then our system is S-controllable from an arbitrary $x_0 \in S$.

If $L_A = L_A'$ (i.e. $p = p'$), then by part i) of Theorem 3.6 the system (2.2) is S-controllable from any $x_0 \in S$. If $p' = p - 1$, then the system has many properties of the hypersurface system (2.1).

Theorem 3.10. *Let U be the region of reachability of the system* (2.2) *from x_0 and assume that $p' = p - 1$. Then ∂U is an integral manifold of the bundle L_A' and f assigns vectors on ∂U which point in the direction of $\overline{U}$ on ∂U and in the direction of U for an open dense subset of ∂U.*

Proof. Since $p' = p - 1$, we have by Theorem 3.9 that ∂U is a C^∞ integral manifold of L_A'. The statement concerning f follows from Theorem 3.7 □

This concludes our discussion of necessary conditions for C^∞ manifolds M. If M is real analytic and if our vector fields on M are real analytic, then statements of the Frobenius Theorem and Chow's Theorem exist (see [1]) which will improve some of our results. We shall not go into this matter in this paper.

IV. Sufficient Conditions

Given $x_0 \in M$ and S containing x_0, sufficient conditions for an open set $O \subset S$ to be the region of reachability from x_0 for a general system like (2.2) are of interest to us. One such result in the literature applies to the system

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)), \qquad x(0) = x_0 \in M. \tag{4.1}$$

Theorem 4.1 [4]. *Let L_A' be the Lie algebra generated by f and g in* (4.1), *and let L_0 be the smallest subalgebra of L_A' which contains g and is closed under Lie bracketing with f. Suppose that for all h in L_0 we have $[h,g] = \alpha_h g$ for some constant α_h. Then the reachable set from x_0 is equal to*

$$\{\exp L_0\}(\exp tf)x_0.$$

For a system that behaves like a hypersurface system as in Theorem 3.10, we have the following theorem.

Theorem 4.2. *Let $x_0 \in M$ and let O be an open subset of $S \subset M$ which contains x_0 in its closure and which is reachable from x_0 for* (2.2). *Suppose that L_A' is a vector bundle with dimension $p' = p - 1$ on M. If ∂O is an integral manifold of the bundle L_A' and if f points in the direction of O on ∂O, then O is the region of reachability U from x_0.*

Proof. Since f points toward O, f and the bundle L_A' must be linearly independent on ∂O. If O is not the region of reachability, then there is a point $x \in \partial O$ and a neighborhood W of x in S which can be reached from x_0. Let $X_1,\ldots,X_{p'}$ be a basis for L_A' in W. Since ∂O is a C^∞ integral manifold of L_A', Lie brackets of the form $[X_i,X_j]$, $1 \leq i,j \leq p', i \neq j$, yield no "new" directions in which to move from $W \cap \partial O$ (i.e. the set $\{X_1,\ldots,X_{p'}\}$ is involutive). Because $f, X_1,\ldots,X_{p'}$ span $T(S)$ on W, brackets like $[f,X_i]$, $i = 1,\ldots,p'$ give us vector fields on W which are linear combinations of $f, X_1,\ldots,X_{p'}$. The same is also true for successive Lie brackets. Not being able to control the drift term f, the only linear combinations which arise from these brackets and which indicate directions that we can move are those with a nonnegative coefficient for the f term (see the proof of Theorem 4.3 in which a more general case is considered). Since f points toward O, and ∂O is an integral manifold of L_A', we are unable to move outside of O, and in particular, reach W. □

The above proof follows that of the sufficient conditions for the hypersurface case given in [5]. These proofs depend on the fact that $[f,X]$, where $X \in L_A'$, near some boundary point of O, must be a linear combination of f and basis elements in L_A'. There are no "new" directions, given to us by some Lie bracket, that we can move.

The following interpretation of the Lie bracket is taken from [3]. Let f and g be vector fields on a manifold M and let $x_0 \in M$. Then $[f,g]$ is the tangent vector at x_0 to the curve segment

$$t \to \exp(-\sqrt{t}\, f)\exp(-\sqrt{t}\, g)\exp(\sqrt{t}\, f)\exp(\sqrt{t}\, g)x_0 \qquad (t \geq 0).$$

We return to our system (2.2) and let x_0, M, S, L_A, and L_A' be as before. Our next theorem is the converse statement of Theorem 3.9.

Theorem 4.3. *Let O be an open subset of $S \subset M$ containing x_0 in its closure and which is reachable from x_0. Suppose that L_A' is a vector bundle with vector space dimension p' on M and ∂O contains the C^∞ p'-dimensional integral manifolds of L_A' that intersect it. If f points in the direction of $\bar{O}$ on ∂O, then O is the region of reachability from x_0 for the system* (2.2).

Proof. If we are to reach an open set in the complement of $\bar{O}$, $\bar{\bar{O}}$, then we must pass through ∂O at some point $x \in \partial O$. It is obvious that an open subset of $\bar{O}$ cannot be reached by using the integral curves of f (in positive time) or the integral curves of any section in the bundle L_A'. Moreover, we cannot reach $\bar{O}$ by

using a linear combination of these with a nonnegative coefficient for the $f(x(t))$ term. Hence we must consider Lie brackets of the form $[f,g_i], 1 \leq i \leq m$, and higher order brackets of this type.

If we travel along an integral curve of f in positive time, then we assume that we can return along this curve until the starting point is reached. We consider $[f,g_i]$ at x, our point in ∂O, for some i, $1 \leq i \leq m$. First we move from x in the g_i direction for $\sqrt{t}$ units of time, and we remain in ∂O since ∂O contains the integral manifolds of L_A'. Next, moving along f for $\sqrt{t}$ units of time, we stay in $\bar{O}$ because f points in the direction of $\bar{O}$ on ∂O. If f left us in ∂O, then using $-g$ for $\sqrt{t}$ units of time keeps us in ∂O. If f left us in O, then we stay there. Finally, moving along $-f$ for $\sqrt{t}$ units of time either leaves us in $\bar{O}$ which is fine, or in $\tilde{\bar{O}}$, which is impossible since we cannot control f. Hence we can start an integral curve at x in the direction of $[f,g_i]$ if and only if $[f,g_i]$ points towards $\bar{O}$. This is true for all $i = 1,\ldots,m$ and also for $[g_i,f]$ at x. By repeating the above argument for successive Lie brackets, we find that we can start an integral curve at x in the direction of a particular bracket if and only if that bracket points in the direction of $\bar{O}$ at x.

Thus the Lie brackets at x will not let us reach an open set in $\tilde{\bar{O}}$, and O is the region of reachability. □

Suppose that we knew an open subset of S could be reached from x_0 which contains x_0 in its closure. Let U be the smallest open subset of S with $x_0 \in \bar{U}$ and satifying the hypotheses of Theorem 4.3. By Theorem 3.9 we can reach U, and we would have the following result.

Conjecture 4.4. *Let $x_0 \in M$ and let U be the smallest open subset of S containing x_0 in its closure and satisfying the hypotheses of Theorem* 4.3. *Then U is the region of reachability from x_0 for system* (2.2).

In the conclusion of Theorem 3.9 and in the hypotheses of Theorem 4.3 we have that f must point towards $\bar{O}$ on ∂O (or $\bar{U}$ on ∂U). Suppose there is an open neighborhood W of $x \in \partial O$ so that all integral curves of f that intersect $W \cap \partial O$ actually are contained in $W \cap \partial O$. Since ∂O contains the integral manifolds of L_A', following integral curves of f and g_i, $1 \leq i \leq m$, in $W \cap \partial O$ leaves us in $W \cap \partial O$. But this contradicts the fact that the vector space dimension of L_A is p and Chow's Theorem.

References

1. R. W. Brockett, Nonlinear systems and differential geometry, *Proc. IEEE* 64, 61–72, (1976).
2. W. L. Chow, Uber Systems von Linearen Partiellen Differentialgleichungen erster Ordnung, *Math Ann.* 177 (1939), 98–105.
3. S. Helgason, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.
4. R. Hirschorn, Topological semigroups, set of generators, and controllability, *Duke Math. J.* 40, 937–947, (1973).
5. L. R. Hunt, Controllability of nonlinear hypersurface systems, in progress.
6. L. R. Hunt, J. C. Polking, and M. J. Strauss, Unique continuation for solutions to the induced Cauchy-Riemann equations, *J. Differential Equations* 23, 436–447 (1977).
7. A. J. Krener, A generalization of Chow's Theorem and the Bang-Bang Theorem to nonlinear control problems, *SIAM J. Control* 12, 43–52, (1974).
8. H. Sussmann, and V. Jurdjevic, Controllability of nonlinear systems, *J. Differential Equations* 12, 95–116, (1972).

Received July 24, 1978 and in revised form December 15, 1978 and January 19, 1979

9. Reprint of "Control Theory for Nonlinear Systems" by L.R. Hunt from the Proceedings of the 1979 International Symposium on the Mathematic of Networks and Systems, T.H. Delft, July 1979, pp. 339-343.

Abstract

Suppose M is a connected paracompact C^∞ n-dimensional manifold and $f, g_1, \ldots, g_m$ are complete C^∞ vector fields on M. We examine the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M,$$

where $u_1, \ldots, u_m$ are real-valued controls. Under the assumptions that the Lie algebra generated by $f, g_1, \ldots, g_m$ and the Lie algebra generated by $g_1, \ldots, g_m$ are vector bundles of dimensions p and p' respectively, with $p' < p$, we characterize the largest open subset of the submanifold S of M (given by Chow's Theorem) which is reachable from x_0 for our system. If M is a real-analytic manifold and $f, g_1, \ldots, g_m$ are complete real-analytic vector fields, then the assumption that the Lie algebra generated by $f, g_1, \ldots, g_m$ is a vector bundle of constant dimension can be removed. In addition the requirement that the Lie algebra generated by $g_1, \ldots, g_m$ is a vector bundle of constant dimension can be weakened.

1. INTRODUCTION

Let M be a connected paracompact C^∞ n-dimensional manifold, and let $f, g_1, \ldots, g_m$ be complete C^∞ vector fields defined on M. If $u_1, \ldots, u_m$ are real-valued controls, we are interested in characterizing the reachable set of the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M \tag{1.1}$$

We assume that the Lie algebra L_A generated by $f, g_1, \ldots, g_m$ and their Lie brackets is a vector bundle of dimension p. By Chow's Theorem we know that the reachable set is contained in a C^∞ p-dimensional submanifold S of M which is the integral manifold of L_A through x_0. (1),(2) For the real analytic case and the C^∞ case we know that an open subset of S is reachable.(8),(7) Krener's proof is quite nice in that he shows that we can go up one dimension at a time until an open subset of S is reached. This parallels the one dimension at a time proof of Greenfield which is concerned with holomorphic extension theory in several complex variables under the assumption of constant dimension of the Levi algebra.(4) Working through Krener's proof we find that it is possible to reach an open subset of S which is arbitrarily close to x_0.

The largest open subset U of S which is reachable from x_0 for the system (1.1) is called the region of reachability. If we assume that the Lie algebra L_A' generated by $g_1, \ldots, g_m$ and their Lie brackets is a vector bundle with vector space dimension $p' < p$, then the region of reachability of (1.1) from x_0 is a connected open subset of S, which we now call

*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.

the domain of reachability. Using published results we show that the domain of reachability U is the smallest open subset of S containing x_0 in its closure and satisfying the following conditions. (6) The boundary of U contains C^∞ p'-dimensional integral manifolds of L_A', and the vector field f assigns vectors on ∂U which point in the direction of $\bar{U}$. If p' = p we know that we reach all points in S. For the real-analytic case we can remove some of our assumptions on the dimensions of L_A and L_A'. (6)

Section 2 of this paper contains definitions, and we prove our result concerning the domain of reachability of the system (1.1) in section 3. In section 4 we assume that M and f, $g_1,\ldots,g_m$ are real-analytic and improve our main result, Theorem 3.1, for this case.

2. DEFINITIONS

Denote by T(M) the tangent bundle to our n-dimensional manifold M. If $x \in M$, then $T_x(M)$ is the tangent space in T(M) at x. For X a vector field on M, a curve α is an integral curve of X if α is a C^∞ mapping from a closed interval $I \subset R$ into M so that

$$\frac{d\alpha(t)}{dt} = X(\alpha(t)) \text{ for all } t \in I.$$

If D is a subset of T(M), then an integral curve of D 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 sections $X_1,\ldots,X_k$ of T(M) in D with the restriction of α to $[t_{i-1},t_i]$ being an integral curve of X_i, for each $i = 1,2,\ldots,k$. For $x_0 \in M$, a point $x \in M$ is D-reachable from x_0 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 D-reachable from x_0 if every point $x \in A$ is D-reachable from x_0.

In the remainder of this article the set D of interest to us is the subset of T(M) given by the system (1.1). Hence we drop the D in the above definitions.

Let S be a p-dimensional submanifold of M, let O be an open subset of S, and let $x \in \partial O$. A vector field f on M points in the direction of $\bar{O}$ (the closure of O in S) at x if their exists an open neighborhood W of x in M so that the integral curve of f starting at x and intersected with W is contained in $\bar{O}$. If this occurs for every point $x \in \partial O$, then f points in the direction of $\bar{O}$ on ∂O.

Given C^∞ vector fields f and g on M, the Lie bracket of f and g is given by

$$[f,g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g,$$

where $\frac{\partial g}{\partial x}$ denotes the Jacobian matrix. Of course we can take successive Lie brackets like [f,[f,g]], [g,[f,g]], etc.

As before L_A is the Lie algebra generated by f, $g_1,\ldots,g_m$ from (1.1) and all successive Lie brackets. If L_A is a vector bundle with vector space dimension p, Chow's Theorem together with results of Krener show that we can reach an open subset of the C^∞ p-dimensional submanifold $S \subset M$ through x_0.(7) The largest open subset U of S which is reachable for (1.1) from x_0 is the region of reachability from x_0. If U = S, we say that the system is S-controllable from x_0.

In the next section we give a characterization of the region of reachability of our system.

3. THE DOMAIN OF REACHABILITY

Our objective is to prove the following result for system (1.1). This theorem can be found elsewhere for the case m = n-1 and f, $g_1,\ldots,g_{n-1}$ linearly independent on M.(5)

Theorem 3.1. Let O be the smallest open subset of S containing x_0 in its closure and satisfying the following properties. Suppose that the Lie algebra L_A' generated by $g_1,\ldots,g_m$ and successive Lie brackets is a vector bundle of vector space dimension $p' < p$ on M and ∂O contains the C^∞ p-dimensional integral manifolds of L_A' that intersect it. If f points in the direction of $\bar{O}$ on ∂O, then O is the region of reachability from x_0 for our system.

The following two results give necessary conditions and sufficient conditions for an open set to be the region of reachability from x_0 for (1.1).(6) In both theorems we assume that L_A' has constant vector

space dimension $p' < p$ on M.

Theorem 3.2. Let U be the region of reachability from x_0 of our system. Then ∂U contains the C^∞ p'-dimensional integral manifolds of L_A' which intersect it and f points in the direction of $\bar{U}$ on ∂U.

Theorem 3.3. Let O be an open subset of $S \subset M$ containing x_0 in its closure and which is reachable from x_0. Suppose ∂O contains the C^∞ p'-dimensional integral manifolds of L_A' which intersect it and f points in the direction of $\bar{O}$ on ∂O. Then O is the region of reachability from x_0 for the system (1.1).

Proof of Theorem 3.1.

As mentioned earlier Krener's proof shows that arbitrarily close to any reachable point is a reachable open set in S.(7) Thus we can reach an open subset V of S which contains x_0 in its closure. From the definition of reachable, it is obvious that we can choose V so that $\bar{V}$ is connected. By the proof of Theorem 3.2 we can assume that ∂V contains the integral manifolds of L_A' which intersect it and f points towards $\bar{V}$ on ∂V.(6)

We first show that V is connected. Suppose there exist two disconnected components V_1 and V_2 of V with a point $x \in \partial V_1 \cap \partial V_2$. Then the unique integral manifold N of L_A' through x is contained in $\partial V_1 \cap \partial V_2$ and f points in the direction of $\bar{V}_1$ and $\bar{V}_2$ on this manifold. If every point on N is an equilibrium point for $\dot{x} = f(x(t))$, then we contradict the fact that the vector space dimension p' of L_A' is less than the vector space dimension p of L_A on N. Thus we may as well assume that the differential equation $\dot{x} = f(x(t))$ has no equilibrium points on N. Each solution of $\dot{x} = f(x(t))$ which starts at a point in N must remain in both ∂V_1 and ∂V_2 since f points towards $\bar{V}_1$ and $\bar{V}_2$ on ∂V_1 and ∂V_2, respectively. Through each point of $\partial V_1 \cap \partial V_2$ we have an integral manifold of L_A'. Hence we have a subset of $\partial V_1 \cap \partial V_2$ which contains both integral curves of f and integral manifolds of L_A'. It is impossible that the vector space dimension of L_A is p on this subset, a contradiction. Since $\bar{V}$ is connected, it follows that V is connected.

By Theorem 3.3 we have that V is the region of reachability U from x_0 for our system. It remains to show that in fact $V = O$. Since $x_0 \in \partial V \cap \partial O$ and ∂O satisfies the conditions concerning L_A' and f, the argument given above shows it is impossible to have $V \cap O = \emptyset$. If $V \neq O$, then there is a point $x \in \partial O \cap \partial V$, and a repeat of the same argument with obvious changes implies a contradiction. Hence O is the connected region of reachability U of (1.1) from x_0. Q.E.D.

Theorem 3.1 completely characterizes the region of reachability U (or more correctly, the domain of reachability) of our system by determining conditions that must be satisfied by ∂U. Of course, it would be nice to develop a computational method for finding integral manifolds of subbundles of the tangent bundle. Research in this direction is currently being done by Michael Freeman in the case that M is a real-analytic manifold and f, $g_1, \ldots, g_m$ are real-analytic vector fields on M. For this reason we prove an improved version of Theorem 3.1 in the real-analytic case.

4. REAL-ANALYTIC CONTROLLABILITY

In this section we assume that M and f, $g_1, \ldots, g_m$ are real-analytic. Thus we can apply the real-analytic version of Chow's Theorem which allows us to remove the assumption that the Lie algebra L_A generated by f, $g_1, \ldots, g_m$ and their Lie brackets is a vector bundle of dimension p on M.(1),(3) If this Lie algebra has dimension p at x_0, then there is a real-analytic p-dimensional submanifold S of M through x_0 which contains the reachable set. With this in mind we state and prove the next result.

Theorem 4.1.

If O is the smallest open subset of S containing x_0 in its closure such that f points in the direction of $\bar{O}$ on ∂O, then O is the region of reachability from x_0 for our system provided one of the following conditions is satisfied.

i) The Lie algebra L_A' generated by $g_1, \ldots, g_m$ and successive Lie brackets is a vector

bundle of dimension $p' < p$ on M and ∂O contains the real-analytic p'-dimensional integral manifolds of L_A' that intersect it.

ii) The Lie algebra L_A' generated by $g_1,\ldots,g_m$ and successive Lie brackets has dimension p at x_0 and ∂O contains the real-analytic integral manifolds of L_A' that intersect it.

Proof. The statement regarding i) follows directly from Theorem 3.1.

Suppose that L_A' has vector space dimension p at x_0. Since our vector fields are real analytic, the set of points in S where this dimension is less than p are nowhere dense in S. Let S' be the largest connected open component of S which contains x_0 and on which the dimension of L_A' is p. By known results we have that we can reach an open neighborhood of any point at which the dimension of L_A' is p.(6) Thus S' is a reachable set which is contained in the region of reachability of our system. If $S' = S$ we have nothing to prove, and we assume that $\partial S' \subset S$ is nonempty. Let x be an arbitrary point in $\partial S'$. Since the dimension of L_A' at x is less than p, by the real-analytic version of Chow's theorem there is a real analytic manifold N (which is the unique integral manifold of L_A') through x of dimension less than p. Because the dimension of L_A' is constant along this manifold, N cannot move into any open subset of S where the dimension of L_A' is p. Thus N must remain in $\partial S'$, and $\partial S'$ consists of the integral manifolds of L_A' that intersect it.

If the integral curve of f starting at some point x in $\partial S'$ moves into the complement of $\bar{S}'$, the same is true for all points in $\partial S'$ near x. From this we conclude that an open subset of the complement of $\bar{S}'$ can be reached by starting at points in S' near ∂S. Thus we assume that f points in the direction of $\bar{S}'$ on $\partial S'$, implying $S' = O$. Since the integral manifolds of L_A' which intersect $\partial S' = \partial O$ are contained in ∂O and since f points towards $\bar{O}$ on ∂O, a repeat of the arguments given elsewhere shows that O is the region of reachability from x_0 for our system.(6) In this case x_0 is actually an interior point of O because the dimension of L_A' at x_0 is p. Q.E.D

For an example we consider the system in R^2,

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)),\ x(0) = x_0 \in R^2, \quad (1.2)$$

where f and g are real-analytic on R^2. Suppose that we want to reach a point x in R^2 from x_0. First we compute the Lie algebra L_A generated by f and g at x_0. If the vector space dimension at x_0 is 1, there is an 1-dimensional integral manifold N of L_A through x_0. If x is not in this integral manifold, there is no hope to reach it. If x is an element of N, then we must check for equilibrium points of g in N. If x_0 is an equilibrium point, then it divides N into two components N_1 and N_2. If f points in the direction of $\bar{N}_1$ at x_0 and $x \in N_2$, or if f points in the direction of $\bar{N}_2$ at x_0 and $x \in N_1$, then x is not in the reachable set from x_0. So we assume that f points in the direction of $\bar{N}_1$ at x_0 and $x \in N_1$. Thus we must check all equilibrium points of g in N_1 between x_0 and x, including x itself. If at any such point, the vector f points in such a direction that the integral curve of f moves in N towards x_0, then we cannot reach x.

Suppose that the dimension of L_A at x_0 is 2. Then there is a unique integral manifold, which we also call N, of L through x_0 that contains the reachable set of our system from x_0. We assume that x_0 and x are not equilibrium points of g, that $x \in N$, and that the equilibrium points of g do not separate N into 2 or more components (remember that this set of equilibrium points is nowhere dense in N). Thus we can delete the set of equilibrium points of g, and we denote by N' the remaining subset of N. This set N' is a connected real-analytic manifold. We have that the dimension of L_A', the Lie algebra consisting of just g itself, is 1 on N'. By Theorem 3.1 the problem of determining if x is reachable from x_0 reduces to the problem of deciding if x is in the region of reachability from x_0 or not. This theorem completely characterizes the region by giving necessary and sufficient conditions on its boundary. Hence we ask if there is an integral curve of g which separates N' into two components, one containing x and the other containing x_0, such that f points in the direction of the closure of

the component containing x_0. If such an integral curve exists, we cannot reach x from x_0. Conversely, if such an integral curve does not exist, then x is reachable from x_0.

REFERENCES

1. R. W. Brockett, Nonlinear systems and differential geometry, Proc. IEEE 64(1967), 61-72.

2. W. L. Chow, Über Systems von Linearen Partiellen Differentialgleichungen erster Ordnung, Math. Ann. 177(1939), 98-105.

3. M. Freeman, Integration of analytic differential systems with singularities and some applications to real submanifolds of $\mathbb{C}^n$, to appear.

4. S. J. Greenfield, Cauchy-Riemann equations in several variables, Ann. Scuola Norm. Sup. Pisa 22(1968), 275-314.

5. L. R. Hunt, Controllability of nonlinear hypersurface systems, to appear.

6. L. R. Hunt, Controllability of general nonlinear systems, Math. Systems Theory, to appear.

7. A. J. Krener, A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems, SIAM J. Control 12(1974), 43-52.

8. H. J. Sussmann and V. Jurdjevic, Controllability of nonlinear systems, J. Differential Equations 12(1972), 95-116.

Professor Hunt is currently in the Department of Mathematics at Texas Tech University. He has been at this institution for approximately ten years. Professor Hunt received his Bachelor of Science degree from Baylor University in 1964 and his Ph.D. from Rice University in 1970. He is a member of the American Math. Society and SIAM.

10. Reprint of "Controllability of Nonlinear Systems" by L.R. Hunt from the Proceeding of the 12th Asilomar Conference on Circuits, Systems, and Computers, Pacific Grove, Ca., November 1978, pp. 466-467.

Let M be a connected $\mathcal{C}^\infty$ real n-dimensional manifold. Given a nonlinear system where the controls enter linearly, we find the largest open subset of M (or the largest open subset of a submanifold of M) which is reachable from some initial point $x_0 \in M$. Assumptions that certain Lie algebras, which are generated by the vector fields of the system, form vector subbundles of M are made in some cases.

Suppose we have the nonlinear system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M,$$

where M is a connected $\mathcal{C}^\infty$ real 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. We are interested in characterizing the reachable set in M of this system. R. W. Brockett's paper contains a nice discussion of this topic.[1]

First we consider the hypersurface case in which $m = n-1$ and $f, g_1, \ldots, g_{n-1}$ are linearly independent on M. It is known that the reachable set of a hypersurface system contains an open set in M. The largest open subset of M which is reachable from x_0 is called the region of reachability U from x_0, and if $U = M$, the system is controllable from x_0. If $U \neq M$ we prove that the boundary of U is a $\mathcal{C}^\infty$ real (n-1)-dimensional submanifold of M which is an integral manifold of the vector fields $g_1, \ldots, g_{n-1}$ and that the vector field f points in the direction of U on ∂U. This leads to a result which gives us the region of reachability of the hypersurface system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M.$$

Theorem 1.[3] Suppose U is the smallest open subset of M with $x_0 \in U$ satisfying ∂U is an integral manifold of $g_1, \ldots, g_{n-1}$ and f assigns vectors on ∂U which point in the direction of U. Then U is the region of reachability from x_0 for our hypersurface system.

If no integral manifolds of $g_1, \ldots, g_{n-1}$ as given in the theorem exist, then the hypersurface system is controllable from any $x_0 \in M$.

The ideas used in proving the above result follow those found in the solution of the problem of uniqueness of analytic continuation for the CR-functions on a $\mathcal{C}^\infty$ real hypersurface in C^n, $n > 1$.[5]

For a general m, $1 \leq m \leq n-1$, and arbitrary $\mathcal{C}^\infty$ vector fields $f, g_1, \ldots, g_m$ in our nonlinear system, we assume that the Lie algebra generated by $f, g_1, \ldots, g_m$ and by taking successive Lie brackets of these vector fields is a vector bundle with constant fiber dimension p on M. By Chow's Theorem there exists a maximal $\mathcal{C}^\infty$ real p-dimensional submanifold S of M containing x_0 with the generated bundle as its tangent bundle.[1,2] It is known that the reachable set from x_0 must contain an open set in S.[6,7] The largest open subset U of S which is reachable from x_0 is called the region of reachability from x_0. If O is an open subset of S which is reachable from x_0, we find necessary conditions and sufficient conditions on the boundary of U in S so that $O = U$. Best results are obtained when it is assumed that the Lie algebra L_A' generated by $g_1, \ldots, g_m$ and their Lie brackets is a vector bundle on M.

Theorem 2.[4] Let U be the region of reachability from x_0 of the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M.$$

If the fiber dimension of L_A' is the constant

CH1369-8/79/0000-0466$00.75 © 1979 IEEE

$p' < p$ on M, then ∂U contains the ∞ p'-dimensional integral manifolds of the bundle L_A' that intersect ∂U. Also the vector fields f always point in the direction of $\bar{U}$ on ∂U.

Theorem 3.[4] Let O be an open subset of $S \subset M$ containing x_0 in its closure and which is reachable from x_0. Suppose that L_A' is a fiber bundle with fiber dimension p' on M and ∂O contains the p'-dimensional integral manifolds of L_A' that intersect it. If f points in the direction of $\bar{O}$ on ∂O, then O is the region of reachability from x_0 for the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M.$$

Although the above described concepts would appear to be abstract, in some cases, especially for a hypersurface system on a low dimensional manifold, they are readily tested via standard analytic techniques. Here, one generates the integral manifolds of the vector fields g_i, if any such manifolds exist, by numerically integrating the equation

$$\dot{x}(t) = \sum_{i=1}^{n-1} u_i(t) g_i(x(t)).$$

Then one determines the direction of flow by evaluating f along these manifolds. If no separating integral manifold admits a unidirectional flow the system is globally controllable, whereas if such a manifold exists it becomes a candidate for the boundary of the reachable set for points on one side of the separating manifold. In the case of a general nonlinear system with linear inputs one must first generate the Lie algebra of the g_i as an intermediary step to the computation of the integral manifolds. As such, the above described controllability conditions are not as readily tested though there are a number of special cases wherein a practical controllability test can be obtained. Indeed, we have investigated a number of examples which have appeared in the literature and found that in each case our controllability criterion has yielded a definitive characterization of the reachable sets in state-space.

Given a bundle of vector fields on an n-dimensional manifold M, current research involving computational methods of finding integral manifolds of the bundle is of obvious interest with regard to our controllability results.

References

1. R. W. Brockett, Nonlinear systems and differential geometry, Proc. IEEE 64 (1976), 61-72.
2. W. L. Chow, Über Systems von Linearen Partiellen Differentialgleichungen erster Ordnung, Math. Ann. 177 (139), 98-105.
3. L. R. Hunt, Controllability of nonlinear hypersurface systems, to appear.
4. L. R. Hunt, Controllability of general nonlinear systems, to appear.
5. L. R. Hunt, J. C. Polking, and M. J. Strauss, Unique continuation for solutions to the induced Cauchy-Riemann equations, J. Differential Equations 23 (1977), 436-447.
6. A. J. Krener, A generalization of Chow's Theorem and the Bang-Bang Theorem to nonlinear control problems, SIAM J. Control 12 (1974), 43-52.
7. H. Sussmann and V. Jurdjevic, Controllability of nonlinear systems, J. Differential Equations 12 (1972), 95-116.

Research supported in part by the National Science Foundation under NSF Grant MCS 76-05267-A01 and by the Joint Services Electronics Program under ONR Contract 76-C-1136.

11. Abstract of "Global Controllability of Nonlinear Systems in Two Dimensions", by L.R. Hunt.

Let M be a connected real-analytic 2-dimensional manifold. Consider the system

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)),\ x(0) = x_0 \in M,$$

where f and g are real-analytic vector fields on M which are linearly independent at some point of $\mathbb{R}^2$, and us is a real-valued control. Sufficient conditions on the vector fields f and g are given so that the system is controllable from x_0. Suppose that every integral curve of g which disconnects M has a point where f and g are linearly dependent, g(p) is nonzero, and that the Lie bracket [f,g] and g are linearly independent at p. Then the system is controllable (with the possible exception of a closed, nowhere dense set which is not reachable) from any point x_0 such that the vector space dimension of the Lie algebra L_A generated by f, g and successive Lie brackets is 2 at x_0. This is a generalization of the linear theory for the system

$$\dot{x}(t) = Ax(t) + u(t)B,\ x(0) = x_0 \in \mathbb{R}^2$$

in that the Lie bracket of Ax and B is the constant vector field AB. Hence if AB and B are linearly independent (i.e. the controllability matrix {B,AB} has rank 2), then the linear system is controllable.

12. <u>Abstract</u> of "Controllability of Nonlinear Hypersurface Systems" by L.R. Hunt.

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 $x_o \in \overline{U}$ 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 of $\overline{U}$. Then U is the region of reachability from x_o for our system. Much of the work is involved in proving a similar result in the more general C^∞ case under the stronger assumption that $f, g_1, .., g_{n-1}$ are linearly independent on the connected C^∞ n-dimensional manifold M.

13. Abstract of "Controllability and Stability" by L.R. Hunt.

Consider the system

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)),\ x(0) - x_0 \in \mathbb{R}^2,$$

where f and g are real-analytic vector fields on $\mathbb{R}^2$. If this is a controllable linear system, then it is well known the system is stablizable by linear feedback. We want to consider a similar problem for nonlinear systems, with emphasis on bilinear systems. Sufficient conditions for the above system to be controllable have been found, and implementation for bilinear systems has been discussed. If a bilinear system is controllable under these conditions, we show that we can move from any point $x_0 \in \mathbb{R}^2 - \{(0,0)\}$ to the origin.

Texas Tech University — Institute for Electronic Science

Joint Services Electronics Program — Research Unit: 3

1. Title of Investigation: Nonlinear Fault Analysis
2. Senior Investigator: R. Saeks — Telephone: (806) 742-3528
3. JSEP Funds: $23,500
4. Other Funds: $15,000*
5. Total Number of Professionals: PI's 1 (1 mo.) RA's 1 (1/2 time)
6. Summary:

The goal of the proposed work unit is to develop computationally efficient fault analysis algorithms which are compatible with the dual mode ATPG/FDS software structure typically employed in a fault diagnosis system. We have previously developed several such algorithms applicable to linear systems, one of which is presently being implemented at the Naval Ocean System Center. This work is represented in this report by several reprints and also serves as the foundation for the on-going research on nonlinear fault analysis.

We have investigated three alternative approaches to the nonlinear fault analysis problem. These include a nonlinear state space approach, an approach which employs nonlinear integral performances measures in lieu of frequency domain information, and an approach in which an affine approximation of a linear system is employed. During the past year our major activity has been in the state space area in which we have two on-going activities. Both of these assume an augmented state model

$$\dot{x} = f(x,r) \qquad 1.$$
$$\dot{r} = 0$$

* ONR funding for a joint program to implement a linear fault analysis algorithm at NOSC.

where x is the state vector for the system and r is the vector of component parameters which we must identify to diagnose a failure. One then measures a subvector for x or an output process $y = g(x,r)$ and applies some type of system identification process to estimate r. To this end we have investigated the possibility of employing nonlinear observers and several possible quasi-linearization algorithms. The former approach was reported in a conference paper which is reprinted in this report while the latter approach is the subject of a Ph.D. dissertation which is in preparation.

In addition to the nonlinear state space approach to the fault analysis problem, we have recently reopened an earlier investigation into the feasibility of employing affine approximations to nonlinear circuits in the fault diagnosis problem. Although the viability of such an approach is still open to question, fault analysis by affinization appears to be the only viable approach to the nonlinear problem which can effectively exploit the dual mode software structure employed in linear analog and digital fault analysis. At the present time a master's thesis in which the viability of the approach will be investigated is in proparation, though we do not yet have definitive results.

7. Publications and Activities:

A. Refereed Journal Articles

1. Sen, N., and R. Saeks, "Fault Diagnosis for Linear Systems Via Multigrequency Measurements", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 457-465, (1979).

2. 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, Vol. CAS-26, pp. 589-594, (1979).

3. Lu, K.S., and R. Saeks, "Failure Preduction for an On-Line Maintenance System in a Poisson Shock Environment", IEEE Trans. on Systems, Man, and Cybernetics, Vol. SMC-9, pp. 356-362, (1979).

4. Saeks, R., "An Approach to Built-in Testing", IEEE Trans. on Aerospace and Electronic Systems, Vol. AES-14, pp. 813-818, (1978).

B. Conference Papers and Abstracts

1. Olivier, P.D., and R. Saeks, "Nonlinear Observers and Fault Analysis", Proc. of the 22nd Midwest Symposium on Circuits and Systems, Univ. of Penn., Philadelphia, June 1979, pp. 535-536.

2. Olivier, P.D., and R. Saeks, "On Large Nonlinear Perturbations of Linear Systems", Proc. of the 12th Asilomar Conf. on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1978, pp. 473-477.

3. Saeks, R., "CAD Oriented Measures of Testability", Proc. of the Industry/Joint Services Automatic Test Conference and Workshop, NSIA, San Diego, April 1978, pp. 71-72.

4. Saeks, R., "An Application of Large-Scale Systems Techniques to the Fault Analysis Problem", Proc. of the 21st Midwest Symp. on Circuits and Systems, Iowa State Univ., Ames, IA., Aug. 1978, p. 314.

C. Theses

1. Hsieh, M., Ph.D. Dissertation, Texas Tech Univ., (in preparation).

2. Ngo, Q. D., M.S. Thesis, Texas Tech Univ., (in preparation).

D. Conferences and Symposia

1. Saeks, R., 21st Midwest Symposium on Circuits and Systems, Iowa State Univ., Ames, Ia., Aug. 1978.

2. Saeks, R., 12th Asilomar Conf. on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1978.

3. Saeks, R., 22nd Midwest Symposium on Circuits and Systems, Univ. of Penn., Philadelphia, June 1979.

4. Olivier, P.D., 22nd Midwest Synposium on Circuits and Systems, Univ. of Penn., Philadelphia, June 1979.

5. Chao, K.S., IEEE Inter. Symp. on Circuits and Systems, Tokyo, July 1979.

E. Lectures

1. Saeks, R., "Fault Diagnosis: The Missing Circuit Theory", Elec. Engrg. Colloq., State Univ. of New York at Stony Brook, Dec. 1979.

8. Reprint of "Fault Diagnosis for Linear Systems Via Multifrequency Measurements, by N. Sen and R. Saeks from the IEEE Transactions on Circuits and Systems, Vol. CAS-26, pp. 457-455, (1979).

Fault Diagnosis for Linear Systems Via Multifrequency Measurements

NEERAJ SEN, MEMBER, IEEE, AND RICHARD SAEKS, FELLOW, IEEE

Abstract—The fault diagnosis problem for a linear system whose transfer function matrix is measured at a discrete set of frequencies is formalized. A measure of solvability for the resultant equations and a measure of testability for the unit under test is developed. These, in turn, are used as the basis of algorithms for choosing test points and test frequencies.

I. Introduction

CONCEPTUALLY, the fault analysis problem for an analog circuit or system amounts to the measurement of a set of externally accessible parameters of the system from which one desires to determine the internal system parameters or equivalently locate the failed components as illustrated in Fig. 1. Here, the measurements m_i may represent data taken at distinct test points or alternatively, data taken at a fixed test point under different stimuli. Similarly, the r_i represent parameters characterizing the various internal system components. Here, a single parameter may characterize an entire component, say a

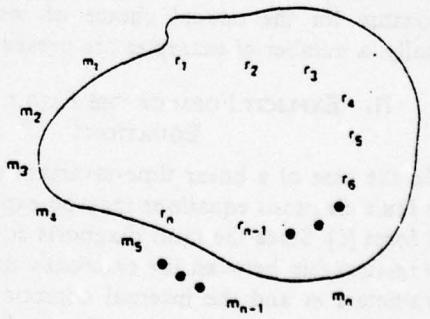


Fig. 1. Conceptual model of fault diagnosis problem.

Manuscript received July 3, 1978; revised April 2, 1979. This research was supported in part by Office of Naval Research Contracts 75-C-0924 and 76-C-1136.

N. Sen was with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX. He is now with the Datapoint Corporation, San Antonio, TX.

R. Saeks is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

0098-4094/79/0700-0457$00.75

resistance, capacitance or inductance. Alternatively, a component may be represented by several parameters: the h parameters of a transistor, the poles and gain of an op-amp, etc. In general, one models a system component by the minimum number of parameters which will allow the failure to be isolated up to a shop replaceable assembly (SRA) with all "allowed" system failures manifesting themselves in the form of some parameter change.

To solve the fault diagnosis problem, one then measures $m = \text{col}\ (m_i)$ and solves a nonlinear algebraic equation

$$m = F(r) \tag{1}$$

for $r = \text{col}\ (r_i)$ to diagnose the fault. The parameters in the resultant r vector which are out of tolerance then indicate the faulty component [6].

The purpose of the present paper is to give an explicit formulation of the *fault diagnosis equations* which arise in the maintenance of linear systems. Here, one measures the system frequency response as observed from a specified set of externally accessible test points at a discrete set of frequencies and it is desired to solve for a vector of internal system parameters r which completely characterize the frequency response matrices of the individual system components; $Z_i(s,r)$, $i = 1,2,\cdots,q$.

In the following section the explicit form for the fault diagnosis equations is derived for a given set of test frequencies. A *measure of solvability* [15] of these equations is then developed in Section III and employed in Section IV in an algorithm for optimally selecting test frequencies. The measure of solvability for the fault analysis equations, given an optimal choice of test frequencies, is then taken as a *measure of testability* [1], [2], [5] for the unit under test (UUT) and is used as the basis of an algorithm for the optical choice of test points [3]–[5]. Finally, a number of examples are presented in Section V.

II. Explicit Form of the Fault Diagnosis Equations

In the case of a linear time-invariant circuit or system, the fault diagnosis equations may be expressed in analytical form [6]. Since the fault diagnosis equations deal with the relationship between the externally measurable system parameters m and the internal component parameters r we adopt a *component connection model* as the starting point for the derivation of the fault diagnosis equations [7], [8]. This is one of several commonly employed large scale system models in which the components and connections in a circuit or system are modeled by distinct equations, thereby permitting one to explicitally deal with the relationship between the individual component parameters and the composite system parameters.

Since the present study is restricted to linear time-invariant systems, we assume that each component is characterized by a transfer function matrix which is dependent on the potentially variable component parameters, $Z_i(s,r)$. For the classical *RLC* components $Z_i(s,r)$ may take the form R, Ls, or $1/sC$ for the case of a resistor, inductor, or capacitor, respectively. More generally, one may model an op-amp by the transfer function $k/(s-p_1)(s-p_2)$ where the parameter vector r now represents the three potentially variable component parameters; k, p_1, p_2; or a delay by ke^{sT}, etc. Although the symbol Z is used, the components are not assumed to be represented by impedance matrices. Indeed, hybrid models are used in most of our examples. For the purpose of analysis, it is assumed that all faults manifest themselves in the form of changes, possibly catastrophic, in the parameter vector r with the frequency characteristics of the components unchanged. Although not universal, this *fault hypothesis* covers the most commonly encountered situations and subsumes the common industrial practice of assuming that all failures in analog circuits and systems take the form of open and short circuited components [9].

Our system components are thus characterized by a set of simultaneous equations

$$b_i = Z_i(s,r)a_i, \qquad i = 1,2,\cdots,q \tag{2}$$

where a_i and b_i denote the component input and output vectors, respectively. For notational brevity, these component equations may be combined into a single block diagonal matrix equation

$$b = Z(s,r)a \tag{3}$$

where $b = \text{col}\ (b_i)$, $a = \text{col}\ (a_i)$, and $Z(s,r) = \text{diag}\ (Z_i(s,r))$.

Although there are many ways to represent the connection in a circuit or system; say, a block diagram, linear graph or signal flow graph, any such representation is simply a graphical means for displaying a set of connection equations: Kirchhoff laws, adder equations, etc. As such, for our component connection model we adopt a purely algebraic connection model in which the connection equations are displayed explicitally without the intermediary of some kind of graphical connection diagram. This takes the form

$$\begin{aligned} a &= L_{11}b + L_{12}u \\ y &= L_{21}b + L_{22}u \end{aligned} \tag{4}$$

where u and y represent the vectors of accessible inputs and outputs which are available to the test system. In simple systems, the connection matrices L_{ij} are usually obtainable by inspection, whereas, in more complex systems, computer codes have been developed for their derivation [7]. Moreover, they are assured to exist in all but the most pathological systems [8].

It is the pair of simultaneous matrix equations (3) and (4) which are termed the component connection model. By combining (3) and (4) to eliminate the component input and output variables a and b one may derive [6], [7] an expression for the transfer function matrix observable by the test system between the test input and output vectors u and y obtaining

$$S(s,r) = L_{22} + L_{21}(1 - Z(s,r)L_{11})^{-1}Z(s,r)L_{12} \tag{5}$$

where

$$y = S(s,r)u. \tag{6}$$

For a linear time-invariant system the transfer function $S(s,r)$ is a complete description of the measurable data about the unit under test available to the test system. Moreover, being rational it is completely determined by its value at a finite number of frequencies. As such, without loss of generality, we may take our measured data to be of the form

$$\text{col}\,[S(s_1,r),S(s_2,r),\cdots,S(s_k,r)]. \tag{7}$$

The *fault diagnosis equations* then take the form

$$\begin{bmatrix} S(s_1,r) \\ S(s_2,r) \\ \vdots \\ S(s_k,r) \end{bmatrix} \Bigg| \begin{matrix} L_{22}+L_{21}(1-Z(s_1,r)L_{11})^{-1}Z(s_1,r)L_{12} \\ L_{22}+L_{21}(1-Z(s_2,r)L_{11})^{-1}Z(s_2,r)L_{12} \\ \vdots \\ L_{22}+L_{21}(1-Z(s_k,r)L_{11})^{-1}Z(s_k,r)L_{12} \end{matrix} \tag{8}$$

Since $S(s,r)$ is, in general, a matrix, the fault diagnosis equations as derived above take the form of a matrix (col $[S(s_i,r)]$) valued function of a vector valued variable r. Computationally, however, we prefer to work with a vector valued function of a vector valued variable and hence, we transform $S(s,r)$ into a column vector via

$$\text{vec}\,[S(s,r)]=\text{col}\,[S^i(s,r)] \tag{9}$$

where $S^i(s,r)$ denotes the ith column of the matrix, $S(s,r)$. With the aid of the identity $\text{vec}[XYZ]=[Z^t\otimes X]\text{vec}\,[Y]$ (8) then transforms into [7], [12]

$$M=\begin{bmatrix} \text{vec}\,[S(s_1,r)] \\ \text{vec}\,[S(s_2,r)] \\ \vdots \\ \text{vec}\,[S(s_k,r)] \end{bmatrix} = \begin{matrix} \text{vec}\,[L_{22}]+\left[L_{12}^t\otimes L_{21}(1-Z(s_1,r)L_{11})^{-1}\right]\text{vec}\,[Z(s_1,r)] \\ \text{vec}\,[L_{22}]+\left[L_{12}^t\otimes L_{21}(1-Z(s_2,r)L_{11})^{-1}\right]\text{vec}\,[Z(s_2,r)] \\ \vdots \\ \text{vec}\,[L_{22}]+\left[L_{12}^t\otimes L_{21}(1-Z(s_k,r)L_{11})^{-1}\right]\text{vec}\,[Z(s_k,r)] \end{matrix} = F(r) \tag{10}$$

which is the form of the fault diagnosis equations with which we desire to work.

III. Solvability of the Fault Diagnosis Equations

For the fault diagnosis equations derived above to be a viable tool of circuit and system diagnosis two fundamental questions remain to be answered: "What test frequencies should be employed to optimize the solvability of the equations?" and "How solvable are the equations given an optimal choice of test frequencies?" Both of these questions, in turn, hinge on the development of some type of *measure of solvability* [15] for the fault diagnosis equations.

For a set of linear equations

$$m=Fr \tag{11}$$

where r is an n vector, m is a p vector, and F is a p by m matrix one may characterize the solvability of the equations in terms of the number of arbitrary parameters in its solution (if a solution exists). As such, $\delta=n-\text{rank}(F)$ is a natural measure of the solvability for (11). Here, $\delta=0$ implies that the equation has a unique solution, $\delta=1$ implies that the solution is determined up to one arbitrary parameter and so on, with increasing values of δ representing decreasing degrees of solvability.

The fault diagnosis equations are, however, nonlinear even for linear systems—hence we must resort to the *implicit function theorem* to obtain a measure of solvability [15], [16] analogous to the above [13]. Indeed, if r_f is a solution to the fault diagnosis equations, then r_f is determined up to a

$$\delta(r_f)=n-\text{rank}\left[\left[\frac{dF}{dr}(r_f)\right]\right] \tag{12}$$

dimensional manifold (of arbitrary parameters) in a neighborhood of r_f. Here dF/dr is the Jacobian matrix of partial derivatives of F with respect to r. With the aid of the matrix identity $d(M^{-1})/dr=-M^{-1}[dM/dr]M^{-1}$, dF/dr can be computed explicitly from (8) and (10) yielding

$$\frac{dF}{dr}(r_f)=\begin{bmatrix} \left\{\left(\left[1+L_{11}(1-Z(s_1,r_f)L_{11})^{-1}Z(s_1,r_f)\right]L_{12}\right)^t\otimes\left(L_{21}(1-Z(s_1,r_f)L_{11})^{-1}\right\}\left[(d\text{vec}\,Z(s_1,r_f)\right]/dr \\ \left\{\left(\left[1+L_{11}(1-Z(s_2,r_f)L_{11})^{-1}Z(s_2,r_f)\right]L_{12}\right)^t\otimes\left(L_{21}(1-Z(s_2,r_f)L_{11})^{-1}\right\}\left[(d\text{vec}\,Z(s_2,r_f)\right]/dr \\ \vdots \\ \left\{\left(\left[1+L_{11}(1-Z(s_k,r_f)L_{11})^{-1}Z(s_k,r_f)\right]L_{12}\right)^t\otimes\left(L_{21}(1-Z(s_k,r_f)L_{11})^{-1}\right\}\left[(d\text{vec}\,Z(s_k,r_f)\right]/dr \end{bmatrix} \tag{13}$$

where "t" denotes matrix transposition and $\otimes$ denotes the matrix Kronecker (or tensor) product.

The difficulty with the implicit function theorem is that it only yields local information valid in a neighborhood of a solution. Fortunately, however, given the special nature of the Jacobian matrix of (13) coupled with an assumption that the component transfer function matrices $Z_i(s,r)$ are rational in r, it is possible to show that the rank of the Jacobian matrix is "almost constant." This, in turn, allows us to transform the local measure of solvability of (12) into a global measure of solvability. For this purpose we adopt the *algebraic geometric* definition for the term "almost constant;" i.e., we say that a function of r_f is *almost constant* if it is constant except possibly for those values of r_f lying in an *algebraic variety* (the solution space of a finite set of nonzero simultaneous polynomial equations in n variables). More generally, we say that a property holds "almost everywhere" or for *almost all* r_f in n space if it is true for all values of r_f except possibly those lying in an algebraic variety. Since the Lebesque measure of an algebraic variety is zero, this definition for the concept "almost everywhere" is consistent with the more common measure theoretic definition and is more natural in the context of our application [14].

Theorem 1

Let $Z_i(s,r)$; $i=1,2,\cdots,q$; be rational in r. Then $\delta(r_f)$ is almost constant.

Note, the assumption that $Z_i(s,r)$ is rational in r is quite minor being satisfied by all of the examples given in Section II except for the delay (which can be approximated by a function which is rational in r). In practice, the component transfer function matrices will also be rational in s though this is not required for the present theorem since F and dF/dr are formulated in terms of specific test frequencies, $s_1,s_2,\cdots,s_k$. Given our assumption on the $Z_i(s,r)$, together with (13), it then follows that $(dF)/(dr)(r_f)$ is also rational in r_f.

Proof of Theorem 1: We begin by showing that an arbitrary polynomial matrix in r, $P(r)$, has almost constant rank. Since rank $P(r)$ is restricted to the finite set of integers $(0,1,2,\cdots,j$; where j is the minimum of the number of rows and columns in $P(r)$, there exists an r_m which maximizes the rank of $P(r)$

$$\text{rank}\,[P(r_m)] > \text{rank}\,[P(r)]. \tag{14}$$

Now, the rank of a matrix is the dimension of its largest nonsingular square submatrix. As such, $P(r)$ admits a square submatrix $M(r)$, whose dimension is equal to the rank $P(r_m)$ and for which

$$\det M(r_m) \neq 0. \tag{15}$$

Now, $\det[M(r)]$ is a polynomial in r which is not identically zero (from (15)) and hence, it is nonzero, almost everywhere. As such,

$$\text{rank}\,[P(r_m)] > \text{rank}\,[P(r)] > \text{rank}\,[M(r)] = \text{rank}\,[P(r_m)] \quad \text{a.e.} \tag{16}$$

showing that rank $[P(r)]$ = rank $[P(r_m)]$ almost everywhere. As such, rank $[P(r)]$ is almost constant.

Now, to verify that rank $[(dF)/(dr)(r_f)]$ is constant we decompose this matrix as

$$\frac{dF}{dr}(r_f) = \frac{P(r_f)}{d(r_f)} \tag{17}$$

where $P(r_f)$ is a polynomial matrix and $d(r_f)$ is a nonzero common denominator. $P(r_f)$ has almost constant rank while $d(r_f)$ is nonzero almost everywhere and hence can effect the rank of $P(r_f)$ only on an algebraic variety (since the division of a matrix by a nonzero scalar does not effect its rank.) As such, our Jacobian matrix has almost constant rank implying that

$$\delta(r_f) = n - \text{rank}\left[\frac{dF}{dr}(r_f)\right] \tag{18}$$

is also almost constant. The proof of the theorem is therefore complete.

Given the theorem, we may now define a *global measure of solvability* for the fault diagnosis equation δ as the *generic value* of $\delta(r_f)$; i.e., the value $\delta(r_f)$ takes on for almost all r_f. This proves to be a natural measure of solvability since it indicates the ambiguity which will result from an attempt to solve the fault diagnosis equations in a neighborhood of almost any failures. Of course, one requires some sort of equation solving algorithm [10], [11] to locate a neighborhood of an actual failure. The δ parameter, however, represents a bound on the performance of any such algorith. Finally, we note that since δ is independent of r_f, the solution of the fault diagnosis equations, it can be computed at the time the system and its test algorithm are developed by evaluating $\delta(r)$ at a randomly chosen generic point, say r_0. In turn, this parameter may then be employed as an aid in the choice of test frequencies and test points.

IV. Test Frequency Selection

Adopting the measure of solvability δ formulated in the preceeding section, it remains to develop an algorithm for choosing a set of test frequencies; $s_1,s_2,\cdots,s_k$; which maximize the solvability of the fault diagnosis equations (i.e., minimize δ). To this end, let $\delta_{\min}$ denote the minimum value achieved by δ for any set of test frequencies; $s_1,s_2,\cdots,s_k$; $k=1,2,\cdots,$. Since the possible values for δ are restricted to the finite set; $\delta=0,1,\cdots,n$; such a minimum is assured to exist.

The following theorem gives an explicit formula for computing $\delta_{\min}$ while its proof yields an algorithm for choosing a set of test points which achieve $\delta_{\min}$. Since the purpose of this theorem is to formulate an algorithm for choosing test frequencies, the theorem is expressed in terms of

$$\text{vec}\,[S(s,r)] = \text{vec}\,[L_{22}] + [L_{12}^t \otimes L_{21}(1 - Z(s,r)L_{11})^{-1}]\,\text{vec}\,[Z(s,r)] \tag{19}$$

and

$$\frac{d\text{vec}\,[S(s,r)]}{dr} = \left\{\left(\left[1 + L_{11}(1 - Z(s,r)L_{11})^{-1}Z(s,r)\right]L_{12}\right)^t \otimes \left(L_{21}(1 - Z(s,r)L_{11})^{-1}\right)\right\} \cdot \left\{d\text{vec}\,[Z(s,r)]/dr\right\} \quad (20)$$

viewed as rational functions in s rather than in terms of the function $F(r)$ which is formulated in terms of an *a priori* choice of test frequencies.

Theorem 2

Let $Z_i(s,r)$; $i = 1,2,\cdots,q$; be rational in s and r. Then

$$\delta_{\min} = n - \text{col-rank}\left[\frac{d\text{vec}\,[S(s,r)]}{dr}\right]$$

where n is the dimension of the parameter vector r and "col-rank" denotes the generic number of linearly independent columns of the rational matrix $[d\text{vec}\,[S(s,r)]/dr]$ over the field of complex numbers. Moreover, $\delta_{\min}$ is achieved by almost any choice of $n - \delta_{\min}$ distinct complex frequencies.

Proof: For the sake of brevity, we will prove the theorem only for the special case where $S(s,r)$ is a scalar transfer function (allowing us to drop the "vec" transformation) though essentially the same proof goes through in the general case modulo some notational complexities [5]. Also, since the rank of the Jacobian matrix is almost constant it suffices to fix the parameter vector r at any generic point, say r_0. This then reduces $[d\text{vec}\,[S(s,r)]/dr]$ to a row vector of rational functions

$$R(s) = \left[\, R_1(s) \quad R_2(s) \quad \cdots \quad R_n(s) \,\right] \quad (21)$$

where

$$R_i(s) = \left[d\text{vec}\,[S(s,r_0)]/dr_i\right] \quad (22)$$

and our problem reduces to the verifications of the fact that the number of linearly independent columns of $R(s)$ over the field of complex scalars is equal to the maximum possible rank of the complex matrix

$$\begin{bmatrix} R(s_1) \\ R(s_2) \\ \vdots \\ R(s_k) \end{bmatrix} = \begin{bmatrix} R_1(s_1) & R_2(s_1) & \cdots & R_n(s_1) \\ R_1(s_2) & R_2(s_2) & \cdots & R_n(s_2) \\ \vdots & \vdots & & \vdots \\ R_1(s_k) & R_2(s_k) & \cdots & R_n(s_k) \end{bmatrix} = \text{col}\,(R(s_i)) \quad (23)$$

over all possible choices of the complex frequencies; $s_1, s_2, \cdots, s_k$; $k = 1,2,\cdots$. Now, clearly if some column of $R(s)$, say the nth, is dependent on the remaining columns, then

$$R_n(s) = \sum_{j=1}^{n-1} c_j R_j(s) \quad (24)$$

for all s. Then by applying (24) individually for each s_i

$$\text{col}\,(R_n(s_i)) = \sum_{j=1}^{n-1} c_j\,\text{col}\,(R_j(s_i)) \quad (25)$$

for any possible number or choice of the s_i. The rank of the matrix of (23), therefore, is less than or equal to the number of linearly independent columns of $R(s)$ over the field of complex numbers.

To prove that equality can be achieved with an appropriate choice of $n - \delta_{\min}$ complex test frequencies s_i we invoke our assumption that $S(s,r)$ is a scalar transfer function. Without loss of generality, we may assume that $R_1(s)$ through $R_q(s)$ are the linearly independent entries in $R(s)$ over the field of complex numbers in which case we must show that there exists complex frequencies $s_1, s_2, \cdots, s_k$ ($k = q$ in this case) which make the first q columns of the matrix of (23) linearly independent.

If $q = 1$, $R_1(s)$ is not identically zero (since otherwise it would be linearly dependent) and hence for almost all s_1, $R_1(s_1) \neq 0$. As such, the columns in this trivial one by one matrix are linearly independent. With this as a starting point, we will use an inductive argument to show that the theorem holds for all values of q. We, therefore, assume that it has been shown that for $q = p$ there exist complex frequencies; $s_1, s_2, \cdots, s_p$; such that the matrix

$$R_p = \begin{bmatrix} R_1(s_1) & R_2(s_1) & \cdots & R_p(s_1) \\ R_1(s_2) & R_2(s_2) & \cdots & R_p(s_2) \\ \vdots & & & \\ R_1(s_p) & R_2(s_p) & \cdots & R_p(s_p) \end{bmatrix} \quad (26)$$

has linearly independent columns and we desire to show that there exists an s_{p+1} such that the matrix

$$\bar{R}_{p+1}(s) = \begin{bmatrix} R_1(s_1) & R_2(s_1) & \cdots & R_p(s_1) & R_{p+1}(s_1) \\ R_1(s_2) & R_2(s_2) & \cdots & R_p(s_2) & R_{p+1}(s_2) \\ \vdots & & & & \\ R_1(s_p) & R_2(s_p) & \cdots & R_p(s_p) & R_{p+1}(s_p) \\ R_1(s) & R_2(s) & & R_p(s) & R_{p+1}(s) \end{bmatrix} \quad (27)$$

has linearly independent columns for $s = s_{p+1}$. By virtue of our assumption that $S(s,r)$ is a scalar both $\bar{R}_p$ and $\bar{R}_{p+1}(s)$ are square and we may test for linear independence of the columns of $\bar{R}_{p+1}(s)$ by computing its determinant. Expanding (27) in cofactors along its bottom row, we obtain

$$\det\left(\bar{R}_{p+1}(s)\right) = \sum_{j=1}^{p+1} (-1)^{p+j+1}\Delta_{p+1,j}R_j(s). \quad (28)$$

Since $\bar{R}_p$ has linearly independent columns $\Delta_{p+1,p+1} \neq 0$, hence, the coefficients in the summation of (28) are not all zero and thus by the linear independence of the $R_i(s)$ the

summation is not identically zero. As such, one can choose almost any s_{p+1} which will make the determinant of $\bar{R}_{p+1}(s_{p+1})$ nonzero thus assuring the $\bar{R}_{p+1}$ has linearly independent columns when its rows are evaluated at the complex frequencies $s_1, s_2, \cdots, s_{p+1}$. The proof of the theorem is thus complete.

Note that the proof of the theorem yields a natural sequential algorithm for choosing test frequencies. Moreover, for the scalar case we have shown that the number of required test frequencies is exactly $n - \delta_{\min}$ (equal to the column rank of the Jacobian matrix). In the general case where $S(s,r)$ is not a scalar, the number of required test frequencies is less than or equal to $n - \delta_{\min}$ [5].

Although the theorem implies that one can randomly choose almost any $n - \delta_{\min}$ test frequencies to maximize the solvability of the fault diagnosis equations, the result does not take cognizence of numerical considerations. Although no theory yet exists for choosing test points with numerical considerations in mind, it has been our experience that the "well posedness" of the fault diagnosis equations is quite sensitive to the choice of test frequencies [5]. In most of our experiments, we have worked with real test frequencies to eliminate the necessity of working in the complex plane. On the other hand, m is most easily measured when values of s_i on the jw axis are employed whereas it has been suggested that test frequencies symetrically spaced around a circle in the complex plane might yield numerically "well posed" equations.

Although the measure of solvability δ for the fault diagnosis equations is dependent on the choice of test frequencies, as well as the properties of the unit under test, $\delta_{\min}$ is determined entirely by the UUT; its components, connections and accessible test points; and is completely independent of the test algorithm employed. As such, $\delta_{\min}$ may be taken as a natural *measure of testability* [1] for the UUT which characterizes the degree to which the fault analysis equations can be solved given an optimal choice of test frequencies and solution algorithm. Moreover, $\delta_{\min}$ may be used as an aid for the optimal selection of test points [3]–[5]. To this end we may choose a set of test points, from several options, so as to minimize $\delta_{\min}$. Alternatively, we may attribute a cost to each input and output test point and then choose the least cost combination of test points which yield a specified $\delta_{\min}$. This latter process reduces to a rather straightforward integer programming problem and is thus readily automated [4], [5]. The technique is illustrated in the examples of the following section.

V. Examples

An initial illustration of the theory consider the *RC* coupled amplifier with inductive load shown in Fig. 2. Here we will take E_i to be the only test input but we will initially allow E_0, i_L, i_C, and V_i to all be taken as test outputs with the measure of testability $\delta_{\min}$ being used to extract a reduced set of test outputs from these options. A

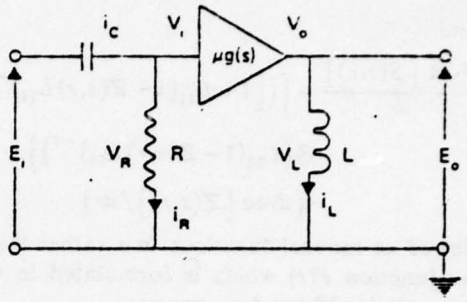


Fig. 2. *RC* coupled amplifier with inductive load.

component connection model for this circuit is given by

$$\begin{bmatrix} V_0 \\ i_L \\ V_C \\ i_R \end{bmatrix} = \begin{bmatrix} \mu g(s) & 0 & 0 & 0 \\ 0 & 1/LS & 0 & 0 \\ 0 & 0 & 1/CS & 0 \\ 0 & 0 & 0 & 1/R \end{bmatrix} \begin{bmatrix} V_i \\ V_L \\ i_C \\ V_R \end{bmatrix} \tag{29}$$

$$\begin{bmatrix} V_i \\ V_L \\ i_C \\ V_R \\ \hline E_0 \\ i_0 \\ i_R \\ V_i \end{bmatrix} = \left[\begin{array}{cccc|c} 0 & 0 & -1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 & 1 \\ \hline 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{array}\right] \begin{bmatrix} V_0 \\ i_L \\ V_C \\ i_R \\ \hline E_i \end{bmatrix}. \tag{30}$$

Taking our vector of potentially variable component parameters to be $r = \text{col}\,(\mu, L, C, R)$ each with unity nominal value, we obtain a nominal transfer function matrix

$$S(s,r) = \begin{bmatrix} \dfrac{s(g(s)+1)+1}{s+1} \\ \dfrac{g(s)}{s+1} \\ \dfrac{s}{s+1} \\ \dfrac{s}{s+1} \end{bmatrix} \tag{31}$$

whereas our Jacobian matrix evaluated at the nominal parameter values is given by

$$\frac{d\text{vec}[S(s,r)]}{dr} = \begin{bmatrix} \dfrac{sg(s)}{s+1} & 0 & \dfrac{sg(s)}{(s+1)^2} & \dfrac{sg(s)}{(s+1)^2} \\ \dfrac{g(s)}{s+1} & \dfrac{-g(s)}{s+1} & \dfrac{g(s)}{(s+1)^2} & \dfrac{g(s)}{(s+1)^2} \\ 0 & 0 & \dfrac{s}{(s+1)^2} & \dfrac{-s^2}{(s+1)^2} \\ 0 & 0 & \dfrac{s}{(s+1)^2} & \dfrac{s}{(s+1)^2} \end{bmatrix} \tag{32}$$

Now, an inspection of this matrix will reveal that it has four independent columns over the field of complex numbers and hence if all four possible outputs are used, we will have $\delta_{min}=0$ implying that the fault diagnosis equations have locally unique solutions. On the other hand, if only two outputs E_0 and i_C are measured, our modified Jacobian matrix will reduce to the first and third rows of the matrix shown in (32) which has column rank 3. As such, if we only use these two test outputs, we obtain $\delta_{min}=1$ and hence the solution to the fault diagnosis equations will be characterized by a single arbitrary parameter.

In this latter case, with only E_0 and i_C taken as test outputs, Theorem 2 implies that dF/dr will have rank 3 for almost any choice of $3=n-\delta_{min}$ test frequencies. Choosing $s_1=1$, $s_2=2$, and $s_3=3$, we obtain

$$\frac{dF}{dr}(r_0)=\begin{bmatrix} g(1)/2 & 0 & g(1)/4 & g(1)/4 \\ 0 & 0 & 1/4 & -1/4 \\ \hline 2g(2)/3 & 0 & 2g(2)/9 & 2g(2)/9 \\ 0 & 0 & 2/9 & -2/9 \\ \hline 3g(3)/4 & 0 & 3g(3)/16 & 3g(3)/16 \\ 0 & 0 & 3/16 & -3/16 \end{bmatrix} \quad (33)$$

which has three linearly independent columns as long as $g(1)\neq 0$, $g(2)\neq 0$, and $g(3)\neq 0$. Indeed, in this example, any two of the three frequencies would have sufficed to yield three linearly independent columns. Note, for scalar transfer functions, Theorem 2 implies that $n-\delta_{min}$ frequencies are actually required but for matrix transfer functions fewer frequencies may suffice.

Of course, for the circuit of Fig. 2, we have a choice of some 15 combinations of the four outputs with which we may choose to work for the diagnosis of the circuit. The resultant δ_{min}'s for the various combinations of outputs are given in Table I [5].

Finally, with the aid of Table I, one may readily develop a test point selection algorithm for our circuit [4], [5]. For instance, if we desire to find the smallest set of outputs which yield a $\delta_{min}<1$ an inspection of the table will reveal that E_0 and i_L, i_L and i_C, or E_0 and i_C are the optimal choices. Of course, if one attributes a cost to the various outputs (determined by the convenience of making the required measurements), then we may further distinguish between these three possibilities. For instance, if voltage measurements are deemed to be easier than current measurements, the combination of i_L and i_C may be excluded with the decision between the remaining two options being dependent on whether it is easier to measure the circuit's input current (i_C) or its load current (i_L).

As a second example, consider the one stage transistor amplifier shown in Fig. 3 with the ac equivalent circuit of Fig. 4. Since it is clearly impossible to distinguish between failures in the two parallel bias resistors, R_a and R_b, these two resistors have been combined into the single resistor, R_a' in the component connection model of (34) and (35). Taking all of the component parameters as potentially faulty, r becomes a 12 vector composed of $C_1, r_x, \cdots, R_L$ and as before, we take all parameters to have the nominal value of unity.

TABLE I
MEASURES OF TESTABILITY FOR THE CIRCUIT OF FIG. 2 USING VARIOUS COMBINATIONS OF TEST OUTPUTS

Outputs	δ_{min}
E_o, i_L, i_C, v_i	0
E_o, i_L, i_C	0
i_L, i_C, v_i	1
i_L, v_i, E_o	1
v_i, E_o, i_C	1
E_o, i_L	1
i_L, i_C	1
i_L, v_i	2
v_i, E_o	2
E_o, i_C	1
i_C, v_i	2
E_o	2
i_L	2
i_C	2
v_i	3

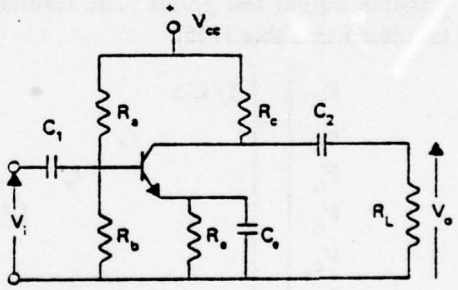


Fig. 3. One stage transistor amplifier.

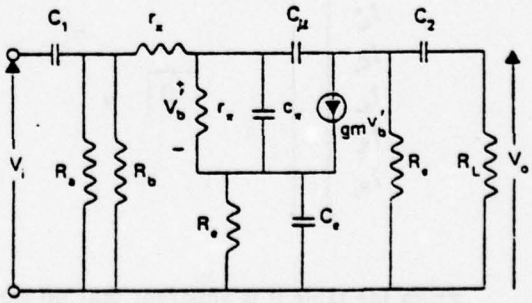


Fig. 4. Amplifier equivalent circuit.

$$\begin{bmatrix} I_{C_1} \\ I_{r_x} \\ I_{r_\pi} \\ I_{c_\mu} \\ I_{C_2} \\ V'_{R_a} \\ V_{R_\pi} \\ V_{c_\pi} \\ V_{C_\pi} \\ V_{g_m} \\ V_{R_c} \\ V_{R_L} \\ \hline V_0 \\ I_{C_1} \\ V_{R'_a} \\ I_e \end{bmatrix} = \left[\begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & V_{C_1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & V_{r_x} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & V_{r_\pi} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & V_{c_\mu} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & V_{C_2} \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & I_{R'_a} \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & I_{R_\pi} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{c_\pi} \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & I_{C_\pi} \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{g_m} \\ -1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & I_{R_c} \\ -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & I_{R_L} \\ \hline -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & V_i \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & \end{array}\right] \begin{bmatrix} V_{C_1} \\ V_{r_x} \\ V_{r_\pi} \\ V_{c_\mu} \\ V_{C_2} \\ I_{R'_a} \\ I_{R_\pi} \\ I_{c_\pi} \\ I_{C_\pi} \\ I_{g_m} \\ I_{R_c} \\ I_{R_L} \end{bmatrix}. \tag{34}$$

Once again we let the input voltage be the only test input for the system and we take V_0, I_{C_1}, $V_{R'_a}$, and I_e, to be possible output test points. The resultant δ_{min} for each of the 15 possible combinations of these output terminals is tabulated in Table II [5].

$$\begin{bmatrix} V_{C_1} \\ V_{r_x} \\ V_{r_\pi} \\ V_{c_\mu} \\ V_{C_2} \\ I_{R'_a} \\ I_{R_\pi} \\ I_{c_\pi} \\ I_{C_\pi} \\ I_{g_m} \\ I_{R_c} \\ I_{R_L} \end{bmatrix} = \begin{bmatrix} 1/C_1 s & & & & & & & & & & & & \\ & r_x & & & & & & & & & & & \\ & & r_\pi & & & & & & & & & \boxed{0} & \\ & & & 1/C_\mu s & & & & & & & & & \\ & & & & 1/C_2 s & & & & & & & & \\ & & & & & 1/R'_a & & & & & & & \\ & & & & & & 1/R_\pi & & & & & & \\ & & & & & & & C_\pi s & & & & & \\ & & & & & & & & C_\pi s & & & & \\ & \boxed{0} & & & & & & g_m & & 0 & & & \\ & & & & & & & & & & 1/R_c & & \\ & & & & & & & & & & & 1/R_L & \end{bmatrix} \begin{bmatrix} I_{C_1} \\ I_{r_x} \\ I_{r_\pi} \\ I_{c_\mu} \\ I_{C_2} \\ V_{R'_a} \\ V_{R_\pi} \\ V_{c_\pi} \\ V_{C_\pi} \\ V_{g_m} \\ V_{R_c} \\ V_{R_L} \end{bmatrix}. \tag{35}$$

From the table it is apparent that no single test output suffices to yield a $\delta_{min}=0$ (perfect testability) though $\delta_{min}=0$ can be achieved using two test outputs; V_0 and I_{C_1} or V_0 and I_e.

VI. Conclusions

Our purpose in the preceeding has been to formulate an analytic theory in support of the intuitive art usually associated with the design of a test algorithm. With the aid of the techniques developed above, we believe that it will be possible to develop an automated test program generation (ATPG) algorithm for linear systems [4], [5]. Indeed, such an algorithm could be readily combined with the same computer-aided design (CAD) algorithm used in the system design process [9]. Given the component connection equations such an algorithm could be employed to automatically (or interactively) choose test points and test frequencies and generate the required set of fault diagnosis equations. These could then be stored on tape and supplied to the automatic test equipment (ATE) in which a faulty system would be tested and the fault diagnosis equations solved.

TABLE II
MEASURE OF TESTABILITY FOR THE CIRCUIT OF FIG. 3 USING VARIOUS TEST OUTPUTS

Outputs	δ_{min}
V_o	3
I_{C_1}	2
$V_{R'_a}$	2
I_e	3
V_o, I_{C_1}	0
V_o, $V_{R'_a}$	1
V_o, I_e	0
I_{C_1}, $V_{R'_a}$	2
I_{C_1}, I_e	1
$V_{R'_a}$, I_e	0
V_o, I_{C_1}, $V_{R'_a}$	0
V_o, I_{C_1}, I_e	0
V_o, $V_{R'_a}$, I_e	0
I_{C_1}, $V_{R'_a}$, I_e	0
V_o, I_{C_1}, $V_{R'_a}$, I_e	0

Although we do not propose to discuss the actual solution of the fault diagnosis equations here, it should be pointed out that by assuming that relatively few components have failed, say $p \ll n$, it is possible to develop specialized algorithms for the solution of the fault diagnosis equations which are far more efficient than standard equation solvers in this application [7], [11], [12]. These are typically derived from the *fault simulation* algorithms used in the diagnosis of digital systems and may naturally be classified into "*simulation before test*" and "*simulation after test*" algorithms. Some of the algorithms are discussed in [7] and [9]–[11].

Finally, we note that as formulated above, the measure of testability δ_{min} assumes that any combination component failure is possible. If, however, we assume that at most $p \ll n$ components fail simultaneously, the ambiguity in the solution of the fault diagnosis equations may actually be less than δ_{min}. For instance, in the example of Fig. 3, with only V_0 taken as an output $\delta_{min} = 3$, yet the fault diagnosis equations can be solved exactly if we assume that only one parameter is out of tolerance [10]. The point, here, is that even though the solution of the fault diagnosis equations in n space has three arbitrary parameters when the solution is restricted to the one dimensional manifold of parameter vectors in which all but one coordinant are nominal it is unique.

REFERENCES

[1] W. J. Dejka, "Measure of testability in device and system design," in *Proc. 20th Midwest Symp. Circuits Syst.* (Lubbock, TX), pp. 39–52, Aug., 1977.

[2] ——, "A review of measurements of testability for analog systems," *Proc. 1977 AUTOTESTCON* (Hyannis, MA), pp. 279–284, Nov. 1977.

[3] N. Sen and R. Saeks, "A measure of testability and its application to test point selection—Theory," *Proc. 20th Midwest Symp. Circuits Syst.* (Lubbock, TX), pp. 576–593, Aug. 1977.

[4] ——, "A Measure of testability and its application to test point selection—Computation," *Proc. 1977 AUTOTESTCON* (Hyannis, MA), pp. 212–219, Nov. 1977.

[5] N. Sen, M.S. thesis, Texas Tech Univ., Lubbock, TX, 1977.

[6] M. N. Ransom and R. Saeks, "A functional approach to fault analysis in linear systems," in *Rational Fault Analysis*. New York: Marcel Dekker, 1977, pp. 124–134.

[7] R. Saeks and R. A. DeCarlo, *Interconnected Dynamical Systems*. New York: Marcel Dekker, to be published.

[8] S. P. Singh and R.-W. Liu, "Existence of state equation representation of linear large-scale dynamical systems," *IEEE Trans. Circuits and Syst.*, vol. CAS-20, pp. 239–246, 1973.

[9] *Proc. of the Workshop on Automatic Test Technology*, NSIA, San Diego, April, 1978.

[10] H. M. S. Chen and R. Saeks, "A search algorithm for the solution of the fault diagnosis equations," Texas Tech Univ., (unpublished notes), 1978.

[11] H. M. S. Chen, M.S. thesis, Texas Tech Univ., 1977.

[12] M. N. Ransom and R. Saeks, "The connection function—Theory and application," *Int. J. Circuit Theory and its Applications*, vol. 3, pp. 5–21, 1975.

[13] W. Fleming, *Functions of Several Variables*. Reading: Addison-Wesley, 1965.

[14] M. Spivak, *Calculus on Manifolds*. Benjamin: Amsterdam, 1966.

[15] R. S. Berkowitz, "Conditions for network element value solvability", *IRE Trans. Circuit Theory*, vol. CT-9, pp. 24–29, 1962.

[16] S. O. Bedrosian and R. S. Berkowitz, "Solution procedure for single-element-kind networks," *IRE Inter. Conv. Record*, vol. 10, Part 2, p. 16, 1962.

Neeraj Sen (M'78) was born in India, on October 25, 1950. He received the B.S. degree in electrical engineering from Panjab University, Chandigarh, India, in 1971, and the M.S. degree in electrical engineering from Texas Tech University, Lubbock, in 1977. He is currently enrolled in a Ph.D. program in electrical engineering at the University of Texas, Austin.

From 1971 to 1976 he worked at BEL in India, in the field of radar engineering. He also spent several months on deputations at Thomson-CSF in France, working on signal processing and data handling for radar systems. From 1976 through 1977 he was a Research Assistant at Texas Tech, working on system theory, fault analytic techniques, and computer applications. Since January 1978, he has been employed by the Datapoint Corporation where he works in large system engineering and is responsible for developing programs for computer systems diagnosis. His present interests are in the area of systems and software engineering.

Mr. Sen is a member of the Eta Kappa Nu and Tau Beta Pi.

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. 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 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 AMS, SIAM, and Sigma Xi.

9. Reprint of "A Search Algorithm for the Solution of the Multifrequency Fault Diagnosis Equations" by H.S.M. Chen and R. Saeks from the IEEE Transactions on Circuits and Systems, Vol. CAS-26, pp. 589-594, (1979).

A Search Algorithm for the Solution of the Multifrequency Fault Diagnosis Equations

H. S. M. CHEN AND RICHARD SAEKS

Abstract—A search algorithm for the solution of the fault diagnosis equations arising in linear time invariant analog circuits and systems is presented. By exploitation of Householder's formula an efficient algorithm whose computational complexity is a function of the number of system failures rather than the number of system components is obtained.

I. Introduction

Conceptually, the fault analysis problem for an analog circuit or system amounts to the measurement of a set of externally accessible parameters of the system from which one desires to determine the internal system parameters. To solve the fault diagnosis problem, one then measures a vector of external parameters, $m = \text{col}\ (m_i)$, and solves a nonlinear algebraic equation

$$m = F(r) \tag{1}$$

for a vector of internal system parameters, $r = \text{col}\ (r_i)$, to diagnose the fault. For linear time-invariant systems the function F can be expressed analytically [14]. More generally, in the nonlinear case, one can evaluate $F(r)$ for any given parameter vector r with a simulator, and thus solve (1) numerically, even though F has no analytic expression.

Although one does not usually formulate the fault diagnosis problem in terms of the above described equation solving notation, this formulation is equivalent to the classical fault simulation concept [9]. Indeed, fault simulation is simply a search algorithm for solving (1). Here, one precomputes $\hat{m} = f(\hat{r})$ for each allowable[1] faulty parameter vector $\hat{r}$ and then compares the measured m with the simulated $\hat{m}$'s, stored in a fault dictionary, to solve (1).

Manuscript received June 1, 1978; revised April 20, 1979. This work was supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

H. S. M. Chen is with the National Semiconductor Corporation, Santa Clara, CA on leave from the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

R. Saeks is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

[1]By allowable faults we mean all possible parameter vectors $\hat{r}$ which satisfy a specified set of fault hypotheses. These typically restrict the maximum number of component parameters which are simultaneously out of tolerance and the type of failure (open circuit, short circuit, small change, etc.).

0098-4094/79/0700-0589$00.75

Although the above described approach to fault simulation has been successful[2] when applied to digital system, there is considerable question surrounding its applicability to analog circuits and systems [1]. The problem here is two fold. First, rather than simply failing as a one or zero, an analog parameter has a continuum of possible failures. Second, unlike a digital system wherein a component is either good or bad, in an analog system, a component parameter is either in tolerance or out of tolerance. As such, for each hypothesized failure, it may prove necessary to do an entire family of Monte Carlo simulations in which the values of the good components are randomly chosen within their tolerance limits. Although, at the present time we have insufficient practical experience to determine the precise number of fault simulations required for analog fault diagnosis, it is estimated that the number of simulations required for an analog system will exceed the number of simulations required for a digital system of similar complexity by a factor ranging between two and six order of magnitude [1]. As such, the fault simulation concept which has proven to be so successful for a digital system may not be applicable in the analog case.

As an alternative to fault simulation, one may adopt one of the more classical equation solving algorithms for the solution of (1) [2], [3]. Here, one first measures m and on the basis of this measurement, makes an initial guess r^0 (usually taken to be nominal parameter vector) at the solution of the equations. One then evaluates $m^0 = F(r^0)$ and compares it with m. If $m^0 = m$, r^0 is the solution to the fault diagnosis equation. If not, one makes a new "educated" guess at the solution r^1 (usually based on the deviation between m and m^0) and repeats the process by evaluating $m^1 = F(r^1)$ and comparing it with m. Hopefully, sequence of component parameter vectors r^i and simulated data vectors, $m^i = F(r^i)$, is obtained which "quickly" converges to r and m, respectively. Since the evaluation of $F(r^i)$ is essentially equivalent to the simulation of the system with the faulty parameter values r^i this technique is really another form of fault simulation. In this case, however, one simulates the system after the data vector has been measured and uses this data to make an educated guess at a (hopefully) small number of parameter vectors at which the system should be simulated. As such, the approach has been termed *simulation after test* [1] to distinguish it from the classical approach wherein all *simulation* is done *before test* [1].

At the time of this writing, both approaches are under study [1], neither of which have been shown to be superior. Fault "simulation after test" requires that one include an efficient simulator in the ATE itself, which can be used for on-line computation of $m^i = F(r_i)$ after the UUT has been measured. On the other hand, simulation after test eliminates the requirement of searching a large fault dictionary for the (approximate) data matches required by "simulation before test." In addition, the complex ATPG requirement for "simulation before test" is eliminated.

To make "simulation after test" feasible, however, an efficient equation solving algorithm is required to obtain convergence of the r^i sequence in a reasonable amount of time. Moreover, since "real world" failures in analog circuits and systems often take the form of open and short circuited components or large parameter deviations from nominal the classical perturbational algorithms a la Newton–Raphson are inapplicable. Fortunately, in the context of the fault diagnosis problem, one can reasonably assume that relatively few component parameters have failed. As such, even though it is not valid to assume that $r - r^0$ (the deviation of r from nominal) is small in norm, it is reasonable to assume that it is small in "rank." The purpose of the present paper is to formulate a search algorithm for the solution of the fault diagnosis equations which exploits such an assumption.

In the remainder of the introduction, the explicit form for the fault diagnosis equations arising in linear time-invariant circuits and systems derived in [3] and [14] is reviewed. Householder's formula [4] is then used to exploit the special form of these equations in combination with an assumption that r differs from r^0 in relatively few coordinants to formulate a search algorithm for the solution of the fault diagnosis equations in which the computational complexity of the simulation process is a function of the number of the failures rather than the number of components. This algorithm is based on a similar algorithm suggested by Temes [5] for "simulation before test" and a large-change sensitivity algorithm first given by Leung and Spence [6]. Finally, examples of the application of the algorithm are presented and a study of the robustness of the algorithm to deviations of the "good" components from their nominal values is presented [7].

In the case of a linear time-invariant circuit or system, the fault diagnosis equations may be expressed explicitly in analytical form [3], [14]. Indeed, it is the explicit nature of this form which makes our simplified solution algorithm possible. Using a "component connection model" as the starting point for the derivation of the fault diagnosis equations [8]. The system components are characterized by a set of simultaneous equations

$$b_i = Z_i(s,r)a_i, \qquad i = 1,2,\cdots,n \tag{2}$$

where a_i and b_i denote the component input and output vectors, respectively, r is our vector of internal system parameters which characterizes the "fault state" of the various components, and s is the complex frequency variable. For notational brevity, these component equations are combined into a single block diagonal matrix equation

$$b = Z(s,r)a \tag{3}$$

where $b = \text{col}\,(b_i)$, $a = \text{col}\,(a_i)$, and $Z(s,r) = \text{diag}\,(Z_i(s,r))$.

The connection equations for the model take the form

$$\begin{aligned} a &= L_{11}b + L_{12}u \\ y &= L_{21}b + L_{22}u \end{aligned} \tag{4}$$

where u and y represent the vectors of accessible test inputs and outputs which are available to the test system. By combining (3) and (4) to eliminate the component input and output variables a and b, one may derive [3], [8], [14] an expression for the transfer function matrix observable by the test system between the test input and output vectors u and y obtaining

$$S(s,r) = L_{22} + L_{21}(1 - Z(s,r)L_{11})^{-1}Z(s,r)L_{12} \tag{5}$$

where

$$y = S(s,r)u. \tag{6}$$

For a linear time-invariant system the transfer function $S(s,r)$ is a complete description of the measurable data about the UUT available to the test system. Moreover, being rational it is completely determined by its value at a finite number of frequencies. As such, without loss of generality we may take our vector of measured data to be of the form

$$m = \text{col}\,[S(s_1,r), S(s_2,r), \cdots, S(s_k,r)]. \tag{7}$$

[2]Most industrial users of ATE obtain satisfactory fault detection in digital circuits via fault simulation techniques but require guided probe techniques in addition to the fault dictionary data for fault diagnosis (isolation).

The fault diagnosis equations then take the form

$$m \triangleq \begin{bmatrix} S(s_1,r) \\ S(s_2,r) \\ \vdots \\ S(s_k,r) \end{bmatrix} = \begin{bmatrix} L_{22}+L_{21}(1-Z(s_1,r)L_{11})^{-1}Z(s_1,r)L_{12} \\ L_{22}+L_{21}(1-Z(s_2,r)L_{11})^{-1}Z(s_2,r)L_{12} \\ \vdots \\ L_{22}+L_{21}(1-Z(s_k,r)L_{11})^{-1}Z(s_k,r)L_{12} \end{bmatrix} \triangleq F(r). \tag{8}$$

In the present context we will assume that $s_1, s_2, \cdots, s_k$ represent sufficiently many frequencies to permit the fault diagnosis equations to be solved. Indeed, algorithms for determining such a set of frequencies when they exist are given in [10]–[12], and [14]. The problem at hand is the development of an efficient algorithm for the solution of these fault diagnosis equations.

II. Householder's Formula and the Search Algorithm

Given the explicit form of fault diagnosis equations of (8), it is apparent that the vast majority of the computation required for the simulation of $F(r)$, either before or after test, is the inversion of the family of matrices; $(1-Z(s_i,r)L_{11})$, $i=1,2,\cdots,k$. Fortunately, given the assumption that relatively few components have failed, i.e., that r differs from its nominal value r^0 in only a small number of coordinents, Householder's formula [4] may be invoked to compute $(1-Z(s_i,r)L_{11})^{-1}$ in terms of $(1-Z(s_i,r^0)L_{11})^{-1}$ together with the inversion of a small dimensional matrix. More precisely, if A, B, C, and D are given matrices of dimension $n\times n$, $n\times n$, $n\times p$, and $p\times n$, respectively, where

$$A = B + CD \tag{9}$$

then

$$A^{-1} = \left[1 - B^{-1}C(1+DB^{-1}C)^{-1}D\right]B^{-1}. \tag{10}$$

As such, once B^{-1} is known, one may compute the inverse of the $n\times n$ matrix A in terms of B^{-1} and the inverse of the $p\times p$ matrix $(1+DB^{-1}C)$. This technique has been used effectively for large change sensitivity analysis [6] and has recently been suggested by Temes for application to fault simulation [5]. This is achieved by exploiting the block diagonal character of $Z(s,r)$. Thus if r differs from r^0 in q coordinents $Z(s,r)$ will differ from $Z(s,r^0)$ only in the $p\times p$ block composed of components which are effected by the faulty parameters.[3] If the rows and columns of $Z(s,r)$ are reordered so that this block appears in the upper left corner of $Z(s,r)$ then,

$$Z(s,r) = Z(s,r^0) + \left[\begin{array}{c|c} \Delta & 0 \\ \hline 0 & 0 \end{array}\right] \tag{11}$$

where Δ is $p\times p$ and $Z(s,r)$ is $n\times n$. We then have

$$(1-Z(s_i,r)L_{11}) = (L-Z(s_i,r^0)L_{11}) + \left[\begin{array}{c} -\Delta(s_i,r) \\ \hline 0 \end{array}\right] L_{11}^{p} \tag{12}$$

where L_{11}^{p} denotes the upper (after reordering) p rows of L_{11}.

[3]Here, p is the sum of the dimensions of all the blocks of $Z(s,r)$ which are dependent on the q coordinents in which r differs from r^0. Typically, $q\approx p$ with the exact relationship depending the block sizes.

Finally, an application of Householder's formula yields

$$(1-Z(s_i,r)L_{11})^{-1} = \left[1 - (1-Z(s_i,r^0)L_{11})^{-1}\left[\begin{array}{c} -\Delta(s_i,r) \\ \hline 0 \end{array}\right]\right. \cdot \left.\left(1 + L_{11}^{p}(1-Z(s_i,r^0)L_{11})^{-1}\left[\begin{array}{c} -\Delta(s_i,r) \\ \hline 0 \end{array}\right]\right)^{-1} L_{11}^{p}\right] \cdot (1-Z(s_i,r^0)L_{11})^{-1}. \tag{13}$$

Although quite complex, the only major matrix computation required for the inversion of $(1-Z(s_i,r)L_{11})$ via 13 is the inversion of the $p\times p$ matrix in parentheses. As such, as long as the number of faulty parameter values remains small, (13) represents an extremely efficient means of carrying out a large number of fault simulations with relatively little computational capacity. Although Temes originally suggested the technique in the context of a "simulation before test" algorithm, the above application of Householder's formula is ideally suited for "simulation after test," wherein, it reduces the computational requirements for the simulation process to well within the capabilities of the minicomputers usually found in modern ATE.

Although Householder's formula yields an efficient means for solving the fault diagnosis equations once the faulty parameters have been determined, it remains to locate the set of faulty parameters. Fortunately, the efficiency of the solution algorithm based on Householder's formula is such that one can justify a search through "all" allowable sets of faulty parameters to locate the actual failures. Indeed, if we denote the "reduced fault diagnosis equations" in which all component values are assumed to be nominal except for q specified parameters; $r_{i(1)}, r_{i(2)}, \cdots, r_{i(q)}$; by $F_{i(1),i(2),\cdots,i(q)}$ then the equation

$$m = F_{i(1),i(2),\cdots,i(q)}(r_{i(1)}, r_{i(2)}, \cdots, r_{i(q)}) \tag{14}$$

will have a solution if and only if the faulty parameter values are among the $r_{i(1)}, r_{i(2)}, \cdots, r_{i(q)}$. As such, if one attempts to solve (14) for each allowable family of faulty parameters, the actual fault will be indicated by the existence of a solution to the equation.

Although such a search algorithm might at first seem to be highly inefficient, when one observes that with the aide of Householder's formula, the evaluation of $F_{i(1),i(2),\cdots,i(q)}$ requires only the inversion of $p\times p$ ($p\approx q$) matrix it is seen that this is not the case. Moreover, if one searches for the most likely failures first, relatively few equations need be solved in practice. In actual implementation in a "simulation after test" algorithm, one can readily search through all possible combinations of one, two, or three simultaneous failures, and commonly encountered combinations of larger numbers of failures, thus locating the far majority of failures in a reasonable amount of ATE time.

An alternative formulation of the search algorithm which alleviates the numerical difficulties associated with the attempt to solve a set of equations which may not have a solution (as is the case whenever one attempts to solve (14) with the wrong choice of faulty parameters) is to employ an optimization algorithm, rather than an equations solver, to minimize

$$J_{i(1),i(2),\cdots,i(q)}(r_{i(1)}, r_{i(2)}, \cdots, r_{i(q)}) = \|m - F_{i(1),i(2),\cdots,i(q)}(r_{i(1)}, r_{i(2)}, \cdots, r_{i(q)})\|^2. \tag{15}$$

Since (15) has a zero minimum if and only if (14) has a solution a search through the minimization of (15) for all allowable sets of faulty parameters will also locate the faulty parameters (indicated by a zero minimum).

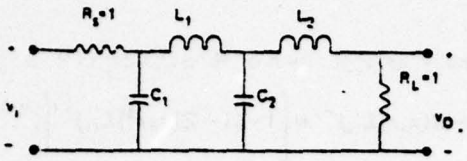


Fig. 1. *LC* filter.

Examples

As a first example, consider the *LC* filter shown in Fig. 1 for which the component connection model takes the form

$$\begin{bmatrix} V_{c1} \\ i_{L1} \\ V_{c2} \\ i_{L2} \end{bmatrix} = \begin{bmatrix} 1/S_{c1} & 0 & 0 & 0 \\ 0 & 1/S_{L1} & 0 & 0 \\ 0 & 0 & 1/S_{c2} & 0 \\ 0 & 0 & 0 & 1/S_{L2} \end{bmatrix} \begin{bmatrix} i_{c1} \\ V_{L1} \\ i_{c2} \\ V_{L2} \end{bmatrix} \quad (16)$$

and

$$\begin{bmatrix} i_{c1} \\ V_{L1} \\ i_{c2} \\ V_{L2} \\ \hline V_0 \end{bmatrix} = \left[\begin{array}{cccc|c} -1 & -1 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 \end{array}\right] \begin{bmatrix} V_{c1} \\ i_{L1} \\ V_{c2} \\ i_{L2} \\ \hline V_i \end{bmatrix}. \quad (17)$$

Since we assume that the source and load resistors are external to the filter and do not fail they have been imbedded into the connection equations and thus do not appear explicitly as components. The filter components are assumed to have the nominal values

$$C_1 = 10 \quad L_1 = 20 \quad C_2 = 30 \quad \text{and} \quad L_2 = 40 \quad (18)$$

and it is assumed that no more than one component fails at a time (though the failure may be catastrophic). Our "simulation after test" fault diagnosis algorithm then requires that we minimize $J_1(C_1)$, $J_2(L_1)$, $J_3(C_2)$, and $J_4(L_2)$. The performance measure with zero minimum then represents the failed component with the minimizing value for that performance measure representing the value of the failed component. All other component values must then be nominal (since it is assumed that only one component fails). Note that the minimizing value for the nonzero J_i's does not correspond with the correct component values for those components.

This filter was simulated with each of its four components out of tolerance (by as much as 100 percent) with the search algorithm being applied to the simulated data. Since only one parameter is assumed to fail at a time and $Z(s,r)$ is diagonal each of the four required minimizations was carried out by purely scalar operations using a Golden Section search. In all four cases the fault was correctly located with the faulty parameter value being determined "exactly." The resultant data is summarized in Table I. Note: in each case the minimum value for J_i for the faulty component is at least three orders of magnitude lower than the minimum value J_i for any nonfaulty component. As such, the failure is easily located and one can expect the algorithm to remain viable in the face of numerical and or approximation error.

As a more sophisticated example, consider the one stage transistor amplifier of Fig. 2 and its wide band equivalent circuit shown in Fig. 3. Note that the parallel resistors R_a and R_b appearing in this model have been lumped together into a single resistance R_s since it is clearly impossible to distinguish between failures in these two components from external measurements.

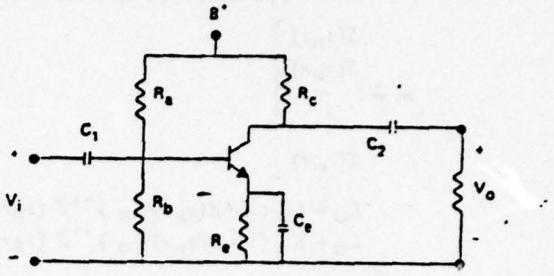


Fig. 2. One stage transistor amplifier.

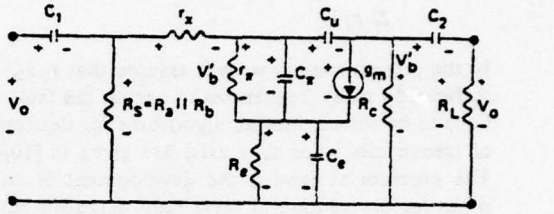


Fig. 3. Amplifier equivalent circuit.

TABLE I
FAULT ANALYSIS FOR *LC* FILTER

Component	C_1	L_1	C_2	L_2
Actual Parameter Values	20	20	30	40
Minimum Value for J_i	1.25×10^{-12}	5.98×10^{-8}	9.46×10^{-9}	1.96×10^{-6}
Minimizing Component Value	20	30.66	39.87	51.89

Component	C_1	L_1	C_2	L_2
Actual Parameter Values	10	40	30	40
Minimum Value for J_i	1.62×10^{-7}	2.10×10^{-14}	2.42×10^{-7}	5.94×10^{-8}
Minimizing Component Value	28.75	40	48.5	62.2

Component	C_1	L_1	C_2	L_2
Actual Parameter Value	10	20	50	40
Minimum Value for J_i	2.91×10^{-6}	2.73×10^{-7}	1.47×10^{-13}	6.87×10^{-6}
Minimizing Component Value	30.27	41.62	50	64.01

Component	C_1	L_1	C_2	L_2
Actual Parameter Value	10	20	30	45
Minimum Value for J_i	3.92×10^{-7}	4.93×10^{-7}	4.13×10^{-7}	1.54×10^{-13}
Minimizing Component Value	14.13	24.46	34.12	45

$$\begin{bmatrix} V_{c1} \\ V_{rx} \\ V_{ry} \\ V_{cu} \\ V_{c2} \\ I_{RS} \\ I_{Re} \\ I_{cv} \\ I_{Ce} \\ I_{gm} \\ I_{Rc} \\ I_{RL} \end{bmatrix} = \begin{bmatrix} 1/C_1 s & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & r_x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & r_y & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/C_u s & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/C_2 s & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/R_S & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/R_e & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_v s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_e s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & g_m & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/R_c & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/R_L \end{bmatrix} \cdot \begin{bmatrix} I_{c1} \\ I_{rx} \\ I_{ry} \\ I_{cu} \\ I_{C2} \\ V_{RS} \\ V_{Re} \\ V_{Cv} \\ V_{Ce} \\ V_{gm} \\ V_{Rc} \\ V_{RL} \end{bmatrix} \tag{19}$$

and

$$\begin{bmatrix} I_{c1} \\ I_{rx} \\ I_{ry} \\ I_{cu} \\ I_{c2} \\ V_{RS} \\ V_{Re} \\ V_{cv} \\ V_{ce} \\ V_{gm} \\ V_{Rc} \\ V_{RL} \\ \hline V_0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} V_{c1} \\ V_{rx} \\ V_{ry} \\ V_{cu} \\ V_{c2} \\ I_{RS} \\ I_{Re} \\ I_{cv} \\ I_{ce} \\ I_{gm} \\ I_{Rc} \\ I_{RL} \\ \hline V_i \end{bmatrix} \tag{20}$$

The component and connection equations for this circuit are given by (19) and (20) and the nominal values for the component parameters are taken to be

$$\begin{array}{llllll} C_1 = 20 & r_x = 10 & r_y = 40 & C_u = 25 & C_2 = 20 & R_S = 75 \\ R_e = 30 & C_v = 15 & C_e = 10 & g_m = 10 & R_c = 10 & R_1 = 20. \end{array} \tag{21}$$

As before, it was assumed that no more than one component failed and $J_i(r_i)$ was minimized for each of the 12 component parameters. Once again the failure was clearly located by the smallest minima with accurate determination of the faulty parameter value. Indeed, in each case, the minimum value of $J_i(r_i)$ for the faulty parameter value is at least five orders of magnitude less than the minima for the remaining $J_i(r_i)$. As such, there is no ambiguity whatsoever in the determination of the faulty component and its value even though the component parameters have been allowed to deviate from their nominal values by as much as 500 percent.

Robustness

Unlike the case of fault diagnosis in a digital system wherein a component is unambiguously good or bad, in an analog circuit or system, a component parameter is either in tolerance or out of tolerance. As such, any fault diagnosis algorithm which makes use of the nominal component parameters must be tested for robustness, i.e., how effective is the algorithm at locating the faulty component(s) when the good components are not precisely equal to their nominal values. As such, our search algorithm for fault diagnosis was applied to the transistor amplifier using simulated measurements in which one component was out of tolerance (taken to be 10 percent) and the remaining component parameters were in tolerance but not equal to their nominal values [7]. Of course, the nominal values are used to define the F_i since the actual value of the good components is unknown. Not surprisingly, this results in some ambiguity in the diagnosis process since $J_i(r_i)$ can never be reduced exactly to zero. As such, our simulation yielded good though not perfect results. In particular, the algorithm correctly located the fault in 71 percent of the trials with an ambiguity group of one in 50 percent of these cases and ambiguity groups of two, three, and four in the remaining cases. Since *all* of the good components in this simulation were taken to be at the limits of their tolerance interval, these results actually represent a worst case situation. As such, we believe that the search algorithm will yield significantly better results in a "real world" situation, wherein most of the components will have near nominal values with relatively few of the good component parameters lying near their tolerance limits.

Hybrid Algorithms

Although the terminology has only recently been formulated [1], most of the algorithms which have been proposed over the years for the solution of the fault analysis problem in analog circuits and systems can naturally be categorized as either "simulation before test" or "simulation after test" algorithms [9]. Although the preceding development has been presented in the context of a "simulation after test" algorithm, many of the

techniques, such as the application of Householder's formula [5] are also applicable to "simulation before test" algorithms. Indeed, the techniques are ideally suited to a hybrid algorithm. Here, one would employ a two-pass diagnostic algorithm wherein the measured data vector m is first compared with presimulated data stored in a fault dictionary. If the fault is so located, the diagnosis process is terminated. If the fault is not located among those which have been presimulated and stored in the fault dictionary, the hybrid algorithm will then revert to a "simulation after test" mode until a sequence of parameter vectors r_i and simulated data vectors m_i have been computed which converge to the solution of the fault diagnosis equations. At the same time the results of each of these "after test" simulations are stored in the fault dictionary for use in future applications of the test algorithm. As such, a fault dictionary is slowly built up which includes simulations of those failures which are most commonly encountered in actual practice. Such a hybrid algorithm would seem to achieve the best of both worlds. Common faults would be found quickly on the first pass, yet the system would still have the "simulation after test" algorithm upon which to fall back when encountering a new failure mode. Moreover, ATPG requirements would be greatly reduced with only the most common faults (say open and short circuits, single failures, etc.), being presimulated and the remainder of the fault dictionary being adaptively generated by the "simulation after test" algorithm as new fault modes are encountered. Such a hybrid scheme alleviates the necessity of determining the fault modes of a system in advance, as required for "simulation before test" while simultaneously eliminating the duplicate simulations of common faults required for "simulation after test."

III. Conclusions

Our purpose in the preceding has been the formulation of a class of techniques which we believe can serve as the basis of an effective algorithm for fault diagnosis in linear analog circuits and systems. These techniques have proven to be effective in the situation where all good component parameters are "near" nominal and give promise of sufficient robustness to cope with the "real world" situation, in which the good component parameters are in tolerance though not nominal.

Although the presentation has been formulated in the context of a "simulation after test" algorithm, the techniques presented are also applicable to "simulation before test" and hybrid algorithms.

References

[1] "Report of the industry-joint services automatic test task force," San Diego, Apr., 1977.

[2] R. Saeks, S. P. Singh, and R. W. Liu, "Fault isolation via components simulation," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 634–640, 1972.

[3] M. N. Ransom and R. Saeks, "A functional approach to fault analysis in linear systems," in *Rational Fault Analysis*, Eds. R. Saeks and S. R. Liberty, Eds. New York: Marcel Dekker, 1977, pp. 124–134.

[4] A. S. Householder "A survey of some closed methods for inverting matrices", *SIAM J. Appl. Math.*, vol. 5, pp. 155–169, 1957.

[5] G. C. Temes, "Efficient methods for fault simulation," *Proc. 20th Midwest Symp. Circuits Syst.* Texas Tech Univ., pp. 191–194, Aug. 1977.

[6] K. H. Leung and R. Spence, "Multiparameter large-change sensitivity analysis and systematic exploration," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 796–804, 1975.

[7] H. S. M., Chen M.S. thesis, Texas Tech University, Lubbock, TX, 1977.

[8] R. Saeks and R. S. DeCarlo, *Interconnected Dynamical Systems*. New York: Marcel Dekker, 1979.

[9] R. Saeks and S. R. Liberty, *Rational Fault Analysis*. New York: Marcel Dekker, 1977.

[10] N. Sen, M.S. thesis, Texas Tech University, Lubbock, TX., 1977.

[11] N. Sen and R. Saeks, "A measure of testability and its application to test point selection—Computation", *Proc. AUTOTESTCON '77*, (Hyannis, MA) pp. 212–219, Nov. 1977.

[12] ——, "A measure of testability and its application to test point selection—Theory", *Proc. 20th Midwest Symp. Circuits Syst.*, Texas Tech Univ., Lubbock, TX, pp. 576–583, Aug. 1977.

[13] H. Trauboth and N. Prasad, "MARSYAS—A software package for digital simulation of physical systems," *Proc. Spg. Joint Comp. Conf.*, pp. 223–235, 1970.

[14] N. Sen and R. Saeks, "Fault diagnosis for linear systems via multifrequency measurements," this issue, pp. 457–465.

10. Reprint of "Failure Prediction for an On-Line Maintenance System in a Poisson Shock Environment" by K.S. Lu and R. Saeks from IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-9, pp. 356-362, (1979).

Failure Prediction for an On-Line Maintenance System in a Poisson Shock Environment

K. S. LU AND R. SAEKS, FELLOW, IEEE

Abstract—**A failure prediction algorithm for application in a periodic on-line maintenance system operating in a Poisson shock environment is described. The system under test is measured at periodic maintenance intervals with the data derived therefrom being used to estimate system lifetime and determine an optimal replacement time. The resultant algorithm is simulated and compared with various fixed replacement schedules.**

I. Introduction

Although considerable effort has been expended during the past decade to develop techniques for fault detection and diagnosis in both analog and digital electronic circuits [10], little attention has been given to the possibility of formulating algorithms for fault prediction. To accurately predict a fault, a device must be tested at periodic maintenance intervals. If the device fails or does not operate correctly, it is replaced immediately. The device may be assumed good if its characteristics are in tolerance. However, if the characteristics are slightly off nominal but the device still operates correctly, one can attempt to predict if the device will fail before the next scheduled maintenance interval. If device failure is predicted, it can be replaced before failure occurs as part of planned preventative maintenance.

Manuscript received April 3, 1978; revised September 18, 1978 and February 1, 1979. This work was supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

R. Saeks is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409

K. S. Lu is with the Texas Instruments Incorporated, Dallas, TX.

0018-9472/79/0600-0356$00.75

With the advent of the low-cost microprocessor, on-line fault prediction is possible and practical [9]. For this purpose, a curve fitting algorithm for on-line fault prediction was first introduced by Saeks, Liberty, and Tung [11]-[13] in 1975. The disadvantage of this algorithm, however, is that the second-order polynomial model employed is too simple to describe the aging curve of a real-world component. Employing the Poisson-shock model for the wear process introduced by Esary, Marshall, and Proschan [1], [2], [6], another curve fitting fault prediction algorithm which overcomes these disadvantages is discussed in the present paper [9].

In the following section a model for the failure dynamics of a system component parameter is formulated. Here it is assumed that the failure is due to the component being subjected to a sequence of Poisson distributed shocks [3], [7], with the measurable parameter being controlled by an unknown difference equation whose underlying discrete "component time" process is defined by the number of shocks to which the component has been subjected. Since both the failure dynamics (i.e., the difference equation) and the relationship between "component time" and real time are unknown, our failure model is doubly stochastic. The third section of the paper is devoted to the formulation of an algorithm for estimating the component failure dynamics, and its "lifetime" is defined to be the number of shocks required to cause component failure. This is followed by the formulation of an "optimal" replacement theory wherein the optimal real time at which to replace a component is computed in terms of its estimated "lifetime." Finally, the results of a simulation of the algorithm in both an ideal and noisy environment are presented and compared with the simulated performance for several fixed replacement schedules.

II. Failure Dynamics

Let $C(N)$ represent values of a particular component parameter, where the "component time" N denotes the number of shocks the component has received. It is assumed that the drifting parameters can be described by a first-order[1] difference equation of the form:

$$C(N+1) = C(N) - a_0 - a_1 N - a_2 N^2 - \cdots - a_h N^h$$
$$C(0) = 1. \tag{1}$$

Here the coefficients and order of the "forcing polynomial" are assumed to be unknown and must be estimated as part of the fault prediction process. A little algebra together with the standard recursive formula for solving a difference equation will reveal that

$$C(N) = 1 - \sum_{j=0}^{N-1} \sum_{i=0}^{h} a_i j^i. \tag{2}$$

Now, if the tolerance limit for the component parameter is normalized to $C = 0$, we may define the lifetime of the component to be the smallest integer N for which $C(N) \le 0$. This integer, which we denote by L, then represents the number of shocks necessary to cause the component to fail.

Since the failure model of (1) is dependent on "component time," i.e., the number of shocks the component has received, rather than real time, it remains to define the relationship between "component time" and real time. Following common practice in reliability theory [1], we assume that this relationship is determined by a Poisson process. Indeed, this is the unique point process which has the scaling properties required for such an application [3]. Here the probability of N shocks occurring in the time interval t is

$$P_N(t) = e^{-kt} \frac{(kt)^N}{N!}, \qquad N = 0, 1, 2, \cdots \tag{3}$$

where k is a given constant representing the average number of shocks per unit time. Therefore, (kt) is the average number of shocks in the time interval t.

III. Estimation of Failure Dynamics and Lifetime

In a periodic maintenance system, the performance of a component is measured at each maintenance interval nT. That is to say, $(C_1, C_2, \cdots, C_g)$ is the performance data taken at maintenance times $(T, 2T, \cdots, gT)$. The estimation problem can be stated as, "Given performance data $(C_1, C_2, \cdots, C_g)$, T and k, estimate the unknown constants $(a_0, a_1, \cdots, a_h)$ of the failure dynamics." Since it is assumed that the system is subjected to Poisson shock with constant k, the expected number of shocks in each maintenance interval is kT.[2] As such, if we assume that C_m is the value of the component parameter at $N = mkT$, then upon substituting $C_m = C(mkT)$ in (2), we obtain

$$\sum_{j=0}^{mkT-1} a_0 j^0 + \sum_{j=0}^{mkT-1} a_1 j^1 + \cdots + \sum_{j=0}^{mkT-1} a_h j^h = 1 - C_m$$

where $m = 1, 2, 3, \cdots, g$, or in the matrix form:

$$JA \triangleq \begin{bmatrix} \sum_{j=0}^{kT-1} j^0 & \sum_{j=0}^{kT-1} j^1 & \cdots & \sum_{j=0}^{kT-1} j^h \\ \sum_{j=0}^{2kT-1} j^0 & \sum_{j=0}^{2kT-1} j^1 & \cdots & \sum_{j=0}^{2kT-1} j^h \\ \vdots & \vdots & & \vdots \\ \sum_{j=0}^{gkT-1} j^0 & \sum_{j=0}^{gkT-1} j^1 & \cdots & \sum_{j=0}^{gkT-1} j^h \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_h \end{bmatrix} = \begin{bmatrix} 1 - C_1 \\ 1 - C_2 \\ \vdots \\ 1 - C_g \end{bmatrix} \triangleq Z. \tag{4}$$

Since the number of data points g is typically much greater than the order of the polynomial assumed in the failure model h, it is not expected that (4) admits an exact solution. Rather, we attempt to solve for a coefficient vector A which minimizes the error between JA and Z. In particular, if one adopts a least squares error criterion, the optimal A is given by

$$A^0 = J^{-G} Z \tag{5}$$

where J^{-G} denotes the generalized inverse of J [8]. Indeed, if as is typically the case, J has full column rank, then $J^{-G} = (J^t J)^{-1} J^t$ where t denotes matrix transposition. As such, we take $A^0 = \text{col}\,(a_0^0, a_1^0, \cdots, a_h^0)$ as our estimate of the coefficients of the difference equation characterizing the failure dynamics of our drifting parameter C as per (1).

To estimate the failure dynamics of a drifting parameter, the proper choice of the order h is, in general, quite difficult and depends upon physical considerations and engineering experience. Once h is preselected, however, coefficients to best approximate the failure dynamics can be readily computed via (5). The

[1] The concepts described herein carry over without modification to the case where the failure model is characterized by higher order difference equations. The first-order model, however, suffices to illustrate the theory and is hence used throughout the present paper.

[2] Although not theoretically necessary, we assume that kT is an integer.

accuracy of the resultant estimate, however, is highly dependent on the choice of the order h and on the number of measurements which are taken g. To find a new set of coefficients for a different combination of h and g, the entire calculation procedure is typically repeated from the very beginning, a process which is impractical in the on-line maintenance system. Fortunately, sequential refinement schemes for obtaining new sets of coefficients without repeating the entire calculation can be developed [5], [8]. As such, it is possible to sequentially update one's estimates of the parameters $a_0, a_1, \cdots, a_h$ as additional measurements are taken and/or to increase the order of the model for the failure dynamics without repetitious matrix inversion. Our algorithm for estimation of the failure dynamics underlying the measured data may thus be readily implemented on-line with the computational power presently available in today's microprocessors. The matrix algebraic details of the required sequential refinement schemes are straightforward [5], [8] and readily available in the literature. As such, they will not be repeated here.

In practice, given g measurements $C_1, C_2, \cdots, C_g$ taken at maintenance intervals $T, 2T, 3T, \cdots, gT$, one sequentially estimates the coefficients of the failure dynamics $a_0, a_1, \cdots, a_h$, increasing h until no further error reduction is achieved. The resultant set of coefficients is then used in (2) to determine the component lifetime L. Upon solving the equation, the resultant estimated lifetime is found to be the smallest integer L such that

$$\sum_{j=0}^{L-1} \sum_{i=0}^{h} a_i j^i \geq 1. \tag{6}$$

Of course, if the measured data is not decaying towards zero, i.e., the component is not failing, this inequality will have no solution, in which case we take L to be infinite [4].

IV. Replacement Theory

Although the algorithm outlined in the preceeding section yields an "optimal" estimate of the number of shocks required to cause failure, the time at which the Lth shock takes place is statistical in nature, and hence, it still remains to determine the optimal (in an appropriate sense) time at which to replace the component. One such criterion is formulated in the following. For this purpose, it is assumed that L has been computed to our satisfaction and we desire to choose a time T_r at which to replace the component as a function of L. Given L and T_r, we denote the resultant probability of on-line failure (i.e., failure before T_r) by P_f. $P_r = 1 - P_f$ then denotes the probability that the component is replaced at time T_r before it fails. Similarly, we let $\hat{T}_f$ denote the expected time to failure for those components which fail on-line, we let $\hat{T}$ denote the expected time to failure for all components, and we let T^* denote the expected time to failure for the components if they were operated to failure without replacement (i.e., $T^* = \hat{T}|_{T_r \to \infty}$). Finally, we let $f_L(t)$ denote the probability density function that the component receives the Lth shock at time t, given that the component fails on-line, whereas $p_i(t)$ represents the density function of the Poisson distribution with parameter (k), and $E_L(t)$ represents the corresponding distribution function; i.e.,

$$p_i(t) = \frac{(kt)^i}{i!} e^{-kt}, \qquad i = 0, 1, 2, \cdots \tag{7}$$

$$E_L(t) = \sum_{i=0}^{L-1} p_i(t). \tag{8}$$

With the aid of some elementary calculus [9], P_f, P_r, $\hat{T}_f$, and $\hat{T}$, as well as their derivatives with respect to T_r, can be computed analytically. As such, upon defining an appropriate cost measure, an explicit formula for determining an "optimal" T_r (given L) can be derived. We begin with the derivation of the explicit formula for the various quantities involved in our replacement theory.

Since a component will be replaced by our algorithm if and only if it is still operating at time T_r, i.e., if it has not yet received L shocks at time T_r, the probability of replacement is just the probability of receiving less than L shocks by time T_r. We thus have the following.

Property 1:

$$P_r = E_L(T_r).$$

Proof:

$$P_r = \sum_{i=0}^{L-1} \frac{(kT_r)^i}{i!} e^{-kT_r} = \sum_{i=0}^{L-1} P_i(T_r) = E_L(T_r). \tag{9}$$

Property 2:

$$P_f = 1 - E_L(T_r).$$

Property 3:

$$\int_0^{T_r} p_i(t)\, dt = (1/k)\{1 - E_{i+1}(T_r)\}.$$

Proof:

$$\int_0^{T_r} p_i(t)\, dt = \int_0^{T_r} \frac{(kt)^i}{i!} e^{-kt}\, dt = \frac{k^i}{i!} \int_0^{T_r} t^i e^{-kt}\, dt. \tag{10}$$

Using the identity

$$\int x^m e^{ax}\, dx = e^{ax} \sum_{r=0}^{m} (-1)^r \frac{m!\, x^{m-r}}{(m-r)!\, a^{r+1}} \tag{11}$$

this becomes

$$\begin{aligned}
\int_0^{T_r} p_i(t)\, dt &= \frac{k^i}{i!} e^{-kt} \sum_{r=0}^{i} (-1)^r \frac{i!\, t^{i-r}}{(i-r)!\, (-k)^{r+1}} \Bigg]_0^{T_r} \\
&= \frac{k^i}{i!} \left| e^{-k \cdot 0} \frac{i!}{k^{i+1}} - e^{-kT_r} \sum_{r=0}^{i} \frac{i!\, T_r^{i-r}}{(i-r)!\, k^{r+1}} \right| \\
&= \frac{1}{k} \left| 1 - e^{-kT_r} \sum_{r=0}^{i} \frac{(kT_r)^{i-r}}{(i-r)!} \right| \\
&= \frac{1}{k} \left| 1 - e^{-kT_r} \sum_{j=0}^{i} \frac{(kT_r)^j}{j!} \right| \\
&= \frac{1}{k} \left| 1 - \sum_{j=0}^{i} p_j(T_r) \right| \\
&= \frac{1}{k} \{1 - E_{i+1}(T_r)\}.
\end{aligned} \tag{12}$$

Property 4:

$$f_L(t) = \frac{p_{L-1}(t)}{1/k(1 - E_L(T_r))}.$$

Proof: To derive this conditional density function we partition the interval $(0, T_r)$ into N segments of length $\Delta = T_r/N$, and we compute the probability that the Lth shock takes place in the ith time interval $((i-1)\Delta, i\Delta]$. Since this can be caused by having $L-1$ shocks before $(i-1)\Delta$ and at least one shock in the interval $((i-1)\Delta, i\Delta]$, or by having $L-2$ shocks before $(i-1)\Delta$ and at

least two shocks in the interval $((i-1)\Delta, i\Delta]$, etc., the probability of failure in the ith interval is given by

$$\sum_{j=1}^{L} p_{L-j}((i-1)\Delta)[1-E_j(\Delta)] = \sum_{j=1}^{L} p_{L-j}((i-1)\Delta) \sum_{q=j}^{\infty} \frac{(\Delta k)^q}{q!} e^{-\Delta k} = \sum_{r=1}^{\infty} \frac{1}{r!} \left[\sum_{s=1}^{r} p_{L-s}((i-1)\Delta)\right] (\Delta k)^r e^{-\Delta k}. \quad (13)$$

Taking the probability density function at a point t in the interval $((i-1)\Delta, i\Delta]$ to be limiting value of this quantity divided by Δ as Δ goes to zero [7], it is observed that the terms of (13) containing powers of (Δk) greater than 1 go to zero in the limit. As such, the probability density function for the Lth shock to take place at time t is given by

$$\lim_{\Delta \to 0} \frac{p_{L-1}((i-1)\Delta)(\Delta k)e^{-k}}{\Delta} = k p_{L-1}((i-1)\Delta). \quad (14)$$

Finally, since we are interested only in the conditional probability density function that the Lth shock will take place at time t, given that the component fails on-line, the quantity of (14) must be normalized, yielding

$$f_L(t) = \frac{k p_{L-1}((i-1)\Delta)}{P_f} = \frac{p_{L-1}(t)}{\frac{1}{k}(1-E_L(T_r))} \quad (15)$$

as was to be shown.

Property 5:

$$\bar{T}_f = \frac{L}{k} \frac{1-E_{L+1}(T_r)}{1-E_L(T_r)}.$$

Proof: Since T_f is the expected lifetime of the components which fail before replacement,

$$T_f = \int_0^{T_r} t f_L(t)\,dt = \int_0^{T_r} \frac{t P_{L-1}(t)}{\frac{1}{k}\{1-E_L(T_r)\}}\,dt = \frac{\int_0^{T_r} t \frac{(kt)^{L-1}}{(L-1)!} e^{-kt}\,dt}{\frac{1}{k}\{1-E_L(T_r)\}} = \frac{\frac{L}{k}\int_0^{T_r} \frac{(kt)^L}{L!} e^{-kt}\,dt}{\frac{1}{k}\{1-E_L(T_r)\}} = \frac{-L\int_0^{T_r} p_L(t)\,dt}{\{1-E_L(T_r)\}}. \quad (16)$$

From Property 3, (6) thus reduces to the desired equality.

Property 6:

$$\hat{T} = \frac{L}{k}\{1-E_{L-1}(T_r)\} + T_r E_L(T_r).$$

Proof:

$$\hat{T} = P_f \bar{T}_f + P_r T_r = \{1-E_L(T_r)\}\frac{L}{k}\frac{1-E_{L+1}(T_r)}{1-E_L(T_r)} + T_r E_L(T_r) = \left|\frac{L}{k}1 - E_{L+1}(T_r)\right| + T_r E_L(T_r). \quad (17)$$

Property 7:

$$T^* = \frac{L}{k}.$$

Property 8:

$$\frac{d(P_f)}{d(kT_r)} = p_{L-1}(T_r) \quad \text{and} \quad \frac{d(P_r)}{d(kT_r)} = -p_{L-1}(T_r).$$

Proof: This result follows simply by differentiating the expressions for P_f and P_r of properties 1 and 2 analytically:

$$P_r = E_L(T_r) = \sum_{i=0}^{L-1} \frac{(kT_r)^i}{i!} e^{-kT_r} = e^{-kT_r} + \sum_{i=1}^{L-1} \frac{(kT_r)^i}{i!} e^{-kT_r}. \quad (18)$$

Thus

$$\frac{d(P_r)}{d(kT_r)} = -e^{-kT_r} + \sum_{i=1}^{L-1} \frac{i(kT_r)^{i-1}}{i!} - \frac{(kT_r)^i}{i!} e^{-kT_r} = \sum_{i=1}^{L-1} \frac{(kT_r)^{i-1}}{(i-1)!} e^{-kT_r} - \sum_{i=0}^{L-1} \frac{(kT_r)^i}{i!} e^{-kT_r} = E_{L-1} - E_L = -p_{L-1}(T_r). \quad (19)$$

Moreover, since

$$P_f = 1 - P_r, \quad (20)$$

$$\frac{d(P_f)}{d(kT_r)} = \frac{d(1-P_r)}{d(kT_r)} = p_{L-1}(T_r). \quad (21)$$

Property 9:

$$\frac{d(k\bar{T}_f)}{d(kT_r)} = L\frac{[1-E_L(T_r)]p_L(T_r) - \{1-E_{L+1}(T_r)\}p_{L-1}(T_r)}{[1-E_L(T_r)]^2}.$$

Proof: From Property 3,

$$k\bar{T}_f = L\frac{1-E_{L+1}(T_r)}{1-E_L(T_r)}. \quad (22)$$

Thus by direct differentiation

$$\frac{d(k\bar{T}_f)}{d(kT_r)} = L\frac{\{1-E_L(T_r)p_L(T_r) - \{1-E_{L+1}(T_r)\}p_{L-1}(T_r)}{\{1-E_L(T_r)\}^2}. \quad (23)$$

Property 10:

$$\frac{d(k\hat{T})}{d(kT_r)} = E_L(T_r).$$

Proof: From Property 6,

$$\hat{T} = \frac{L}{k}\{1-E_{L-1}(T_r)\} + T_r E_L(T_r). \quad (24)$$

Hence

$$k\hat{T} = L\{1 - E_{L+1}(T_r)\} + kT_r E_L(T_r). \tag{25}$$

Thus by direct differentiation

$$\frac{d(k\hat{T})}{d(kT_r)} = L\{p_L(T_r)\} + (kT_r(-p_{L-1}(T_r)) + E_L(T_r)$$

$$= Lp_L(T_r) - kT_r p_{L-1}(T_r) + E_L(T_r). \tag{26}$$

Since

$$Lp_L(T_r) = L\frac{(kT_r)^L}{L!}e^{-kT_r}$$

$$= (kT_r)\frac{(kT_r)^{L-1}}{(L-1)!}e^{-kT_r}$$

$$= kT_r p_{L-1}(T_r), \tag{27}$$

this reduces to

$$\frac{d(k\hat{T})}{d(kT_r)} = E_L(T_r) \tag{28}$$

as required.

Given the above statistics for replacement, on-line failure, and expected time to failure of a component with estimated lifetime L and assumed replacement time T_r, we desire to choose T_r (given L) which minimizes some appropriate cost function. Intuitively, this cost function should represent both the cost of on-line failure and the cost of wasted component lifetime due to replacing components before failure [12], [13]. We therefore adopt the cost functional

$$\text{cost} = C_f P_f + C_w(kT^* - k\hat{T}). \tag{29}$$

Here C_f and C_w are appropriate weight factors representing the cost of on-line failure and the cost of component lifetime wastage, respectively. Thus the first term in the cost functional represents the expected cost of a failure (i.e., the probability of an on-line failure times the cost of such a failure), whereas the second term in the cost functional represents the expected cost of wasted component lifetime (i.e., the expected lifetime reduction times the cost per unit time for such a lifetime reduction, with k serving as a normalizing factor).

To minimize the cost functional of (29), one simply substitutes the values for $P_f(T_r)$, T^*, and $\hat{T}(T_r)$ computed in the preceding pages, differentiating the cost with respect to kT_r and setting it equal to zero. This then results in the equality [4]

$$0 = C_f p_{L-1}(T_r) - C_w E_L(T_r) \tag{30}$$

where $d(P_f)/d(kT_r)$ is given by property 9 and $d(k\hat{T})/d(kT_r)$ is given by property 10. Thus the choice of an optimal T_r (given L) is reduced to the solution of a single nonlinear equation in one unknown. The solutions of this equation are plotted in Fig. 1 for a number of values of L and C_f/C_w. Indeed, it can be readily shown that (30) has exactly one solution for $T_r > 0$. Moreover, the function

$$R_L(t) = C_f p_{L-1}(t) - c_w E_L(t) \tag{31}$$

takes on negative values for $0 < t < T_r$ and positive values for $T_r < t$; hence in an on-line maintenance system one need not even solve (30). Rather, one simply evaluates $R_L(t)$ at the time of the next scheduled maintenance. If this results in a negative number, the next scheduled maintenance precedes the optimal replacement time, and hence we should wait at least until the next scheduled maintenance (when we will have more data) to replace the component. On the other hand, if the evaluation of $R_L(t)$ at the next scheduled maintenance time results in a positive value, the optimal replacement time will have passed by the next scheduled maintenance, and hence the component should be replaced at the present maintenance interval.

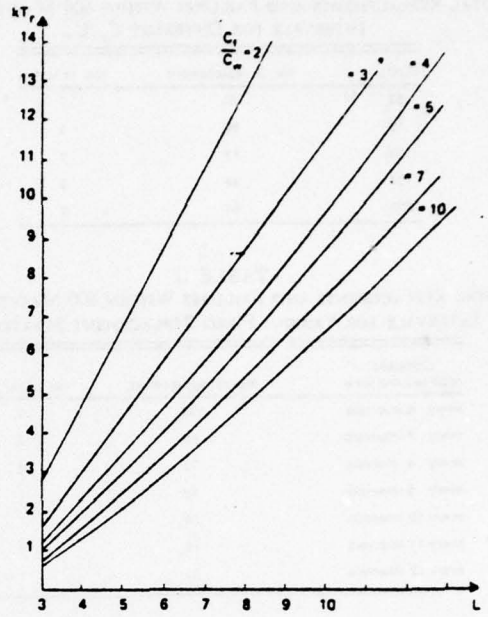


Fig. 1. Replacement time (kT_r) versus Lifetime L with different weight constant.

V. The Algorithm

Summarizing the on-line maintenance algorithm resulting from the above theory takes the following form. At the gth scheduled maintenance interval (at time gT) one measures the component parameter C_g. If C_g is already out of tolerance, the component is simply replaced, and no further analysis is required. If, however, C_g is in tolerance ($C_g > 0$ in our notation), it is used together with the values of the component parameter measured at the previous maintenance intervals to estimate the dynamics of the failure model for the component. Here sequential refinement schemes may be used both to include the effect of C_g on the estimates made at the $(g-1)$st maintenance interval and to increase the order of the polynomial used to represent the component failure dynamics. Once the component failure dynamics have been satisfactorily estimated, one evaluates (31) to estimate whether or not to replace the component. If $R_L((g+1)T) \geq 0$, the component is replaced, whereas if $R_L((g+1)T) < 0$, the component is not replaced, and the system is returned to service until the next scheduled maintenance.

VI. Simulations

A computer simulation of an on-line periodic maintenance system based on the above described algorithm was performed for 600 maintenance intervals with a new component replacing the old component after each replacement decision and/or on-line failure [4]. The system was subjected to computer-generated Poisson shocks with constant $k = 0.1$ shocks per hour and a maintenance interval of $T = 20$ h. The simulation was first run using identical components with $L = 28$ (expected lifetime of 14 maintenance intervals) and then repeated using random components and noisy data measurements.

TABLE I
TOTAL REPLACEMENTS AND FAILURES WITHIN 600 MAINTENANCE INTERVALS FOR DIFFERENT C_f/C_w

C_f/C_w	No. of replacement	No. of failure
50	48	7
75	56	1
100	52	2
150	54	2
200	54	2

TABLE II
TOTAL REPLACEMENTS AND FAILURES WITHIN 600 MAINTENANCE INTERVALS FOR VARIOUS FIXED REPLACEMENT STRATEGIES

Constant replacement time	No. of replacement	No. of failure
every 6 intervals	100	0
every 7 intervals	85	0
every 8 intervals	75	0
every 9 intervals	65	1
every 10 intervals	59	1
every 11 intervals	48	6
every 12 intervals	39	11

TABLE III
OVERALL COST WITH DIFFERENT METHODS AND DIFFERENT C_f/C_w

Cost: C_f/C_w / Methods	50	75	100	150	200
every 6 intervals	1600	1600	1600	1600	1600
every 7 intervals	1096	1096	1096	1096	1096
every 8 intervals	900	900	900	900	900
every 9 intervals	698	723	748	798	848
every 10 intervals	530	555	580	630	680
every 11 intervals	612	762	912	1212	1512
every 12 intervals	750	1025	1300	1850	2400
the algorithm	590	471	512	568	768

TABLE IV
TOTAL REPLACEMENTS AND FAILURES WITHIN 600 MAINTENANCE INTERVALS FOR VARIOUS FIXED REPLACEMENT STRATEGIES AND THE ALGORITHM AT DIFFERENT NOISE LEVELS

noise level / method	20%		30%		40%		60%	
	No. of replace	No. of fail	No. of replace	No. of fail	No. of replace	No. of fail	No. of replace	No. of fail
every 6 intervals	100	0	100	0	100	0	94	6
every 7 intervals	85	0	85	0	84	1	78	8
every 8 intervals	75	0	72	3	71	4	64	12
every 9 intervals	64	2	63	3	60	7	52	17
every 10 intervals	56	4	51	9	45	15	45	18
every 11 intervals	45	10	45	10	45	10	39	20
every 12 intervals	36	15	35	16	36	17	31	23
the algorithm	56	3	55	5	55	5	50	11

TABLE V
OVERALL COST FOR DIFFERENT METHODS AT DIFFERENT NOISE LEVELS

cost: noise level / method	0%	20%	30%	40%	60%
every 6 intervals	1600	1600	1600	1600	2200
every 7 intervals	1096	1096	1096	1280	2008
every 8 intervals	900	900	1200	1300	2128
every 9 intervals	748	848	948	1376	2432
every 10 intervals	580	880	1380	1980	2364
every 11 intervals	912	1340	1340	1340	2452
every 12 intervals	1300	1728	1828	1984	2612
the algorithm	512	752	980	752	1608

For the case where identical components were employed, Table I gives the total number of replacements and failures resulting from the application of the algorithm over the 600 simulated maintenance intervals with different values of C_f/C_w. For comparison, Table II shows the total number of replacements and failures which would have resulted from a fixed replacement strategy ranging from six to twelve maintenance intervals. Since the cost function is

$$\text{cost} = C_f P_f + C_w(kT^* - k\hat{T}) \tag{32}$$

the overall cost can be expressed as

$$\text{cost} = \frac{C_f}{C_w}(\text{number of failures}) + 0.1\,(280^*(\text{number of components used}) - 12000). \tag{33}$$

The overall costs resulting from the application of our algorithm and the various fixed replacement schedules may be computed from the data in Tables I and II. The resultant costs for different values of C_f C_w are given in Table III.

Note that since $L = 28$ for each component in this simulation, an optimal replacement strategy of approximately ten maintenance intervals can be computed from (30) without estimating L. As such, it is not surprising that this fixed replacement strategy resulted in lower costs than the algorithm. It should, however, be noted that the algorithm did not have the advantage of an *a priori* knowledge of L, and yet it still outperformed all of the fixed replacement strategies except the optimal strategy (that is, optimal once L is known).

In our second simulation, random noise was added to the data to simulate both the effects of imperfect measurement and the effect of components with random failure characteristics. Various simulations were run as before for 600 maintenance intervals each, with $k = 0.1$ and $T = 20$ and with noise levels ranging between 20 and 60 percent of the tolerance interval. The results of these simulations are given in Tables IV and V. Except for a single case, which we believe to be anomolous, the algorithm outperformed any fixed replacement strategy.

VII. CONCLUSION

In the preceeding pages, we have described a curve fitting algorithm for the prediction of failures in analog devices. The algorithm was tested in a variety of situations and found to be surprisingly effective in predicting failures with relatively little wastage of component lifetime and on-line failure cost.

REFERENCES

[1] J. D. Esary, A. W. Marshall, and F. Proschan, "Shock models and wear processes," *Ann. Probability*, vol. 1, 1973.
[2] J. D. Esary, Ph.D. dissertation, Univ. of California, Berkeley, 1957.
[3] N. I. Johnson, *Discrete Distribution*. New York: Wiley.
[4] K. S. Lu, Ph.D. dissertation, Texas Tech University, Lubbock, TX, 1977.
[5] ——, M.S.E.E. thesis, Texas Tech University, Lubbock, TX, 1973.
[6] A. W. Marshall, "Some comments on the hazard gradient," *J. Stoch. Proc. Applic.*, vol. 3, 1975.
[7] A. Papoulis, *Probability, Random Variable and Stochastic Processes*. New York: McGraw-Hill, 1965.
[8] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*. New York: Wiley, 1971.
[9] R. Saeks, "An approach to built-in testing," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-14, pp. 813-818, 1978.
[10] R. Saeks and S. R. Liberty, *Rational Fault Analysis*. New York: Marcel Dekker, 1977.
[11] L. Tung, M.S. thesis, Texas Tech University, Lubbock, TX, 1975.
[12] L. Tung and R. Saeks, "An experiment in fault prediction," in *Proc. 4th Int. Symp. Reliability Electronics*, Budapest, Oct. 1977, pp. 249-257.
[13] L. Tung, S. R. Liberty, and R. Saeks, "Fault prediction towards a mathematical theory," in *Rational Fault Analysis*. New York: Marcel Dekker, 1977, pp. 135-142.

11. Reprint of "An Approach to Built-In Testing" by R. Saeks from the IEEE Transactions on Aerospace and Electronic Systems, Vol. AES-14, pp. 813-818, (1978).

An Approach to Built-In Testing

R. SAEKS, Fellow, IEEE
Texas Tech University

Abstract

The architecture and justification for an approach to built-in testing (BIT) in electronic circuits and systems is presented. The proposed system is capable of on-line fault detection and prediction up to the shop replaceable assembly (SRA) level and is designed to interface with external automatic test equipment (ATE) for off-line fault diagnosis within the SRA. The constituent parts of the BIT system have been extensively simulated and the approach appears to be suitable for hardware implementation both with respect to computational and economic considerations.

I. Introduction

An approach to built-in testing (BIT) for electronic circuits and systems is outlined. The approach assumes a two-level hierarchical architecture in which a central microprocessor controls and coordinates the testing of a number of subsystems each of which has built-in test equipment (BITE) such as sensors and a nanoprocessor for preprocessing the test data prior to transmission to the central microprocessor. The approach allows for on-line fault detection and prediction up to the level of a shop replaceable assembly (SRA) and off-line fault diagnosis within the various SRA's

Manuscript received February 10, 1977; revised April 18, 1978.

Author's address: Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409

0018-9251/78/0900-0813 $00.75 © 1978 IEEE

Section II is devoted to a description of the BIT system architecture. This two-level structure has been formulated to be applicable either at the printed circuit board level in which the SRA's represent individual devices (IC chips, elementary components, etc.) or at the level of an entire electronics system in which the SRA's represent printed circuit boards.

Section III is devoted to a study of the fault diagnosis problem. In either the case of a linear or nonlinear circuit it is shown that this problem can be reduced to the solution of a set of simultaneous nonlinear algebraic equations. In the proposed BIT architecture a linearization of these equations is used on-line for fault detection and prediction whereas the full set of nonlinear equations are used off-line for fault diagnosis within the SRA.

Two algorithms for fault prediction are described in Section IV. Both are essentially curve fitting algorithms implemented on the central test microprocessor in a time multiplexed mode. Here the microprocessor periodically receives test data from the various SRA's and extrapolates this data to determine whether or not the SRA is likely to fail in the near future. The final section of the paper is devoted to a discussion of the concept of self-testing, in particular, the possibility of self-testing in a predictive mode.

At the time of this writing the approach to BIT described has yet to be fully implemented. It is, however, predicated on several years of research in the area and each of its constituent subsystems have been extensively simulated [17, 18, 19]. At the present time the hardware implementation of the various algorithms is under investigation [15, 16] and it is hoped that an entire BIT system will be in operation in the near future.

II. BIT Architecture

The basic BIT architecture is a two-level hierarchical structure illustrated in Fig. 1. Intuitively, the overall system may represent a printed circuit board while the subsystems represent various SRA's such as integrated circuits, power supplies, vaccuum tubes, etc. Alternatively, the overall system may represent an entire electronics system with the SRA's being its constituent printed circuit boards. In either case the SRA's may be throw-away units or units intended for off-line repair with BITE designed to detect and/or predict faults in the SRA. For those units intended

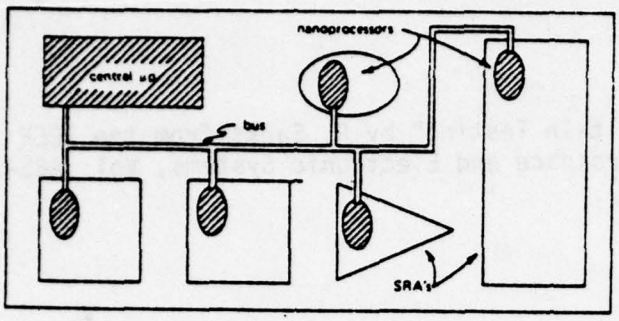


Fig. 1. Two-level BIT architecture.

for off-line repair the BITE may also be used as an interface with an external test stand but will not be capable of isolating the failure within the SRA.

This structure is motivated by several years of basic research into the relative computational complexity of the three fundamental problems of fault analysis: fault detection, fault diagnosis, and fault prediction [9]. The latter problem requires considerable computational power [18, 19] but need only be carried out at widely spaced test intervals, say one test per hour (minute, second, ?). As such, a single central microprocessor can be time-multiplexed through the testing of a large number of subsystem parameters thereby achieving the required computational power for the fault prediction algorithm while still holding the amount of dedicated test equipment within reasonable bounds [10].

While fault diagnosis can be carried out with considerable success the process requires significant computational power (at least a mini by today's standards) and lengthy computer runs [7, 11, 17]. As such, fault diagnosis within an SRA is done off-line on an external test stand containing the required mini (or maxi) computer. Each SRA, however, will include sufficient BITE, say a nanoprocessor, to collect and condition test data on the SRA to be periodically communicated to the central microprocessor for purposes of fault prediction and detection.

Fortunately, both of these endeavors may be achieved using a model of the SRA linearized about its nominal values and hence can be implemented with relatively little computational power built into the SRA [12]. For fault prediction in particular, one is interested in tracking various internal parameters of the SRA as they drift from nominal to their tolerance limit. Since the tolerance interval is typically only a few percent this can be achieved with a linearized model. For catastrophic errors a linearized model may be used to detect failures even though it is not sufficiently accurate for fault diagnosis. As such, the BITE within an SRA may be kept within reasonable bounds while still delivering sufficient data to the central microprocessor for its fault prediction and fault detection tasks. If needed, fault diagnosis within an SRA can, however, be done off-line with the BITE simply serving as an interface between the SRA and an external test stand.

A final aspect of the BIT architecture is the communication link between the SRA's and the central microprocessor. Here one desires to keep the wiring between the SRA's and the central microprocessor at a minimum and simultaneously have all data transmitted to the central microprocessor in a uniform format to permit interchangeability of component parts within the system. Although the details of this communications link have yet to be formalized, the existence of an active computing capability in each SRA gives one considerable flexibility. As such, we believe that it will be possible to work with a single test bus [16]. Here the central microprocessor requests data from the individual SRA's by transmitting a signal on the bus. This signal is received by the built-in nanoprocessor in the SRA which, in turn, transmits appropriately conditioned test data back to the central microprocessor on the same bus.

The BIT architecture just described would seem to achieve most of the requirements for a BIT system:

1) continuous on-line fault prediction and detection to an SRA is achieved;
2) the system includes an interface for off-line fault diagnosis within an SRA;
3) dedicated test equipment represents a small percentage of the total system;
4) busing is minimized and test data is transmitted to the central microprocessor in a uniform format thereby facilitating component interchangeability.

III. Fault Diagnosis

For the purposes of doing fault diagnosis we work with a component connection model for the circuit or system under test which takes the form

$$b_i = Z_i(s,r)\,a_i, \quad i = 1, 2, \ldots, n \tag{1}$$

$$a = L_{11}\,b + L_{12}\,u$$

$$y = L_{21}\,b + L_{22}\,u \tag{2}$$

in the frequency domain [6, 11, 12]. Here $Z_i(s,r)$ is the transfer function of the ith circuit or system component where $R = \mathrm{col}(r_i)$ is the vector of unknown component parameters and s is the complex frequency variable. Typically, the unknown component parameters take the form of amplifier gains and cutoff frequencies, pole and zero positions, resistances, inductances, etc. In particular, it is assumed that enough parameters are employed to completely characterize the performance of the device. The L_{ij} are known connection matrices, $a=\mathrm{col}(a_i)$ and $b=\mathrm{col}(b_i)$ are composite vectors of component inputs and outputs, respectively, and u and y are the test input and output signals, respectively. In the nonlinear case the component equations are replaced by the state models

$$x_i = f_i(X_i, a_i, r)$$

$$b_i = g_i(X_i, a_i, r), \quad i = 1, 2, \ldots, n \tag{3}$$

with the connection equations remaining as in (2). Although these component connection models for a circuit or system are nonclassical they are widely used in large-scale system simulation and computer-aided circuit design and are readily amenable to the "computer speed-up techniques" developed for these applications [12]. As such, they are ideally suited for the fault diagnosis problem.

Combining (1) and (2) yields the fault diagnosis equation [7]

$$S^m = L_{22} + L_{21}\,[1 - Z(s,r)\,L_{11}]^{-1} Z(s,r)\,L_{12} \tag{4}$$

where $Z(s,r) = \mathrm{diag}[Z_i(s,r)]$ and S^m is the measured transfer function relating the input test signal u to the output test signal y. The solution of the fault diagnosis problem therefore amounts to the solution of (4) for the parameters vector r, given S^m and the connection matrices. Although it is possible to give an analytic description of all possible solutions to this equation [12, 13] given any fixed value for the complex frequency variable s in a "real world" situation the number of unknowns greatly exceeds the number of equations and, as such, the analytic representation of the solution manifold proves to be of little value. This difficulty is alleviated via a multifrequency diagnosis algorithm wherein one writes the set of simultaneous equations:

$$\begin{aligned}
S(s_1,r) &= L_{22} + L_{21}\,[1 - Z(s_1,r)L_{11}]^{-1} Z(s_1,r)L_{12}\\
S(s_2,r) &= L_{22} + L_{21}\,[1 - Z(s_2,r)L_{11}]^{-1} Z(s_2,r)L_{12}\\
&\vdots \qquad\qquad\qquad \vdots\\
S(s_k,r) &= L_{22} + L_{21}\,[1 - Z(s_k,r]^{-1} Z(s_k,r)L_{12}
\end{aligned} \tag{5}$$

where k different complex frequencies are used in (4) simultaneously. The interesting and somewhat surprising result is that the additional equations in (5) may be independent thus increasing the number of fault diagnosis equations without increasing the number of its unknowns [7]. While the set of simultaneous equations (5) often has a unique solution, no analytic solution technique is known and we must resort to time-consuming numerical solution procedures carried out off-line.

Although the multifrequency fault diagnosis equations of (5) do not admit an analytic solution their numerical solution can be significantly speeded up by careful analysis of the equations. In particular, a little algebra [6, 12] will reveal that

$$\begin{aligned}
dS^m/dr_j = L_{21}\,&[1 - Z(s_i,r)L_{11}]^{-1}\,[dZ(s_i,r)/dr_j]\\
&\cdot \{1 + L_{11}\,[1 - Z(s_i,r)L_{11}]^{-1}\}L_{12}
\end{aligned} \tag{6}$$

showing that one can compute the partial derivatives required for the numerical solution to (5) analytically. Moreover, if one observes that the inverse matrix required to compute the partial derivatives in (6) is precisely the same inverse matrix required to evaluate the multifrequency fault diagnosis equations (5) it is seen that the partial derivative information is obtained at virtually no computational cost above that required for the evaluation of the equations. In a similar vain one can reduce the computation required to compute the inverses at different complex frequencies by integrating the differential equation

$$\begin{aligned}
d\,[1 - Z(s,r)L_{11}]^{-1}/ds = &[1 - Z(s,r)L_{11}]^{-1}\\
&\cdot [dZ(s,r)/ds]L_{11}\,[1 - Z(s,r)L_{11}]^{-1}
\end{aligned} \tag{7}$$

using the inverse computed at one particular frequency as a starting point [2, 14]. Although of extremely high dimension this equation is easily integrated without the requirement for matrix inversions. With the aid of these observations it is possibe to carry out an entire iteration of a Newton–Raphson algorithm for the solution of the multifrequency fault diagnosis equations with the aid of only a single matrix inversion.

Although one does not have a "neat" set of equations such as those described above for the solution of the fault diagnosis problem in a nonlinear circuit or system, surprisingly similar computational techniques can be invoked in the nonlinear case. The key to these techniques is the replacement of the multifrequency information of the linear case by a family of integral performance measures on the test signals u and y. These play exactly the same role in nonlinear fault diagnosis as played by the frequency information in the linear case, allowing one to formulate multiple independent fault diagnosis equations from the same test signals.

In the nonlinear case, the sparse tableau algorithm [3, 12] is used to evaluate the fault diagnosis equations at each iteration of a Newton–Raphson algorithm. As in the linear case this algorithm allows one to compute the derivative required for the Newton–Raphson algorithm with essentially no additional computational cost above that required for the evaluation of the equations [3, 4, 12]. It is possible to obtain significant computational gains in the solution of the fault diagnosis equations in the nonlinear case as well as in the linear case, by optimally exploiting the computational efficiencies inherent in the sparse tableau formulation for an electronic circuit or system.

Even using computational efficiencies which are possible for solving the fault diagnosis equations, this method is still long and tedious and not well suited to on-line implementation in a BIT system. It is thus recommended that linearization of the fault diagnosis equations be used instead. Although far less accurate than the solution of the full set of fault diagnosis equations [12], we believe that in the context of the previously described BIT architecture, linearization of the fault diagnosis equations will prove to be viable. From the point of view of fault prediction one is interested only in tracking the unknown parameter vector r from its nominal value to its tolerence limit (a few percent deviation from nominal). This is a region in which the solution of the linearized fault diagnosis equations

should be quite accurate. On the other hand, if a catastrophic fault occurs, solution of the linearized equations will detect the fault though it may fail to accurately diagnosis it. In this case, however, the linearized test data and its associated BITE may be employed as an interface between the SRA and an external test stand. As such, the use of linearized fault diagnosis equations will suffice in the context of our BIT architecture.

From the point of view of on-line analysis in a BIT system the solution of the linearized fault diagnosis equations is computationally reasonable. Since the linearization is done about the nominal value, it may be precomputed [via (6) in the linear case and the corresponding equation in the nonlinear case] and its inverse may be precomputed. Thus, the implementation of an algorithm for the solution of the linearized fault diagnosis equations requires only a single matrix multiplication, the matrix having been precomputed off-line and stored in a read-only memory.

IV. Fault Prediction

In the context of the previously described BIT architecture the primary role of the central microprocessor is to periodially collect data from the individual SRA's characterizing their internal parameter vectors r. This data is then used to detect and predict failures of the SRA. When a failure is detected, the central microprocessor signals this fact and the SRA is replaced and/or taken to an external test stand for repair. If no failure is detected, the role of the central microprocessor is to compare the present data with previously measured values in an endeavor to predict whether or not failure is imminent. In this instance predicted failure of the SRA would be signaled in an effort to replace the device before its actual on-line failure [20].

For any particular device one can collect statistical data on which to base a fault prediction algorithm. However, in a practical BIT setting where the same fault prediction algorithm is multiplexed through the testing of many different SRA's, it is necessary to use an algorithm which is independent of the specific properties of the parameter under test. As such, for our BIT system, we expect to employ a curve fitting algorithm [20]. Although less accurate than a statistically based algorithm, we have shown by simulation [18, 19] that such an algorithm can be employed as a satisfactory fault predictor. Such algorithms are computationally simple thus permitting a single central microprocessor to be multiplexed through the testing of a large number of SRA's [15]. Moreover, if one assumes that the data delivered by the SRA to the central microprocessor has been uniformly normalized, the fault prediction algorithm will be completely independent of the parameter under test. As such, one is in a position to completely standardize the central microprocessor in a BIT system so that changes in an SRA do not demand corresponding changes in the fault prediction algorithm.

Over the past several years we have investigated several approaches to the fault prediction problem [5, 15, 16, 18, 19, 20]. The first is extremely naive but has yielded surprisingly effective results in simulation [18, 19, 20]. Basically, one collects data at periodic intervals, fits the data with a second-order polynomial, and solves the quadradic equation to estimate the time at which the parameter will go out of tolerance. The success of this algorithm is due to the fact that one is not really interested in the accuracy of the failure time estimate but only the accuracy of the binary decision (based on this estimate) whether or not to replace the SRA. Moreover, this binary decision is only made when failure is expected in the near future, a region of time in which a polynomial extrapolation is reasonably accurate; i.e., if failure is estimated to take place in 3 years even if the estimate is off by 90 percent, the decision not to replace the SRA at this time will still be correct.

A fault prediction algorithm based on the above mentioned second-order polynominal extrapolation has been extensively studied by Tung and Saeks on some 10 000 complete simulated operations of the algorithm [18, 19]. Most of these simulations were carried out on artificial data generated by a library of special functions to which a noise term was added. These special functions included some highly complex nonmonotonic curves. Additionally, curves based on the empirical drift formula for thin film resistors were studied [$R(t) = At^a$ where a lies between 0.3 and 0.5] [18, 19]. In both cases, random noise with amplitudes of up to 25 percent of the tolerance interval was added to the data. The result of these simulations, which we believe to represent an environment which is more extreme than the real world, was that 99.5 percent of all SRA's were replaced before on-line failure at a cost of about 10 percent of their lifetime.

At the present time a somewhat more sophisticated fault prediction algorithm is under development [5]. This is still essentially a curve fitting algorithm though one in which a failure model (founded in modern reliability theory [1]) is employed. The basic idea for this algorithm is as follows. The drifting SRA parameter r is assumed to satisfy a difference equation

$$r(k+1) = r(k) + f(k) \tag{8}$$

where the "component time" k represents the number of shocks the SRA has received (e.g., switching processes, electrons boiling off a cathode, etc.). The relation between component time, i.e., the number of shocks received, and real time is assumed to be a Poisson-distributed random variable in which the probability of the SRA receiving n shocks in a time interval of length t is

$$P_n(t) = (ct)^n e^{-ct} / \underline{|n} \tag{9}$$

It is assumed that the value of the parameter r is known for a fixed set of points in real time; $r(t_1), r(t_2), \ldots r(t_m)$.

Using this data we desire to estimate the unknown failure dynamics $f(k)$ for the SRA parameter. This is then used in (8) to compute the number of shocks required to cause failure, i.e., the smallest value of k for which $r(k)$ is out of tolerance. Finally, this estimate is used to compute the optimal real time at which to replace the SRA to minimize the cost functional

$$J = c_f P_f + c_w W. \tag{10}$$

Here P_f is the probability of on-line failure, W is the average percentage of SRA lifetime which is wasted by replacing the SRA before its actual failure, and c_f and c_w are weighting factors.

Note that the implementation of the Poisson-shock-based fault prediction algorithm just described requires that we deal simultaneously with two unknown phenomena: the failure dynamics $f(k)$, and the random relationship between "real time" and "component time" given by the Poisson distribution. Although the required analysis is complex, a surprisingly tractable (and optimal in an appropriate sense) fault prediction algorithm can be formulated. The properties of the Poisson distribution are used to estimate the number of shocks which the SRA has received in the time intervals $[t_i, t_{i-1}]$, $i=1,2,\ldots,m$, in combination with a generalized inverse algorithm to estimate $f(k)$. $f(k)$ is approximated by a jth-order polynomial and one must compute the generalized inverse of an m by j matrix. Fortunately, the algorithm is ideally suited to a sequential least squares technique [8] and no matrix inversions need be carried out on-line. Once $f(k)$ has been estimated to a satisfactory level of accuracy (by increasing the order of the approximating polynomial until the estimation error is reduced to a prescribed level), it is used with (8) to compute the number of shocks required for the parameter to go out of tolerance. Finally, this value is used in conjunction with the Poisson distribution to determine the optimal real time at which to replace the SRA. Although apparently complex, this latter optimization can be reduced by analytic techniques to the solution of a single nonlinear equation in one variable [5]. As such, the entire fault prediction algorithm may be easily implemented, on-line, in a BIT system. Unlike the second-order curve fitting algorithm, the Poisson shock algorithm for fault prediction is still under development and its simulation on real world data is just beginning.

From the point of view of our BIT architecture where the central microprocessor is dedicated exclusively to the fault prediction job (plus bookkeeping and control of the test communications link), we anticipate little difficulty in implementing either of our fault prediction algorithms. The key to the viability of the concept, however, is to make the algorithm fast enough so that a single central microprocessor can be time-multiplexed to test a large number of parameters. In an effort to verify the feasibility of such an approach, we are presently in the process of implementing the second-order curve fitting algorithm on an F8 microprocessor [15]. Although the implementation has yet to be completed, most of the subprograms have been written and tested and it appears that the program will require about 500 bytes of memory and execute in about 30 ms. As such, the central microprocessor would be able to cycle through the testing of about 2000 parameters at 1-min intervals.

V. Self-Testing

An interesting side effect of running a BIT system in a predictive mode is that it opens the possibility of reliable self-testing. The key observation is that to do fault prediction in a digital device one must test analog parameters such as rise time, power supply voltage, clock rate, pulsewidths, etc., since digital parameters are either right or wrong and have no gray region from which to extrapolate trends. One may therefore use a microprocessor to predict its own failure by extrapolating the values of its analog parameters. As long as the prediction is made before these parameters go out of tolerance, the digital performance of the microprocessor is still exact. The point is that in a predictive mode the microprocessor is still working at the time it predicts its own failure and hence may be used reliably in a self-testing mode. Of course, once the analog parameters of the microprocessor have exceeded their tolerance limits, it may no longer be trusted as a digital signal processor and hence the device cannot be used to diagnose its own faults after failure.

Although the self-testing concept just described is purely conceptual and has yet to be implemented or even simulated, it is indicative of the potential of fault prediction in a BIT system. Indeed, if one can reliably predict failure before it actually takes place, such concepts as self-repair move into the realm of feasibility, since at the time a replacement decision is made the device under test is still working.

VI. Conclusions

Our purpose in the preceding has been to outline an approach for designing a BIT system applicable to electronic circuits and systems. Although not yet implemented in hardware, each of the constituent parts of the BIT system has been extensively simulated and we believe that a hardware implementation is feasible both computationally and economically. At the present time, we are implementing the polynomial curve fitting algorithm for fault prediction in hardware and are in the preliminary stages of implementing the entire system in a high-voltage power supply.

References

[1] R.E. Barlow and F. Prochan, *Statistical Theory of Reliability and Life Testing: Probability Methods.* New York: Holt, Rinehard, and Winston, 1975.

[2] K.-S. Chao and R. Saeks, "Continuation methods in circuit analysis," *IEEE Proc.*, vol. 65, pp. 1187-1194, 1977.

[3] G.D. Hachtel, R.K. Brayton, and F.G. Gustavson, "The sparce tableau approach to network analysis and design," *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 101-113, 1971.

[4] G.D. Hachtel and R.A. Rohrer, "Techniques for the optimal design synthesis of switching circuits," *IEEE Proc.*, vol. 55, pp. 1864-1877, 1967.

[5] K.-S. Lu, Ph.D. Thesis, Texas Tech Univ., Lubbock, Tex., 1977.

[6] M.N. Ranson and R. Saeks, "The connection function – Theory and application," *Int. J. Circuit Theory Its Application*, vol. 3, pp. 5-21, 1975.

[7] ———, "A functional approach to fault analysis," in *Rational Fault Analysis.* New York: Marcel Dekker, 1977, pp. 124-134.

[8] C.R. Rao and S.K. Mitra, *Generalized Inverse of Matrices and Its Applications.* New York: Wiley, 1971.

[9] R. Saeks et al., "Fault analysis in electronic circuits and systems," Texas Tech Univ., Lubbock, Tex., Tech. Rept., Nov. 1975.

[10] R. Saeks, "An experiment in fault prediction II," *Proc. 1976 IEEE AUTOTESTCON*, Arlington, Tex., Nov. 1976, p. 53 (abstract only).

[11] R. Saeks, S.P. Singh, and R.-W. Liu, "Fault isolation via components simulation," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 634-640, 1972.

[12] R. Saeks and R.A. DeCarlo, *Interconnected Dynamical Systems.* New York: Marcel Dekker, to be published.

[13] R. Saeks, G. Wise, and K.-S. Chao, "Analysis and design of interconnected dynamical systems," in *Large-Scale Dynamical Systems.* N. Hollywood: Point Lobos Press, 1976, pp. 59-79.

[14] R. Saeks and K.-S. Chao, "Continuations approach to large change sensitivity analysis," *IEEE (London) J. Electron. Circuits Syst.*, vol. 1, pp. 11-16, 1976.

[15] S. Sangani, Unpublished Notes, Univ. Detroit, Detroit, Mich., 1976.

[16] N. Sen, M.S. Thesis, Texas Tech Univ., Lubbock, Tex., 1975.

[17] Y.-L. Tsui, M.S. Thesis, Texas Tech Univ., Lubbock, Tex., 1975.

[18] L. Tung, M.S. Thesis, Texas Tech Univ., Lubbock, Tex., 1975.

[19] L. Tung and R. Saeks, "An experiment in fault prediction," *Proc. 4th Symp. Reliability in Electronics*, Budapest, pp. 249-257, Oct. 1977.

[20] L. Tung, S.R. Liberty, and R. Saeks, "Fault prediction - Towards a mathematical theory," in *Rational Fault Analysis.* New York: Marcel Dekker, 1977, pp. 135-142.

Richard Saeks (S'59–M'65–SM'74–F'77) was born in Chicago, Ill., 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, Ill., Colorado State University, Fort Collins, and Cornell University, Ithaca, N.Y., respectively, all in electrical engineering.

He is presently Professor of Electrical Engineering and Mathematics at Texas Tech University, Lubbock, Tex., 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.

12. Reprint of "Nonlinear Observers and Fault Analysis" by P.D. Olivier and R. Saeks from the Proceedings of the 22nd Midwest Symposium on Circuits and Systems, University of Pennsylvania, Philadelphia, June 1979, pp. 535-536.

Abstract

A fault analysis algorithm appropriate for time varying and nonlinear systems, is developed. The algorithm essentially constructs an observer for a nonlinear system that is intimately related to the system under test.

INTRODUCTION

Given enough time and computing capability brute force searches will identify possible fault sets. The real problem in fault analysis is to construct algorithms that, in some sense, locate the fault sets "efficiently". "Efficiently", in this context, means that the fault isolation must be done relatively quickly and with limited on site computing. Such techniques have been developed to handle linear time invariant and digital systems.[2,3,4] These, however, make heavy use of the defining properties for these systems, and don't generalize. The purpose of this paper is to show that an observer for an appropriate nonlinear differential equation can be utilized, on line, to determine the values of the system parameters. A technique, based on optimal control theory, for constructing such observers is also presented.

OBSERVERS AND FAULT ANALYSIS

Consider testing a system that is described by the nonlinear state equations

$$\dot{x}_1 = f(x_1,a,u,t)$$

$$y = g(x_1)$$

where x_1 is the dynamical state vector, a is the vector of parameters to be estimated (they are assumed constant over the test time), and u is the input used in the test procedure. If we want to estimate a we need to include it in the state vector, i.e., we want to build an observer for the augmented differential equation

$$\begin{bmatrix} \dot{x}_1 \\ \dot{a} \end{bmatrix} = \begin{bmatrix} f(x_1,a,u,t) \\ 0 \end{bmatrix}$$

$$y = g(x_1).$$

If it is possible to build an observer that will observe the subvector a we have solved the fault analysis problem. It would then be necessary to justify our solution in terms of time and computation requirements.

AN OBSERVER DESIGN

We chose to design an observer with the following structure

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{a}} \end{bmatrix} = \begin{bmatrix} f(\hat{x},\hat{a},u,t) \\ 0 \end{bmatrix} + H(\hat{y} - y) \text{ (H time invariant)}$$

$$\hat{y} = g(\hat{x}_1).$$

we term such an observer as a Model reference

linear time invariant observer. The term "linear time invariant" is used because the residuals enter in a linear time invariant fashion. The problem is now that of choosing H. To avoid involved stability considerations (at least initially) we choose H so that it minimizes the following function

$$J(H) = \int_{t_0}^{t_1} [(x_1-x)^2+(a-a)^2]\ dt$$

and hope that the stability takes care of itself. The construction of H can now be done by solving the following optimal control problem[1]

$$\min_{H\ R^2} J(H) \quad \left(J(H) = \int_{t_0}^{t_1} X^T Q X dt, \quad Q = \begin{bmatrix} I & -I \\ -I & I \end{bmatrix} \right)$$

subject to the Differential equations constraints

$$\dot{X} = \begin{bmatrix} x_1 \\ a \\ \vdots \\ \dot{x}_1 \\ \vdots \\ \dot{a} \end{bmatrix} = \begin{bmatrix} f(x_1,a,u,t) \\ 0 \\ f(\hat{x}_1,\hat{a},u,t) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ H \\ \end{bmatrix} (\hat{y}-y)$$

$$X(t_0) = [x_1(t_0)^T, a(t_0)^T, \hat{x}_1(t_0)^T, \hat{a}(t_0)^T]^T$$

Note. 1) H will be dependent on the $X(t_0)$ used in its construction, so when it is used to estimate a (when a differs from $a(t_0)$) it has little chance of being the optimal H. So even though we use optimization techniques to construct H, it will not, in general, be optimal. 2) Several observers may need to be constructed, each one convergent for a in a different region.

Experience indicates that only a few components fail at a time. Because of this a reasonable approach is to construct an observer for each component, (thereby minimizing the dimension of the augmented state vector) and estimate the parameters for each component in parallel. Observers can also be built for the common two and three element faults.

References

1. D. E. Kirk, Optimal Control Theory, Prentice Hall, Englewood Cliff's, New Jersey, 1970.
2. R. Saeks, and S. R. Liberty, eds., Rational Fault Analysis, Marcel Dekker, Inc., New York, 1977.
3. R. Saeks, N. Sen, H.M.S. Chen, K. S. Lu, S. Sangani, and R. A. DeCarlo, "Fault Analysis in Electronic Circuits & Systems II." Technical Report, Texas Tech University, 1978.
4. N. Sen, M.S. Thesis, Texas Tech University, Lubbock, Tex., 1975.

13. Reprint of "On Large Nonlinear Perturbations of Linear Systems" by P.D. Olivier and R. Saeks from the Proceedings of the 12th Asilomar Conference on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1978, pp. 473-477.

Abstract

This paper generalizes the classical Householder's formula to certain nonlinear operators. This class of nonlinear operators is shown to be common in circuit theory. Several examples are provided that show where these operators occur and the result is applied.

1. Introduction

The purpose of this paper is to present a technique for analyzing lumped analog systems with some linear and some nonlinear elements. It is shown that such a system is described by an operator of the form

(1.1) $B+Y\circ D$

where B and D are linear and Y is nonlinear. Since no assumptions are made about the nature of the nonlinearities, it is impossible to view the operator $Y\circ D$ as small in any sense, hence, $Y\circ D$ has to be viewed as a large nonlinear perturbation of the linear operator B.

The technique to be presented is based on a theorem that allows us to invert (1.1) in two steps. First, invert the linear operator B. If there are N_L linear elements and N_N nonlinear elements, B will be an $(N_L+N_N)\times(N_L+N_N)$ matrix. Second, invert a nonlinear operator of rank N_N. That such a result exists, is not surprising. Those experienced in solving equations involving such operators apply Gaussian elimination until there are N_N nonlinear equations in N_N unknowns. Another way to see that this segregation can be accomplished is to view the nonlinear elements as a "load" on an appropriate linear circuit, in much the same way as a circuit with one nonlinear element is analyzed by viewing that element as the load and finding the Thevenin's Equivalent circuit that it sees.

The main result of this paper is obtained by generalizing to operators of the form $B+Y\circ D$, a classical theorem concerning linear operators known as Householder's Formula. This classical result and its generalization are stated and proven in section 2. In section 3, we show such operators do, indeed, occur in circuit theory and then two examples are presented. The results are summarized in section 4.

Section 2

The classical Householder's formula [1] provides a means of calculating the inverse of the matrix B+CD in terms of B^{-1} and $(I+DB^{-1}C)^{-1}$ If B^{-1} is known and if the dimensions of C and D are appropriate, then a great savings in time and effort can be realized using this technique.

Theorem 1: (Classical Householder's Formula) If B is an NxN matrix, C is an NxP matrix, and D is a PxN matrix, then

$$(B+CD)^{-1}=B^{-1}-B^{-1}C(I+DB^{-1}C)^{-1}DB^{-1}.$$

In the nonlinear extension, the linear operator C is replaced by the nonlinear operator Y. The proof of this extension looks, at first glance, like the proof of a linear rather than a nonlinear theorem. To see that this is indeed a nonlinear result, the differences between the nonlinear and linear operator algebra will be reviewed by giving two basic definitions.

Definition 1: (Operator Addition) Let f and g be two operators (linear or nonlinear) with the same domain, then the operator f+g is defined by the following

$$(2.1)\quad (f+g)(x)\triangleq f(x)+g(x)$$

Definition 2: (Linearity) an operator f is linear if for all x and y in its domain and all scalars α and β

$$(2.2)\quad f(\alpha x+\beta y)=f(\alpha x)+f(\beta y)$$
$$=\alpha f(x)+\beta f(y).$$

The argument distributes to the left for all operators, but the operator distributes to the right only for linear operators. With this distinction in mind, we are ready to state and and prove our main result which is a closed form expression for $(B+YD)^{-1}$ in terms of B^{-1} and $(I+DB^{-1}Y)^{-1}$ (of course, operator multiplication is to be interpreted as composition i.e. $YD\triangleq Y\circ D$.

Theorem 2: If i)B and D are linear operators, ii)B^{-1} exists, iii)y is an arbitrary operator and iv)B+YD is defined, then

$$(2.3)\quad (B+YD)^{-1}=B^{-1}-B^{-1}Y(I+DBY)^{-1}DB^{-1}.$$

Proof: Consider the operator X+XYX where X is linear and Y is possibly nonlinear.

$$X+XYX=X(I+YX)=(I+XY)X.$$

If (I+YX) and (I+XY) are both invertible (it can be shown [2] that one is invertible if and only if the other is) we have

$$(2.4)\quad (I+XY)^{-1}X=X(I+YX)^{-1}.$$

Now consider the identity

$$I=(I+YX)(I+YX)^{-1}=I(I+YX)^{-1}+YX(I+YX)^{-1}$$
$$=(I+YX)^{-1}+Y(I+XY)^{-1}X$$

where we have used (4). Solving for $(I+YX)^{-1}$ yields

$$(5)\quad (I+YX)^{-1}=I-Y(I+XY)^{-1}X.$$

Finally, consider the operator B+YD.

$$(B+YD)^{-1}=[(I+YDB^{-1})B]^{-1}=B^{-1}(I+YDB^{-1})^{-1}.$$

Letting $X=DB^{-1}$ in (5) yields

$$(B+YD)^{-1}=B^{-1}[I-Y(I+DB^{-1}Y)^{-1}DB^{-1}]$$
$$=B^{-1}-B^{-1}Y(I+DB^{-1}Y)^{-1}DB^{-1}.$$

To see how this result is useful, consider the case where B+YD is an N^{th} order nonlinear operator, D a linear operator that maps $\mathbb{R}^N\to\mathbb{R}^P, P<N$ and Y a nonlinear operator that maps $\mathbb{R}^P\to\mathbb{R}^N$. This result allows the solution of N nonlinear equation in N unknowns to be replaced by the solution of N linear equations in N unknowns and also the solution of P (recall P<N)

nonlinear equations in P unknowns. Thus, we have, via a closed form expression, ordered our equations and unknowns properly to make maximum use of linear techniques and minimum use of nonlinear techniques.

It should be noted that the proof of Theorem 2 relied on the fact that B and D were linear operators and allowed Y to be arbitrary. B and D were not assumed to be matrices and Y was not assumed to map $|R^P \to |R^N$. Any or all of the operators could be differential operators and the result would still be valid. Regardless of whether the operators are differential or functional, we have succeeded in breaking it up into a linear portion and a nonlinear portion. If there are few nonlinear components in comparison with the number of linear components, the nonlinearities can be viewed as a perturbation on the linear system.

Section 3

The purpose of this section is to show that operators of the form B+YD occur in circuit analysis problems and to apply Theorem 2 to two examples.

This type of operator arises naturally in nonlinear network analysis. Consider the Node analysis of a network with reduced incidence matrix A, [3]. Kirchoff's Laws are

(KCL) $A\underline{j}=0$,

(KVL) $\underline{v}=A^T\underline{e}$.

The branch equations might be

$$\underline{j}=G\underline{v}+\underline{j}_s-G\underline{v}_s+f(\underline{v})$$

where

$\underline{j} \triangleq$ the branch current vector;

$\underline{v} \triangleq$ the branch voltage vector;

$\underline{j}_s \triangleq$ the current source vector;

$\underline{v}_s \triangleq$ the voltage source vector;

$\underline{e} \triangleq$ the mode-to-datum voltage vector;

G is assumed to be an "invertible" matrix of differential operators, f is a nonlinear differential operator, and all branches are voltage controlled. If

$$\underline{i} \triangleq AGV_s - A\underline{j}_s$$

then Kirchoff's Laws and the branch equations can be combined to yield

$$(3.1)\quad (AGA^T)\underline{e}+Af(A^T\underline{e})=\underline{i}.$$

Letting $AGA^T=B$, and $Af(\cdot)=Y$, we see that this operator is of the desired form. The typical situation is for f to be a function of only a few ($p<n$) linear combinations of the components of $\underline{v}$ then (3.1) can be rewritten in the form

$$(3.2)\quad B+Y(CA^T)\ \underline{e}=i$$

which is precisely the type of operator that is ammenable to the results of Theorem 2.

We now apply Theorem 2 to the problem of solving two nonlinear simultaneous equations in two unknowns. Consider the following nonlinear equations

$$f(X)=z=\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}=\begin{bmatrix} -\frac{1}{2}X_1-1X_2+X_1^3 \\ \frac{1}{2}X_1+\frac{1}{2}X_2 \end{bmatrix}=\begin{bmatrix} 24 \\ 3 \end{bmatrix}.$$

In order to apply Theorem 2, f(x) must be put into the form

$$(B+YD)(X)$$

where B is an invertible matrix. One way to do this is

$$f(X)=(B+YD)x=\frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}+Y([1,0]X)$$

where

$$Y(.)=\begin{bmatrix}(.)^3-(.)\\ 0\end{bmatrix}.$$

Theorem 2 says that

$$X=B^{-1}Z-B^{-1}Y(I+DB^{-1}Y)^{-1}DB^{-1}Z=X_L-X_{NL}$$

where

$$X_L=B^{-1}Z=\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}\begin{bmatrix}24\\ 3\end{bmatrix}=\begin{bmatrix}27\\ -21\end{bmatrix},$$

and

$$\begin{aligned}X_{NL}&=B^{-1}Y(I+DB^{-1}Y)^{-1}DB^{-1}Z\\ &=B^{-1}Y(I+DB^{-1}Y)^{-1}DX_L\\ &=B^{-1}Y(I+DB^{-1}Y)^{-1}27.\end{aligned}$$

Now

$$(I+DB^{-1}Y)^{-1}27=u$$

is equivalent to

$$(I+DB^{-1}Y)u=27.$$

$$(I+DB^{-1}Y)u=u+[1,0]\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}\begin{bmatrix}u^3 & -u\\ 0\end{bmatrix}$$

$$=u+[1\ \ 1]\begin{bmatrix}u^3-u\\ 0\end{bmatrix}=u+u^3-u=u^3=27$$

which implies

$$u=3.$$

Now

$$X_{NL}=B^{-1}Y(3)=\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}\begin{bmatrix}(3)^3-(3)\\ 0\end{bmatrix}=\begin{bmatrix}24\\ -24\end{bmatrix}.$$

So

$$X=X_L-X_{NL}=\begin{bmatrix}27-24\\ -21+24\end{bmatrix}=\begin{bmatrix}3\\ 3\end{bmatrix}$$

The reason for choosing a functional example is that for large circuits or systems, the differential equations are solved numerically so at each iteration, an operator of the form B+YD must be inverted. To see that this is indeed the case, consider discretizing the differential equations obtained from the component connection model [2] of a system. The component equations are assumed (here) to be given in state form

$$\begin{aligned}\dot{X}&=f(X,a)\\ b&=g(X,a).\end{aligned}$$

where a is the vector of component inputs, b is the vector of component outputs, and x is the state vector of the components. The connection equations (KVL and KCL equations) are given by

$$\begin{bmatrix}a\\ y\end{bmatrix}=\begin{bmatrix}L_{11} & L_{12}\\ L_{21} & L_{22}\end{bmatrix}\begin{bmatrix}b\\ u\end{bmatrix}$$

where u is the vector of system inputs and y is the vector of system outputs. If we order the entries in all of the vectors correctly, we can partition the vectors in the following manner

$$a=\begin{bmatrix}a^N\\ a^L\end{bmatrix},\quad b=\begin{bmatrix}b^N\\ b^L\end{bmatrix},\quad \text{and } x=\begin{bmatrix}x^N\\ x^L\end{bmatrix}$$

where the superscript N(L) denotes entries associated with the nonlinear (linear) components. The descretized equations that the computer is to solve have the form

$$\sum_{i=0}^{r} d_i X^N_{k-i}=f^N(X^N_k,a^N_k),$$

$$\sum_{i=0}^{r} d_i X^L_{k-i}=AX^L_k+Ba^L_k,$$

$$b_k^N=g^N(X_k^N,a_k^N),$$

$$b_k^L=CX_k^L+Da_k^L,$$

$$a_k^N=L_{11}^{NN}b_k^N+L_{11}^{NL}b_k^L+L_{12}^N u_k,$$

$$a_k^L=L_{11}^{LN}b_k^N+L_{11}^{LL}b_k^L+L_{12}^L u_k,$$

$$Y_K=L_{21}^N b_k^N+L_{22}^L b_k^L+L_{22}u_k.$$

The last equation is just the output equation and is not used during the iterations. These equations can be put in the following form

$$TW_k+Y(DW_k) =$$

$$\begin{bmatrix} -Ido & 0 & 0 & 0 & -I & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A-doI & B & 0 & 0 \\ 0 & 0 & C & D & 0 & -I \\ 0 & I & 0 & 0 & -L_{11}^{NN} & -L_{11}^{NL} \\ 0 & 0 & 0 & I & -L_{11}^{NN} & -L_{11}^{LL} \end{bmatrix} \begin{bmatrix} x_k^N \\ a_k^N \\ x_k^L \\ a_k^L \\ b_k^N \\ b_k^L \end{bmatrix}$$

$$+ \begin{bmatrix} f^N(x_k^N, a_k^N) \\ g^N(x_k^N, a_k^N) \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{r} d_i x_{r-i}^N \\ 0 \\ \sum_{i=1}^{r} d_i x_{r-i}^L \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where $D=[I,0]$, and I is conformable with

$$\begin{bmatrix} x_k^N \\ a_k^N \end{bmatrix}$$

4. Conclusion

The classical Householder's Formula has been generalized to certain nonlinear operators. It was shown that these nonlinear operators occur in circuit theory, both in the differential equations that describe the circuit and in the discretized equations that are used in the computer aided analysis of these circuits. It is hoped that this result will be as useful a tool in the fault analysis of nonlinear circuits as the classical result turned out to be in the fault analysis of linear circuits.

References

1. A.S. Householder, "A Survey of Some Closed Methods for Inverting Matrices," SIAM Jour. on Appl. Math., Vol. 5, pp. 155-169, (1957).
2. Saeks, R., and R.A. DeCarlo, Interconnected Dynamical Systems, New York, Marcel Dekker, (to appear).
3. Desoer, C.A. and S.E. Kuh, Basic Circuit Theory, McGraw-Hill, New York, 1969, pp. 423-425.

14. Reprint of "CAD Oriented Measures of Testability" by R. Saeks from the Proceedings of the Industry/Joint Services Test Conference and Workshop, NSIA, San Diego, April 1978, pp. 71-72.

ABSTRACT

Measures of testability for both analog and digital systems which can be incorporated into a computer-aided design package are surveyed. The application of these measures for evaluating and improving system diagnosability is discussed.

Although maintenance related questions have historically been given low priority in system design with the advent of integrated circuit technology during the past decade the cost of maintenance has become a dominating factor in determining the system life-cycle costs. As such, considerable interest in the design of systems which are readily testable has developed along with a concomittment interest in the development of a quanitative measure of testability.[1,2] The latter may be used to aid in the design of readily testable systems. Moreover, such a measure of testability can be employed as a means of specifying system testability for acquisition purposes.

Two classes of testability measures have been proposed both of which have applicability. The first is a coarse measure of testability which can be computed by hand from an inspection of the system. This might include numbers of input and output test points, number and complexity of system components, memory complexity, etc. Alternatively, one might choose to adopt a more sophisticated measure of testability using computer aid for its evaluation. Indeed, with the growing prevalence of CAD packages in the design houses of subrouting for computing a measure of testability during the design process could be incorporated into an existing CAD package with little difficulty.

Although at the time of this writing no clear criterion for defining a measure of testability has yet to emerge several approaches are presently under investigation.[1,2] One such approach is based on the concepts of controllability and observability intuitively deeming a system to be "more testable" if it is "more controlable and observable". Since controllability and observability are measures of one's ability to exercise the internal system elements via external inputs and observations such a viewpoint is quite reasonable. Unfortunately, as classically defined, controllability and observability only measure one's ability to exercise the system memory elements and hence some type of extension of the concept is required if the resultant measure of testability is to include combinational information as well.

An alternative approach primarily intended as a measure of testability for analog circuits has recently been proposed by Sen and the author.[3,4,5] Here, one uses the implicit function theorem to estimate the dimension of the manifold of arbitrary parameters resulting from the solution of a set of "fault diagnosis equations" with a lower dimensional ambiguity set indicating a "more testable" systems. Unfortunately, this approach to the formulation of a measure of testability is heavily predicted on the differentiable properties of analog systems with considerable work still required for its extension to the digital case.

References

1. Dejka, W.J., "Measure of Testability in Device and System Design", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, Tx., Aug., 1977, pp. 39-52.

2. Dejka, W.J., "A Review of Measurements of Testability for Analog Systems", Proc. of the 1977 AUTOTESTCON, Hyannis, Mass., Nov. 1977, pp. 279-284.

3. Sen, N., M.S. Thesis, Texas Tech Univ., 1977.

4. Sen, N., and R. Saeks, "A Measure of Testability and its Application to Test Point Selection - Theory", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, Tx., Aug. 1977, pp. 576-583.

5. Sen, N. and R. Saeks, "A Measure of Testability and its Application to Test Point Selection - Computation", Proc. of the 1977 AUTOTESTCON, Hyannis, Mass., Nov. 1977, pp. 212-219.

6. Hartman, R., Unpublished Notes, Pacific Applied Systems, Corp., 1977.

Texas Tech University — Institute for Electronic Science

Joint Services Electronics Program — Research Unit: 4

1. Title of Investigation: Qualitative Analysis of Large Scale Systems
2. Senior Investigator: Kwong-shu Chao — Telephone: (806) 742-3469
3. JSEP Funds: $23,500
4. Other Funds: $18,710*
5. Total Number of Professionals: PI's 2 (2 mo.) RA's 1 (1/2 time)
6. Summary:

In the analysis of large scale systems it is often necessary to solve a continuously parameterized family of numerical problems; inversion of a family of sparse matrices, computation of the roots of a family of polynomials or nonlinear equations, solution of the eigenvalue problem for a family of sparse matrices, etc. The goal of the present work unit is to develop a class of continuation algorithms for the solution of such problems in which one formulates a nonlinear differential equation whose trajectories represent the solutions to the given family of problems as a function of the underlying parameter. One then computes the solution to the given problem at one parameter value by classical techniques and numerically integrates the differential equation using this value as an initial condition to obtain solutions to the entire family of problems. We believe that these techniques are far more efficient than the classical technique of discretizing the parameter value and applying standard numerical techniques at each point. Indeed, this has been born out by our preliminary experience in applying continuation algorithms to large scale systems problems.

* NSF grant for related work applied to the computer-aided design problem.

The research may naturally be subdivided into two areas; the formulation of analysis and design techniques for large scale systems and the development of numerical methods for their solution. The former area includes system simulation algorithms, large change sensitivity analysis algorithms, a multivariate Nyquist theory and several root locus algorithms. In the numerical area we have developed continuation algorithms for inverting sparse matrices and for the solution of the eigenvalue problem in a family of sparse matrices. Additionally, we have formulated several root locus algorithms and a method for tracking the solutions of a parameterized family of nonlinear equations.

The major result obtained during the year has been the formulation of several continuation algorithms for the solution of the eigenvalue problem in a family of sparse matrices. Although a continuation algorithm for the solution of the eigenvalue problem has been known for a number of years the existing algorithm uses the eigenvectors as auxiliary variables. As such, since the matrix of eigenvectors for a sparse matrix is non-sparse this algorithm fails to preserve the sparseness of the given matrix. We have therefore developed three alternative continuation algorithms in which the auxiliary variables take the form of appropriate similarity transformations for the given family of matrices which are assured to be sparse if the given family of matrices is sparse.

In other areas we have developed a continuation algorithm which is employed to find multiple roots of a polynomial with applications to root locus problems. The latter application is also represented by reprints of two papers in which various root locus algorithms are formu-

lated. Additionally, we have included a reprint of a paper from the IEEE Proceedings in which a continuation algorithm for sparse matrix inversion is developed and reprints of two conference papers on the solution of parameterized families of nonlinear equations.

7. Publications and Activities:

A. Refereed Journal Articles

1. Pan, C.T., and K.S. Chao, "A Computer-Aided Root-Locus Method", IEEE Trans. on Automatic Control, Vol. AC-23, pp. 856-860. (1978).

2. DeCarlo, R.A., and R. Saeks, "A Root Locus Technique for Interconnected Systems", IEEE Trans. on Systems, Man, and Cybernetics, Vol. SMC-9, pp. 53-55, (1979).

3. Saeks, R., "A Continuation Algorithm for Sparse Matrix Inversion", IEEE Proc., Vol. 67, pp. 682-683, (1979).

4. Pan, C.T., and R. Saeks, "Multiple Solutions of Nonlinear Equations: Roots of Polynomials", IEEE Trans. on Circuits and Systems (to appear).

B. Conference Papers

1. Pan, C.T., and K.S. Chao, "Multiple Solutions of a Class of Nonlinear Equations", Proc. of the 1979 IEEE Inter. Symp. on Circuits and Systems, Tokyo, July 1979, pp. 577-580.

2. Pan, C.T., and K.S. Chao, "A Continuation Method for Finding the Roots of a Polynomial", Proc. of the 22nd Midwest Symp. on Circuits and Systems, Univ. of Pennsylvania, Philadelphia, June 1979, pp. 428-431.

C. Preprints

1. Green, B., "Continuation Algorithms for the Solution of the Eigenvalue Problem", (preliminary draft).

D. Theses

1. Green, B., "Continuation Algorithms for the Solution of the Eigenvalue Problem", M.S. Thesis, Texas Tech Univ., 1979.

E. Conferences and Symposia

1. Chao, K.S., 21st Midwest Symposium on Circuits and Systems, Iowa State Univ., Aug. 1978.

2. Chao, K.S., 22nd Midwest Symposium on Circuits and Systems, Univ. of Pennsylvania, June 1979.

3. Chao, K.S., 1979 IEEE Inter, Symp. on Circuits and Systems, Tokyo, July 1979.

4. Green, B., Texas Systems Workshop, Dallas, March 1979.

5. Chao, K.S., Texas Systems Workshop, Dallas, March 1979.

6. Chao, K.S., Inter. Colloq., on Circuits and Systems, Taipei, July 1979.

8. Reprint of "A Computer-Aided Root-Locus Method", by C.T. Pan and K.S. Chao from the IEEE Transactions on Automatic Control, Vol. AC-23, pp. 856-860. (1979).

A Computer-Aided Root-Locus Method

C. T. PAN AND K. S. CHAO, MEMBER, IEEE

Abstract—An efficient computer-aided root-locus method is described. The approach is based on the concept of continuation methods in which the solution of a parameterized family of algebraic problems is converted into the solution of a differential equation. The root-locus plot is obtained in a systematic manner by numerical integration. Singularities are analyzed and classified according to the properties of higher order derivatives. Depending on their classification, singular points on the root loci are taken care of accordingly.

I. Introduction

The root-locus method is one of the important design techniques for linear time-invariant feedback systems. In addition to yielding frequency response information of the system, it also provides a powerful tool for solving problems in the time domain. The basic idea of the root-locus method is to determine the closed-loop pole configuration as a function of the gain from the configuration of the open-loop poles and zeros. A great deal of information is available in texts and literature on the method for the construction of root loci. The graphical method using certain elementary geometric properties of the locus is probably the most commonly used approach (see e.g. [1]–[3]). Other approaches [4]–[7] employ either analytic or semi-analytic representations that involve the use of equations of the loci. Although analytic approaches enable one to obtain accurate plots along with certain qualitative features of the root paths, the point-to-point plotting is just a formidable task. Besides, investigations for higher order systems are virtually impractical.

It is the purpose of this paper to develop a computer-aided method for plotting root loci in a systematic manner. The approach to be presented in Section II is based on the concept of continuation methods [8]–[10]. The basic idea is to convert the solution of a parameterized family of algebraic problems into the solution of a set of associated differential equations. Section III is concerned with the existence and classification of singular points on the root loci. In Section IV the results obtained are illustrated by means of examples.

II. The Root-Locus Method

Consider the closed-loop system shown in Fig. 1. Let the open-loop transfer function be expressed by

$$G(s)H(s)=K\frac{A(s)}{B(s)}\triangleq K\frac{(s-z_1)(s-z_2)\cdots(s-z_m)}{(s-p_1)(s-p_2)\cdots(s-p_n)} \tag{1}$$

where K is the open-loop gain and $m<n$. The closed-loop transfer function is

$$T(s)=\frac{G(s)}{1+G(s)H(s)}=\frac{G(s)B(s)}{B(s)+KA(s)}. \tag{2}$$

The root-locus plot of the closed-loop transfer function $T(s)$ is defined as the locus of the poles of $T(s)$ when K varies from zero to infinity. This plot consists of a set, denoted by l, of points in the s-plane such that

Manuscript received December 13, 1977; revised April 28, 1978. Paper recommended by E. Polak, Chairman of the Computational Methods and Discrete Systems Committee. This work was supported in part by the NSF under Grant ENG-75-09074/ENG-77-22991 and the ONR under Grant 76-C-1136.

C. T. Pan was with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409. He is now with the Department of Electrical and Power Engineering, National Tsing Hua University, Hsinchu, Taiwan.

K. S. Chao is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

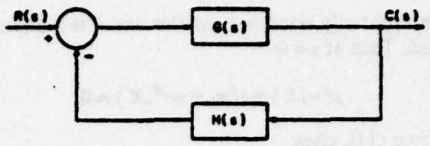


Fig. 1. A closed-loop feedback system.

$$g(s,K)\triangleq B(s)+KA(s)=0, \tag{3}$$

i.e.,

$$l=\{s\,|\,g(s,K)=0\}. \tag{4}$$

Instead of solving the roots of (3) directly for each K, a system of two simultaneous differential equations

$$\frac{d}{dt}g(s(t),K(t))=-g(s(t),K(t)),g(s(0),K(0))=0$$

$$\frac{d}{dt}K(t)=\pm 1,\qquad K(0)=K_0 \tag{5}$$

is considered where $s(0)=s_0$ is a root of [3] corresponding to an initial gain K_0, and t is a dummy variable. Application of the chain-rule to (5) results in

$$\frac{ds}{dt}=-\left(g\pm\frac{\partial g}{\partial K}\right)\Big/\frac{\partial g}{\partial s},\qquad s(0)=s_0$$

$$\frac{dK}{dt}=\pm 1 \tag{6}$$

or equivalently,

$$\frac{ds}{dt}=-\frac{(B(s)+KA(s))\pm A(s)}{B'(s)+KA'(s)},\qquad s(0)=s_0$$

$$\frac{dK}{dt}=\pm 1,\qquad K(0)=K_0. \tag{7}$$

Equation (7) can now be solved by any numerical integration technique. For example, using Euler's method, (6) reduces to

$$s_{k+1}=s_k-h\frac{B(s_k)\pm K_kA(s_k)+A(s_k)}{B'(s_k)+K_kA'(s_k)} \tag{8}$$

$$K_{k+1}=K_k\pm h.$$

It is seen from the solution of (5)

$$g(s,K)=g(s(0),K(0))e^{-t}=0e^{-t}\equiv 0$$

$$K=\pm t$$

that for any admissible pair K_0 and s_0 satisfying (3), the corresponding trajectory resulted from (5) will remain on the solution curve $g(s,K)=0$ as K changes. The + or − sign is chosen depending on whether one would like to increase or decrease K. Since the computed trajectory may not satisfy (3) exactly, the minus sign in front of g in (5) is used to ensure that the computed trajectory does not diverge away from the locus.

It is a well-known fact that the root-locus plot for $T(s)$ contains n branches starting from the open-loop poles at $K=0$. Therefore, in the case where the open-loop poles are distinct, the n initial conditions for (7) are selected at $K(0)=0$ and $s(0)=p_i$, $i=1,2,\cdots,n$. In the case where the open-loop transfer function contains repeated poles, the term $(\partial g/\partial s)$ becomes zero when evaluated at the repeated poles. As a result, the selection of starting points cannot be made at $K=0$ for the repeated poles. Approximate starting points, however, can easily be obtained by analyzing the properties of the root loci in the neighborhood of the repeated pole. Suppose the open-loop gain has a repeated pole p_i with

0018-9286/78/1000-0856$00.75

multiplicity r and the corresponding $g(s,K)$ has the form

$$g(s,K)=B(s)+KA(s)$$
$$=(s-p_1)(s-p_2)\cdots(s-p_k)^r\cdots(s-p_{n-r})+KA(s)=0. \quad (9)$$

Let

$$w=p_k+\Delta s=p_k+\epsilon e^{j\theta} \quad (10)$$

where ϵ is an arbitrarily small real number and θ is a phase angle yet to be determined. Thus at $s=w$

$$g(w,K)=g(p_k+\epsilon e^{j\theta},K)=0. \quad (11)$$

Solving Δs from (11), gives

$$\Delta s=\epsilon e^{j\theta}=(K\rho e^{j\phi})^{1/r} \quad (12)$$

where

$$\rho e^{j\phi}=-\frac{A(s)(s-p_k)^r}{B(s)}\bigg|_{s=p_k}. \quad (13)$$

Therefore r approximate starting points in the neighborhood of the repeated pole p_k and the corresponding open-loop gain can be evaluated from (12) as

$$w_i=p_k+\epsilon e^{j\left(\frac{\phi+2\pi i}{r}\right)}, \qquad i=0,1,2,\cdots,r-1$$

and

$$K=\frac{1}{\rho}\epsilon^r. \quad (14)$$

With the proper choice of n starting points, the n branches of the root-locus plot can be traced in a continuous manner by numerical integration. In computing the root locus, care must be exercised when approaching a singular point on the locus.

III. Singular Points

A point s^* satisfying

$$\frac{dK}{ds}\bigg|_{s=s^*}=\frac{d}{ds}\left(-\frac{B(s)}{A(s)}\right)\bigg|_{s=s^*}$$
$$=-\frac{A(s^*)B'(s^*)-B(s^*)A'(s^*)}{A^2(s^*)}=0 \quad (15)$$

in the complex plane is called a singular point. Since the numerator of (15) is a $(n+m-1)$th order polynomial of real coefficients, there are $(n+m-1)$ singular points in the s-plane. Only those singular points that are located on the root loci will be considered. In view of the fact that $A(s^*)$ cannot be zero for a finite K, it follows that on the root locus, the condition

$$A(s^*)B'(s^*)-B(s^*)A'(s^*)=A(s^*)(B'(s^*)+KA'(s^*))=0 \quad (16)$$

implies

$$\frac{\partial g}{\partial s}\bigg|_{s=s^*}=B'(s^*)+KA'(s^*)=0. \quad (17)$$

Hence, (7) is not valid at $s=s^*$ and modifications must be made to handle these singular points. Condition (15) does include the conventional break-in and break-away points at which K is either a local maxima or a local minimum on the real axis, respectively. In general, (dK/ds) is a complex quantity if s is not located on the real axis and it does not make sense to talk about local extremal values without proper modification. Now, since K is a real-valued function of s for all s on the root locus l, the directional derivative (dK/dl) together with its higher order derivatives along the tangential direction of the locus are well defined. It is thus possible to consider local extremal values of K along l using the notion of directional derivatives. Let

$$K(s)=-\frac{B(s)}{A(s)}\triangleq U(s,y)+jV(x,y) \quad (18)$$

where U and V are real-valued functions of x and y, and

$$s=x+jy. \quad (19)$$

The directional derivative of K at a point $s\in l$ in the unit direction $v=e^{j\theta}=v_x+jv_y$ tangential to the root locus is

$$\frac{dK}{dl}=\nabla U(x,y)\cdot v=\frac{\partial U}{\partial x}v_x+\frac{\partial U}{\partial y}v_y. \quad (20)$$

The above equation can also be written in the form

$$\frac{dK}{dl}=\mathrm{Re}\left[\left(\frac{\partial U}{\partial x}-j\frac{\partial U}{\partial y}\right)(v_x+jV_y)\right], \qquad s\in l. \quad (21)$$

Making use of the Cauchy–Riemann condition, (21) reduces to

$$\frac{dK}{dl}=\mathrm{Re}\left[\left(\frac{\partial U}{\partial x}+j\frac{\partial V}{\partial x}\right)e^{j\theta}\right]=\mathrm{Re}[K'(s)e^{j\theta}] \quad (22)$$

where $K'=dK/ds$.

Similarly, higher order directional derivatives of K with respect to l are related to the higher order derivatives by

$$\frac{d^mK}{dl^m}=\mathrm{Re}[K^{(m)}(s)e^{jm\theta}]. \quad (23)$$

Thus, it is seen from (23) that along l, K and its directional derivatives are all real-valued functions. The following theorem which plays an important role in singularity classification will be proved.

Theorem: Suppose $p(s)$ is an analytic function such that

$$p(s^*)=a, \qquad \text{for a real } a\neq 0,$$
$$p^{(k)}(s^*)=0, \qquad k=1,2,\cdots,q-1,$$

and

$$p^{(q)}(s^*)\neq 0, \qquad 2<q<n$$

at some point s^* located on $\mathrm{Im}p(s)=0$. Let

$$R_p=\{s|\mathrm{Im}p(s)=0\}.$$

Then, in the neighborhood of s^*, R_p consists of q branches, $R_{p1}, R_{p2}, \cdots, R_{pq}$, and $R_{p1}\cap R_{p2}\cap\cdots\cap R_{pq}=s^*$. Furthermore, for each i, $1<i<q$, $\mathrm{Re}p(s)|_{R_{pi}}$ is either a local maximum or a local minimum at s^* if q is even; it is either an increasing function or a decreasing function if q is odd.

Proof: Without loss of generality s^* can be assumed to be zero. Before considering the general case, the theorem is proved for

$$h(s)=a+s^q.$$

Identifying the set R_h from

$$R_h=\{s|\mathrm{Im}h(s)=0\}$$

yields

$$R_h=\{re^{j\theta}|r^q\sin q\theta=0\}=\{re^{j\theta}|\theta=i\pi/q, \qquad i=0,1,2,\cdots,q-1\}$$

where θ is restricted in the upper half of the s-plane and r assumes negative values in the lower-half plane. Thus, R_h consists of q intersecting branches R_{hi}, $i=0,1,2,\cdots,q-1$. The intersection occurs at $r=0$. For each i,

$$\mathrm{Re}h(s)|_{R_{hi}}=a+r^q\cos(q\theta_i)=a+r^q\cos(i\pi).$$

Therefore if q is even, r^q is always nonnegative, and

$$\mathrm{Re}h(s)|_{R_{hi}}\gtrless a \quad \begin{matrix}\text{when } \cos i\pi=1\\ \text{when } \cos i\pi=-1,\end{matrix}$$

i.e., $h(s)|_{R_{hi}}$ is either a local maximum or a local minimum. Now if q is odd, then for each i,

$$\mathrm{Re}h(s)|_{R_{hi}}=a+r^q\cos(i\pi).$$

Hence, $(h(s)|_{R_{hi}}-a)$ will change sign either from plus to minus as r

increases from negative value to positive value or vice versa, i.e., $h(s)|_{R_\omega}$ is a monotonic function.

Returning to the general case, $p(s)$ can be expanded into a Taylor series around s° as

$$p(s)=p(s^\circ)+\sum_{i=q}^{\infty} c_i(s-s^\circ)^i=a+(s-s^\circ)^q\sum_{k=0}^{\infty} c_{q+k}(s-s^\circ)^k.$$

Since the summation in the above equation is an analytic function and has no zero in a small disc around s°, from a theorem in the theory of complex variables (see e.g. [11]), there exists an analytic function $u(s)$ such that

$$\sum_{k=0}^{\infty} c_{k+q}(s-s^\circ)^k=\exp(u(s)).$$

Let $v(s)=\exp(u(s)/q)$. Then

$$p(s)=a+[(s-s^\circ)v(s)]^q \triangleq a+[f(s)]^q$$

where $f(s^\circ)=0$, $f'(s^\circ)\neq 0$. Thus, $f(s)$ is a local homeomorphism. Now

$$R_p=\{s|\operatorname{Im}(a+(f(s))^q)=0\}=\{s|f(s)\in R_h\}.$$

Let

$$R_{p_i}=\{s|f(s)\in R_{h_i}\},\qquad i=0,1,2,\cdots,q-1.$$

If q is even, then $s\in R_{p_i}$ implies $f(s)\in R_{h_i}$. It follows that

$$\operatorname{Re}h(f(s))\gtrless \operatorname{Re}h(0)=\operatorname{Re}h(f(s^\circ)),\quad \begin{matrix}\text{when } \cos i\pi=1\\ \text{when } \cos i\pi=-1,\end{matrix}$$

i.e.,

$$\operatorname{Re}p(s)|_{R_{p_i}}\gtrless \operatorname{Re}p(s^\circ),\quad \begin{matrix}\text{when } \cos i\pi=1\\ \text{when } \cos i\pi=-1.\end{matrix}$$

Therefore $\operatorname{Re}p(s)|_{R_{p_i}}$ is either a local maximum or a local minimum at s°. Similarly, if q is odd, then it can be shown that $\operatorname{Re}p(s)|_{R_{p_i}}$ is either an increasing function or a decreasing function.

As a direct consequence of the above theorem, the following corollary is deduced.

Corollary: Suppose s° is a singular point on l such that

$$K(s^\circ)\neq 0$$

$$\left.\frac{d^kK}{ds^k}\right|_{s=s^\circ}=0,\qquad k=1,2,\cdots,q-1$$

$$\left.\frac{d^qK}{ds^q}\right|_{s=s^\circ}\neq 0,\qquad 2<q<n$$

where $K(s)=-B(s)/A(s)$, then there are q branches intersecting at $s=s^\circ$. Furthermore, if q is even, then along each branch of the intersecting root loci at $s=s^\circ$, $K(s^\circ)$ is either a local maximum or a local minimum; otherwise $K(s^\circ)$ is a monotonic function of s on that branch in the neighborhood of s°.

Proof: Let $f(s)=-B(s)/A(s)$. Then $f(s)$ is an analytic function. It follows that

$$B(s)+f(s)A(s)=0.$$

Comparing the above equation to (3), it is obvious that

$$K=\operatorname{Re}f(s)$$

$$0=\operatorname{Im}f(s),\qquad \text{for all } s\in l.$$

Application of the above theorem to $f(s)$ completes the proof.

According to the above corollary, singular points are characterized by the properties of higher order derivatives. It is noted that

$$\frac{d^kK}{ds^k}=\frac{d^k}{ds^k}[-B(s)/A(s)]=-\frac{\partial^k g}{\partial s^k}/A(s). \tag{24}$$

Since $g(s)$ is an nth-order polynomial with real coefficients, $\partial^k g/\partial s^k$ can easily be generated. Furthermore, q can at most be equal to n, since

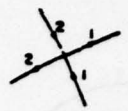


Fig. 2. An even singular point.

Fig. 3. An odd singular point.

$$\frac{d^nK}{ds^n}=-\frac{n!}{A(s)}\neq 0. \tag{25}$$

With the above corollary, the conventional break-in and break-away points defined on the real axis can now be generalized as follows.

Definition: In the theorem, the singular point is called an even singular point if q is even; otherwise, it is said to be an odd singular point.

It is clear from the corollary that on the root locus an even singular point is either a local maximum or a local minimum along the branch as defined in the theorem. The conventional break-in and break-away points are just special cases of even singular points of order 2 (i.e., $q=2$) which are located on the real axis. In general, if a singular point is located off the real axis, the concept of directional derivatives can be used to characterize the singular point. From (23) and the corollary it is obvious that at an qth-order singular point, $d^kK/d^kl=0$ for $k=1,2,\cdots,q-1$ and $d^qK/dl^q\neq 0$. There are q branches of the root loci intersecting at the singular points.

In generating the root-locus plot, each branch is plotted separately as K increases. In the neighborhood of an odd singular point, since K is a monotonic function of s on the locus, the first-order directional derivative is a continuous function. Thus, when approaching an odd singular point, it is necessary to jump over the singular point by adding a small variation $|\Delta s|$ along the tangential direction of the locus. For even singular points, such procedure is invalid. Since on the root-locus plot an even singular point is either a local maximum or a local minimum, such construction does not give rise to an increasing K. Thus, in order to continue the plotting of the root locus as a function of increasing K, it is necessary to change the direction of the locus when an even singular point is approached. Depending on the order q of the even singular point, Δz, the change in direction, is chosen as

$$\Delta z=\Delta s e^{-j\pi/q} \tag{26}$$

where Δs is a sufficiently small vector in the tangential direction of the locus when approaching the singular point. The factor $e^{-j\pi/q}$ can be viewed as a rotational operator, which rotates the direction clockwise by π/q. On a branch so constructed, the even singular point no longer has the characteristics of a local extremum.

It is apparent from the foregoing root-locus construction that the q branches will directly intersect each other at an odd singular point and that the modified q root loci will touch each other at an even singular point. For obvious reasons, an odd singular point is known as an intersecting point while an even singular point is called a touching point. The graphical illustrations for these two types of singularities are shown in Fig. 2 and Fig. 3 for $q=2$ and $q=3$, respectively. The branches are numbered and the arrows are pointed in the direction of increasing K.

After the change of direction at a touching point or the jump over an intersecting point, it may be necessary to make corrections if the point selected is not close enough to the locus. This can usually be achieved within a few steps by using the Newton iteration

$$s_{k+1}=s_k-\frac{B(s_k)+KA(s_k)}{B'(s_k)+KA'(s_k)}. \tag{27}$$

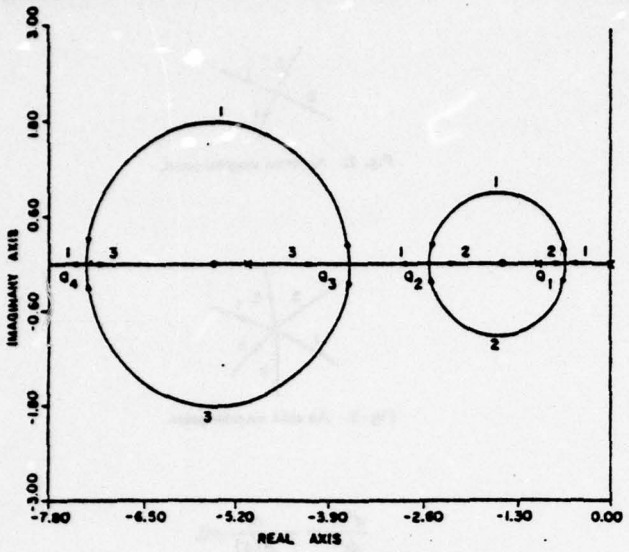


Fig. 4. $G(s)H(s)=K\frac{(s+1.5)(s+5.5)}{s(s+1)(s+5)}$.

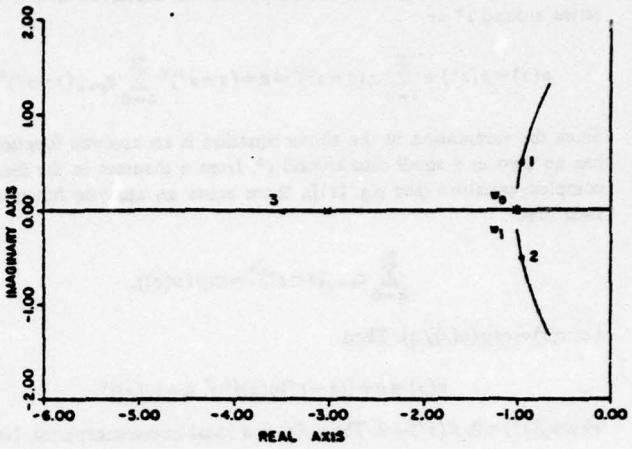


Fig. 5. $G(s)H(s)=\frac{K}{(s+3)(s+1)^2}$.

The gain K which corresponds to the corrected point can be evaluated either from

$$K=-\mathrm{Re}\,B(s)/\mathrm{Re}\,A(s) \tag{28}$$

or

$$K=-\mathrm{Im}\,B(s)/\mathrm{Im}\,A(s). \tag{29}$$

Equations (28) and (29) are derived by taking the real and the imaginary parts of $g(s)=0$, respectively.

IV. Examples

In this section a number of examples are presented to illustrate the proposed algorithm for obtaining the root-locus plot. Although any integration technique can be used to solve (7), only the Euler's method with variable step size is used for illustration.

Example 1: Consider a linear feedback system whose open-loop transfer function is given by

$$G(s)H(s)=K\frac{(s+1.5)(s+5.5)}{s(s+1)(s+5)}.$$

There are three simple poles at 0, -1, and -5, and two finite zeros at -1.5 and -5.5. Application of (7) with starting points 0, -1, and -5 at $K=0$ leads to three root loci shown in Fig. 4. It turns out that all four singular points are located on the root-locus plot and they are all classified as even singular points with $q=2$. It is thus necessary to change the direction of the root locus when each singular point is approached. For branch 1, when Q_1 is approached, the tangential direction $\Delta s=-\epsilon$ and Δz_1 is chosen as $(-\epsilon)e^{-j\pi/2}=+\epsilon j$. Similarly, when branch 1 approaches Q_2, Q_3, and Q_4, the changes in direction are chosen as $\Delta z_2=(-j\epsilon)e^{-j\pi/2}$, $\Delta z_3=(-\epsilon)e^{-j\pi/2}$, and $\Delta z_4=(-j\epsilon)e^{-j\pi/2}$, respectively. Other branches, denoted by 2 and 3, are obtained in a similar manner.

Example 2: As an example of the case with multiple poles, consider

$$G(s)H(s)=\frac{K}{(s+3)(s+1)^2}$$

where $s=-1$ is a repeated pole with multiplicity $r=2$. From (14)

$$w_0=1+\epsilon e^{j\phi/2}$$

$$w_1=1+\epsilon e^{j(\phi+2\pi)/2}$$

$$K=\frac{1}{\rho}\epsilon^2$$

where

$$\rho e^{j\phi}=-\frac{1}{s+3}\bigg|_{s=-1}=0.5e^{j\pi}$$

and ϵ is chosen as 0.2 for illustration. Thus, the two approximate starting points are

$$w_0=-1+j0.2, \qquad K=0.08$$
$$w_1=-1-j0.2, \qquad K=0.08.$$

The root loci are obtained by using (8) with $s(0)=w_0, w_1, -3$ and $K(0)=0.08, 0.08, 0$, respectively. The results are shown in Fig. 5 where the root loci are plotted up to $K=5$.

Example 3: In this example, the open-loop transfer function is assumed to have one real pole, and two complex conjugate poles:

$$G(s)H(s)=\frac{K}{s(s+3+j\sqrt{3})(s+3-j\sqrt{3})}.$$

There is only one odd singular point with $q=3$ located on the root loci. It is thus necessary to jump over the singular point when it is approached and Δz is chosen in the tangential direction of the locus. The complete root-locus plot is shown in Fig. 6.

Example 4: The final example demonstrates the case where the singular points are located off the real axis. Consider

$$G(s)H(s)=K\frac{A(s)}{B(s)}=\frac{K}{(s+1)^2(s+1+j\sqrt{18})(s+1-j\sqrt{18})}.$$

It is seen that

$$g(s)=B(s)+KA(s)=s^4+4s^3+24s^2+40s+19+K.$$

Setting the derivative of g with respect to s to zero, yields three singular points, namely, $s_1=-1$, $s_2=-1+j3$, and $s_3=-1-j3$. Since $s_1=-1$ is a repeated open-loop pole, it can be taken care of as a starting point. At s_2 and s_3, it is easily verified that

$$\mathrm{Im}\,g(s_2)=\mathrm{Im}\,g(s_3)=0$$

and

$$\frac{\partial^2 g}{\partial s^2}\bigg|_{s=s_2\text{ or }s_3}\neq 0.$$

This indicates that both s_2 and s_3 are located on the root loci and, furthermore, they are classified as even singular points with $q=2$. Ap-

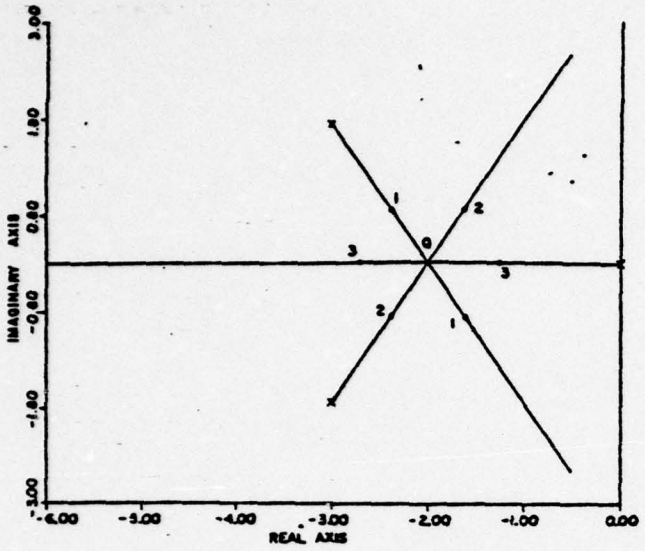


Fig. 6. $G(s)H(s)=\dfrac{K}{s(s+3+j\sqrt{3})(s+3-j\sqrt{3})}$.

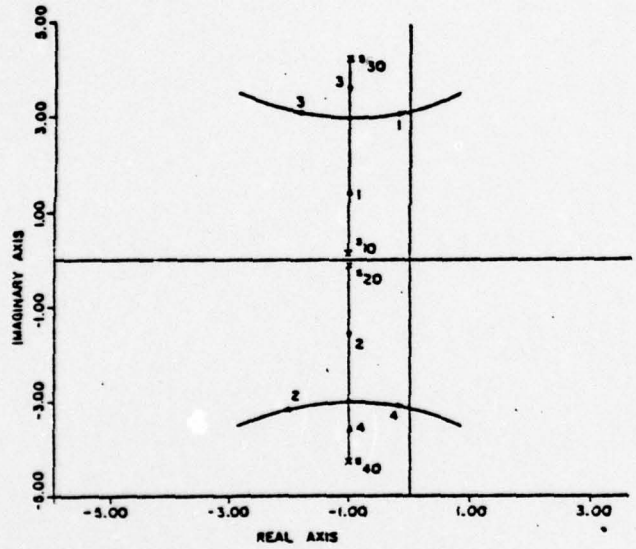


Fig. 7. $G(s)H(s)=\dfrac{K}{(s+1)^2(s+1+j\sqrt{18})(s+1-j\sqrt{18})}$.

plication of the proposed root-locus plotting procedure with four starting points

$$s_{10}=-1+j\epsilon, \qquad K=0.18,\ \epsilon=0.1$$
$$s_{20}=-1-j\epsilon, \qquad K=0.18,\ \epsilon=0.1$$
$$s_{30}=-1+j\sqrt{18}, \qquad K=0$$
$$s_{40}=-1-j\sqrt{18}, \qquad K=0$$

leads to the complete root-locus plot shown in Fig. 7. It is clear from this example that a necessary condition for the existence of complex singular points is that the order of the open-loop transfer function be greater than or equal to four.

V. Conclusion

An algorithm for generating the root-locus plot has been presented. Classification of singular points has also been discussed in detail. It is shown that the conventional break-in and break-away points are just special cases of even singular points. The computer-aided method successfully solves the problem of discontinuity of the direction of the locus at singular points and enables one to plot the root loci without missing or repeating any branch.

References

[1] W. R. Evans, *Control-System Dynamics*. New York: McGraw-Hill, 1954.
[2] H. Raven, *Automatic Control Engineering*. New York: McGraw-Hill, 1961.
[3] C. Gupta and L. Hasdorff, *Fundamentals of Automatic Control*. New York: Wiley, 1970.
[4] V. Krishnan, "Semi-analytic approach to root locus," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 102–108, Jan. 1966.
[5] K. Steiglitz, "An analytic approach to root loci," *IRE Trans. Automat. Contr.*, vol. AC-6, pp. 326–332, Sept. 1961.
[6] C. K. Wojcik, "Analytical representation of root locus," *J. Basic Eng.*, pp. 37–43, Mar. 1964.
[7] B. P. Bhattacharyya, "Root locus equations of the fourth degree," *Int. J. Contr.*, vol. 1, no. 6, pp. 533–556, 1965.
[8] F. H. Branin, "Widely convergent method for finding multiple solutions of simultaneous nonlinear equations," *IBM J. Res. Dev.*, vol. 16, no. 5, pp. 504–522, Sept. 1972.
[9] K. S. Chao, D. K. Liu, and C. T. Pan, "A systematic search method for obtaining multiple solutions of simultaneous nonlinear equations," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 748–753, Sept. 1975.
[10] K. S. Chao and R. Saeks, "Continuation methods in circuit analysis," *Proc. IEEE*, vol. 65, no. 8, pp. 1187–1194, Aug. 1977.
[11] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 1966.

V. Conclusion

An algorithm for generating the root-locus plot has been presented. Classification of singular points has also been discussed in detail. It is shown that the conventional break-in and break-away points are just special cases of even singular points. The computer-aided method successfully solves the problem of discontinuity of the direction of the

9. Reprint of "A Root Locus Technique for Interconnected Systems" by R.A. DeCarlo and R. Saeks from the IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-9, pp. 53-55, (1979).

A Root Locus Technique for Interconnected Systems

RAYMOND A. DECARLO, MEMBER, IEEE, AND
R. SAEKS, FELLOW, IEEE

Abstract—**This note presents a numerically feasible technique for computing composite system eigenvalues from component/subsystem eigenvalues in the component connection model context. The technique is a natural extension of previous artificial methods of computing system eigenvalues in a state model having a perturbed A matrix. The present technique allows one to trace the movement of component eigenvalues as coupling is introduced. Furthermore, the technique permits investigation of eigenvalue movement as a function of interconnection gains. This is useful in analyzing short/open circuit phenomena as well as other system characteristics. Finally, the technique is useful for determining and understanding composite system stability in terms of component and connection information.**

I. Introduction

This note describes a root locus technique in which the "root locus" begins at the eigenvalues of the individual component state equations and traces out trajectories as coupling between components is continuously and uniformly introduced. The component connection model [5], [14], [16], [12], [13], [4], [18] is the natural vehicle for executing the approach as opposed to distantly related perturbation schemes [6], [7], [11].

Numerical implementation is discussed through a continuations approach [2], [3], [19], [10], [1] which in this case is a set of coupled differential equations characterizing the interplay between the eigenvalues/eigenvectors of the appropriate composite system matrix as a function of an underlying parameter r. The coupled differential equations are given in the Appendix. The idea is to initialize these equations at the eigenvalues/eigenvectors of the components and integrate to obtain those of the composite system.

The technique is applicable to the study of short/open circuit phenomena as well as the investigation of the effect on composite system eigenvalues of increasing/decreasing coupling gains between components.

It is assumed that each component has a completely controllable and observable state model:

$$\dot{x}_i = A_i x_i + B_i a_i$$
$$b_i = C_i x_i + D_i a_i \tag{1}$$

where a_i, b_i, and x_i are vectors of component inputs, outputs, and states, respectively. Combining the component descriptions defines a *composite component state model* as

$$\dot{x} = Ax + Ba$$
$$b = Cx + Da \tag{2}$$

where $a = [a_1, \cdots, a_n]^t$, $b = [b_1, \cdots, b_n]^t$, $x = [x_1, \cdots, x_n]^t$, $A = \text{diag}\,[A_1, \cdots, A_n]$, etc. All vectors are assumed conformable, and for each time instant they possess values in the appropriately dimensioned Euclidean space. The *system* description is completed with the connection equations, which are

$$\left[\begin{array}{c} a \\ \hline y \end{array}\right] = \left[\begin{array}{c|c} L_{11} & L_{12} \\ \hline L_{21} & L_{22} \end{array}\right] \left[\begin{array}{c} b \\ \hline u \end{array}\right] \tag{3}$$

where L_{ij} are real matrices accounting for KVL, KCL, and/or other conservation laws; y and u are vectors of *system* outputs and inputs, respectively. Equations (2) and (3) constitute the component connection model. Except for theoretical analysis and/or describing relationships between classical models and the component connection model, one never combines (2) and (3) into a single set of equations. All relevant simulation and analysis is possible without combining the equations [16]–[18], [5], [2], [4], [12]–[14], [3].

Under general considerations [16], [5], [18] the composite component state vector x and the *system* state vector may be chosen to coincide, so that a valid composite system state model exists as follows:

$$\dot{x} = Fx + Gu$$
$$y = Hx + Ju \tag{4}$$

where F, G, H, and J can be expressed in terms of the matrices of (2) and (3). In particular,

$$F = A + B(I - L_{11}D)^{-1}L_{11}C. \tag{5}$$

Since G, H, and J are unrelated to the discussion of this note, their specific form is omitted. The component interconnection matrix L_{11} is explicit in (5). This explicitness motivates and permits the root locus technique.

II. The Root Locus Technique

Clearly the composite system eigenvalues are the roots of the polynomial $\det[\lambda I - F]$ where F is defined in (5). To consider the effect of coupling/interaction among components, replace L_{11} by rL_{11} for a scalar r to obtain

$$F(r) = A + B[I - (rL_{11})D]^{-1}(rL_{11})C. \tag{6}$$

Clearly, $F(0) = A$, as in (2), and $F(1) = F$, the relevant composite system matrix.

A continuations approach to computing the root locus of the eigenvalues of $F(r)$, $0 \le r \le 1$, uses (21)–(23) from the Appendix, which require knowledge of $\dot{F}(r)$. Unfortunately, the eigenvalues of $F(r)$ must be distinct, since the coupled differential equations become singular otherwise. Using the well-known matrix identities,

$$[\dot{M}^{-1}] = -M^{-1}\dot{M}M^{-1} \tag{7}$$
$$X(I - YX)^{-1} = (I - XY)^{-1}X, \tag{8}$$

whenever either inverse exists, and

$$(I - XY)^{-1} = (I + X(I - YX)^{-1}Y) \tag{9}$$

results in

$$\dot{F}(r) = B(I - rL_{11}D)^{-2}L_{11}C \tag{10}$$

where the superscript -2 indicates the inverse squared.

By initializing (21)–(23) at the component (subsystem) eigenvalues/eigenvectors, one integrates the coupled differential equations to obtain the root locus—i.e., the trajectories of the eigenvalues of the system as coupling is introduced. If any of the

Manuscript received May 11, 1978; revised August 28, 1978

R. A. DeCarlo is with the Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

R. Saeks is with the Department of Electrical Engineering and Mathematics, Texas Tech University, Lubbock, TX 79409.

AD-A077 652 TEXAS TECH UNIV LUBBOCK INST FOR ELECTRONICS SCIENCE F/G 9/4
ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICE ELECTRONICS P--ETC(U)
OCT 79 R SAEKS , K S CHAO , J WALKUP N00014-76-C-1136

UNCLASSIFIED NL

2 OF 2
ADA
077652

END
DATE
FILMED
1 -80
DDC

eigenvalue trajectories cross, it is necessary to perturb r around the point of intersection and reinitialize the algorithm.

When the composite state model D-matrix is zero, $\dot{F}(r) = BL_{11}C$, a constant. When $D \neq 0$, it is necessary to compute $(I - rL_{11}D)^{-1}$ at each step of the integration. Typically, rank $(D) \ll$ dim (I). Viewing $-rL_{11}D$ as a low-rank perturbation on I, Householder's formula [9], [5] provides a convenient and quick means for computing $(I - rL_{11}D)^{-1}$.

Note that (6) is *not* an *artificial* linear perturbation of the composite system F matrix to account for stray capacitances/inductances on eigenvalue movement or ease computing the eigenvalues of F [6], [7], [11]. Indeed the QR-algorithm is much more efficient and accurate. However, this formulation is natural in the component connection model context; it is numerically implementable through a continuations approach as above or through an iterative calling of a QR-algorithm; it identifies components giving rise to particular composite system eigenvalues by tracing the eigenvalue locus of $F(r)$; and finally, by replacing L_{11} by $L_{11} + rP$ (P is of low rank), it offers a means of investigating coupling gains on composite system eigenvalue locations as follows.

Assume the composite system eigenvalues/eigenvectors are known and distinct. To investigate the effects of gain changes, replace L_{11} in (5) by $(L_{11} + rP)$, where P is a low-rank matrix characterizing the gain variation and r varies over some relevant finite interval containing zero. This produces

$$F(r) = A + B(I - L_{11}D - rPD)^{-1}(L_{11} + rP)C. \tag{11}$$

Computing a root locus via a continuations approach requires $\dot{F}(r)$, which can be derived as

$$\dot{F}(r) = [I - L_{11}D - rPD]^{-1}P[I + D(I - L_{11}D - rPD)^{-1}(L_{11} + rPD)]. \tag{12}$$

Of course if $D = 0$, $\dot{F}(r) = DPC$, which is constant, sparse, and of low rank. Although (12) seems formidible, it is possible to view rPD as a low-rank perturbation on $(I - L_{11}D)$ and use Householder's formula to compute the necessary inverses during each step of the integration of (21)–(23). Of course, stability or instability is known simply by noting the position of the terminal points of the root locus.

In accounting for effects of stray capacitances/inductances, the component connection model is superior to some other approaches [6], [7], [11]. Consider that the A matrix in (2) is block diagonal (predominately diagonal for circuits) describing *only* component information. Suppose a typical entry is $1/C$, characterizing a capacitance. Clearly $\partial A/\partial C$ is zero, except for the particular C-dependent diagonal entry. If C_0 is the nominal C, using $(\partial A/\partial C)|_{C_0}$ in (21) gives the approximation $(\partial \lambda_i/\partial C)|_{C_0}$. Considering $\Delta C(\partial \lambda_i/\partial C)|_{C_0}$ for each i gives a good first-order approximation to direction and magnitude of eigenvalue movement relative to small perturbations in C.

III. Conclusions

We have described a technique for determining eigenvalue movement of a system in terms of connection information. Several applications were described. It is especially useful for determining what *components* give rise to particular interconnected system eigenvalues. This in turn provides a means of reduced order modeling without changing the system structure. Suppose a component (as identified by the root locus technique) gives rise to a composite system eigenvalue corresponding to *slow* mode of the system. This component can then be replaced by a constant gain amplifier. This reduces the number of state variables in the system model without changing the interconnected structure of the system which is represented by the connection matrices.

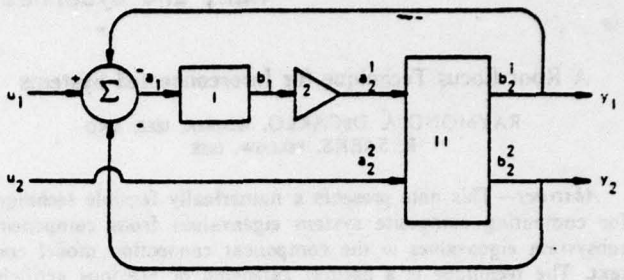


Fig. 1. Block diagram of interconnected system.

IV. An Example

Consider the system configuration of Fig. 1. There are two components (subsystems) as indicated by the roman numerals I and II. The triangular block indicates an amplifier whose gain is two. For convenience, the action of this amplifier is reflected in the connection equations.

Suppose component I has the following state model:

$$\dot{x}_1 = -x_1 + a_1$$
$$b_1 = x_1. \tag{13}$$

Let the state model for component II be given by

$$\begin{bmatrix}\dot{x}_2^1\\ \dot{x}_2^2\end{bmatrix} = \begin{bmatrix}-1 & -2\\ 2 & -1\end{bmatrix}\begin{bmatrix}x_2^1\\ x_2^2\end{bmatrix} + \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}a_1^2\\ a_2^2\end{bmatrix}$$
$$\begin{bmatrix}b_2^1\\ b_2^2\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_2^1\\ x_2^2\end{bmatrix} \tag{14}$$

Defining $x = [x_1, x_2^1, x_2^2]^t$, $a = [a_1, a_2^1, a_2^2]^t$, and $b = [b_1, b_2^1, b_2^2]^t$, the composite component state model is

$$\dot{x} = \begin{bmatrix}-1 & 0 & 0\\ 0 & -1 & -2\\ 0 & 2 & -1\end{bmatrix}x + \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}a$$
$$b = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}x. \tag{15}$$

By inspection of Fig. 1, the connection equations are

$$\begin{bmatrix}a\\ \hline y\end{bmatrix} = \left[\begin{array}{ccc:cc} 0 & -1 & -1 & 1 & 0\\ 2 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1\\ \hline 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\end{array}\right]\begin{bmatrix}b\\ \hline u\end{bmatrix} \tag{16}$$

where $y = [y_1, y_2]^t$, and $u = [u_1, u_2]^t$. Equations (15) and (16) constitute the complete system model—i.e., the component connections model for the system of Fig. 1.

Define $F(r)$ as per (6), in which case

$$F(r) = \begin{bmatrix}-1 & 0 & 0\\ 0 & -1 & -2\\ 0 & 2 & -1\end{bmatrix} + \begin{bmatrix}0 & -r & -r\\ 2r & 0 & 0\\ 0 & 0 & 0\end{bmatrix}. \tag{17}$$

The derivative of $F(r)$ clearly is

$$\dot{F}(r) = \begin{bmatrix}0 & -1 & -1\\ 2 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \tag{18}$$

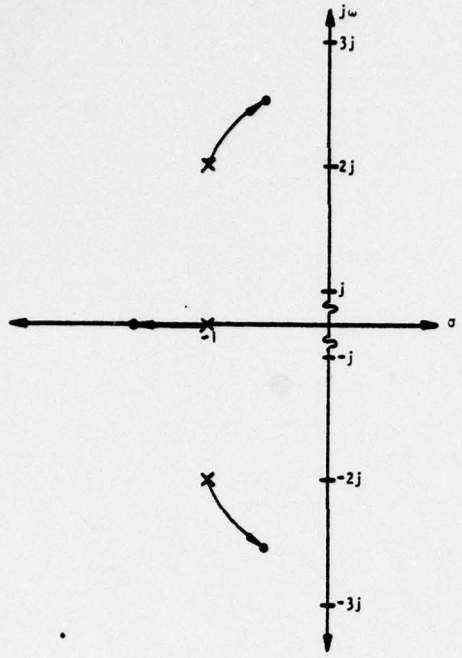


Fig. 2. Plot of "root locus."

Implementing the root locus via the continuation equations (21)–(23) results in the root locus plotted in Fig. 2.

Appendix
Eigenvalue Dynamics

Let $F(r)$ be an $n \times n$ matrix with possibly complex entries depending on the parameter r. Let $F(r)^*$ be its *adjoint* matrix. Note $F(r)^*$ is the unique matrix satisfying

$$\langle F(r)x, y\rangle = \langle x, F(r)^* y\rangle \tag{19}$$

for all complex n-vectors, x and y, where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product defined as

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i \bar{y}_i, \tag{20}$$

where $\bar{y}_i$ is the complex conjugate of the ith entry of the column vector. The essential theorem here is the following.

Theorem 1: Let $F(r)$ and its adjoint $F(r)^*$, have eigenvector trajectories $e_i(r)$ and $\bar{e}_i(r)$, and eigenvalue trajectories $\lambda_i(r)$ and $\bar{\lambda}_i(r)$, respectively, for $i = 1, 2, \cdots, n$. Then for any value of r where the eigenvalues of $F(r)$ are *all distinct*,

$$\frac{d\lambda_i}{dr} = \frac{\left\langle \frac{dF}{dr} e_i, \bar{e}_i \right\rangle}{\langle e_i, \bar{e}_i\rangle}, \qquad i = 1, 2, \cdots, n \tag{21}$$

$$\frac{de_i}{dr} = \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left\langle \frac{dF}{dr} e_i, \bar{e}_j \right\rangle}{(\lambda_i - \lambda_j)\langle e_j, \bar{e}_j\rangle} e_j \tag{22}$$

$$\frac{de_i}{dr} = \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left\langle e_i, \frac{dF}{dr} \bar{e}_j \right\rangle}{(\lambda_i - \lambda_j)\langle e_j, \bar{e}_j\rangle} e_j. \tag{23}$$

The eigenvalues of $F(r)$ and $F(r)^*$ are complex conjugates of each other, whereas the eigenvectors $e_i(r)$ and $\bar{e}_i(r)$ are not. This theorem is motivated by [8], where (21) is implicitly expressed and the necessary tools for proving the theorem are made clear.

The proof of the theorem can be found in [5], [4]. A derivation of (21) can be found in [7] as well as [8].

References

[1] K. S. Chao, D. K. Liu, and C. T. Pan, "A systemic search method for obtaining multiple solutions of simultaneous nonlinear equations," *IEEE Trans. Circuits and Syst.*, vol. CAS-22, pp. 748–753, 1975.

[2] K. S. Chao and R. Saeks, "Continuations approach to large-change sensitivity analysis," *Electron. Circuits Syst.*, vol. 1, no. 1, Sept. 1976.

[3] ——, "Continuations methods in circuit analysis," *Proc. IEEE*, to appear.

[4] R. Decarlo and R. Saeks, "A 'root locus' technique for large interconnected dynamical systems," *Proc. 1978 Int. Symp. Circuits and Systems*, New York, 1978.

[5] ——, *Interconnected Dynamical Systems*. New York: Marcel Dekker, 1978.

[6] C. A. Desoer, "Network design by first order predistortion," *IRE Trans. PGCT*, vol. 3, pp. 167–170, Sept. 1957.

[7] ——, "Perturbations of eigenvalues and eigenvectors of a network," 5th Allerton Conf., 1967.

[8] D. K. Faddeev and V. N. Fadeva, *Computational Methods of Linear Algebra*. San Francisco: Freeman, 1963.

[9] A. S. Householder, "A survey of some closed methods for inverting matrices," *SIAM J. Appl. Math.*, vol. 5, pp. 155–169, 1957.

[10] C. T. Pan, Ph.D. dissertation, Texas Tech Univ., Lubbock, 1977.

[11] A. Papoulis, "Perturbations of the natural frequencies and of the eigenvectors of a network," *IEEE Trans. Circuit Theory*, vol. CT-13, pp. 188–195, June 1966.

[12] M. N. Ransom and R. Saeks, "The connection function-theory and application," *Inter. J. Circuit Theory and Its Applications*, vol. 3, 1975.

[13] ——, "Fault isolation with insufficient measurements," *IEEE Trans. Circuit Theory*, vol. CT-20, pp. 416–417, 1973.

[14] R. Saeks, G. Wise, and K. S. Chao, "Analysis and design of interconnected dynamical systems," in *Large Scale Systems*, R. Saeks, Ed. Los Angeles: Point Lobos, 1976.

[15] R. Saeks, *Large Scale Dynamical Systems*, R. Saeks, Ed. Los Angeles: Point Lobos, 1976.

[16] S. P. Singh and R. W. Liu, "Existence of state equations representations of large-scale dynamical systems," *IEEE Trans. Circuit Theory*, vol. CT-20, pp. 399–346, 1973.

[17] H. Trauboth and W. McCallum, "MARSYAS users manual," tech. rep. A1-34812, Computation Lab. NASA/MSFC, 1973.

[18] H. Trauboth and S. P. Singh, "MARSYAS I and II," *IEEE Circuits and Systems Soc. Newsletter*, vol. 6, no. 3, 1973.

[19] E. Wasserman, "Numerical solutions by the continuations methods," *SIAM Rev.*, vol. 15, pp. 89–119, 1973.

10. Reprint of "A Continuation Algorithm for Sparse Matrix Inversion" by R. Saeks from the IEEE Proceedings, Vol. 67, pp. 682-683, (1970).

A Continuations Algorithm for Sparse Matrix Inversion

RICHARD SAEKS

In the various algorithms used for the analysis and design of large-scale circuits and systems, the problem of inverting a continuously parameterized family of sparse matrices $M(r)$ is often encountered [1]-[5]. In frequency domain analysis, this might represent a transfer function matrix which one must invert over a specified frequency range [3] while in time-domain analysis, such an $M(r)$ arises in the form of the Jacobian matrix for the system equations [1] which is dependent on some potentially variable parameter r. Typically, one inverts $M(r)$ at a discrete set of points r_i, $i = 1, 2, \cdots, n$; using a sparse matrix algorithm. Indeed, the more efficient algorithms exploit the fact that the matrices $M(r_i)$ have a common sparsity structure allowing much of the computational overhead to be shared by the n inversions [1].

An alternative to repeated inversion is the *continuations algorithm* [5] wherein one integrates the differential equation

$$\dot{Z}(r) = -Z(r)(dM/dr)\,Z(r) \quad Z(0) = M(0)^{-1} \tag{1}$$

to obtain $M(r)^{-1} = Z(r)$. While the integration of (1) is far more efficient than repeated matrix inversion for small matrices, it fails to take advantage of the sparseness of $M(r)$, thereby rendering the technique inapplicable in a large-scale systems context. The purpose of the present note is to present an alternative continuation algorithm which combines the LU factorization technique of sparse matrix inversion with (1).

Recall the standard sparse matrix inversion technique [6] wherein one factors a matrix into the form $M = LU$ where L is lower triangular and U is upper triangular with ones along the diagonal. We then represent the inverse matrix in the form $M^{-1} = U^{-1}L^{-1}$. The key to the technique is that both L and U and their inverses will be sparse if M is sparse (though, in general, M^{-1} is not sparse). As such, one may store and manipulate the inverse of a sparse matrix via its sparse upper and lower triangular factors U^{-1} and L^{-1}, even though the inverse matrix itself is nonsparse. These ideas are combined with the continuation algorithm concept in the following theorem [7]. Here, the notation ${}^{u}[M]$ is used to denote the strictly upper triangular matrix obtained from M by setting all of the entries of M on or below the diagonal to zero. Similarly, ${}^{l}[M]$ denotes the lower triangular matrix obtained from M by setting all of the entries above the diagonal to zero.

Theorem: Let $X(r)$ and $Y(r)$ be solutions of the matrix differential equation

$$\begin{aligned} \dot{X} &= -X^{u}[Y(dM/dr)X], \qquad X(0) = U(0)^{-1}. \\ \dot{Y} &= -{}^{l}[Y(dM/dr)X]\,Y, \qquad Y(0) = L(0)^{-1}. \end{aligned} \tag{2}$$

Then, $X(r) = U(r)^{-1}$ and $Y(r) = L(r)^{-1}$ where $M(r)^{-1} = U(r)^{-1}L(r)^{-1}$ is the LU factored form of $M(r)^{-1}$. Note, if $M(r)$ and dM/dr are sparse then every matrix involved in the integration of (2) will be sparse. Moreover, the integration may be carried out with the aid only of a matrix multiplication algorithm plus a simple procedure for extracting the upper and lower triangular submatrices of $Y(dM/dr)X$.

Proof: First, we observe that if $Y(0)$ is lower triangular, then $\dot{Y}$ will be lower triangular and so will $Y(r)$ for all r. Similarly, if $X(0)$ is upper triangular with ones on the diagonal, then $\dot{X}$, being the product of an upper triangular and strictly upper triangular matrix, will be strictly upper triangular. As such, $X(r)$ will be upper triangular with ones on the diagonal for all r. Thus $X(r)$ and $Y(r)$ have the correct form and it remains to verify the equality $M(r)^{-1} = X(r)\,Y(r)$. Here

$$\begin{aligned} X(r)\,Y(r) &= X(0)\,Y(0) + \int_0^r [X(q)\,Y(q)]\,dq \\ &= X(0)\,Y(0) + \int_0^r [\dot{X}(q)\,Y(q) + X(q)\,\dot{Y}(q)]\,dq \\ &= X(0)\,Y(0) + \int_0^r \{-X(q)^{u}[Y(q)(dM/dq)\,X(q)]\,Y(q) \\ &\quad - X(q)^{l}[Y(q)(dM/dq)\,X(q)]\,Y(q)\}\,dq \\ &= X(0)\,Y(0) + \int_0^r \{-X(q)[Y(q)(dM/dq)\,X(q)]\,Y(q)\}\,dq \\ &= X(0)\,Y(0) + \int_0^r [X(q)\,Y(q)](dM/dq)[X(q)\,Y(q)]\,dq. \end{aligned} \tag{3}$$

Differentiation of both sides of (3) with respect to r then results in

$$[X(r)\,\dot{Y}(r)] = [X(r)\,Y(r)](dM/dr)[X(r)\,Y(r)]. \tag{4}$$

Finally, a comparison of (4) and (1) reveals that $X(r)\,Y(r) = M(r)^{-1}$ since both $X(r)\,Y(r)$ and $M(r)^{-1}$ satisfy the same differential equation.

Consider the family of matrices

$$M(r) = \begin{bmatrix} 1 & r \\ -1 & 1 \end{bmatrix}. \tag{5}$$

Here, $M(0)$ is lower triangular and, hence, has the trivial LU-factorization

$$M(0) = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = L(0)\,U(0) \tag{6}$$

while

$$\dot{M}(r) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{7}$$

As such, we have

$$L(0)^{-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{8}$$

and

$$U(0)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{9}$$

Now, upon using an Euler integration formula $[Z(h) \approx Z(0) + h\dot{Z}(0)]$, we may estimate $U(0.1)^{-1}$ and $L(0.1)^{-1}$ via the equalities

$$\begin{aligned} U(0.1)^{-1} &\approx U(0)^{-1} + (0.1)\,\dot{U}(0)^{-1} \\ &= U(0)^{-1} - (0.1)\,U(0)^{-1}{}^{u}[L(0)^{-1}\dot{M}(0)\,U(0)^{-1}] \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0.1 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1/10 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{10}$$

Manuscript received July 27, 1978.
The author is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

0018-9219/79/0400-0682$00.75

and

$$L(0.1)^{-1} \approx L(0)^{-1} + (0.1)\,L(\dot{0})^{-1}$$
$$\doteq L(0)^{-1} - {}^{1}[L(0)^{-1}\dot{M}(0)\,U(0)^{-1}]L(0)^{-1}$$
$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0.1 & 0.1 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 \\ 9/10 & 9/10 \end{bmatrix}. \tag{11}$$

Multiplying these estimates then yields

$$M(0.1)^{-1} = U(0.1)^{-1}L(0.1)^{-1}$$
$$\approx \begin{bmatrix} 91/100 & -9/100 \\ 9/10 & 9/10 \end{bmatrix} \tag{12}$$

which compares favorably with the exact inverse

$$M(0.1)^{-1} = \begin{bmatrix} 10/11 & -1/11 \\ 10/11 & 10/11 \end{bmatrix}. \tag{13}$$

The error here is due to the approximation inherent in the numerical integration process and can be reduced by use of a more accurate integration procedure. Of course, the result of the theorem is exact and the computed value for $M(r)^{-1}$ will be as accurate as the integration process employed.

References

[1] G. Hachtel, R. K. Brayton, and F. Gustavson, "The sparse Tabeau approach to network analysis and design," *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 101–113, 1971.

[2] L. O. Chua and P. M. Lin, *Computer-Aided Analysis of Electronic Circuits*. Englewood Cliffs, NJ, Prentice-Hall, 1975.

[3] A. M. Erisman and G. Spies, "Exploiting problem characteristic in sparse matrix approach to frequency domain analysis," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 260–264, 1972.

[4] R. Saeks and K. S. Chao, "Continuations approach to large-change sensitivity analysis," *IEE (London) J. Electron. Circuits and Syst.*, vol. 1, pp. 11–16, 1976.

[5] K. S. Chao, and R. Saeks, "Continuation methods in circuit analysis," *Proc. IEEE*, vol. 65, pp. 1187–1194, 1977.

[6] R. P. Tewerson, *Sparse Matrices*. New York: Academic Press, 1972.

[7] R. Saeks and R. A. DeCarlo, *Interconnected Dynamical Systems*. New York: Marcel Dekker (to appear).

11. Reprint of "Multiple Solutions of a Class of Nonlinear Equations" by C.T. Pan and K.S. Chao from the Proceedings of the 1979 IEEE International Symposium on Circuits and Systems, Tokyo, July 1979, pp. 577-580.

ABSTRACT

A search method is presented for obtaining multiple solutions of a system of n nonlinear equations whose first (n-1) equations do not necessarily define a unique space curve. In particular, the approach is used to find all the roots of a complex polynomial. Singularities on the space curve are analyzed and properly classified according to their high order derivatives. Depending on the nature of singularities, the rules for a sign change in the algorithm are determined so that the root-finding procedure can be continued.

I. INTRODUCTION

An important problem in the analysis and design of non-linear circuits and systems is the determination of multiple solutions of a nonlinear equation

$$f(x) = 0 \tag{1}$$

where f is a continuously differentiable function from R^n into itslef. Several methods [1] - [5] have been proposed for finding multiple solutions. In [4], Chao, Liu and Pan developed a systematic search method for solving multiple solutions by numerical integration of a set of differential equations of the form

$$\dot{f}_i[x(t)] = -f_i[x(t)], \quad f_i[x(0)] = 0,$$

$$\dot{f}_n[x(t)] = \pm f_n[x(t)], \quad f_n[x(0)] = f_{no} \tag{2}$$

$$i = 1,2,\ldots,n-1$$

or in the x-space

$$\dot{x} = (\partial f/\partial x)^{-1}(-f_1 \; -f_2 \; \ldots \; \pm f_n)^T, \; x_o \in \ell \tag{3}$$

along the space curve, ℓ, defined by the intersection of the solution manifolds for

$$f_i[x(t)] = 0, \quad i = 1,2,\ldots,n-1. \tag{4}$$

The transition in sign of f_n should be made at the solution points and points where the Jacobian determinant changes sign. The method is capable of finding all solutions provided that the intersection is a simple curve, i.e., a continuous, differentiable curve which does not intersect itself.

The purpose of this paper is to generalize the above method to cases where the intersection is multi-branched and may indeed intersect itself. In Section II, properties of nonlinear equation with multi-branched and intersecting solution curves are discussed. The method is then applied to the computation of all the roots of a polynomial in Section III. In Section IV, the results obtained are illustrated by means of an example.

II. THEORECTICAL BASIS

In section I, a systematic search method has been outlined. Success in finding all solutions depends heavily on whether or not ℓ defines a unique simple curve. If it does, then a complete traversal of it enables one to find all solutions. On the other hand if ℓ is multi-branched, the application of the method may lead only to those solutions that lie on the branch containing the starting point. As will be seen later, multiple branches do exist for some classes of functions. Therefore in order to find all solutions a starting point on each branch of ℓ must be initiated.

On a continuous, differentiable solution curve, ℓ, the sign change of the method is indicated by the fact that the directional derivative of $f_n(x)$ in the tangential direction of ℓ changes sign if and only if the corresponding Jacobian of f on ℓ changes sign [4]. However, difficulty arises when ℓ does intersect itself. Since the directional derivative is not defined at the point of intersection, the Jacobian can no longer be used for judging the monotonicity of f_n along ℓ at that point. The situation, however, can be described by the following theorem.

This work was supported in part by the National Science Foundation under grant ENG-77-22991 and the Joint Service Electronic Program under ONR Contract 76-C-1136.

Theorem 1:

Let $f(x)$: $R^n \to R^n$, $n \geq 2$, be a C^1 function. If

$$\nabla f_n = v \triangleq (\Delta_{n1}, \Delta_{n2}, \ldots, \Delta_{nn})^T \tag{5}$$

where

Δ_{ni} is the (ni)th cofactor of $\partial f/\partial x$, then

$$|J| = ||v||^2 \geq 0. \tag{6}$$

Furthermore, if $f \in C^2$, then f_n also satisfies the Laplace's equation

$$\nabla^2 f_n = \frac{\partial^2 f_n}{\partial x_1^2} + \frac{\partial^2 f_n}{\partial x_2^2} + \ldots + \frac{\partial^2 f_n}{\partial x_n^2} = 0. \tag{7}$$

Proof: Expansion of the Jacobian determinant along the nth row results in

$$\det J = \sum_{i=1}^{n} \frac{\partial f_n}{\partial x_i} \Delta_{ni} = ||v||^2 \geq 0.$$

The proof of the second part although complicated is quite straightforward and is therefore omitted.

On a continuous, differentiable solution curve, the directional derivative is given by [4]

$$\frac{df_n}{d\ell} = |J|/||v||. \tag{8}$$

Thus, under the conditions of Theorem 1, if ℓ does not intersect itself, then on a given branch of ℓ, the directional derivative would never change sign and this also implies that there exists at most one solution on that branch. From this and Theorem 1, the following theorem can be deduced.

Theorem 2:

Let $f(x)$: $R^n \to R^n$, $n \geq 2$, be a C^1 function and $f = 0$ has more than one solution. If ℓ_i, defined by

$$f_i(x) = 0, \quad i = 1,2,\ldots,n-1,$$

is a unique, simple curve, then $\nabla f_n \neq v$.

For $n = 2$, condition (5) reduces to

$$\frac{\partial f_2}{\partial x_1} = -\frac{\partial f_1}{\partial x_2}, \qquad \frac{\partial f_2}{\partial x_2} = \frac{\partial f_1}{\partial x_1}. \tag{9}$$

If f_1 and f_2 represent both the real part and the imaginary part of an analytic function, respectively, then the condition $\nabla f_n = v$ in the two-dimensional case is essentially equivalent to the Cauchy-Riemann conditions. For polynomials of a complex variable, the relationship between the Cauchy-Riemann conditions and the determinant of the Jacobian matrix has previously been discussed by Branin [6], [7]. Branin also described a procedure, called the method of signatures, for computing all the roots of a polynomial. However, the algorithm fails in cases where singular points do exist on the search trajectory. In the following section, a systematic method for obtaining all the solutions of a complex polynomial with complex coefficients is presented.

III. ROOTS OF A POLYNOMIAL

The Basic Algorithm

In view of the foregoing comments, the method for obtaining all solutions can now be formulated for the nth order polynomial equation

$$g(z) = f_1(x_1,x_2) + jf_2(x_1x_2) = 0. \tag{10}$$

Application of the method described in Section I to $f = 0$ leads to

$$\frac{df_1}{dt} = -f_1, \quad f_1(x_{10}^i, x_{20}^i) = 0$$
$$\frac{df_2}{dt} = \pm f_2, \quad f_2(x_{10}^i, x_{20}^i) = f_{20}, \tag{11}$$

where the initial conditions are such that the starting point x_o^i lies on or close to each branch ℓ_i of ℓ defined by $f_1 = 0$. By using the chain-rule of differentiation and assuming $\det J \neq 0$, (10) can be rewritten as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \frac{1}{\det J} \begin{bmatrix} -f_1 \dfrac{\partial f_1}{\partial x_1} \mp f_2 \dfrac{\partial f_2}{\partial x_1} \\ f_1 \dfrac{\partial f_2}{\partial x_1} \mp f_2 \dfrac{\partial f_1}{\partial x_1} \end{bmatrix}, \tag{12}$$

$$x_o^i \varepsilon \ell_i, \quad i = 1,2,\ldots,n.$$

Let

$$\frac{dg}{dz} \triangleq g'(z) = h_1(x_1x_2) + jh_2(x_1x_2). \tag{13}$$

An expression in the complex domain is obtained as

$$\dot{z} = \frac{1}{(h_1^2 + h_2^2)} [(-f_1h_1 \pm f_2h_2) + j(f_1h_2 \mp f_2h_1)]$$

$$z_o^i \varepsilon \ell_i, \quad i = 1,2,\ldots,n. \tag{14}$$

Further simplification by using the notion of complex conjugate(*) leads to the following compact form:

$$\dot{z} = -g(z)/g'(z), \quad \text{if the minus sign is chosen}$$
$$\dot{z} = -g^*(z)/g'(z), \quad \text{if the plus sign is chosen} \tag{15}$$
$$z_o^i \varepsilon \ell_i, \quad i = 1,2,\ldots,n.$$

The $\pm$ sign in (14) or the complex conjugate sign in (15) must be switched at the solution points and at the even singular points to be defined later.

Even though the algorithm is derived from f it is seen from (15) that analytic expressions for f_1, f_2, h_1 and h_2 are not required explicitly. The algorithm is thus most suited for finding all the roots of a complex polynomial.

Basis for the Sign Change

Due to the nature of isolated singularities, the existence of (n-1) such points in the denominator of (15) does not pose any problem to the root-finding. The sign change at the solution points has been discussed in [4] and it will not be repeated here. The sign change at certain singular points where the Jacobian vanishes can be judged from the following theorem.

Theorem 3: Suppose g(z) is an analytic function such that

$$g(w) = ja, \quad \text{where a is real and } a \neq 0$$

$$g^{(q)}(w) = 0, \quad q = 1,2, \ldots, m-1$$

$$g^{(m)}(w) \neq 0, \quad m \geq 2$$

at some point w located on Re g(z) = 0. Let $R_g = \{z | \text{Re } g(z) = 0\}$. Then in the neighborhood of w, R_g consists of m branches $R_{g1}, R_{g2}, \ldots, R_{gm}$ and $R_{g1} \cap R_{g2} \cap \cdots \cap R_{gm} = \{w\}$. Furthermore from each i, $1 \leq i \leq m$, Im g(z)$|R_{gi}$ is either a local maximum or a local minimum at w if m is even; it is either an increasing or a decreasing function if m is odd.

For convenience, a singular point in Theroem 3 is called an even singular point if m is even; otherwise it is said to be an odd singular point. The criterion for the sign change at singularities follows directly from the above theorem and is formally stated in the following theorem.

Theorem 4: If g(z) is an nth-order complex polynomial and w is not a solution point of g(z) = 0, then the directional derivative of Im g(z) in the tangential direction of ℓ defined by Re g(z) = 0 changes sign when passing through a point w if and only if w is an even singular point.

In view of the previous theorem, it is clear that the sign must be changed when passing through an even singular point even though the Jacobian does not change its sign. For an odd singular point the sign must be kept unchanged when passing through it since $f_2(x)$ does not change its monotonicity. Due to the very nature of an odd singular point, it is clear that a sign change at such a singular point will cause the algorithm to oscillate. Since there are only two possibilities, higher order derivative tests can be avoided in the actual implementation. One can always initiate a sign change whenever a singular point is passed. If oscillation results, one should proceed without any sign change.

In the formulation of the present problem, it is assumed that $f_1(x)$ = Re g(z) and $f_2(x)$ = Im g(z). Thus the root-searching is along the trajectory ℓ defined by $f_1 = 0$. As such, the previous two theorems are stated in a manner compatible to this particular formulation. The same conclusion can also be drawn when the trajectory of $f_2 = 0$ (Im g(z) = 0) is used for searching all the solutions of a complex polynomial. A theorem similar to that of Theorem 3 has been proved in [8] for plotting the root locus of a feedback control system.

Starting Points

In order to prevent unnecessary search in finding multiple solutions, it is important to estimate the upper bound inside which all the roots of a polynomial will lie. Once the absolute bound is obtained, the search can then be confined to the circle of radius M. Several methods for computing such bounds are available [9]. One of such root bounds derived from the Gershgorin circle is given in the following theorem.

Theorem 5: If x^* is a solution of f(x) = 0 where $f = (f_1, f_2)^T$, f_1 = Re g(z) and f_2 = Im g(z), then

$$||x^*|| \leq M = \max\{|a_n|, 1+|a_k|, k=1,2,\ldots,n-1\} \quad (16)$$

where a_i's are the coefficients of the corresponding monic polynomial

$$g(z) = z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n. \quad (17)$$

As pointed out previously, the algorithm requires n starting points located on or close to each branch of ℓ defined by $f_1 = 0$. This can be accomplished with the aid of Theorem 5 and the properties of a polynomial for large z. It is easily shown that for large z, or $|z| >> M$, the trajectories of $f_1 = 0$ approach straight lines with constant phases

$$\theta_k = \frac{k\pi}{2n}, \quad k = 1,3,5, \ldots, 4n-1 \quad (18)$$

and the asymptotes for the nth-order monic polynomial (17) intersect at the centroid

$$z_c = -\frac{a_1}{n}. \quad (19)$$

The starting point can now be obtained from (18) and (19) as

$$z_o^k = z_c + Re^{j\theta_k}, \quad k = 1,3,5,\ldots,4n-1 \quad (20)$$

for an arbitrary $R >> M$.

Although there are 2n points in (20), only half of them are used. The other half are just the end points of each trajectory and they can easily be identified by checking their phases. The choice of R depends on the accuracy required for the starting point. Since the root bound M is known, R need not be much larger than M. A point on ℓ can usually be obtained quite accurately within a few steps from a rough estimate of (20) by using the Newton iteration

$$z_{k+1} = z_k - \text{Re } g(z)/g'(z) \qquad (21)$$

IV. NUMERICAL EXAMPLE

In this section an example is presented to illustrate the proposed algorithm. For simplicity only the Euler's method is used for illustration. In practice, more efficient integration techniques may be used to integrate the proposed equations.

Example:

A third order polynomial equation

$$g(z) = z^3 + 3z^2 + 28z + 26 = 0$$

is considered where all the singular points are located on one branch of Re $g(z) = 0$. The trajectories of both Re $g(z) = 0$ and Im $g(z) = 0$ are shown in the Figure. Trajectories of $f_1 = \text{Re } g(z) = 0$ are used to search for the solutions with the bound, $M = 29$ the centeroid, $z_c = -1$ and $R = 50$. Tracing along ℓ_{13} results in three solutions $z_1 = -1 + j5$, $z_2 = -1$ and $z_3 = -1 + j5$ while the other two branches, ℓ_{11} and ℓ_{12}, contain no solution at all.

V. CONCLUSION

A systematic search method has been developed for computing multiple solutions of a nonlinear equation with multi-branched solution curves. In particular, the approach has been applied to a class of functions derived from both the real part and the imaginary part of a complex polynomial. It turns out that the algorithm, expressed in its complex form, is most suited for finding all the roots of a polynomial. Analytic expressions for both the real and the imaginary parts of a polynomial are not required explicitly. The key to the continuation of the root-finding procedure at the singularities on the solution curve is the sign change associated with the numerical algorithm. It is shown that the sign must be changed whenever an even singular point is encountered along the search trajectory and no such change is allowed when passing through an odd singular point. Although the method is formulated in such a way as to find all solutions for a class of functions from R^2 into itself, it is conceivable that the approach may be generalized to higher-dimensional functions satisfying the conditions as described in Theorem 1 provided that a starting point for each branch of the solution curve can be determined.

REFERENCES

[1] K. M. Brown and W. B. Gearhart, "Deflation techniques for the calculation of further solutions of a nonlinear system," Numer. Math. Vol. 16, pp. 334-342, 1971.

[2] F. H. Branin, Jr., "Widely convergent method for finding multiple solutions of simultaneous nonlinear equations," IBM J. Res. Dev., vol. 16, No. 5, pp. 504-522, Sept. 1972.

[3] K. S. Chao and R. J. P. de Figueiredo, "Optimally controlled iterative schemes for obtaining the solution of a nonlinear equation," Int. J. Control, Vol. 18, pp. 377-384, 1973.

[4] K. S. Chao, D. K. Liu and C. T. Pan, "A systematic search method for obtaining multiple solutions of simultaneous nonlinear equations," IEEE Trans. Circuits Syst., Vol. CAS-22, pp. 748-753, Sept. 1975.

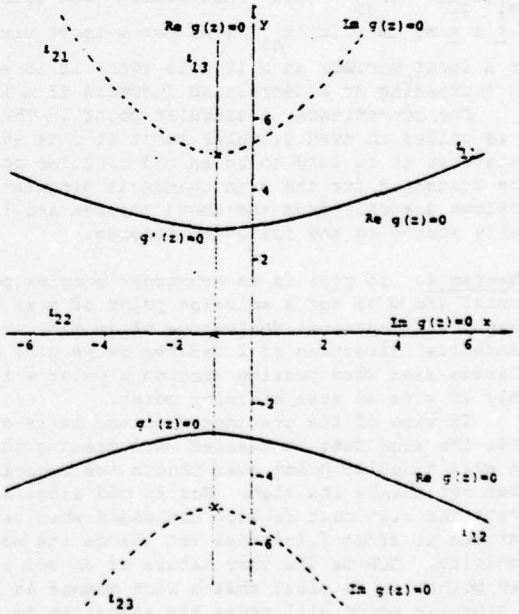
[5] L. O. Chua and A. Ushida, "A switching-parameter algorithm for finding multiple solutions of nonlinear resistive circuits," Circuit Theory & Applications, Vol. 4, 215-239 (1976).

[6] F. H. Branin, Jr., "A systematic method for finding all the roots of a polynomial," Proc. 19th Midwest Symp Circuits and Systems, pp. 59-62, Aug. 1976.

[7] F. H. Branin, Jr., "Poles and zeroes, eigenvalues of matrices and roots of polynomials by the method of signatures," Proc. 1977 IEEE International Symp. on Circuits and Systems, pp. 89-94, April 1977.

[8] C. T. Pan and K. S. Chao, "A computer-aided root-locus method," to appear in the IEEE Trans. on Automatic Control, Oct. 1978.

[9] A. Van der Sluis, "Upperbounds for roots of polynomials," Numer. Math. 15, pp. 250-262, 1970.



Figure

12. Reprint of "A Continuation Method for Finding the Roots of a Polynomial" by C. T. Pan and K. S. Chao from the Proceedings of the 22nd Midwest Symposium on Circuits and Systems, University of Pennsylvania, Philadelphia, June 1979, pp. 428-431.

Abstract

A continuation method is presented for finding all the roots of a polynomial. Each root is obtained systematically by numerical integration. Selection of starting points and the existence of singular points are discussed. Moreover, transformations may be applied to reduce the computational effort

1. INTRODUCTION

The solution for the roots of real or complex polynomials is a fundamental requirement in many areas of applied mathematics as well as in engineering. This is particularly so for the application of Laplace transform theory to linear autonomous systems. Although it is well-known that an nth-order polynomial has exactly n roots with multiplicity counted, the evaluation of all the roots is not at all a simple task. The difficulty is primarily due to its nonlinear nature.

The objective of this paper is to propose a systematic approach for finding all the roots of a polynomial. The algorithm is based on continuation methods [1] - [4]. The original polynomial is first embeded into a new equation by introducing a parameter r. This will result in n-branch root loci as r varies continuously. The root finding procedure then becomes a matter of tracing along these loci up to the desired roots..

2. THE CONTINUATION METHOD

Consider the problem of finding all the roots of the polynomial equation

$$P(s) = s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_{n-1} s + a_n = 0 \qquad (1)$$

where s is a complex variable and a_is are complex coefficients. It is well known [1], [2] that such a problem can be solved by using continuation methods. For example, the roots of (1) can be obtained by solving the continuation equation [1]

$$F(s,r) = (1-r)Q(s) + rP(s) = 0 \qquad (2)$$

where r is a positive real number and Q(s) is any nth-order polynomial whose n roots are known. It is seen that when r = 1, (2) reduces to (1), while for r = 0 (2) becomes

$$F(s,0) = Q(s) = 0 \qquad (3)$$

whose n roots are already known. Thus as r varies from zero to one continuously the trajectories of these roots comprise n branches of root loci. Each locus starts from a known root of Q(s) and terminates on a desired root of P(s). Therefore by tracing along these trajectories all roots of (1) can be located.

The advantage of this type of formulation

lies in the fact that one can easily obtain the approximate roots provided there exists no singular point on the root trajectory. It can therefore be used as a means for obtaining sufficiently close initial guesses for other methods which have rapid rate of convergence such as Newton's method. However, for any arbitrary polynomial equation $P(s) = 0$ and a given polynomial $Q(s)$, there is no guarantee that the root loci will not intersect each other. Unless singular points can be handled properly, one would have to try a different $Q(s)$. In what follows, an algorithm given in [4] for computing root-locus plot, and hence the solution of (2) is described. The problem of singularities on the trajectory will also be discussed.

Consider the set of differential equations

$$\frac{d}{dt}F(s(t),r(t)) = -F(s(t),r(t)), \qquad F(s(0),r(0)) = 0$$

$$\frac{d}{dt}r(t) = 1, \qquad r(0) = 0 \tag{4}$$

where t is a dummy variable. Application of the chain rule to (4) yields

$$\frac{ds}{dt} = \frac{F(s,r)+(\partial F/\partial r)}{\partial F/\partial s}, \quad s(0) = s_o$$

$$\frac{dr}{dt} = 1, \qquad r(0) = 0 \tag{5}$$

where s_o is a root of $Q(s)$. Equivalently, (5) can be rewritten as

$$\frac{ds}{dt} = \frac{-rQ(s)+(1+r)P(s)}{(1-r)Q'(s)+rP'(s)}, \quad s(0) = s_o$$

$$\frac{dr}{dt} = 1 \qquad r(0) = 0 \tag{6}$$

or

$$\frac{ds}{dt} = \frac{-tQ(s)+(1+t)P(s)}{(1-t)Q'(s)+tP'(s)}, \quad s(0) = s_o, \quad t\in[0,1] \tag{7}$$

where Q' and P' denote the derivatives of Q(s) and P(s) with respect to s, respectively. Equation (7) can now be integrated by using any numerical technique until t=1 is reached. This will result a root-locus plot for $0\le t\le 1$ which contains n branches. On the root locus, it is possible that the denominator of (7) may become zero. A point s* such that

$$D(s) \triangleq (1-r)Q'(s^*)+rP'(s^*) = 0, \tag{8}$$

is called a singular point. From (8) it is obvious that there can be at most (n-1) isolated singular points located on the trajectories. This corresponds to the situation that more than one root loci intersect at s*. Singular points can be classified according to their higher order derivatives. An qth-order singular point is defined as a singular point such that

$$\left.\frac{d^k r}{ds^k}\right|_{s=s^*} = 0, \quad k = 1,2,\ldots q-1$$

$$\left.\frac{d^q r}{ds^q}\right|_{s=s^*} \neq 0, \quad 2\le q\le n. \tag{9}$$

Depending on whether q is odd or even, singular points can further be classified as odd or even singular points. It is shown in [4] that whenever an odd singular point is encountered on the trajectory, it is necessary to jump over it by adding a small variation $|\Delta s|$ along the tangential direction of the locus: for an even singular point, the direction of the locus at s* must be changed by a vector

$$\Delta z = \Delta s\, e^{-j\pi/q} \tag{10}$$

where Δs is a sufficiently small variation in the tangential direction of the locus when approaching s*.

3. STARTING ROOTS

The proposed method requires n starting roots of $Q(s) = 0$. Theoretically Q(s) can be any nth-order polynomial so long as its roots are known. For practical purpose, Q(s) should be as simple as possible. The choice of

$$Q(s) = s^n - M^n \tag{11}$$

is made where M is assumed to be positive

and real. The n corresponding roots are

$$s_{ok} = M \exp(j2\pi k/n), \quad k = 0,1,2,\ldots,n-1. \tag{12}$$

In order to reduce unnecessary computational effort M should be chosen properly. Hence it is important to estimate the bound inside which all the roots may lie. One such bound is given in the following theorem [5].

Theorem. Let

$$h(s) = s^n + a_1 s^{n-1} + \ldots + a_{n-1}s + a_n$$

be a monic polynomial. If s^* is a root of $h(s) = 0$, then $|s^*| \leq N$ where

$$N = \max\{|a_n|, 1 + |a_j|, j = 1,2,\ldots n-1\}.$$

From the above theorem, it is perhaps logical that M should not be chosen to be greater than N. In view of the fact that N is usually much larger than the least upper bound for the roots, two transformations given in [6] may be used to modify (2). Using the transformation

$$y = s + \frac{a_1}{n} \tag{13}$$

equation (2) reduces to

$$y^n + b_2 y^{n-2} + \ldots + b_n = 0. \tag{14}$$

A second transformation

$$y = \sqrt[n]{(b_n)}\ z \tag{15}$$

is then used to convert (14) into a polynomial equation in z. The roots in the transformed z-plane are more uniformly distributed, and as a result, the computational effort may be reduced considerably.

4. EXAMPLE

As an illustration of the approach presented, consider the polynomial equation

$$P(s) = s^3+(1-j3)s^2+(23+j32)s+(-37-j185)$$

which has three known roots at $s_1=3+j2$, $s_2=-5+j7$ and $s_3 = 1-j6$. The root bound $N=188.7$ Application of (7) and (12) using M=5 along with a fourth order Range Kutta method with a step size of 0.005 results in three roots

$$s_1^* = 2.997 + j2.001$$

$$s_2^* = -5.001 + j6.998$$

$$s_3^* = 1.005 - j6.000.$$

From this example it is obvious that N is much larger than the least upper bound for the roots. After using transformations (13) and (15), the normalized equation becomes

$$z^3+(0.780729+j1.03421)z$$

$$-(0.4169435+j0.908944) = 0.$$

It is seen that the new root bound $N'=2.296$. Application of the proposed method yields three roots in the z-plane

$$z_1 = 0.58136 + j0.17441$$

$$z_2 = -0.8139 + j1.04643$$

$$z_3 = 0.23255 - j1.22085$$

where the same integration technique with a step size of 0.02 is used. Transforming back to the s-plane gives

$$s_1 = 2.99999 + j2.00000$$

$$s_2 = -5 + j6.99994$$

$$s_3 = 1.00004 - j5.99996.$$

The root loci in both the s-plane and the z-plane are shown in Figs. 1 and 2, respectively.

5. CONCLUSION

A continuation method is presented for finding all the roots of a polynomial. Singular points along the trajectory can be properly handled so that the root-finding procedure can be continued for any given imbedding polynomial.

REFERENCES

[1] J. M. Ortega and W. C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[2] E. Wasserstrom, "Numerical solutions by the continuation method," SIAM Review, Vol. 15, No. 1, pp. 89-119, January 1973.
[3] K. S. Chao and R. Saeks, "Continuation methods in circuit analysis," Proc. IEEE, Vol. 65, No. 8, pp 1187-1194, Aug., 1977.
[4] C. T. Pan and K. S. Chao. "A computer-aided root-locus method," IEEE TRANS. ON AUTOMATIC CONTROL, Vol. AC-23, No. 5, pp. 856-860, October, 1978.
[5] Morris Marden, The Geometry of the Zeros of a Polynomial in a Complex Variable, American Mathematical Society, New York, 1949.
[6] C. T. Chen and N. R. Straden, "A normalized multidimensional Newton-Raphson method, Int. J. Control, Vol. 12, No. 2, pp. 273-279, 1970.

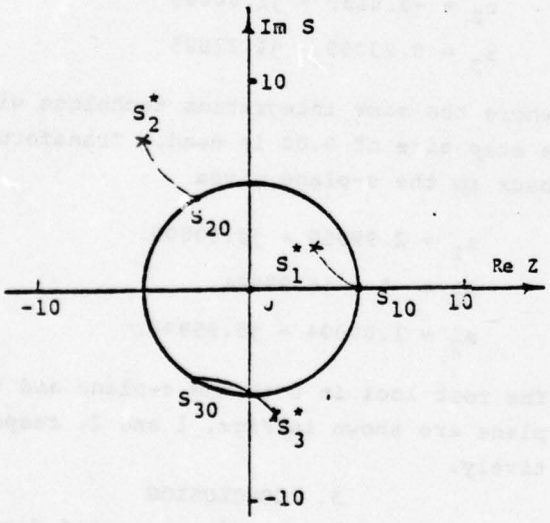


Fig. 1

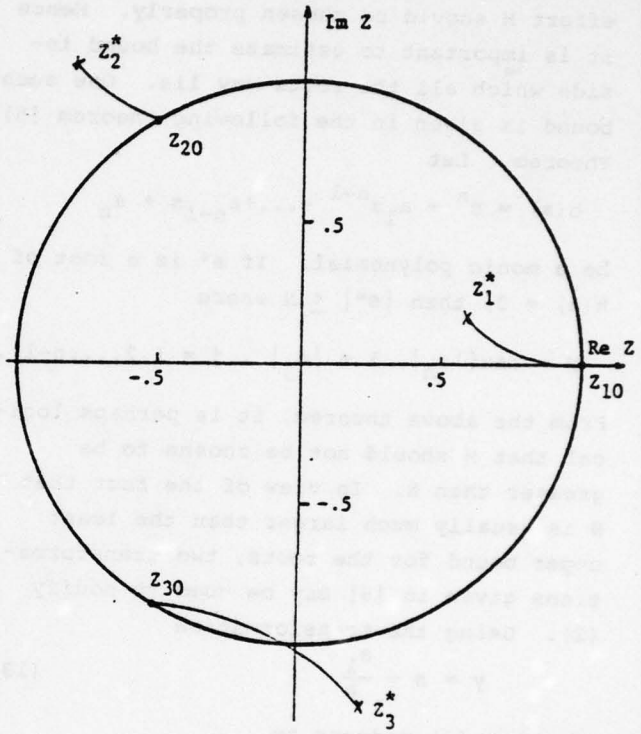


Fig. 2

14. Abstract of "Continuation Algorithms for the Solution of the Eigenvalue Problem", by B. Green.

A continuations algorithm for tracking the eigenvalues of a large sparse system of equations is presented. This is achieved by constructing a family of similarity transformations which triangularize the given system as a function of the underlying parameter. Since both the resultant triangular matrix and the similarity transformations themselves retain the sparseness of the given system of equations, the resultant algorithm proves to be quite efficient when applied to our large scale system problems.

Texas Tech University — Institute for Electronic Science

Joint Services Electronics Program — Research Unit: 5

1. Title of Investigation: Multidimensional System Theory
2. Senior Investigator: J. Murray — Telephone: (806) 742-3506
3. JSEP Funds: $23,500
4. Other Funds:
5. Total Number of Professionals: PI's 1 (3 mo.) RA's _____
6. Summary:

The goal of the work unit is the application of the techniques of the theory of functions in several complex variables to several problems in circuit and system theory which are modeled by rational functions in two or more complex variables. Possibly the most important of these is the analysis and design problem for multidimensional digital filters in which a multidimensional z-transform is employed. The investigation, however, also includes a study of the stability problem for mixed lumped/transmission line systems and a study of the multivariable passive synthesis problem.

Our major activity during the past year has been an investigation of the design problem for two-dimensional digital image processing filters. Since the filter design problem has historically been inextricably intermingled with the spectral factorization problem, this study began with an investigation of the fundamental limitations on the existence of spectral factors for two-dimensional transfer functions. Since the resulting conditions for the existence of a quarter-plane stable spectral factorization proved to be extremely stringent, we turned our attention to the design of half-plane stable digital filters. These filters have far less stringent existence conditions and we

are presently developing a general purpose design procedure for such filters.

In an effort to alleviate the need for artificially imposing any "causality" structure on the image processing problem, we have also initiated a study of the class of periodically varying discrete-time systems which naturally model the actual scanning process used in a "real world" image processing system. Although these systems are time-varying, they represent the only known class of time-varying systems which admit a "viable" frequency domain theory. As such, we believe that it will be possible to formulate a viable frequency domain theory for two-dimensional image processing filters in terms of the physical scanning process actually employed, thereby alleviating many of the difficulties hitherto encountered in two-dimensional filtering theory, which are actually due to the artificiality of the model rather than the physical problem.

The above work is represented by two reprints; a paper which appeared in the IEEE Transactions on Circuits and Systems, which summarizes our study of the existence criteria for two-dimensional spectral factors and a reprint of a conference paper describing the frequency domain theory for periodically varying discrete time-systems and some of its generalizations.

7. Publications and Activities:

A. Refereed Journal Articles

1. Murray, J., "Spectral Factorization and Quarter-Plane Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-25, pp. 586-592, (1978).

B. Conference Papers and Abstracts

1. Murray, J., "Semidirect Products and the Stability of Time-Varying Systems", Proc. of the Inter. Symp. on the Mathematics of Networks and Systems, Vol. 3, T.H. Delft, Delft, July 1979, pp. 121-125.

2. Saeks, R., and J. Murray, "Stability and Homotopy II", Proc. of the 1979 Joint Automatic Control Conf., Denver, June 1979, p. 358, (abstract only).

C. Conferences and Symposia

1. Murray, J., 12th Asilomar Conf. on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1978.

2. Murray, J., 1979 Inter. Symp. on the Mathematics of Networks and Systems, T.H. Delft, Delft, July 1979.

3. Murray, J., 1979 Joint Automatic Control Conf., Denver, June 1979.

8. Reprint of "Spectral Factorization and Quarter-Plane Digital Filters" by J. Murray from the IEEE Transactions on Circuits and Systems, Vol. CAS-25, pp. 586-592, (1978).

Spectral Factorization and Quarter-Plane Digital Filters

JOHN J. MURRAY

Abstract—Two sets of necessary conditions are derived for the existence of a rational spectral factorization of a given rational function of two complex variables: partial converses of these results are given, and the implications of these conditions for the design of minimum-phase FIR filters and stable IIR filters are discussed. In particular, it is shown that these conditions are closely related to the difficulties encountered in the stabilization problem for two-dimensional IIR filters.

Introduction

THE SUBJECT of two-dimensional digital filters has received considerable attention of late: in particular, two-dimensional spectral factorization has been treated in a number of papers—it is considered in great detail in [1]. The major problem which arises is that, in general, the spectral factors of a rational transfer function are not rational: some further processing, such as truncation and smoothing, is usually employed to yield approximate rational factors. It is, therefore, somewhat surprising that the class of rational functions for which a rational spectral factorization exists does not seem to have been investigated. In this paper, we give two sets of conditions which must be satisfied by such functions (Theorems 1 and 3); a converse is given which may be applied to the numerator and denominator polynomials separately. Now, the polynomial spectral factors (when they exist) of a given polynomial are minimum- and maximum-phase polynomials; conversely, every such polynomial gives rise to trivial spectral factors. Motivated by this, we apply the results of Theorems 1 and 3 to the particular case of minimum-phase polynomials (i.e., polynomials without zeros in the unit polydisc).

In this context, the main consequences of the results of this paper may be broadly outlined as follows:

i) A given polynomial has *exactly* the same amplitude response as a minimum-phase polynomial if *and only if* the classical one-variable method (of factoring the original polynomial into a product of two polynomials devoid of zeros in certain regions) can be applied. (This result is in fact implicit in [1], but does not appear to have been explicitly stated in the literature.) The corresponding statement for minimum-phase stable rational functions is false, however.

ii) If the conditions given in Theorems 1 and/or 3 are not satisfied, then not only is there no minimum-phase, stable rational function having exactly the same amplitude response as the original, but the original amplitude response can not even be approximated arbitrarily well by minimum-phase stable rational functions. This follows from the fact that the conditions in Theorems 1 and 3 are conditions on the amplitude response which are preserved under any reasonable kind of convergence.

iii) The conditions in Theorem 3 are easily visualized and surprisingly stringent: they require essentially that the gain of the filter, averaged over certain directions in the frequency plane, have *no* variation in a perpendicular direction. (See the discussion following Theorem 3.) This gives extremely severe restrictions on the amplitude response of minimum-phase FIR filters, minimum-phase stable IIR filters, and the denominator polynomial of arbitrary stable IIR filters.

iv) It has been pointed out by Bose [9] and Woods [10], and again is implicit in [1], that there exist purely recursive filters whose amplitude responses are not realizable as the amplitude response of any *stable* purely recursive filter, and that consequently any stabilization method which attempts to match the amplitude response of the original filter is doomed to failure. The restrictions referred to in iii) reinforce this conclusion and identify the precise properties of the examples in [9] and [10] which make stabilization impossible.

Definitions and Notation

Our notation will follow that used in [2]; we repeat it here for convenience. For simplicity we restrict ourselves throughout to two dimensions, although there does not appear to be any difficulty in extending the results to higher dimensions. Thus all functions are assumed throughout to be rational functions of two complex variables unless otherwise stated; we further exclude the zero function. Two-dimensional complex space will be denoted by C^2, i.e., $C^2 = \{(Z_1, Z_2) | Z_1 \text{ and } Z_2 \text{ are complex numbers}\}$. The open unit polydisc will be denoted by U^2, i.e.,

$$U^2 = \{(Z_1, Z_2) \in C^2 | |Z_1| < 1 \text{ and } |Z_2| < 1\}$$

and its closure will be denoted by $\overline{U^2}$:

$$\overline{U^2} = \{(Z_1, Z_2) \in C^2 | |Z_1| \leq 1 \text{ and } |Z_2| \leq 1\}.$$

Manuscript received May 17, 1977. This work was supported by the Air Force Office of Scientific Research under Grant 74-2631.

The author is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

0098-4094/78/0800-0586$00.75

The distinguished boundary of the unit polydisc will be denoted by T^2:

$$T^2 = \{(Z_1, Z_2) \in \mathbb{C}^2 \mid |Z_1| = 1 \text{ and } |Z_2| = 1\}.$$

The frequency response of the filter whose transfer function is $f(Z_1, Z_2)$ is simply the restriction of f to T^2. We will find it convenient to denote this restriction by f^*.

The one-dimensional sets corresponding to the above are

$$U = \{Z \in \mathbb{C} \mid |Z| < 1\}$$
$$\bar{U} = \{Z \in \mathbb{C} \mid |Z| \leq 1\}$$
$$T = \{Z \in \mathbb{C} \mid |Z| = 1\}.$$

We need one further subset of $\mathbb{C}^2$:

$$V^2 = \{(Z_1, Z_2) \in \mathbb{C}^2 \mid |Z_1| > 1 \text{ and } |Z_2| > 1\}.$$

By the Fourier coefficients of a function $h(\theta_1, \theta_2)$ defined on T^2 we mean the numbers

$$a_{mn} = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} h(\theta_1, \theta_2) e^{-j(m\theta_1 + n\theta_2)} \, d\theta_1 \, d\theta_2.$$

Finally, let us state precisely what we mean by the term spectral factorization. Several different forms of spectral factorization are treated in [1]; here we will be concerned only with the simplest form: if f is a rational function, it will be said to have a (rational, quarter-plane) spectral factorization if $f = f_1 f_2$, where f_1 and f_2 are rational functions, f_1 has no poles or zeros in U^2, and f_2 has no poles or zeros in V^2. Several comments are in order concerning this definition:

i) By "rational" we mean only "finite-order;" i.e., the functions are assumed to be expressible as the quotient of two (finite-order) polynomials.

ii) The quarter-plane property enters only in connection with the regions in which the factors are assumed to be zero- and pole-free: in particular, if f has no poles or indeterminacies on T^2, and has a quarter-plane spectral factorization, then there is a quarter-plane causal, stable filter whose amplitude response is equal to $|f^*|$.

iii) It would possibly be more natural to work with $\bar{U}^2$ and $\bar{V}^2$ rather than U^2 and V^2 (especially when considering stability). However, to do so would complicate the statements of the theorems considerably, and it is usually clear whether or not the results will hold with $\bar{U}^2$ and $\bar{V}^2$ in place of U^2 and V^2. (One needs only to check for zeros and poles on T^2). In general, if the "closed" version is not obvious, it is not true: $1 - Z_1 Z_2$ will serve as a counterexample in all such cases.

iv) To simplify the statements of the theorems, the definition has been given in terms of the rational function f itself, rather than the spectral function $|f^*|^2$; however, the conditions given in the theorems actually involve only $|f^*|^2$.

v) We note that V^2 is defined to be a subset of $\mathbb{C}^2$: thus the behavior of functions at infinity is irrelevant to our purposes.

Spectral Factorization

Our first criterion for the existence of rational spectral factors is very much in the spirit in which spectral factorization is treated in [1]; it is a trivial consequence of Theorem 5.4.7 in [2].

Theorem 1

If a rational function f on $\mathbb{C}^2$ has a rational spectral factorization, then the Fourier coefficients a_{mn} of $\log |f^*|$ are zero for all pairs of integers (m,n) such that $m \neq 0$, $n \neq 0$, and m and n have different signs—that is, for all integer points in the second and fourth quadrants. The converse is true for polynomial f.

As mentioned above, this criterion involves only the absolute value of f; it follows that the existence of spectral factors imposes restrictions on the amplitude response of a two-dimensional filter—in contrast with the situation in one dimension. The above criterion, however, does not present these restrictions in an easily visualized form. For instance, it is difficult to gauge exactly how severe the restrictions are. For this reason, we next present conditions which are stated in terms of the log-amplitude response itself, rather than its Fourier coefficients. This result takes an approach which seems to differ substantially from those previously known; it gives easily visualized necessary conditions on those rational functions which admit a rational spectral factorization. Before we state this theorem, however, we first present a simple result which will be used in the proof, and is also of separate interest; one of its consequences is that when rational spectral factors exist, the usual one-dimensional stabilization method (for unstable denominator polynomials) can be used.

Theorem 2

If the rational function f admits a rational spectral factorization, then there is a rational function $\tilde{f}$ (with $\deg \tilde{f} \leq \deg f$) such that

$$|\tilde{f}^*| = |f^*|$$

and $\tilde{f}$ has no poles or zeros in U^2.

Again, the converse holds for polynomial f.

Thus if the denominator polynomial of an unstable filter has polynomial spectral factors, there is a stable filter of at most the same order with the same amplitude response (provided the polynomial has no zeros on T^2).

Again, most of the proof is contained in [2]; we fill in the details here: suppose f has rational spectral factors, then $f = f_1 P / Q$, where f_1 has no poles or zeros in U^2, and P and Q are polynomials without zeros in V^2. Let

$$\tilde{P} = Z_1^m Z_2^n \bar{P}(1/Z_1, 1/Z_2), \qquad Z_2 \neq 0, \quad Z_1 \neq 0$$

where m is the degree of P in Z_1, n is the degree of P in Z_2, and $\bar{P}$ is the polynomial whose coefficients are the complex conjugates of the coefficients of P. Clearly $\tilde{P}$ is a polynomial of degree less than or equal to the degree of P, and so is also defined for $Z_1 = 0$ and $Z_2 = 0$. Now if

$\tilde{P}(Z_1, Z_2) = 0$, for $Z_1 \neq 0$ and $Z_2 \neq 0$: then $\tilde{P}(1/Z_1, 1/Z_2) = 0$; this implies that either

$$|1/Z_1| < 1 \text{ or } |1/Z_2| < 1 \text{ (since } P \text{ has no zeros in } V^2)$$

and so either

$$|Z_1| > 1 \text{ or } |Z_2| > 1$$

i.e.,

$$(Z_1, Z_2) \notin U^2.$$

Thus the only possible zeros of $\tilde{P}$ in U^2 are for $Z_1 = 0$ or $Z_2 = 0$. But by standard results in the theory of several complex variables [8], if the zero-set were nonempty, this would imply that either Z_1 or Z_2 was a factor of $\tilde{P}$, which is impossible by our choice of m and n. Thus $\tilde{P}$ has no zeros in U^2. Finally, on T^2

$$|\tilde{P}(Z_1, Z_2)| = |Z_1^m Z_2^n \bar{P}(1/Z_1, 1/Z_2)| = |\bar{P}(\bar{Z}_1, \bar{Z}_2)| = |P(Z_1, Z_2)|.$$

$\tilde{Q}$ is defined similarly and has similar properties. Then

$$\tilde{f} = f_1 \frac{\tilde{P}}{\tilde{Q}}$$

clearly has the required properties.

Conversely, suppose f is any polynomial for which there is a rational function $\tilde{f}$ without poles or zeros in U^2 such that

$$|\tilde{f}^*| = |f^*|$$

then $f/\tilde{f}$ is rational and analytic in U, and

$$|(f/\tilde{f})^*| = 1.$$

Thus by Theorems 5.2.5 and 5.2.6 in [2], $f/\tilde{f} = P/Q$, where P and Q are polynomials, P has no zeros in V^2, and Q has no zeros in U^2. Then

$$f = P\tilde{f}/Q$$

gives a rational (in fact, polynomial) spectral factorization of f.

The Second Criterion

Our second set of conditions for the existence of a rational spectral factorization is given in the following.

Theorem 3

If a rational function f on $\mathbb{C}^2$ admits a rational spectral factorization, then

$$\frac{1}{2\pi}\int_0^{2\pi} \log |f(e^{j m\theta}, e^{j(n\theta+\psi)})|\, d\theta$$

is a constant independent of ψ, $(0 \leq \psi < 2\pi)$, for all integers $m > 0$ and $n > 0$.

Again, these conditions depend only on the amplitude response of f. The simplest condition is that for $m = 1$ and $n = 1$: it can be easily visualized by drawing two adjacent squares in the $\theta_1\theta_2$ plane on which the amplitude response is defined (the frequency response extends to the entire $\theta_1\theta_2$ plane by periodicity), and drawing lines L_t with slope 1 and length $2\pi\sqrt{2}$ on these squares (see Fig. 1).

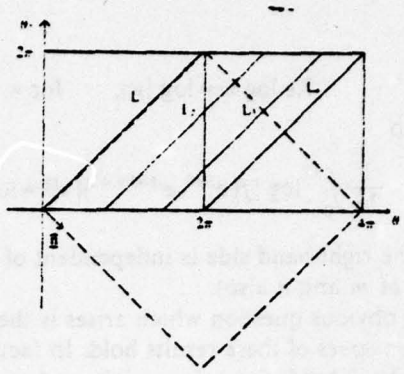


Fig. 1.

Then the condition for $m = 1$, $n = 1$ can be restated as: the "average" amplitude of the function f along the line L_t, is a constant—that is, it is independent of the particular line L_t chosen. ("Average" here is to be understood as the geometric mean of the amplitude, or the arithmetic mean of the log-amplitude). Alternatively, we may say that the average level of the amplitude over any line of slope 1 and of length $2\pi\sqrt{2}$ is independent of the position of the line in the $\theta_1\theta_2$ plane. (For example, we could vary the L_t over the dotted square in the direction $\hat{n}$.) The conditions for higher m and n have a similar interpretation, with a slope of n/m instead of 1, and length $2\pi\sqrt{m^2+n^2}$ instead of $2\pi\sqrt{2}$; clearly, if m and n are not relatively prime, the corresponding condition is superfluous.

This theorem then gives a striking limitation on the amplitude response of a rational function which admits a rational spectral factorization: even the simplest of the conditions (that for $n = m = 1$) implies that such a function cannot accurately approximate an amplitude which has large variations in overall level in the direction $\hat{n}$ shown in Fig. 1.

Proof: In view of Theorem 2, it suffices to prove this under the assumption that f has no poles or zeros in U^2. This assumption implies that f has a holomorphic logarithm in U^2. Then, for any integers $m > 0$, $n > 0$, and any real number ψ

$$\log f(Z^m, Z^n e^{j\psi})$$

is a holomorphic function of one complex variable for $Z \in U$. Thus

$$\mathrm{Re}(\log f(Z^m, Z^n e^{j\psi}))$$

is a harmonic function in U, and so by the mean-value property of harmonic functions

$$\frac{1}{2\pi}\int_T \mathrm{Re}(\log f(Z^m, Z^n e^{j\psi}))\, d\theta = \mathrm{Re}(\log f(0^m, 0^n e^{j\psi}))$$

i.e.,

$$\frac{1}{2\pi}\int_0^{2\pi} \mathrm{Re}(\log f(e^{j m\theta}, e^{j(n\theta+\psi)}))\, d\theta = \mathrm{Re}(\log f(0,0))$$

but

$$\text{Re } \log w = \log |w|, \qquad \text{for } w \neq 0$$

and so

$$\frac{1}{2\pi}\int_0^{2\pi} \log |f(e^{jm\theta}, e^{j(n\theta+\psi)})|\, d\theta = \log |f(0,0)|$$

and the right-hand side is independent of ψ (and, incidentally, of m and n also).

An obvious question which arises is the extent to which the converses of these results hold. In fact, the converse of Theorem 3 holds for polynomials, and modified converses of both Theorems 1 and 3 hold even for rational functions. The modification takes the following form: if the Fourier coefficients of $\log |f^*|$ (where f is a rational function) vanish for $mn < 0$, then there is a rational function $\tilde{f}$ with rational spectral factors (equivalently, a rational function without poles or zeros in U^2), such that $|\tilde{f}^*| = |f^*|$. (A similar statement holds for Theorem 3.) However, the proofs of these converses involve some technical analytic details, and so are given in the Appendix.

The modification in the above converses lies, of course, in the fact that we cannot conclude that f itself has rational spectral factors: thus there are some rational functions which can be stabilized without changing the amplitude response but to which the classical 1-variable factorization technique cannot be applied. A simple example of this is the function

$$f(Z_1, Z_2) = \frac{Z_1 + Z_2 - 1}{Z_1 + Z_2 - Z_1 Z_2}.$$

Here, $|f^*|$ is identically 1, and so has trivial spectral factors; but f itself clearly does not.

Although the converses of Theorems 1 and 3 are proved in the Appendix, there is another result related to the converse of Theorem 3: by strengthening the condition for $m = n = 1$ alone, we can get a stronger converse for polynomials. Before we state this converse, however, we first give a stability criterion (used in the proof of the converse) which, although previously known [3], has not appeared in the engineering literature. Although not as sharp (in terms of dimension) as some other known criteria [4], it has two advantages which make it useful for theoretical purposes: first, it is given in terms of a one-parameter family of discs without the lower dimensional test in [5]; and second, unlike most other stability tests, which conclude the nonvanishing of a polynomial on $\overline{U}^2$ from its nonvanishing on some subset of $\overline{U}^2$ which contains T^2, this test allows the polynomial to vanish at some points in T^2, but concludes only that the polynomial does not vanish on U^2. The criterion is the following.

Theorem 4

Suppose a polynomial f has no zeros in the set

$$\{(Z_1, Z_2) \in U^2 \mid |Z_1| = |Z_2|\}$$

then f has no zeros in U^2.

This is proved in a much more advanced context in [3]; however, it can also be easily proved by applying one of the criteria in [4] to the polydiscs

$$\overline{U}_r^2 = \{(Z_1, Z_2) \in \mathbb{C}^2 \mid |Z_1| \leq r, |Z_2| \leq r\}, \qquad \text{for } 0 < r < 1.$$

The hypotheses imply that f has no zeros on the distinguished boundary $\overline{U}_r^2$, for $0 < r < 1$, and none on the set

$$\{(Z_1, Z_2) \in C^2 \mid Z_1 = Z_2\} \cap \overline{U}_r^2.$$

Thus by Theorem 5 in [4], f has no zeros in $\overline{U}_r^2$, for any $r < 1$, and so f has no zeros in U^2.

We can now state and prove the partial converse to Theorem 3.

Theorem 5

If f is a polynomial with the property that

$$\frac{1}{2\pi}\int_0^{2\pi} \log |f(e^{j\theta}, e^{j(\theta+\psi)})|\, d\theta = \log |f(0,0)|,$$

$$\text{for } 0 \leq \psi < 2\pi$$

then f has no zeros in U^2.

Thus we strengthen the condition for $m = 1$ and $n = 1$ in Theorem 3 by specifying that the constant in question is to be $\log |f(0,0)|$: it then follows not only that f has rational spectral factors, but that it is actually zero-free in U^2.

Proof: By Theorem 4, it suffices to prove that f has no zeros in the set

$$\{(Z_1, Z_2) \in U^2 \mid |Z_1| = |Z_2|\}$$

but this set is the union of the open discs

$$\{(Z_1, Z_2) \mid Z_2 = e^{j\psi} Z_1, |Z_1| < 1\}, \qquad \text{for } 0 \leq \psi < 2\pi.$$

We therefore wish to prove that f has no zeros in any of these discs: or equivalently that the function f_ψ of one variable defined by $f_\psi(Z) = f(Z, Ze^{-j\psi})$ has no zeros in the open unit disc. Applying Jensen's formula [6, p. 299] for the unit disc to f_ψ, we get

$$\frac{1}{2\pi}\int_0^{2\pi} \log |f_\psi(e^{j\theta})|\, d\theta = \log |f_\psi(0)| - \sum \log |Z_i|$$

where the summation is over all the zeros (counted with multiplicity) of f_ψ in the unit disc. Expressing this in terms of f:

$$\frac{1}{2\pi}\int_0^{2\pi} \log |f(e^{j\theta}, e^{j(\theta+\psi)})|\, d\theta = \log |f(0,0)| - \sum \log |Z_i|$$

and so $\sum \log |Z_i| = 0$.

Since for any Z_i in the open unit disc $\log |Z_i| < 0$, the conclusion follows. (It is clear from the proof that we always have

$$\frac{1}{2\pi}\int_0^{2\pi} \log |f(e^{j\theta}, e^{j(\theta+\psi)})|\, d\theta \geq \log |f(0,0)|.$$

It follows from this that in fact the apparently weaker condition

$$\frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}\log|f(e^{j\theta_1},e^{j\theta_2})|\,d\theta_1\,d\theta_2=\log|f(0,0)|$$

is sufficient to guarantee that f is zero-free in U^2. See [2, p. 73].)

Stable IIR Filters and Minimum-Phase FIR Filters

The very close relationship of spectral factorization to the nonvanishing of polynomials in U^2, and thereby to stable IIR filters (via the denominator polynomial) and minimum-phase filters (via the numerator polynomial) is already clear from the previous sections. The force of Theorem 2 is that purely from the point of view of amplitude response, transfer functions having rational spectral factors are equivalent to those whithout poles or zeros in U^2. Thus the restrictions on amplitude response in Theorems 1 and 3 apply to the denominator polynomial of any stable IIR filter; the contribution of the denominator polynomial to the overall amplitude response of the filter (in the case of an all-pole filter, the entire amplitude response) must satisfy the restrictions imposed by Theorems 1 and 3. We have, therefore, identified the properties of the amplitude response which make it impossible to stabilize a filter; if the original amplitude response has large overall variation in the "wrong" directions, attempting to find a stable filter which closely matches this response is futile. Close matching of the amplitude *forces* instability. This has already been shown by example by Bose [9] and Woods [10]: we now see that it is the variations in the amplitude response in the "wrong" directions in their examples which account for their behavior.

It is also of interest to note that, in the Shanks procedure of minimizing

$$\int\int|fg-1|^2\,d\theta_1\,d\theta_2$$

over all polynomials f of given degree (where g is the original polynomial), if the allowable f's were restricted to those which have polynomial spectral factorizations, the procedure would yield a polynomial devoid of zeros in U^2. It does not appear that this observation can be used as the basis for a workable stabilization method, however, since the condition that f have polynomial spectral factors is intractably nonlinear in the coefficients of f: and further, in many cases this procedure would yield an f which was only marginally stable. For the same reasons, restricting oneself throughout the design procedure to polynomials which satisfy the condition in Theorem 5 does not appear to be a feasible method of ensuring stability.

Examples and Comments

An example of the behavior of those polynomials not possessing polynomial spectral factors has already appeared in the literature, although in a different context; we repeat this example here:

$$A(Z_1,Z_2)=1-0.75Z_1+0.9Z_1^2+1.5Z_2-1.2Z_1Z_2+1.3Z_1^2Z_2+1.2Z_2^2+0.9Z_1Z_2^2+0.5Z_1^2Z_2^2.$$

This polynomial was studied in [7]; the associated Shanks polynomial was found to be stable but to have a substantially different amplitude response from that of A (for more details, see [7]). The fact that A does not have polynomial spectral factors was established by checking the condition in Theorem 3, for $m=n=1$ and $\psi=0$, $\psi=\pi$, with the following results (correct to nine decimals):

$$\frac{1}{2\pi}\int_0^{2\pi}\log|A(e^{j\theta},e^{j\theta})|\,d\theta=0.696570700$$

$$\frac{1}{2\pi}\int_0^{2\pi}\log|A(e^{j\theta},e^{j(\theta+\pi)})|\,d\theta=1.134686936.$$

As an example of a polynomial with rational spectral factors, we have

$$B(Z_1,Z_2)=1+2.25Z_1+2.25Z_2+0.5Z_1^2+0.5Z_2^2-6.5Z_1Z_2-Z_1^2Z_2-Z_1Z_2^2-4Z_1^2Z_2^2.$$

This factors into $(1+0.25Z_1+0.25Z_2+0.5Z_1Z_2)(1+2Z_1+2Z_2-8Z_1Z_2)$, where the first factor has no zeros in U^2, the second has none in U^2: reversing the second factor gives a polynomial without zeros in U^2:

$$\tilde{B}(Z_1,Z_2)=(1+0.25Z_1+0.25Z_2+0.5Z_1Z_2)\cdot(-8+2Z_2+2Z_1+Z_1Z_2)$$
$$=-8-2Z_1Z_2+0.5Z_1^2+0.5Z_2^2+1.25Z_1^2Z_2+1.25Z_1Z_2^2+0.5Z_1^2Z_2^2$$

and $\tilde{B}$ has the same amplitude response as B.

In order to gain some idea of the stringency of the conditions in Theorem 3, let us consider the case of an ideal bandpass filter. By an ideal bandpass filter we mean a filter whose amplitude response is equal to 1 on some subset, A, of the square $0\le\theta_1<2\pi$, $0\le\theta_2<2\pi$, and equal to $K\ll 1$ on the complement of A (of course this specification continues over the whole plane by periodicity). This of course is not the amplitude response of any rational function, but in practice for certain shapes of the set A, one may wish to approximate such a response by a rational function. One easily sees that up to a scale factor, the averages in Theorem 3 are in this case merely the fraction

$$\frac{\text{length of the line } L_i \text{ lying in the complement of } A}{\text{total length of the line } L_i}.$$

It is easily seen from this that there are very few passband shapes of practical interest which satisfy even the first of these conditions (where $n=1$ and $m=1$); in other words, there are very few which can be accurately approximated by transfer functions having rational spectral factors. (This is not to imply that one would in practice be restricted to such filters: the above discussion is meant solely as an indication of the severity of the restrictions on the amplitude of such filters.)

Finally, we remark that there does not seem to be any difficulty in extending the results in this paper to higher dimensions, and to multidimensional systems other than digital filters.

APPENDIX

The Converses to Theorems 1 and 3

These converses involve some technical ideas and results from [2]; the most important ideas are those of inner function [2, p. 105], outer function [2, p. 72], Poisson integral [2, p. 17], and the classes $N(U^2)$ [2, p. 44] and $N_+(U^2)$ [2, p. 44].

We will also use the following notation from [2] (here f is an analytic function on U):

i) $$f^*(e^{j\theta_1}, e^{j\theta_2}) \triangleq \lim_{r\to 1^-} f(re^{j\theta_1}, re^{j\theta_2})$$

will denote the radial limit of f (this is clearly consistent with our previous use of f^*).

ii) For $w=(w_1,w_2)\in T^2$, $f_w(Z)$ will denote the one-variable function defined by

$$f_w(Z) \triangleq f(Zw_1, Zw_2).$$

iii) If ϕ is a function defined on T^2 which is absolutely integrable there:

$$\hat{\phi}(m,n) \triangleq \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} \exp(-jm\theta_1 - jn\theta_2)\phi(\theta_1,\theta_2)\, d\theta_2\, d\theta_1$$

will denote the Fourier coefficients of ϕ.

iv) For any function ϕ on T^2

$$\frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} \phi(\theta_1,\theta_2)\, d\theta_2\, d\theta_1$$

will be denoted by

$$\int_{T^2}\phi\, dm \quad \text{or} \quad \int_{T^2}\phi(w)\, dm(w).$$

We will first prove the converse to Theorem 1, and from this derive the converse to Theorem 3. First of all, however, we need the following lemma (which is given as a problem in [2]).

Lemma A1

If ϕ is a real-valued function defined on T^2 such that

$$\phi \in L^1(T^2)$$

i.e.,

$$\int_{T^2}|\phi|\, dm < \infty$$

and

$$\hat{\phi}(m,n) = 0, \qquad \text{for } mn < 0$$

then there is an outer function f on U^2 such that

$$P[\phi] = \log|f|$$

(where $P[\ ]$ denotes "Poisson integral of").

Proof: Let

$$a_{mn} = \begin{cases} \hat{\phi}(m,n), & (m,n)\neq(0,0) \\ 1/2\,\phi(m,n), & (m,n)=(0,0) \end{cases}$$

and let

$$g(Z_1,Z_2) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a_{mn} Z_1^m Z_2^n.$$

This series clearly converges uniformly on compact subsets of U^2, and so defines an analytic function there. If we let $f=e^g$, then f is analytic in U^2, and

$$\begin{aligned}\log|f| &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a_{mn} r_1^m r_2^n \exp(jm\theta_1 + jn\theta_2) \\ &\quad + \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \overline{a_{mn}}\, r_1^m r_2^n \exp(-jm\theta_1 - jn\theta_2) \\ &= \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \hat{\phi}(m,n) r_1^{|m|} r_2^{|n|} \exp(jm\theta_1 + jn\theta_2) \\ &= P[\phi] \qquad \text{[2, p. 17].}\end{aligned}$$

Next we prove that f is outer; we have (for $0<r<1$)

$$\begin{aligned}\int_{T^2}\log^+|f(rw)|\, dm(w) &\le \int_{T^2}|\log|f(rw)||\, dm(w) \\ &= \int_{T^2}|P[\phi](rw)|\, dm(w) \\ &\le \int_{T^2}|\phi(w)|\, dm(w) \\ &\qquad \text{[2, Theorem 2.1.3(c)]} \\ &< \infty\end{aligned}$$

and so $f\in N(U^2)$.

Now f^* exists almost everywhere on T^2 [2, Theorem 3.3.5] and $\log|f^*|=\phi$ almost everywhere on T^2 [2, Theorem 2.2.1]; thus $\log|f| = P[\log|f^*|]$ and so $f\in N_+(U^2)$ [2, Theorem 3.3.5], and $\log|f(0)| = \int_{T^2}\log|f^*(w)|\, dm(w)$. Thus f is outer. Q.E.D.

We can now prove the converse to Theorem 1.

Theorem A2

Let $f(Z_1,Z_2)$ be a rational function ($\not\equiv 0$), and let

$$\phi = \log|f^*|.$$

If $\hat{\phi}(m,n)=0$, for $mn<0$, then there is a rational function g without poles or zeros in U^2 such that $|g^*|=|f^*|$.

Proof: By Lemma A1, there is an outer function g such that

$$\log|g| = P[\log|f|].$$

This implies

$$\log|g^*| = \log|f^*|$$

almost everywhere on T^2. Therefore, for almost all $w\in T^2$

$$\log|g_w^*(Z)| = \log|f_w^*(Z)|$$

for almost all $Z\in T$ [2, Lemma 3.3.2]; and g_w is outer for almost all $w\in T^2$ [2, Lemma 4.4.4].

For any such w, let $Z_1, \cdots, Z_n$ denote the poles, and $Z_{n+1}, \cdots, Z_m$ denote the zeros of $f_w(Z)$ in U, and let

$$\tilde{f}_w(Z) = \prod_{k=1}^{n} \frac{Z - Z_k}{\bar{Z}_k Z - 1} \prod_{k=n+1}^{m} \frac{\bar{Z}_k Z - 1}{Z - Z_k} f_w(Z).$$

Then $\tilde{f}_w$ has no poles or zeros in U and is rational; hence, $\tilde{f}_w$ is outer. Since g_w is outer, we have $\tilde{f}_w / g_w$ is outer. Also $|\tilde{f}_w^*| = |f_w^*|$, and so $|\tilde{f}_w^*| = |g_w^*|$, for almost all $Z \in T$. Thus $\tilde{f}_w / g_w$ is inner. But a function which is both outer and inner is a constant of modulus 1, and so

$$g_w = e^{j\psi} \tilde{f}_w, \qquad \text{for some real } \psi.$$

Thus g_w is rational for almost all $w \in T^2$, and so g_w is rational for all $w \in E$, where $E \subseteq T^2$ is a compact set of positive measure (by the inner regularity of the measure). It follows by [2, Theorem 5.2.2] that g is rational (since the vanishing of a polynomial P on a set of positive measure in T^2 would imply

$$\log |P^*| \notin L^1(T^2)$$

and so $P \equiv 0$.)

Thus g is a rational function without poles or zeros in U^2, and

$$|g^*| = |f^*|, \qquad \text{almost everywhere in } T^2$$

and so, since g and f are both rational

$$|g^*| = |f^*|, \qquad \text{on } T^2. \qquad \text{Q.E.D.}$$

We next prove the converse to Theorem 3.

Theorem A3

Let $f(Z_1, Z_2)$ be a rational function ($\not\equiv 0$) and let

$$\phi = \log |f^*|.$$

If $1/2\pi \int_0^{2\pi} \phi(m\theta, n\theta + \psi)\, d\theta$ is a constant independent of ψ for each pair (m, n), with $m > 0$ and $n > 0$, then there is a rational function g without poles or zeros in U^2 such that $|g^*| = |f^*|$.

Proof: Let $m > 0$, $n > 0$, and let $l \neq 0$ be an integer. Then Then

$$\int_0^{2\pi} e^{jlm\psi} \int_0^{2\pi} \phi(m\theta, n\theta + \psi)\, d\theta\, d\psi = 0$$

$$\Rightarrow \int_0^{2\pi} \int_0^{2\pi} e^{jlm\psi} \phi(m\theta, n\theta + \psi)\, d\theta\, d\psi = 0.$$

Making the change of variables defined by

$$\theta = \frac{1}{m} \theta_1$$

$$\psi = \theta_2 - \frac{n}{m} \theta_1$$

we get

$$\frac{1}{m} \int_0^{2\pi} \int_{n/m\theta_1}^{n/m\theta_1 + 2\pi} \exp(jlm\theta_2 - jln\theta_1) \phi(\theta_1, \theta_2)\, d\theta_2\, d\theta_1 = 0$$

and since the integrand is periodic in θ_1 and θ_2

$$\int_0^{2\pi} \int_0^{2\pi} \exp(jlm\theta_2 - jln\theta_1) \phi(\theta_1, \theta_2)\, d\theta_2\, d\theta_1 = 0$$

and so

$$\hat{\phi}(-ln, lm) = 0, \qquad \text{for all } l \neq 0,\ m > 0, \quad \text{and } n > 0$$

that is

$$\hat{\phi}(m, n) = 0, \qquad \text{for all } m, n, \quad \text{with } mn < 0.$$

The result now follows from Theorem A2. Q.E.D.

Finally, we note that if f in Theorem A3 is a polynomial, then the converse in Theorem 2 implies that f has polynomial spectral factors. Thus we have the full converse of Theorem 3 for polynomials.

REFERENCES

[1] M. P. Ekstrom and J. W. Woods, "Two-dimensional spectral factorization with application to recursive digital filtering," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-24, pp. 115–128, Apr. 1976.

[2] W. Rudin, *Function Theory in Polydiscs*. New York: Benjamin, 1969.

[3] W. Stoll, *Holomorphic Functions of Finite Order in Several Complex Variables* (CBMS Regional Conference Series in Mathematics). Providence, RI: AMS, 1973.

[4] R. A. DeCarlo, J. Murray, and R. Saeks, "Multivariable Nyquist theory," to be published.

[5] T. S. Huang, "Stability of two-dimensional recursive filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 158–163, June 1972.

[6] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 1966.

[7] J. L. Shanks, S. Treitel, and J. H. Justice, "Stability and synthesis of two-dimensional recursive filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 115–128, June 1972.

[8] R. C. Gunning, and H. Rossi, *Analytic Functions of Several Complex Variables*. Englewood Cliffs, NJ: Prentice-Hall, 1965.

[9] N. K. Bose, "Problems in stabilization of multidimensional filters via Hilbert transform," *IEEE Trans. Geosci. Electron.*, vol. GE-12, pp. 146–147, Oct. 1974.

[10] J. W. Woods, *IEEE Trans. Geosci. Electron., (Corresp.), vol. GE-12, p. 104, July 1974.*

John J. Murray 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, TX. His principal research interests are in the areas of several complex variables, multidimensional system theory, and time-varying systems.

9. Reprint of "Semidirect Products and the Stability of Time-Varying Systems" by J. Murray from the Proceedings of the 1979 International Symposium on the Mathematics of Networks and Systems, T.H. Delft, Delft, July 1979, pp. 121-125.

Abstract

It is shown that time-varying systems may be modelled in terms of semidirect product algebras, and that the known theory of induced representations for these algebras in many cases enables one to give sharp criteria for stability of such systems. Finally, an example is given in which a previously known result is proved using these techniques.

1. INTRODUCTION

In this paper the theory of semidirect product algebras is proposed as an approach to the problem of the stability of time-varying systems. We first describe these algebras, and then show how a large class of time-varying systems may be modelled in terms of them. In the third paragraph, an approach to the problem of stability in general Banach-algebraic terms is described, and in the fourth, the special structure and known properties of semidirect product algebras are shown to be particularly suited to this approach. Finally, the theory is applied to a particular situation to derive some results very similar to those proved by Davis[1,2] by other methods.

2. SEMIDIRECT PRODUCT ALGEBRAS

Given a locally compact Abelian group G and a separable C* - algebra A with identity on which the group G acts as a group of isometric*-automorphisms, we define the set $L^1(G,A)$ to be the Banach space of Bochner-integrable functions

$$f: G \to A$$

(To be more accurate, we assume that

$$T: G \to A$$

is a continuous homomorphism of G into the set of isometric* - automorphisms of A with the strong topology; we normally suppress T and consider the elements of G to be automorphisms of A).

The product on $L^1(G,A)$ is defined by

$$(fh)(x) = \int_G f(y)[y(h(x-y))]\, d_\mu(y)$$

$$\forall x,y \in G, \quad f,h \in L^1(G,A) \qquad (1)$$

(The Abelian group G is written additively, and d_μ denotes the Haar measure on G).

The involution on $L^1(G,A)$ is defined by

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

$$f^*(x) = x(f(-x^*).$$

With these definitions, $L^1(G,A)$ becomes a Banach*-algebra, called the twisted group algebra on G with values in A. The enveloping C*-algebra of $L^1(G,A)$,[3], will be denoted by $C^*(G,A)$. The above is a simplified version of a more general construction given in [4].

Finally, we note that an alternative approach to *semidirect* product algebras is to define them as algebras of sections of Banach*-algebraic bundles[5,6]. We will not use this concept here, but we note (in connection with 4. Stability and Primitive Ideas) that Banach*-algebraic bundles were introduced as a powerful tool for calculating certain representations - the induced representations - of Banach*-algebras.

3. ALGEBRAS OF TIME-VARYING SYSTEMS AS SEMIDIRECT PRODUCTS.

In order to simplify the exposition and to ensure that our algebras $L^1(G,A)$ have an identity, we will limit ourselves from now on to discrete-time systems, that is, we will assume $G=\mathbb{Z}$. It should be clear, however, that the corresponding theory will hold good for continuous-time systems.

We take the algebra A to be any complete algebra of bounded functions on $\mathbb{Z}$ which contains the identity. Multiplication is defined pointwise, and the norm is taken to be the sup-norm. Since every such algebra is a commutative C*algebra with identity, it is isomorphic to C(X) for some compact Hausdorff space X.

The group $\mathbb{Z}$ acts on A in the obvious way:

$$(g(a))(n) = a(n-g) \text{ for } a\in A,\ n,g\in\mathbb{Z}.$$

Now the operator describing the input-output mapping of a scalar-input, scalar-output system with coefficients in A may be written formally as

$$\sum_{g\in\mathbb{Z}} a_g(.)g; \quad a_g\in A,\ g\in\mathbb{Z} \tag{2}$$

This transforms input sequences x(n) to output sequences y(n) by

$$y(m) = \sum_{g\in\mathbb{Z}} a_g(m)x(m-g),$$

which has the obvious physical interpretation as a sum of delayed inputs with time-varying weights. It is clear that the formal expressions in (2) are just functions from $\mathbb{Z}$ to A, and so can be regarded as elements of $L^1(\mathbb{Z},A)$; the only non-obvious formal property is that the cascade connection of two such systems is represented by the product (1). This follows formally from:

$$\begin{aligned}
&\Big(\sum_{h\in\mathbb{Z}} b_h(.)h\Big)\Big(\sum_{g\in\mathbb{Z}} a_g(.)g\Big)\,x(n)\\
&= \Big(\sum_{h\in\mathbb{Z}} b_h(.)h\Big)\Big(\sum_{g\in\mathbb{Z}} a_g(n)\,x(n-g)\Big)\\
&= \sum_{h\in\mathbb{Z}}\ \sum_{g\in\mathbb{Z}} b_h(n)\,a_g(n-h)\,x(n-g-h)\\
&= \sum_{h\in\mathbb{Z}}\ \sum_{k\in\mathbb{Z}} b_h(n)\,h(a_{k-h}(n))\,x(n-k)\\
&= \sum_{k\in\mathbb{Z}}\Big\{\sum_{h\in\mathbb{Z}}\big[\,b_h\,h(a_{k-h})\big](n)\Big\}\,x(n-k).
\end{aligned}$$

The analytic details of the above formal manipulations are easily checked and will not be considered here.

4. STABILITY AND PRIMITIVE IDEALS

The utility of transform methods in determining the stability of time-invariant linear systems stems from the fact that in a commutative Banach algebra, an element is invertible if and only if its Gelfand transform is. It is therefore natural to

seek an analogue of this in the non-commutative case, and the most obvious analogue is the representation of a Banach algebra as a subdirect product over the set of primitive ideals. (See [7], where also the terminology used in this section is defined). Here we have the property that an element is invertible if and only if its image in the quotient algebra by every primitive ideal is invertible. Thus, in order to determine invertibility (and hence stability) it is necessary to determine all primitive ideals in our algebra. Since a primitive ideal is by definition the kernel of an irreducible representation, this problem is related to the problem of determining the irreducible representations of an algebra. It is precisely these problems which have received the most attention in the literature on semidirect products. In particular, the primitive ideals are studied in [8,9,10, 11], especially with reference to a conjecture made in [8] that all primitive ideals of a semidirect product algebra can be found as kernels of representations induced from the isotropy subgroups of the group action of G on X, where the original C*-algebra A is given as C(X). In this connection, it is interesting to note, as mentioned before, that semidirect product algebras can be realized as algebras of sections of Banach*-algebraic bundles, which were introduced precisely for the purpose of extending the idea of induced representation to Banach algebras. There is thus some hope that the problem of determining all primitive ideals can be solved by known techniques. Rather than give a theoretical discussion of these concepts, however, which would take us too far afield and occupy too much space, we will simply give an example of their application to derive a previously known result in the next section.

5. EXAMPLE

Let $A = \{f:\mathbb{Z}\longrightarrow\mathbb{C} \mid f(2n)=f(0)$ and $f(2n+1) = f(1), \forall n\}$ which gives us the class of periodically time-varying systems whose period is twice the sampling period.

$$A \simeq C(X) \text{ , where } X = \{0,1\}.$$

$\mathbb{Z}$ acts on X by

$$g = \text{identity, } g \text{ even}$$

$$g(0) = 1 \text{ and } g(1) = 0, g \text{ odd.}$$

Thus X is the unique orbit, and $H = \{2n \mid n\in\mathbb{Z}\}$ is the unique isotropy subgroup.

Now, using the results of [10], we see that every primitive ideal is obtained by inducing from the irreducible representations of H. Since $H \simeq \mathbb{Z}$, the latter are just the usual frequencies, whose representing measures are given by

$$(e^{jnw})_{n=-\infty}^{\infty} \quad \text{for } -\pi < w \leq \pi.$$

We will calculate the induced representations following [8]. An induced measure on $\mathbb{Z} \times X$ is given by

$$\tilde{u} : f \longmapsto \sum_{n=-\infty}^{\infty} f(2n,0)\, e^{jnw}$$

(Here we are making the natural identification of

$$f: \mathbb{Z} \longrightarrow C(X) \text{ with } f: \mathbb{Z}\times X \longrightarrow \mathbb{C}).$$

Now

$$(f^*g)(n,0) = \sum_{m=-\infty}^{\infty} \bar{f}(-2m,0)g(n-2m,0) + \sum_{m=-\infty}^{\infty} \bar{f}(-2m-1,1)g(n-2m-1,1) \qquad (n \text{ even})$$

and

$$(f^*g)(n,0) = \sum_{m=-\infty}^{\infty} \bar{f}(-2m,0)g(n-2m,1)$$

$$+ \sum_{m=-\infty}^{\infty} \bar{f}(-2m+1,1)g(n-2m-1,o) \qquad (n \text{ odd})$$

from which it follows that

$$\tilde{u}(f^*f) = \left| \sum_m f(2m,o)e^{jmw} \right|^2 + \left| \sum_m f(2m-1,1)e^{jmw} \right|^2$$

and so we have a two-dimensional representation

$$g \longmapsto \left(\sum_m g(2m,o)e^{jmw} \ , \ \sum_m g(2m-1,1)e^{jmw} \right)$$

on which the action of f can be represented by left multiplication by the matrix

$$\begin{pmatrix} \sum_m f(2m,o)e^{jmw} & , \sum_m f(2m+1,o)e^{jmw} \\ \sum_m f(2m-1,1)e^{jmw} & , \sum_m f(2m,1)e^{jmw} \end{pmatrix}$$

If the system is finite, this will be a rational matrix in $Z=e^{jw}$; and the original operator will be invertible if this matrix is nonsingular on the unit circle. A homotopy-type argument then easily proves that if the original operator was causal and bounded, it will have a bounded, causal inverse if the determinant of the matrix has zero winding number about the origin.

In this connection, we note that the theory of semidirect products has been developed entirely for *-algebras, so that if we wish to apply it to causal systems, we must either redevelop the entire theory for algebras without involution, or develop the topological (and other) criteria which give conditions for a causal, invertible operator to have a causal inverse. The latter course seems preferable.

Finally, we note that the above example is a special case of the results of Davis[1].

REFERENCES

1. J. H. Davis, Stability Conditions derived from Spectral Theory: Discrete Systems With Periodic Feedback, SIAM Journal on Control, 10(1972) 1-13.

2. J. H. Davis, Fredholm Operators, Encirclements and Stability Criteria, SIAM Journal on Control, 10(1972),608-628.

3. J. Dixmeier, Les C*-Algebres et Leurs Representations, Cahiers Scientifiques, Fasc. 29, Gauthnier-Villers, Paris 1964.

4. R. C. Busby and H. A. Smith, Representations of Twisted Group Algebrás, Trans. Amer. Math. Soc. 149(1970), 503-537.

5. J. M. G. Fell, An Extension of Mackey's Method of Banach*-Algebraic Bundles, Mem. Amer. Math. Soc. No. 90(1969).

6. J. M. G. Fell, Induced Representations and Banach*-Algebraic Bundles, Lecture Notes in Mathematics, Vol. 582, Springer-Verlag, New York. 1977.

7. C. E. Rickart, General Theory of Banach Algebras, Van Nostrand, Princeton, 1960.

8. E. G. Effros and F. Hahn, Locally Compact Transformation Groups and C*-Algebras, Mem. Amer. Math. Soc. No. 75(1967).

9. G. Zeller-Meier, Produits croises dans une C*-algebre par un groupe d'automorphismes, C. R. Acad. Sci. Paris Ser. A-B 263 (1966) A20-A23.

10. E. C. Gootman, Primitive Ideals of C*-algebras Associated With Transformation Groups, Trans. Amer. Math. Soc. 170 (1972), 97-108.

11. O. L. Britton, Primitive Ideals of Twisted Group Algebras, Trans. Amer. Math. Soc. 202(1975), 221-241.

Texas Tech University — Institute for Electronic Science

Joint Services Electronics Program — Research Unit: 7

1. Title of Investigation: Optical Noise
2. Senior Investigator: J.F. Walkup — Telephone: (806) 742-3528
3. JSEP Funds: $23,500
4. Other Funds:
5. Total Number of Professionals: PI's 1 (1 mo.) RA's 1 (1/2 time)
6. Summary:

The goal of the work unit is the development of detection and estimation algorithms for application in image processing. Unlike the detection and estimation algorithms which have been developed for use in a communications context, if an algorithm is to be employed in image processing, it must be designed to cope with the nonlinear character of the optical noise phenomena encountered in such applications. In our research we have emphasized the detection and estimation problem in film-grain noise and have developed optimal MAP and ML estimators for these applications. In addition we have computed bounds on the performance of various classes of estimators in a nonlinear noise environment.

In addition to the nonlinear noise phenomena, two additional factors are encountered in the image processing problem which are not encountered in the classical detection and estimation problems. First, the noise phenomena is typically space-variant, often with widely different characteristics in one part of an image than another. Secondly, the two-dimensional nature of the image processing problem greatly increases the computational requirements for the detection and estimation problem. As such, the optimal detection and

estimation algorithms which we have developed cannot be efficiently implemented in an image processing system. In an effort to alleviate this difficulty, much of our research during the past year has been directed at the development of suboptimal detection and estimation algorithms for image processing. To this end we have investigated the possibility of employing a "modified signal MAP estimator", a "noise cheating" algorithm and a "modified Stein's estimator". We believe that these algorithms have the potential of achieving near-optimal performance with greatly reduced computational effort.

Our research, on bounds, for the detection and estimation problem in image processing and the derivation of the optimal MAP and ML estimators was recently reported in two papers which are reprinted herein. The work on suboptimal estimators is still in progress and will be the subject of a forthcoming Ph.D. dissertation.

7. Publications and Activities:

A. Refereed Journal Articles

1. Froehlich, G.K., Walkup, J.F., and R.B. Asher, "Optimal Estimation in Signal-Dependent Noise", Jour. of the Optical Soc. of Am., Vol. 68, pp. 1665-1671, (1978).

B. Conference Papers and Abstracts

1. Froehlich, G.K., Walkup, J.F., and R.B. Asher, "Optimal Estimation in Signal-Dependent Film-Grain Noise", Proc. of the 11th Inter. Commission for Optics Conf., Madrid, Sept. 1978, pp. 367-369.

2. Froehlich, G.K., Walkup, J.F., and R.B. Asher, "Estimation in Signal-Dependent Noise", 1978 Annual Meeting of the Optical Soc. of Am., San Francisco, Nob. 1978, (abstract in the Jour. of the OSA, Vol. 68, p. 1385A).

C. Theses

1. Froehlich, G.K., Ph.D. Dissertation, Texas Tech Univ., (in preparation).

D. Conferences and Symposia

1. Walkup, J.F., SPIE Technical Symposium, San Diego, Aug. 1979.

2. Walkup, J.F., 1978 IEEE Inter. Optical Computing Conference, London, Sept. 1978.

3. Walkup, J.F., 11th Inter. Commission for Optics Conf., Madrid, Sept. 1978.

E. Lectures

1. Walkup, J.F., Space-Variant Optical Processing", Elec. Engrg. Seminar, Georgia Inst. of Tech., June 1979.

2. Walkup, J.F., "Space-Variant Optical Processing", Elec. Engrg. Colloquim, Virginia Tech, June 1979.

9. Reprint of "Optimal Estimation in Signal-Dependent Noise" by G.K. Froehlich, J.F. Walkup, and R.B. Asher from the Journal of the Optical Society of America, Vol. 68, pp. 1665-1671, (1978).

Optimal estimation in signal-dependent noise

Gary K. Froehlich, John F. Walkup, and Robert B. Asher*

Department of Electrical Engineering, Texas Tech University, Lubbock, Texas 79409

(Received 16 January 1978)

Optimal estimators are derived for a class of signal-dependent noise processes. Such processes are of interest in optics because phenomena, such as film grain noise, are often modeled in this manner. This paper demonstrates that when one ignores the presence of signal-dependent noise and instead assumes only signal-independent noise models, the resulting estimators may pay a severe penalty in performance. This "mismatch" problem is explored, with the results of Monte Carlo simulations of the performances of both optimum and mismatched estimators being presented. The Cramér-Rao lower bounds on the mean-square estimation errors for unbiased estimators are evaluated and compared with the lower bounds derived for the signal-independent noise case. Overall, the results indicate that improved performance will, in most cases, offset the increased complexity inherent in estimators designed for the signal-dependent noise model.

INTRODUCTION

In contrast to the signal-independent additive noise models traditionally encountered in statistical communication theory,[1,2] many physical noise processes are inherently signal-dependent. Common examples from optical processing include film-grain noise, encountered in image processing, and photoelectronic shot noise, which is sometimes dominant when imaging at low-light levels with photoemissive detectors.[3,4] An example of a nonoptical noise source which is effectively signal dependent is magnetic tape recording noise.[5] A study of these particular examples indicates that studies of optimum estimation in signal-dependent noise processes would have applications to a broad class of signal-processing problems in modern optics and in other fields.

To date, the majority of the work dealing with signal-dependent noise has been concentrated on rather specialized examples and applications. Using a Poisson point-process noise model, Goodman and Belsher[6] have considered the restoration of atmospherically degraded images using linear minimum mean-square error filters. Walkup and Choens[4] modified the familiar Wiener filter for various additive, Gaussian signal-dependent noise models, and Naderi[7] has done considerable additional work on this problem. Additionally, Hunt[8] has derived a nonlinear maximum *a posteriori* (MAP) estimator, based on a different model than the one considered here, which can accommodate both signal-dependent and signal-independent noise cases, and they have applied this MAP estimator to restoring noise-degraded images. For such applications, and in the special case where the images of interest exhibit extremely low contrasts, conventional restoration techniques perform rather poorly. Thus, heuristic algorithms, such as the so-called "noise cheating" algorithm for film-grain noise suppression,[9] have been developed. Other algorithms, which explicitly include the signal dependence of the noise, as well as incorporating pertinent properties of the human visual system, have also been investigated.[7,10,11]

The purpose of this paper, then, is twofold. First, several fundamental properties of signal-dependent noise are investigated in order to better understand when consideration of signal-dependence is warranted and when it can be ignored. To this end, the mean-square estimation error is first considered for both the signal-dependent and signal-independent cases. In addition, the mean-square estimation error for a mismatched case is evaluated. The mismatch case considered is one in which the signal-dependent measurement model is valid but is ignored for purposes of simplification. Secondly, optimal estimators are derived for several cases of both signal-dependent and signal-independent models. The Cramér-Rao lower bound on mean square estimation error is also determined, in order to find the lowest error possible for both signal-dependent and signal-independent estimators. The results of Monte Carlo simulations of the performance of the

 0030-3941/78/6812-1665$00.50

various optimal estimators previously derived are presented for several values of the model parameters and for various prior signal probability densities.

PROBLEM STATEMENT

To motivate the investigation of signal-dependent noise processes, it is necessary first to define the models to be used. The signal-dependent measurement model to be used is given by

$$r = s + kf(s)n_1 + n_2, \tag{1}$$

where n_1 and n_2 are signal-independent random noise processes; s is the underlying signal to be estimated which is assumed to have probability density $p(s)$; n_1, n_2, and s are assumed to be mutually statistically independent; $f(s)$ is any function of the signal; k is a scalar constant; and r is the noisy measurement. The signal-dependent noise term in Eq. (1) is, of course, the term $kf(s)n_1$. It is often physically reasonable to assume that both n_1 and n_2 are zero mean and have unimodal probability densities. Further, note that substitution of $k = 0$ in Eq. (1) yields

$$r = s + n_2, \tag{2}$$

which is just the familiar textbook additive, signal-independent noise model.[1,2] In both Eq. (1) and Eq. (2), the arguments of all of the variables have been dropped for simplification. It should be remembered that these arguments may depend on time, position, or both.

It will be shown repeatedly that the model of Eq. (2) yields far simpler estimators than does Eq. (1), as would be expected. The following example serves to illustrate why it may prove worthwhile to employ the more complex estimators resulting from Eq. (1).

When the observations are actually of the type given by Eq. (2), it can be shown[2] that simply using the received value as the estimate results in a minimum-variance unbiased estimate, i.e.,

$$\hat{s} = r, \tag{3}$$

where the circumflex denotes the estimate. The average error is then given by

$$E\{\hat{s} - s\} = E\{r - s\} = E\{n_2\} = 0. \tag{4}$$

The estimator is said to be unbiased since the mean error is zero. A measure of the performance of this estimator, conditioned on the signal value, is given by the conditional mean-square error, and is found to be

$$E\{(\hat{s} - s)^2|s\} = E\{n_2^2\} = \sigma_2^2, \tag{5}$$

which is simply the variance of the additive noise process n_2. This estimator is obviously simple from an implementation point of view.

With this in mind, consider a case in which the observations are actually of the type given by the signal-dependent model of Eq. (1). For ease of implementation it is decided to use the estimate given in Eq. (3), which was designed for the signal-independent noise process. This represents a mismatched situation, where an estimator based upon an incorrect measurement model (corresponding to ignoring the signal-dependency) is used. Once again, the average estimation error is zero, due to n_1 and n_2 being assumed zero mean and to the assumed mutual statistical independence of n_1, n_2, and s.

However, assuming $\hat{s} = r$, the mean-square estimation error for this mismatched case is given by

$$E\{(\hat{s} - s)^2|s\} = k^2\sigma_1^2 E\{[f(s)]^2\} + \sigma_2^2. \tag{6}$$

For convex $f(s)$, i.e., $f''(s) \geq 0$ for all s, Jensen's inequality[12] states that $E\{f(s)\} \geq f\{E[s]\}$, where $E\{\cdot\}$ denotes the expected value. This inequality may be used to find a lower bound for the mean-square estimation error for the mismatched case. Thus, recalling Eqs. (5) and (6),

$$\sigma_2^2 \leq k^2\sigma_1^2\{f[E(s)]\}^2 + \sigma_2^2 \leq k^2\sigma_1^2 E\{[f(s)]^2\} + \sigma_2^2. \tag{7}$$

Note that this gives a lower bound (the middle term) on the mean-square estimation error of the mismatched estimator, and that this bound contains a function of the signal's mean. The left-most term of Eq. (7) is the mean-square estimation error given by Eq. (5). Note that the mismatched mean-square estimation error is in general greater than the error for the same estimator when used in the presence of signal-independent noise. We next consider an illustration of the significance of Eq. (7).

A commonly used model in image processing when the observed quantity is the photographic density is given by Eq. (1) with $f(s)$ given by s^p.[4,10] From Eq. (6), then, the mismatched mean-square estimation error becomes

$$E\{(\hat{s} - s)^2|s\} = k^2\sigma_1^2 E\{s^{2p}\} + \sigma_2^2, \tag{8}$$

where k is a scanning constant relating the scanning aperture area to the mean area of a film grain. A typical value of p used for characterizing photographic film-grain noise is $p = 1/2$, though $p = 1/3$ has also been used.[4,10] Thus, Eq. (8) becomes (for $p = 1/2$)

$$E\{(\hat{s} - s)^2|s\} = k^2\sigma_1^2 E\{s\} + \sigma_2^2, \tag{9}$$

which is greater than the variance of Eq. (5) by the addition of a term which is proportional to the signal mean. Note that in the particular case of $p = 1/2$, the equality holds between the last two terms in Eq. (7), but that for general p this is not the case. Here, the lower bound on the mean-square estimation error given by Eq. (7) becomes

$$E\{(\hat{s} - s)^2|s\} \geq k^2\sigma_1^2[E(s)]^{2p} + \sigma_2^2. \tag{10}$$

The lower bound given by Eq. (10) may be visualized with the aid of Figs. 1 and 2, for various values of k, p, and $E(s)$. In all cases, the plane upon which the surfaces rest is not the zero plane, but rather represents a height of σ_2^2, the left-most term of Eq. (7), which results when the estimator of Eq. (3) is properly matched [to the signal-independent noise process of Eq. (2)]. In Fig. 1, p is fixed at a value of 1/2, σ_1^2 and σ_2^2 are set equal to 1 for illustration, with k and $E(s)$ being varied. In Fig. 2, k is fixed (at $k = 1/2$) and p is varied. It should be noted that, for film-grain-noise applications, common values of k are in the range of from about 0.3 to about 0.7. These figures illustrate the marked deviation from the variance achieved by a properly matched signal-independent estimator. Also, it should be remembered that these surfaces represent *lower bounds* on the mean-square estimation error of the mismatched estimator, and there is no guarantee, in general, that even this measure of performance can be achieved.

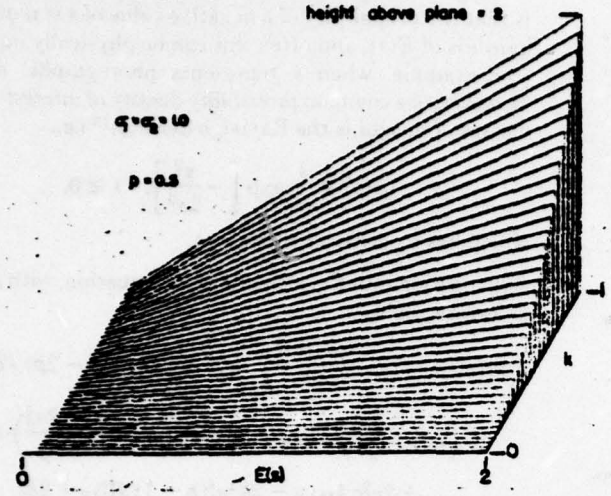


FIG. 1. Mean-square estimation-error lower bound for the mismatched case, $p = 1/2$.

Thus, optimal estimators based on the proper noise model are needed. These estimators are derived in the following sections.

MAP ESTIMATION

An appropriate optimal estimate when the signal is random and its probability density function is known *a priori* is the maximum *a posteriori* probability (MAP) estimate.[2] This estimate, $\hat{s}_{MAP}$, is defined to be that value of s which maximizes the *a posteriori* density $p(s|r)$. In other words, given the observation r, the signal value $\hat{s}_{MAP}$ maximizes the probability of that value of r having been received. Maximizing $p(s|r)$ is equivalent to maximizing $p(r|s)p(s)$, or alternately the logarithm of this product. This follows from the facts that (i)

$$p(s|r) = p(r|s)p(s)/p(r), \tag{11}$$

(ii) the denominator is not a function of s, and (iii) because monotonic transformations (such as the logarithm) preserve maxima and minima.

As an example of the calculation of a MAP estimate, assume that n_1 and n_2 are both zero mean, normally distributed random variables having variances σ_1^2 and σ_2^2, respectively. In this case, the conditional probability density $p(r|s)$ is also normal, with a mean of s and variance $v(s)$, given by

$$v(s) = k^2\sigma_1^2[f(s)]^2 + \sigma_2^2. \tag{12}$$

It can then be shown that the MAP estimate is a solution of the equation

$$(r - \hat{s}_{MAP})^2 v'(\hat{s}_{MAP}) + 2(r - \hat{s}_{MAP})v(\hat{s}_{MAP}) - v'(\hat{s}_{MAP})v(\hat{s}_{MAP}) + 2[v(\hat{s}_{MAP})]^2 \frac{\partial}{\partial \hat{s}_{MAP}} \ln p(\hat{s}_{MAP}) = 0, \tag{13}$$

where the prime denotes the partial derivative with respect to $\hat{s}_{MAP}$.

For the class of situations where $f(s) = s^p$, and assuming s is distributed normally with mean μ_s and variance σ_s^2, Eq. (13), the MAP equation becomes

$$\left(2k^2\sigma_1^2(p-1) - \frac{4k^2\sigma_1^2\sigma_2^2}{\sigma_s^2}\right)\hat{s}_{MAP}^{2p+1} + \left(2rk^2\sigma_1^2(1-2p) + \frac{4k^2\sigma_1^2\sigma_2^2\mu_s}{\sigma_s^2}\right)\hat{s}_{MAP}^{2p} - \left(2\sigma_2^2 + \frac{2\sigma_2^4}{\sigma_s^2}\right)\hat{s}_{MAP} + [2pk^2\sigma_1^2(r^2 - \sigma_2^2)]\hat{s}_{MAP}^{2p-1} + \left(2\sigma_2^2 r + \frac{2\sigma_2^4\mu_s}{\sigma_s^2}\right) - [2pk^4\sigma_1^4]\hat{s}_{MAP}^{4p-1} + \frac{2k^4\sigma_1^4\mu_s}{\sigma_s^2}\hat{s}_{MAP}^{4p} - \frac{2k^4\sigma_1^4}{\sigma_s^2}\hat{s}_{MAP}^{4p+1} = 0. \tag{14}$$

The MAP estimate, $\hat{s}_{MAP}$, is a solution of Eq. (14). For the specific case where $p = 1/2$, Eq. (14) reduces to the cubic equation

$$\frac{2k^4\sigma_1^4}{\sigma_s^2}\hat{s}_{MAP}^3 + \left(\frac{4k^2\sigma_1^2\sigma_2^2 - 2k^4\sigma_1^4\mu_s}{\sigma_s^2} + 2k^2\sigma_1^2\right)\hat{s}_{MAP}^2 + \left(\frac{2\sigma_2^4 - 4k^2\sigma_1^2\sigma_2^2\mu_s}{\sigma_s^2} + k^4\sigma_1^4 + 2\sigma_2^2\right)\hat{s}_{MAP} + k^2\sigma_1^2(\sigma_2^2 - r^2) - 2\sigma_2^2 r - \frac{2\sigma_2^4\mu_s}{\sigma_s^2} = 0. \tag{15}$$

Substitution of $k = 0$ into Eq. (14) or Eq. (15) yields the MAP estimate for the signal-independent noise case of Eq. (2), namely

$$\hat{s}_{MAP} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_2^2} r + \frac{\sigma_2^2}{\sigma_s^2 + \sigma_2^2}\mu_s. \tag{16}$$

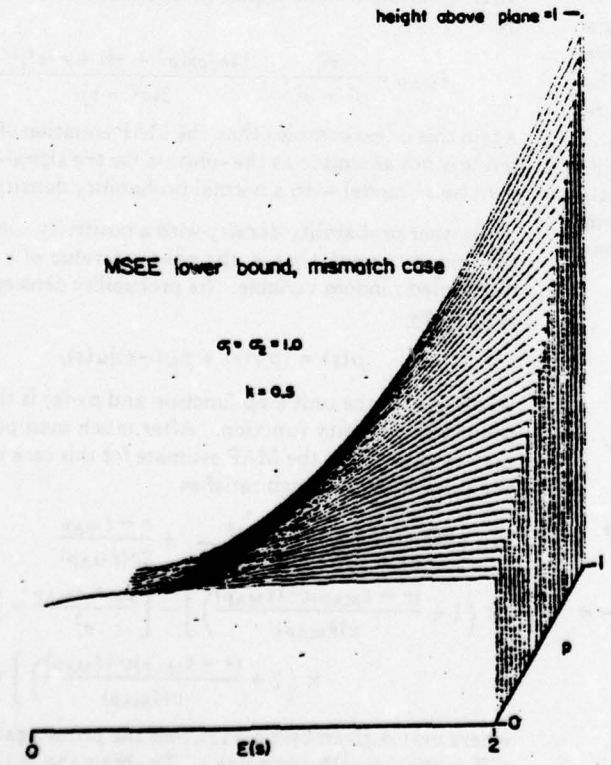


FIG. 2. Mean-square estimation-error lower bound for the mismatched case, $k = 1/2$.

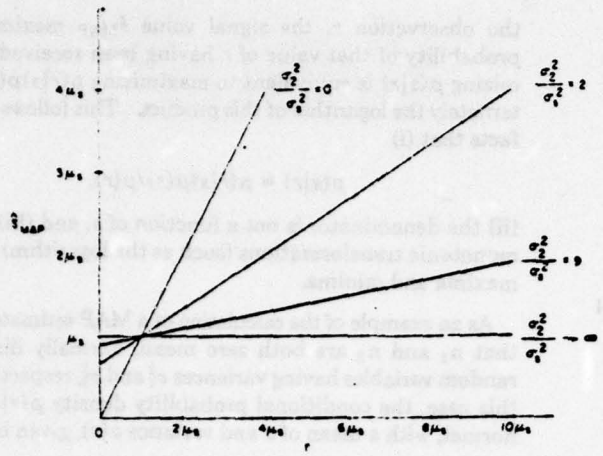


FIG. 3. MAP estimator structures for the signal-independent noise process.

Comparison of Eqs. (14) and (16) demonstrates the much greater complexity of the estimator structure when signal-dependent noise processes are taken into account.

This comparison can also be seen graphically. Fig. 3 represents Eq. (16) plotted for various parameter values. The ratio σ_2^2/σ_s^2 can approach zero in either of two ways: either the noise variance is zero or the signal variance is infinite. In the former case there is no noise, so $\hat{s} = r$; in the latter, the MAP estimate becomes a *ML* estimate, $\hat{s} = r$, as discussed in the next section. The other extreme is for the ratio σ_2^2/σ_s^2 to increase without bound. Here the noise variance becomes infinite or the signal variance becomes zero, and hence the best estimate is the (assumed known) signal mean, μ_s.

For the signal-dependent noise case, Fig. 4 shows quite similar estimator structures. These curves represent the solution of Eq. (15) plotted for various parameter values. The most apparent difference from Fig. 3 is the nonlinear nature of the curves in Fig. 4. Note, however, that the ratio $\sigma_2^2/(k\sigma_1)^2$ is considered while σ_s^2 is fixed for illustration. This corresponds to various degrees of dominance by either the signal-independent noise term or the signal-dependent noise term of Eq. (1). Recall that in both figures, n_1 and n_2 were considered to be Gaussian random variables.

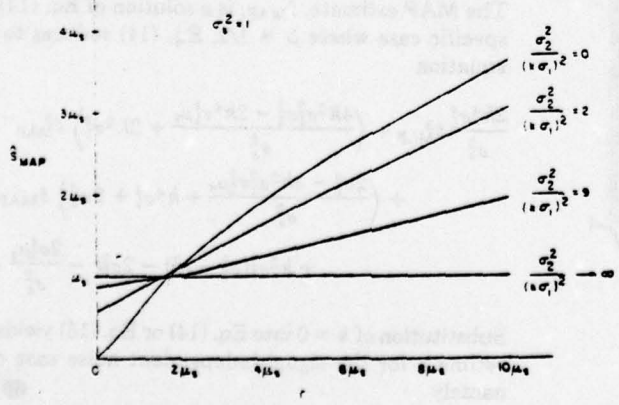


FIG. 4. MAP estimator structures for the signal-dependent noise process.

A shortcoming of the Gaussian density as a model for $p(s)$ is that the probability of a negative value of s is nonzero, regardless of $E(s)$, and often this can be physically impossible (for example, when s represents photographic density). However, one common probability density of interest for some classes of images is the Rayleigh density,[13] i.e.,

$$p(s) = \frac{s}{\sigma^2}\exp\left[-\frac{s^2}{2\sigma^2}\right], \quad s \geq 0. \tag{17}$$

which has a positivity constraint.

Substitution into Eq. (13), the MAP equation, with $f(s) = s^p$ as before, yields

$$\left(2k^2\sigma_1^2(p-1) - \frac{4k^2\sigma_1^2\sigma_2^2}{\sigma^2}\right)\hat{s}_{\mathrm{MAP}}^{2p+2} + 2k^2\sigma_1^2 r(1-2p)\,\hat{s}_{\mathrm{MAP}}^{2p+1}$$
$$-\left(2\sigma_2^2 + \frac{2\sigma_2^4}{\sigma^2}\right)\hat{s}_{\mathrm{MAP}}^2 + 2pk^2\sigma_1^2\left(r^2 - \sigma_2^2 + \frac{2\sigma_2^2}{p}\right)\hat{s}_{\mathrm{MAP}}^{2p}$$
$$+ 2\sigma_2^2 r\,\hat{s}_{\mathrm{MAP}} - 2k^4\sigma_1^4(p-1)\,\hat{s}_{\mathrm{MAP}}^{4p} + 2\sigma_2^4$$
$$- \frac{2k^4\sigma_1^4}{\sigma^2}\hat{s}_{\mathrm{MAP}}^{4p+2} = 0. \tag{18}$$

This equation is quite similar to Eq. (14) with $\mu_s = 0$, but each term is greater in Eq. (18) by degree one. Thus, when $p = 1/2$, the MAP estimate $\hat{s}_{\mathrm{MAP}}$ is a solution of the quartic

$$\frac{2k^4\sigma_1^4}{\sigma^2}\hat{s}_{\mathrm{MAP}}^4 + k^2\sigma_1^2\left(1 + \frac{4\sigma_2^2}{\sigma^2}\right)\hat{s}_{\mathrm{MAP}}^3$$
$$+\left(2\sigma_2^2 + \frac{2\sigma_2^4}{\sigma^2} - k^4\sigma_1^4\right)\hat{s}_{\mathrm{MAP}}^2 - [k^2\sigma_1^2(r^2 + 3\sigma_2^2) + 2r\sigma_2^2]\,\hat{s}_{\mathrm{MAP}}$$
$$- 2\sigma_2^4 = 0. \tag{19}$$

As before, substitution of $k = 0$ into Eq. (19) then yields the MAP estimate for the signal-independent noise case, given by

$$\hat{s}_{\mathrm{MAP}} = \frac{\sigma^2}{\sigma^2 + \sigma_s^2}r + \frac{[4\sigma^2\sigma_2^2(\sigma^2 + \sigma_s^2) + r^2\sigma^4]^{1/2}}{2(\sigma^2 + \sigma_s^2)}. \tag{20}$$

Again this is less complex than the MAP equation of Eq. (18), but it is not as simple as the solution for the signal-independent noise model with a normal probability density for s.

Another probability density with a positivity constraint is the folded normal, that is, the absolute value of a normally distributed random variable. Its probability density function is given by

$$p(s) = [p_N(s) + p_N(-s)]u(s), \tag{21}$$

where $u(s)$ is the unit step function and $p_N(s)$ is the normal probability density function. After much manipulation, it may be shown that the MAP estimate for this case is given by the value of $\hat{s}_{MAP}$ which satisfies

$$\exp\left(\frac{2\mu_s\hat{s}_{\mathrm{MAP}}}{\sigma_s^2}\right)\left[\frac{\mu_s - \hat{s}_{\mathrm{MAP}}}{\sigma_s^2} + \frac{r - \hat{s}_{\mathrm{MAP}}}{2v(\hat{s}_{\mathrm{MAP}})}\right.$$
$$\left.\times\left(1 + \frac{(r - \hat{s}_{\mathrm{MAP}})v'(\hat{s}_{\mathrm{MAP}})}{v(\hat{s}_{\mathrm{MAP}})}\right)\right] - \left[\frac{\mu_s + \hat{s}_{\mathrm{MAP}}}{\sigma_s^2} - \frac{r - \hat{s}_{\mathrm{MAP}}}{2v(\hat{s}_{\mathrm{MAP}})}\right.$$
$$\left.\times\left(1 + \frac{(r - \hat{s}_{\mathrm{MAP}})v'(\hat{s}_{\mathrm{MAP}})}{v(\hat{s}_{\mathrm{MAP}})}\right)\right] = 0, \tag{22}$$

where $v(s)$ is given by Eq. (12), and the prime again denotes differentiation with respect to s. To obtain the MAP estimate for the signal-independent measurement model of Eq. (2), k

= 0 is substituted into Eq. (22) to obtain

$$\exp\left(\frac{2\mu_s \hat{s}_{MAP}}{\sigma_s^2}\right)\left(\frac{\mu_s - \hat{s}_{MAP}}{\sigma_s^2} + \frac{r - \hat{s}_{MAP}}{2\sigma_2^2}\right) - \left(\frac{\mu_s + \hat{s}_{MAP}}{\sigma_s^2} - \frac{r - \hat{s}_{MAP}}{2\sigma_2^2}\right) = 0. \quad (23)$$

Neither of these equations lend themselves to straightforward solution; however, it is once again obvious that the signal-independent noise model yields a much simpler solution.

MAXIMUM-LIKELIHOOD ESTIMATION

Another commonly used estimator is the maximum-likelihood (ML) estimator.[2] The ML estimate is employed when no prior knowledge of the signal is assumed, and it is found by maximizing $p(r|s)$ over s. In other words find a value of s, such that given s, the most probable observation r which would result is the value observed. Using the signal-dependent measurement model of Eq. (2), and still assuming n_1 and n_2 are zero mean normal random variables with variances σ_1^2 and σ_2^2, respectively, the ML estimate, $\hat{s}_{ML}$, is a solution of the equation

$$(r - \hat{s}_{ML})^2 v'(\hat{s}_{ML}) + 2(r - \hat{s}_{ML}) v(\hat{s}_{ML}) - v'(\hat{s}_{ML}) v(\hat{s}_{ML}) = 0, \quad (24)$$

where $v(\hat{s}_{ML})$ and $v'(\hat{s}_{ML})$ are as defined previously. Again, considering the special case $f(s) = s^p$, the ML equation becomes

$$2k^2\sigma_1^2(p-1)\,\hat{s}_{ML}^{2p-1} + 2rk^2\sigma_1^2(1-2p)\,\hat{s}_{ML}^{2p} + 2pk^2\sigma_1^2(r^2 - \sigma_2^2)\,\hat{s}_{ML}^{2p-1} - 2\sigma_2^2\,\hat{s}_{ML} - 2pk^4\sigma_1^4\,\hat{s}_{ML}^{4p-1} + 2\sigma_2^2 r = 0. \quad (25)$$

This equation is at worst no more complex than the MAP equation (14). In fact, for $p = 1/2$, Eq. (25) becomes the quadratic equation

$$k^2\sigma_1^2\,\hat{s}_{ML}^2 + (2\sigma_2^2 + k^4\sigma_1^4)\hat{s}_{ML} + k^2\sigma_1^2(\sigma_2^2 - r^2) - 2\sigma_2^2 r = 0. \quad (26)$$

This has as its positive root the ML estimate

$$\hat{s}_{ML} = \left[r^2 + \left(\frac{k^2\sigma_1^2}{2}\right)^2 + \frac{2r\sigma_2^2}{k^2\sigma_1^2} + \left(\frac{\sigma_2^2}{k^2\sigma_1^2}\right)^2\right]^{1/2} - \frac{k^2\sigma_1^2}{2} - \frac{\sigma_2^2}{k^2\sigma_1^2}. \quad (27)$$

The ML estimate for the signal-independent model of Eq. (2) is found by letting $k = 0$ in any of Eqs. (24)–(26), and is given by

$$\hat{s}_{ML} = r. \quad (28)$$

Note that this is the minimum variance unbiased estimate used in Eq. (3) for the mismatched example, for which we earlier found the mean-square estimation error.

A graphical comparison of Eq. (27) and Eq. (28) for various parameter values is shown in Fig. 5. In this figure, the dashed line represents Eq. (28), while the solid lines represent Eq. (27). Note especially the estimator structure when the signal-independent noise term is comparable to or greater than the signal-dependent noise term. In this case, the estimator takes the form $\hat{s}_{ML} = r - b$, where b is an approximately constant bias determined by the ratio $\sigma_2^2/(k\sigma_1)^2$. Note that as this ratio increases without bound, b asymptotically approaches 0.5.

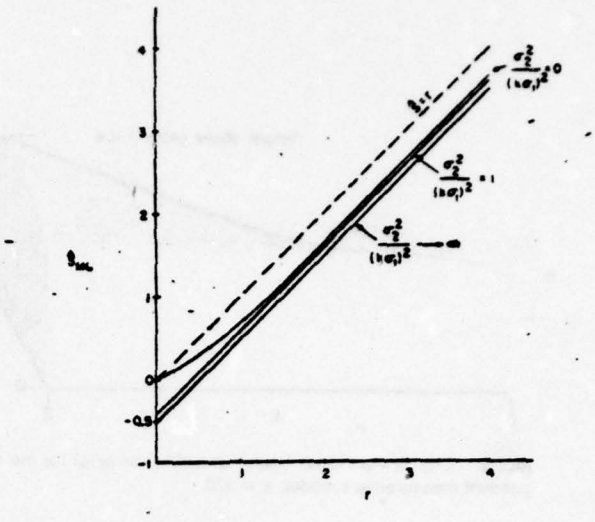


FIG. 5. ML estimator structures.

Another point worthy of note is the similarity between Eq. (13), the general MAP equation, and the ML equation, Eq. (24). These expressions differ only by an additional term in Eq. (13), and it is this term which contains all of the prior knowledge about s. This term vanishes when $\ln p(s)$, and hence $p(s)$, is constant. In other words, if s is distributed uniformly over all of its space of definition (a worst case), then knowledge of its value in no way affects the maximum of $p(s|r) = p(r|s)p(s) \approx cp(r|s)$. Thus, the ML estimator can be viewed as a worst case of the MAP estimator. Because the MAP estimator embodies *a priori* information about s that is not present in the formation of the ML estimate, it would seem reasonable to assume that the MAP estimate would exhibit a smaller mean-square estimation error than the ML estimate. It will be seen that this is indeed the case. In the next section, bounds on the variances of these estimates will be found.

CRAMÉR-RAO LOWER BOUNDS

A well-known lower bound on the variance of any unbiased estimate for a fixed but unknown s is the Cramér-Rao error bound.[2] Given the conditional density $p(r|s)$, the Cramér-Rao bound is given by

$$\text{var}\,[\hat{s} - s|s] \geq \left[-E\left(\frac{\partial^2 \ln p(r|s)}{\partial s^2}\right)\right]^{-1}. \quad (29)$$

For n_1 and n_2 normal with zero mean, Eq. (29) reduces to

$$\text{var}\,[\hat{s} - s|s] \geq \frac{2[v(s)]^2}{2v(s) + [v'(s)]^2}, \quad (30)$$

where $v(s)$ and $v'(s)$ are as given by Eq. (12). For the signal-independent noise model, which is the result of letting $k = 0$ in Eq. (30), the Cramér-Rao bound is given by

$$\text{var}\,[\hat{s} - s|s] \geq \sigma_2^2, \quad (31)$$

which is the variance actually achieved by the ML estimate of Eq. (28) for the signal-independent noise case. When equality holds in Eq. (29), the estimate $\hat{s}$ is said to be efficient.[2]

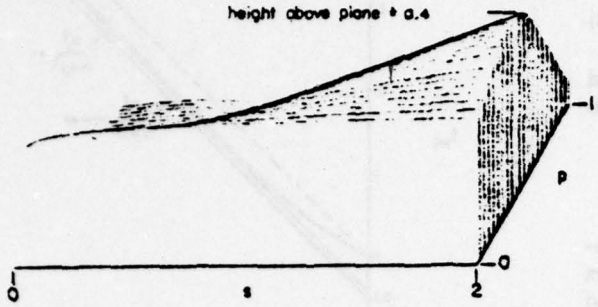


FIG. 6. Cramér-Rao lower bound on estimation error for the signal-dependent measurement model, $k = 1/2$.

Thus, the signal-independent ML estimate is efficient when the measurement is actually of the form given by Eq. (2).

For $f(s) = s^p$, as before, Eq. (30) becomes

$$\text{var}\,[\hat{s} - s|s] \geq \frac{k^4\sigma_1^4 s^{4p} + 2k^2\sigma_1^2\sigma_2^2 s^{2p} + \sigma_2^4}{k^2\sigma_1^2 s^{2p} + 2p^2k^4\sigma_1^4 s^{4p-1} + \sigma_2^2}, \quad (32)$$

which for $p = 1/2$ reduces to

$$\text{var}\,[\hat{s} - s|s] \geq \frac{k^4\sigma_1^4 s^2 + 2k^2\sigma_1^2\sigma_2^2 s + \sigma_2^4}{\left(k^2\sigma_1^2 + k^4\sigma_1^4/2\right)s + \sigma_2^2}. \quad (33)$$

Although it is not obvious by inspection, the bound given by Eq. (32) may actually be smaller than the bound given by Eq. (31). In other words, there are potentially cases where the estimators designed for the signal-dependent measurement model may actually out perform (in a mean-square-estimation-error sense) the estimators designed for the signal-independent measurement model. To better illustrate

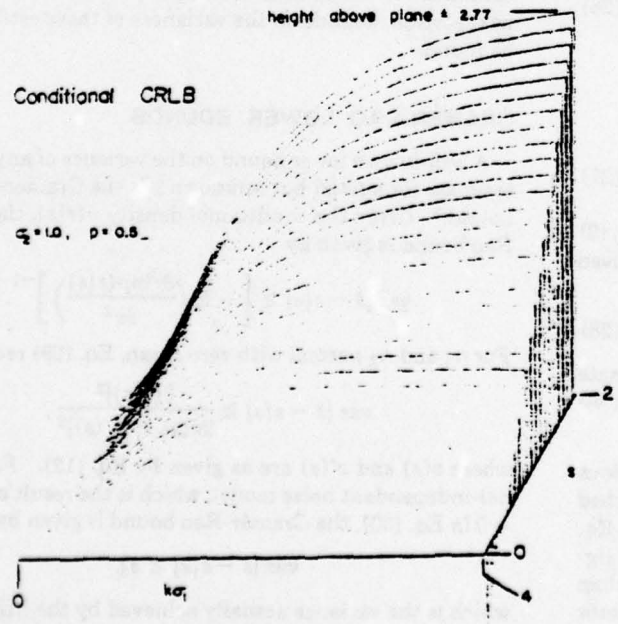


FIG. 7. Cramer-Rao lower bound on estimation error for the signal-dependent measurement model, $p = 1/2$.

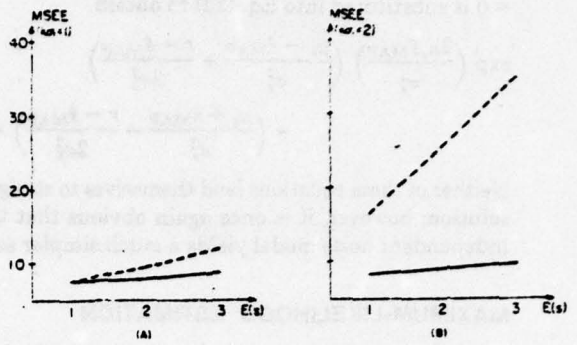


FIG. 8. Mean-square estimation error for the MAP estimator, as a function of the signal mean $E(s)$, with (a) $k\sigma_1 = 1$ and (b) $k\sigma_1 = 2$. The solid line is the signal-dependent estimator error and the dashed line is the mismatched estimator error, and $\sigma_2^2 = 1$.

this. Eq. (32) is plotted in Figs. 6 and 7. In the first of these k is fixed at 1/2, σ_1^2 and σ_2^2 at one, and s and p are varied. As in Figs. 1 and 2, the plane upon which the surface rests is not the zero plane, but rather is the Cramér-Rao lower bound given by Eq. (31), namely σ_2^2. In Fig. 7, p is fixed at 1/2 and $k\sigma_1$ is allowed to vary. Now it is worth noting that in all of the previous equations, when $k \neq 0$, k and σ_1 always appear together. Thus varying $k\sigma_1$ is tantamount to fixing either one and varying the other. Note that in Fig. 7, for certain values of $k\sigma_1$ and s, the Cramér-Rao bound of Eq. (32) dips below the Cramér-Rao bound of Eq. (31), that is, it dips below the plane σ_2^2. This is, of course, the region mentioned above, where the inclusion of signal-dependence in the measurement model may potentially result in improved estimator performance. The values of s and $k\sigma_1$ which result in this region are given by

$$0 \leq s \leq (\sigma_2^2/2)\,[1 - 2/(k\sigma_1)^2], \quad (34)$$

where $k\sigma_1$ must then satisfy

$$k\sigma_1 \geq \sqrt{2}. \quad (35)$$

Recall that these equations are derived for the $p = 1/2$ case.

To get a feeling for the actual mean-square estimation error achieved by the estimators derived above, Monte Carlo simulations were performed, with the results presented in the next section.

MONTE CARLO SIMULATIONS

The performance of each of the estimators derived in the previous sections was evaluated by Monte Carlo simulations to determine the mean-square estimation error. The results for each of the various signal probability densities were so similar that only one case is presented. The Gaussian case was chosen since, for the MAP estimate, it represents the minimum achievable mean-square estimation error (see Appendix). Figure 8 shows the mean-square estimation error (MSEE) of the MAP estimate plotted as a function of the signal mean $E(s)$. In Fig. 8(a), $k\sigma_1 = 1$, while in Fig. 8(b), $k\sigma_1 = 2$. The solid line is the MSEE for the MAP estimator of Eq. (15) and the dashed line is the MSEE for the mismatched case, that is, for the MAP estimate of Eq. (16) when applied to the signal-dependent measurement. Inclusion of signal depen-

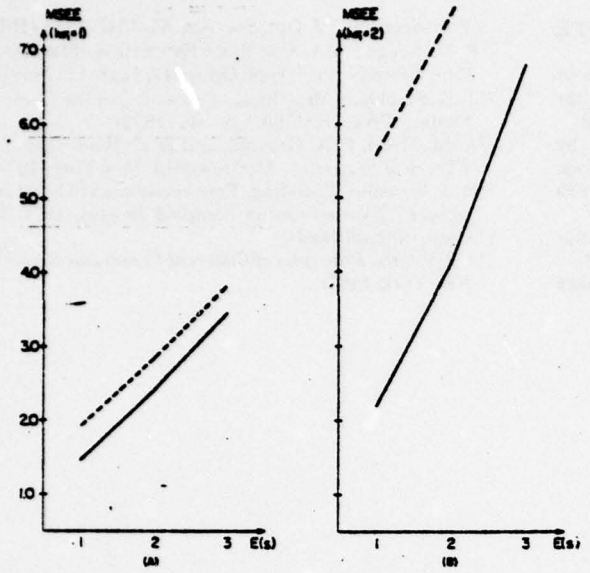


FIG. 9. Mean-square estimation error for the ML estimator, as a function of the signal mean $E(s)$, with (a) $k\sigma_1 = 1$ and (b) $k\sigma_1 = 2$. The solid line is the signal-dependent estimator error and the dashed line is the mismatched estimator error.

dence in the estimator structure is seen to yield estimates of the signal which, on the average, have smaller error than would be the case when signal dependence is ignored. It should be noted that for sufficiently small $k\sigma_1$ and small signal means the signal-dependent noise term is negligible. This results in the estimates for the mismatched case being very nearly equal to those which include the signal-dependence.

Figure 9 presents the results of simulations of the ML estimators. As before, the solid line represents the signal-dependent estimator MSEE and the dashed line represents the MSEE for the mismatched case. Once again, inclusion of signal dependence is seen to yield better estimates on the average. Since the ML estimates include no prior knowledge of the signal statistics, their performance is markedly inferior to the MAP estimates, but as previously discussed, the ML estimate represents a worst case. As before, for small $k\sigma_1$ and small $E(s)$, the estimates are very nearly equal regardless of the inclusion of signal dependence in the estimator structure.

CONCLUSION

Many physical processes are described by a signal-dependent observation model. It has been shown that, in such cases, ignoring the signal dependence for purposes of designing estimators of the signal may result in severe penalties in terms of estimation error. Therefore, optimal estimators which include the signal-dependent structure were derived. Specifically, these were ML estimates, which include no prior knowledge of signal statistics, and MAP estimates, which assume prior knowledge of the signal probability density. The latter estimate was derived for the Gaussian, Rayleigh, and folded Gaussian density functions. The performance of these estimators was then investigated by Monte Carlo simulation. As expected, inclusion of signal dependence in the estimator structure resulted in improved estimator performance.

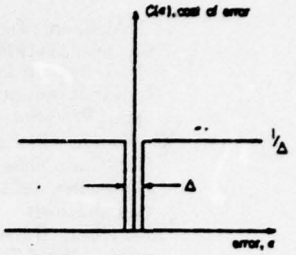


FIG. 10. The uniform cost function.

It should be noted that there may exist suboptimal estimators which prove to be more desirable in terms of implementation or other practical considerations. For example, no mention has been made of how to obtain the signal statistics necessary in the formulation of MAP estimators. However, the purposes of this paper were to demonstrate the potentially severe consequences of ignoring signal dependence and to derive optimal estimators for the signal-dependent noise case.

ACKNOWLEDGMENT

This work was supported in part by the Joint Services Electronics Program at Texas Tech University under Office of Naval Research Contract No. 76-C-1136.

APPENDIX

Bayesian estimators are those estimators which serve to minimize the Bayes risk, where the Bayes risk is the expected cost of estimation based on some cost function. For example, minimum mean-square error is achieved when the cost is proportional to the square of the estimation error, i.e., when the cost function is a parabola. The MAP estimator is a Bayesian estimator based on the uniform cost function shown in Fig. 10.[2] The cost for no error is zero (as it is for some Δ region about no error), and the cost of any other error is uniform (all errors are weighted equally).

It can be shown[2] that, under certain conditions, the optimal Bayes estimate is invariant for a variety of cost functions, and is equal to the minimum mean-square error estimate. These conditions are: (i) the cost function is convex, (ii) the cost function is symmetrical, (iii) the *a posteriori* probability density, $p(s|r)$, is symmetrical, and (iv) $\lim_{s\to\infty} C(s)p(s|r) = 0$, where $C(s)$ is the cost function with argument s. Condition (iv) is simply a requirement that the *a posteriori* density goes to zero faster than the cost function increases. Viterbi[14] has shown that the uniform cost function satisfies these conditions. When the prior signal density, $p(s)$, is assumed Gaussian, then clearly $p(s|r)$ is symmetrical, as required in condition (iii). Thus, for this case we have the optimal Bayesian estimate, and it is the estimate which yields the minimum mean-square estimation error.

* Present address: ORINCON Corp., 3366 N. Torrey Pines Ct., Suite 320, LaJolla, Calif. 92037.

[1]J. B. Thomas, *An Introduction to Statistical Communication Theory* (Wiley, New York, 1969).
[2]H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part 1* (Wiley, New York, 1968).
[3]H. C. Andrews and B. R. Hunt, *Digital Image Restoration* (Prentice-Hall, Englewood Cliffs, New Jersey, 1977).
[4]J. F. Walkup and R. C. Choens, "Image Processing in Signal-Dependent Noise," Opt. Eng. 13, 258–266 (1974).

[7]J. C. Mallison, "Tutorial Review of Magnetic Recording," Proc. IEEE 64, 196–223 (1976).
[8]J. W. Goodman and J. F. Belsher, "Fundamental Limitations in Linear Invariant Restoration of Atmospherically Degraded Images," Proc. Soc. Photo-Opt. Instrum. Eng. 75, 141–154 (1976).
[9]F. Naderi, "Estimation and Detection of Images Degraded by Film-Grain Noise," Ph.D. thesis (University of Southern California, September, 1976). USC Image Processing Institute report 690 (unpublished).
[10]B. R. Hunt, "Bayesian Methods in Nonlinear Digital Image Restoration," IEEE Trans. on COMPUTERS C-26, 219–229 (1977).
[11]H. J. Zwieg, E. B. Barrett and P. C. Hu, "Noise-Cheating Image Enhancement," J. Opt. Soc. Am. 65, 1347–1353 (1975).
[12]F. Naderi and A. A. Sawchuk, "Estimation of Images Degraded by Film Grain Noise," Appl. Optics 17, 1228–1237 (1978).
[13]T. G. Stockham, Jr., "Image Processing in the Context of a Visual Model," Proc. IEEE 60, 828–842 (1972).
[14]A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics* (McGraw-Hill, New York, 1974).
[15]R. J. Arguello, "Encoding, Transmission and Decoding of Sampled Images," Symposium on Sampled Images, 1971, Perkin-Elmer Corp. (unpublished).
[16]A. J. Viterbi, *Principles of Coherent Communication* (McGraw-Hill, New York, 1966).

10. Reprint of "Optimal Estimation in Signal-Dependent Film-Grain Noise" by G.K. Froehlich, J.F. Walkup, and R.B. Asher from the Proceedings of the 11th International Commission for Optics Conferences, Madrid, September 1978, pp. 367-369.

OPTIMAL ESTIMATION IN SIGNAL-DEPENDENT FILM-GRAIN NOISE

G. K. Froehlich, J. F. Walkup and R. B. Asher*

Dept. of Electrical Engineering, Texas Tech University, Lubbock, Tx., 79409

INTRODUCTION

Many physical noise processes are signal-dependent. One well known example is film-grain noise (1-3).

In this paper optimal estimators for images in signal-dependent film-grain noise are presented.

THE MODEL

A versatile model incorporating both signal-independent additive noise and signal-dependent noise is utilized. This model is given in Eq. (1),

$$r = s + kf(s)n_1 + n_2, \qquad (1)$$

where r is the observed photographic density, s is the original uncorrupted image density, k is the scanning constant, f(s) is some function of s, and n_1 and n_2 are signal independent noise processes. Thus, the middle term on the right-hand side of Eq. (1) is the signal-dependent noise term.

It is assumed that n_1, n_2, and s are mutually statistically independent. To apply the model to film-grain noise problems, let $f(s) = s^p$, where p is usually taken to be 1/2 or 1/3 (1-3).

In this paper, we let p = 1/2 and we assume n_1 and n_2 are zero mean Gaussian random variables, with variances σ_1^2 and σ_2^2, respectively. Further, s is assumed to be a Gaussian random variable with mean μ_s and variance σ_s^2.

THE ESTIMATOR STRUCTURES

The maximum likelihood (ML) estimate is found by maximizing p(r/s) over s (3). For the model of Eq. (1), the estimate is found to be

*Present address: ORINCON Corp., 3366 N. Torrey Pines Ct LaJolla, CA. 92037.

$$\hat{s}_{ML} = \left[r^2 + \left(\frac{k^2\sigma_1^2}{2}\right)^2 + \frac{2r\sigma_2^2}{k^2\sigma_1^2} + \left(\frac{\sigma_2^2}{k^2\sigma_1^2}\right)^2\right]^{1/2} - \frac{k^2\sigma_1^2}{2} - \frac{\sigma_2^2}{k^2\sigma_1^2}, \tag{2}$$

as compared to the simple estimate

$$\hat{s}_{ML} = r \tag{3}$$

which results when the signal-dependent noise term of Eq. (1) is omitted.

The maximum a posteriori probability (MAP) estimate is found by maximizing $p(s/r)$ over s(3). For the model of Eq. (1) and the above assumptions, the estimate $\hat{s}_{map}$ is found to be the solution of

$$\left[\frac{2k^4\sigma_1^4}{\sigma_s^2}\right]\hat{s}^3 + \left[\frac{4k^2\sigma_1^2\sigma_2^2 - 2k^4\sigma_1^4\mu_s}{\sigma_s^2} + 2k^2\sigma_1^2\right]\hat{s}^2$$

$$+ \left[\frac{2\sigma_2^4 - 4k^2\sigma_1^2\sigma_2^2\mu_s}{\sigma_s^2} + k^4\sigma_1^4 + 2\sigma_2^2\right]\hat{s}$$

$$+ \left[k^2\sigma_1^2(\sigma_2^2 - r^2) - 2\sigma_2^2 r - \frac{2\sigma_2^4\mu_s}{\sigma_s^2}\right] = 0. \tag{4}$$

Again, omission of the signal-dependent noise term in Eq. (1) results in a comparatively simplified estimate,

$$\hat{s}_{MAP} = \frac{\sigma_s^2}{\sigma_s^2+\sigma_2^2} r + \frac{\sigma_2^2}{\sigma_s^2+\sigma_2^2} \mu_s. \tag{5}$$

Because this MAP estimate includes prior information about the image, it should give superior performance. In fact, under the above assumptions it can be shown that the MAP estimator minimizes the mean square estimation error (3).

RESULTS

Figure 1 is the original image of an archer. Figure 2 is the noisy image generated digitally according to the model of Eq. (1). The image in Fig. 3 is the estimate found by the solution of the MAP equation, Eq. (4), with μ_s and σ_s^2 taken to be the sample mean and variance of the original image.

One factor severely affecting estimator performance is violation of the assumption that the image statistics are

Gaussian. For a discussion of this, see the paper by Froehlich et.al. (3).

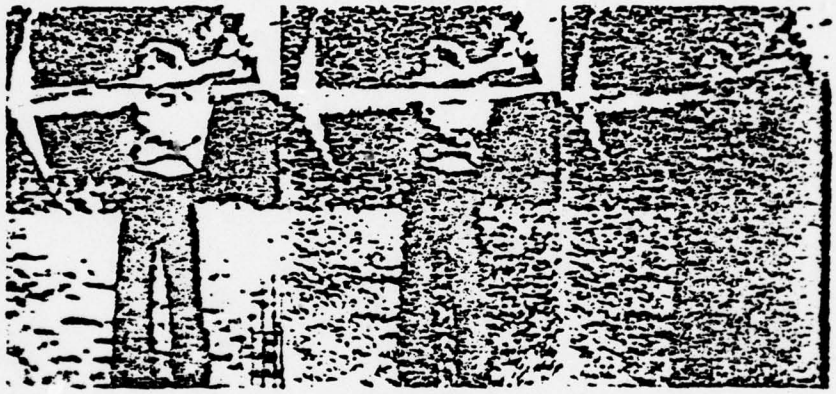


Fig. 1. The original image k = 0.5

Fig. 2. The noisy image, σ_1=0.4, σ_2=0.1, σ_s=0.2

Fig. 3. The estimate.

ACKNOWLEDGEMENTS

This work was supported in part by the Joint Services Electronics Program at Texas Tech University under Office of Naval Research Contract No. 76-C-1136.

The authors are also grateful for the assistance of Gus Oliver, who contributed to this paper while an Undergraduate Research Participant sponsored by the National Science Foundation at Texas Tech University.

REFERENCES

1. J. Walkup and R. Choens, Optical Engineering 13, 258-266 (1974).
2. F. Naderi and A. Sawchuk, Appl. Optics 17, 1228-1237 (1978).
3. G. Froehlich, J. Walkup and R. Asher, to appear J. Opt. Soc. Am., Oct. 1978.

Texas Tech University — Institute for Electronic Science

Joint Services Electronics Program — Research Unit: 7

1. Title of Investigation: Pattern Recognition
2. Senior Investigator: T.G. Newman — Telephone: (806) 742-3528
3. JSEP Funds: $23,500
4. Other Funds: $15,000*
5. Total Number of Professionals: PI's 1 (1 mo.) RA's 1 (1/2 time
6. Summary:

The goal of this work unit is the development of pattern recognition algorithms for moving objects. The motions which an object may undergo are modeled by a Lie group (of translations, magnifications, rotations, etc.) and we are attempting to formulate pattern recognition algorithms which use group invariants to discriminate among patterns. In addition we are investigating several algorithms which can track the motion of a pattern and we hope to develop pointing and tracking algorithms in which both the camera and the object are in motion.

Our main activity during the past year has been the development of an algorithm for identifying multiple moving objects in a scene. The key to this algorithm is the use of an equation of motion which is formulated relative to the coordinate system of the Lie group, rather than a Euclidian coordinate system. In this coordinate system, every point in a rigid body is moving with the same velocity and, as such, if one numerically computes a velocity profile for a scene from photographic data a piecewise constant profile will result with distinct levels corresponding to distinct objects. An algorithm based on these ideas has

*ONR support to permit Professor Newman and a graduate student to work on a program related to this work unit at WSMR during the summer of 1979. (This program was administered as an add-on to JSEP.)

been experimentally implemented at WSMR with considerable success and several papers have been prepared which describe various aspects of the theory and its numerical implementation.

In parallel with the above work we have also continued our investigation of pattern recognition algorithms which use group invariants and relative invariants to discriminate patterns. This work has resulted in several M.S. reports and the publication of a conference paper which is reprinted in this report.

7. Publications and Activities:

A. Conference Papers and Abstracts

1. Newman, T.G., "A Group Theoretic Approach to Invariance and Pattern Recognition", Proc. of the IEEE Conf. on Pattern Recognition and Image Processing", Chicago, Aug. 1979. pp. 407-412.

B. Preprints

1. Newman, T.G., "An Inverse Problem Related to Video Tracking", submitted for publication.

2. Newman, T.G., "Lie Groups and Lie Algebras in Video Tracking", submitted for publication.

3. Fredricks, G., and T.G. Newman, "Results in Differential Geometry with Applications to Video Tracking", submitted for publication.

4. Newman, T.G., and D.A. Demus, "Lie Theoretic Methods in Video Tracking", submitted for publication.

C. Theses

1. Cunningham, D., "Pattern Matching over the Rotation Group", M.S. Report, Texas Tech Univ., 1979.

2. Gimarc, R.L., "Optimization of Sums of Squares without Derivatives", M.S. Thesis, Texas Tech Univ., 1979.

3. Bullard, G., M.S. Report, Texas Tech Univ., (in preparation).

4. Zlobec, L., M.S. Report, Texas Tech Univ., (in preparation).

D. Conferences and Symposia

1. Newman, T.G., Texas Systems Workshop, Southern Methodist Univ., April 1979.

2. Newman, T.G., IEEE Computer Society Conf. on Pattern Recognition and Image Processing, Chicago, August 1979.

3. Newman, T.G., SIAM Fall Meeting, Knoxville, October 1979.

4. Newman, T.G., 6th Annual Computer Science Conference, Denton, April 1979.

E. Lectures

1. Newman, T.G., "A Group Theoretic Approach to Pointing and Tracking", IEEE/AIAA Section Meeting, Holloman AFB, August 1979.

8. Reprint of "A Group Theoretic Approach to Invariance and Pattern Recognition" by T.G. Newman from the Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing, Chicago, August 1979, pp. 407-412.

Abstract

In this paper we consider a class of patterns which are subject to the action of a group of transformations. We are particularly concerned with the existence of measurements or features which are invariant with respect to transformation. A concept of relative invariance is also introduced and explored in depth. In a very general sense, it is shown that every invariant (and relative invariant) is a suitable average over the relevant group of transformations. Finally, invariant means of bounded functions are used to explore existence of pattern invariants. Suggestions for further research are also given.

Key Words and Phrases: Pattern, group, transformation, feature, invariant, relative invariant, group average, invariant mean, representation, linear transformation.

1. Introduction

The importance of group theory as a tool to be exploited in modelling a variety of perceptual phenomena has been demonstrated by a number of writers[2,10,11,16]. Although the influence of group theory is implicit in much of the literature on pattern recognition[1,6,12,15,18], relatively few instances can be found in which explicit utilization of group theory is the central theme[4,5,7,8,17]. Without exception, group theory has been used to effectively model some aspect or feature which is invariant under transformation and to exploit this invariance in performing the recognition function. However, no definitive study has been made of transformational invariance and no general model has been introduced which attempts to formalize the concept of invariance as it relates to pattern recognition. This is indeed strange in view of the relatively advanced state of the theory of invariants within group theory[9,14,19].

In the following we formulate a general model in which many problems in pattern recognition may be cast in a natural fashion. We discuss representations of patterns as functions defined on a group and proceed to investigate the existence of invariant functionals

2. The Model

Let Ω denote a set of objects called patterns and assume that G is a group of transformations which act on Ω on the left. For $\omega \in \Omega$ and $g \in G$ we denote by $g\omega$ the image of ω under the transformation g. Also, for $g_1, g_2 \in G$ we denote their product or composition by $g_1 g_2$. Action on the left is then given by the identity

$$(g_1 g_2)\omega = g_1(g_2\omega), \tag{1}$$

for $g_1, g_2 \in G$ and

We now assume that our ability to "recognize" and/or otherwise "classify" patterns is obtained via measurements performed upon individual patterns. Such measurements can take values of a quite general nature, although the usual situation will result in a vector of real numbers. Accordingly, we define a measurement function to be a mapping $R: \Omega \to V$, where V is a suitable set of permissible values. We shall later assume that V is a real finite-dimensional vector space. We say that a measurement function $R: \Omega \to V$ is invariant provided that $R(g\omega) = R(\omega)$ for all $\omega \in \Omega$ and $g \in G$. Observe that an invariant measurement does not distinguish between the various members of an orbit $[\omega] = \{g\omega \mid g \in G\}$, being constant on each such orbit.

More generally, we say that a measurement $R: \Omega \to V$ is relatively invariant provided that $R(g\omega) = \rho(g)R(\omega)$. Here ρ is a homomorphism of G into a group of transformations on V and is called the modulus of R. As a matter of practice, we are interested in the case in which V is a finite dimensional vector space and ρ is a representation of G in the group $GL(V)$ of invertible linear transformations on V. Note that a relative invariant not only depends upon the orbit of ω but is also sensitive to "position" within the orbit.

In applications one must solve simultaneous equations $R_\alpha(\omega) = R_\alpha^m$ involving a number of invariants $\{R_\alpha\}$ and associated actual measurements $\{R_\alpha^m\}$ to classify the orbit of ω and then solve similar equations $\rho_\beta(g)R_\beta(\omega) = R_\beta^m$ involving relative invar-

+ This work has been conducted under the auspices of the Associate Joint Services Electronics Program at Texas Tech University. The Office of Naval Research is gratefully acknowledged for support under contract N0014-76-C-1136.

iants to determine position within the orbit. Hence we see that the question of existence of invariants and relative invariants become of paramount importance.

3. Representations

In order to pursue the question of existence of invariants we find the need of considerably more structure than we have assumed at this point. It is somewhat surprising that this additional structure can be imposed on the transformation group and need not involve restrictive assumptions about the space of patterns. Since the transformation groups that are typically encountered are quite rich in structure, we find ourselves in an advantageous situation.

Let us briefly digress. Suppose that X is any set and that the group G acts on X on the left. Let $f: X \to Y$ be a mapping of G to some set Y. Then for any $g \in G$ we may define a new mapping $f: X \to Y$ given by

$$(gf)(x) = f(g^{-1}x), \quad x \in X. \tag{2}$$

Note that appearance of g^{-1}, rather than g, is a convenience which makes certain formulae more natural for later use. We easily verify that

$$g_1(g_2 f) = (g_1 g_2)f \tag{3}$$

so that if F is a set of functions such that $gf \in F$ for all $f \in F$, then (2) defines an action of G on the left of F.

Now let $R: \Omega \to V$ be a given measurement function. For each $\omega \to \Omega$ we may define a function $\omega^r: \Omega \to V$ as follows:

$$\omega^r(x) = R(x^{-1}\omega), \quad x \in G. \tag{4}$$

The correspondence $r: \omega \to \omega^r$ thus defines a mapping of Ω into the set F(G,V) of functions from G to V. Now for fixed $g \in G$ we see that for all $x \in G$, $(g\omega^r)(x) = \omega^r(g^{-1}x) = R((g^{-1}x)^{-1}\omega) = R((x^{-1}g)\omega) = R(x^{-1}(g\omega)) = [(g\omega)^r](x)$. That is,

$$g\omega^r = (g\omega)^r, \tag{5}$$

for all $\omega \in \Omega$ and $g \in G$. Equation (5) establishes the desired connection between our patterns and the V-valued functions on G. We define a representation of Ω on V to be a map $r: \Omega \to F(G,V)$ which satisfies (5). Such a representation allows a <u>concrete</u> interpretation of patterns as suitable functions defined on the group.

We have the following:

<u>Theorem 1</u>. The representations r of Ω on V correspond one-to-one to the measurement functions $R: \Omega \to V$. The correspondence is given via $r \leftrightarrow R$ if and only if

$$\omega^r(x) = R(x^{-1}\omega), \quad x \in G, \ \omega \in \Omega.$$

Proof: We have already seen that each measurement R defines a representation. Now, let $\omega \to \tilde{\omega}$ be a representation of Ω on V and let us set $R(\omega) = \tilde{\omega}(1_G)$, where 1_G is the identity element of G. We must show that ω^r as defined by (4) satisfies $\omega^r = \tilde{\omega}$. But for $x \in G$ we have $\tilde{\omega}(x) = (x^{-1}\tilde{\omega})(1_G) = R(x^{-1}\omega) = \omega^r(x)$, from which $\tilde{\omega} = \omega^r$, as desired.

Let us point out that an invariant measurement is characterized by the condition that each ω^r is a constant function, which on the surface seems somewhat uninteresting. This is a deceptive simplification, however, as will be apparent later. Similarly, if R is a relative invariant with modulus ρ, we see that $\omega^r(x) = \rho(x^{-1})R(\omega)$.

Before proceeding to pursue the existence of invariants, it seems appropriate to further accent the importance of relative invariants by demonstrating one of their fundamental properties. Let $R: \Omega \to V$ be a relative invariant with modulus ρ. Suppose that $\omega_1, \omega_2 \in \Omega$ and that $R(\omega_1) = R(\omega_2)$. Then for any $g \in G$ we have $R(g\omega_1) = \rho(g)R(\omega_1) = \rho(g)R(\omega_2) = R(g\omega_2)$. Thus

$$R(\omega_1) = R(\omega_2) \text{ implies } R(g\omega_1) = R(g\omega_2).$$

It is somewhat interesting to note that the condition above is a complete characterization of relative invariants as is shown in the following:

<u>Theorem 2</u>. In order that $R: \Omega \to V$ be relatively invariant it is necessary and sufficient that

$$R(\omega_1) = R(\omega_2) \text{ implies } R(g\omega_1) = R(g\omega_2)$$

for all $g \in G$.

Proof: Necessity has already been shown. Conversely, suppose that $R: \Omega \to V$ satisfies the stated condition. We must construct a homomorphism of G into the group Sym(V) of transformations on V. For $v = R(\omega) \in V$ and $g \in G$, let us define $\rho(g)v = R(g\omega)$. We note that this definition does not depend on ω for if also $v = R(\omega')$ then $R(g\omega') = R(g\omega)$ by the property of R. If $v \in V$ is <u>not</u> of the form $v = R(\omega)$ then set $\rho(g)v = v$. We easily verify that each $\rho(g) \in \text{Sym}(V)$. Also, $\rho(g_1)\rho(g_2)v = \rho(g_1)\rho(g_2)R(g\omega) = \rho(g_1)R(g_2\omega) = R(g_1g_2\omega) = \rho(g_1g_2)v$ in case $v = R(\omega)$ and $\rho(g_1)\rho(g_2)v = v = \rho(g_1g_2)v$ otherwise. Thus, $\rho(g_1)\rho(g_2) = \rho(g_1g_2)$ so that ρ is indeed a homomorphism.

Finally, by definition of $\rho(g)$ we see that $R(g\omega) = \rho(g)R(\omega)$ for all $g \in G$ and $\omega \in \Omega$ so that R is a relative invariant with ρ as modulus.

4. Invariants and Relative Invariants

As previously stated, we may impose additional structure by invoking restrictions on the transformation group. Henceforth, we assume that V is a real (or complex) vector space of finite dimension and that G is a locally compact topological group.

Such a group admits a left invariant integral, called left Haar measure and the integration theory for such groups is well established[14].

The fundamental technique for construction of invariants will be the computation of average values over the entire group G. This technique was exploited by Pitts and McCulloch[16] in their classic work on the perception of audio and visual forms. It appears also in the classical theory of group representations[19] and is prevalent in modern analysis[9,14]. Group averaging has been used as a tool in pattern recognition in a relatively few instances, for example implicitly in [1,12] and explicitly in [5,7,8].

Now, let μ denote the left Haar measure of G and let $f: G \to V$. We define the mean value of f, provided it exists, by

$$M(f) = \lim_{K \uparrow G} \frac{1}{\mu(K)} \int_K f \, d\mu, \tag{6}$$

where K is a compact subset of G and the limit is taken as K increases. Note that K compact implies that $g^{-1}K$ is compact and that $\mu(K) = \mu(g^{-1}K)$. This together with the fact that $\int_K gf\,d\mu = \int_{g^{-1}K} f \, d\mu$, $g \in G$, shows the following:

Lemma 1. If M(f) exists then for any $g \in G$, M(gf) exists and M(f) = M(gf).

We denote by L(V), or simply L, the set of all $f: G \to V$ for which M(f) exists. We have the following:

Lemma 2. L(V) is a linear space on which G acts as the left as a group of linear transformations. Moreover, M is an invariant linear transformation of L(V) into V.

More generally, let ρ be a representation of G in the group GL(V) of invertible linear transformations on V. We form the weighted average of $f: G \to V$, provided the limit exist, as follows:

$$M_\rho(f) = \lim_{K \uparrow G} \frac{1}{\mu(K)} \int_K \rho(x) f(x) d\mu(x), \tag{7}$$

where K is compact, as above. The set of functions for which $M_\rho(f)$ exists will be denoted by $L(V,\rho)$, or simply $L(\rho)$. Note that by the substitution $y = g^{-1}x$ we obtain $\int_K \rho(x) gf(x) d\mu(x) = \int_K \rho(x) f(g^{-1}x) d\mu(x) = \int_{g^{-1}K} \rho(g) f(y) d\mu(y) = \rho(g) \int_{g^{-1}K} \rho(y) f(y) d\mu(y)$. It follows immediately that

$$M_\rho(gf) = \rho(g) M_\rho(f), \tag{8}$$

for all $g \in G$, $f \in L(V,\rho)$. We conclude:

Theorem 3. $L(V,\rho)$ is a linear space on which G acts as a group of linear transformations. Also, M_ρ is a linear mapping of $L(V,\rho)$ into V which is relatively invariant with modulus ρ.

Let us define $\rho': G \to GL(V)$ by $\rho'(x) = \rho(x^{-1}) = (\rho(x))^{-1}$. We have $\rho'(xy) = \rho'(y)\rho'(x)$, so that ρ' is a dual homomorphism. For $f \in L(V)$ we may consider the product $\rho' f$ given by $(\rho' f)(x) = \rho'(x) f(x)$, $x \in G$. Since $\rho(x)\rho'(x) = 1_V$ we see that $\rho' f \in L(V,\rho)$ whenever $f \in L(V)$. Similarly, $f \in L(V)$ implies $\rho f \in L(V,\rho)$.

We evidently then can use ρ and ρ' as multipliers to pass back and forth between L(V) and $L(V,\rho)$. Thus:

Lemma 3: The map $f \to \rho' f$ is a linear isomorphism from L(V) onto $L(V,\rho)$. Moreover, $M(f) = M_\rho(\rho' f)$ for all $f \in L(V)$.

Although this shows that the linear structure of L(V) and $L(V,\rho)$ are no different, it is important to observe that they are quite different with respect to the action of G.

We may now state sufficient conditions for the existence of invariants and/or relative invariants for the pattern space Ω. Quite simply, if $R: \Omega \to V$ is such that each $\omega^r \in L(V,\rho)$ then we obtain a relative invariant $\bar{R}$ by defining

$$\bar{R}(\omega) = M_\rho(\omega^r). \tag{9}$$

We see that $\bar{R}(g\omega) = M_\rho((g\omega)^r) = M_\rho(g\omega^r) = \rho(g) M_\rho(\omega^r) = \rho(g)\bar{R}(\omega)$, as desired. We obtain the corresponding result for invariants in the special case in which ρ is the trival representation, $\rho(g) \equiv 1_V$.

Recall that if $R: \Omega \to V$ is relatively invariant with modulus ρ, then we may write $\omega^r(x) = \rho(x^{-1})R(\omega)$ for all $\omega \in \Omega$, $x \in G$. Thus, we have for each compact subset K of G, $\int_K \rho(x)\omega^r(x)d\mu(x) = \int_K R(\omega)d\mu(x) = \mu(K)R(\omega)$. Comparison with (7) shows that $M_\rho(\omega^r)$ exists and is equal to $R(\omega)$. This shows that each $\omega^r \in L(V,\rho)$ and shows as well the identity $M_\rho(\omega^r) = R(\omega)$. We have thus shown:

Theorem 4. If $R: \Omega \to V$ is such that each $\omega^r \in L(V,\rho)$, then $\bar{R}(\omega) = M_\rho(\omega^r)$ defines a relative invariant $\bar{R}$ with modulus ρ. Conversely, every relative invariant is precisely of this form, since if R is relatively invariant with modulus ρ, then each $\omega^r \in L(V,\rho)$ and $\bar{R} = R$.

The above may be paraphrased by saying that the construction of relative invariants with a given modulus ρ is equivalent to the construction of a representation $\omega \to \omega^r$ of Ω on V such that each $\omega^r \in L(V,\rho)$. Observe that, in particular, we have shown in a strict sense that every relative invariant is a weighted average over the entire group G.

The result in Theorem 4 gives valuable insight to the nature of invariants and relative invariants. Nevertheless, it is less than satisfying in cer-

tain ways. In the first place, it gives no clue as to how to construct a suitable R, although it can certainly eliminate a number of choices. Consequently, it is not a true existence theorem in the sense that for a given application it does not actually produce an invariant. Moreover, there are many examples of invariants which occur in natural ways but are not presented in the form given above (although they are necessarily equivalent to such a form).

5. Existence of Invariants

The consideration of the group average in the preceding section led to the existence of relative invariants and is applicable in any situation where each ω^r belongs to a class of functions for which such an average exists. This is the case, for example, when the class of functions is almost periodic (in the sense of J. von Neumann [9]). It is, however, applicable in a wider variety of cases, namely those in which set B(G) of bounded real valued functions admits an invariant mean, in the following sense:

Definition: An invariant mean M on the class B(G) of bounded real valued functions on G is a real linear functional M on B(G) which is invariant under the action of G on B(G) and satisfies

$$\inf f \leq M(f) \leq \sup f, \quad f \in B(G). \tag{10}$$

Let V* denote the dual space of V and let B(G,V) denote the bounded functions from G to V. For each $f \in B(G,V)$ and for each $v^* \in V^*$ we observe that $v^* \circ f \in B(G)$.

Lemma 4. Let $M_0 \in (B(G))^*$, the dual of B(G). There exists a unique linear transformation $M: B(G,V) \to V$ such that

$$v^* \circ M = M_0 \circ v^* \tag{11}$$

for all $v^* \in V^*$.

Proof: It is clear that any M satisfying (11) is unique. Let $v_1, v_2, \ldots, v_n$ be a basis in V and let $v_1^*, v_2^*, \ldots, v_n^*$ be a dual basis in V*, so that $\langle v_i^*, v_j \rangle = \delta_{ij}$. Let us define $M: B(G,V) \to V$ by

$$M(f) = \sum_{i=1}^{n} M_0(v_i^* \circ f) v_i, \quad f \in B(G,V). \tag{12}$$

Then for any $v^* \in V^*$ we have $v^* \circ M(f) = \langle v^*, M(f) \rangle = \sum_{i=1}^{n} \langle v^*, v_i \rangle M_0(v_i^* \circ f) = M_0\left(\sum_{i=1}^{n} \langle v^*, v_i \rangle v_i^* \circ f\right) = M_0(v^* \circ f)$. That $v^* \circ M = M_0 \circ v^*$, as desired.

Let us observe that for any linear transformation $A: V \to V$ and $f \in B(G,V)$ we may define the composite Af so that B(G,V) may be considered as a (left) module over the ring Lin(V,V) of linear transformations on V. With this in mind, we observe:

Lemma 5. The linear map $M: B(G,V) \to V$ defined by (11) above is a morphism of B(G,V) to V considered as modules over Lin(V,V).

Proof: For any $A \in \mathrm{Lin}(V,V)$, $v^* \in V^*$ and $f \in B(G,V)$, we have $v^* \circ M(Af) = M_0(v^* \circ Af) = M_0((A^*v^*) \circ f) = A^*v^* \circ M(f) = v^* \circ AM(f)$. Hence, $M(Af) = AM(f)$, completing the proof.

Lemma 6: If $M_0 \in (B(G))^*$ is invariant under G then so is the map M as defined by (11).

Proof: M_0 invariant implies that for any $f_0 \in B(G)$ and $f \in G$, we have $M_0(gf_0) = M_0(f_0)$. Thus, if $v^* \in V^*$, $f \in B(G,V)$ then for any $g \in G$ we have $v^* \circ M(gf) = M_0(v^* \circ f) = M_0(g(v^* \circ f)) = M_0(v^* \circ f) = v^* \circ M(f)$. Then by Lemma 4, $M(gf) = M(f)$, as desired.

We have thus shown how to "lift" invariant linear functionals on B(G) to invariant linear maps from B(G,V) to V.

Corollary: If G admits an invariant mean M_0 and $R: \Omega \to V$ is such that each $\omega^r \in B(G,V)$, then

$$\bar{R}(\omega) = M(\omega^r) \tag{13}$$

defines an invariant measurement, where M is given by (11).

Proof: $\bar{R}(g\omega) = M((g\omega)^r) = M(g\omega^r) = M(\omega^r) = \bar{R}(\omega)$.

We may obtain relative invariants in a similar fashion. However, to remain within the bounded functions, we restrict our attention to unitary representations of G.

Let us suppose that M_0 is a given invariant mean on the class of bounded function on G and that M is the lifted map defined by (11) above. Also, let ρ be a given unitary representation of G in GL(V). Observe then that for each $f \in B(G,V)$ we have also $\rho f \in B(G,V)$, where $(\rho f)(g) = \rho(g)g(g)$, $g \in G$. Now, let $R: \Omega \to V$ be a given measurement function such that each $\omega^r \in B(G,V)$. Since this simply means that the values of R on the orbit of ω are bounded, this is not deemed to be a serious restriction.

With this in mind, let us note that $\rho(g\omega)^r = \rho(g)g(\rho\omega^r)$ for all $g \in G$, $\omega \in \Omega$. To see this, we have, at any $x \in G$, $[\rho(g\omega^r]](x) = \rho(x)(g\omega^r)(x) = \rho(x)\omega^r(g^{-1}x) = \rho(g)\rho(g^{-1}x)\omega^r(g^{-1}x) = \rho(g)[g(\rho\omega^r)](x)$. Also, let us observe that, for fixed g, $\rho(g) \in \mathrm{Lin}(V,V)$ and that M is a morphism of Lin(V,V)-modules. We now define $\bar{R}: \Omega \to V$ by the formula

$$\bar{R}(\omega) = M(\rho\omega^r), \quad \omega \in \Omega. \tag{14}$$

Recalling the invariance of M, and the facts above, we see that for $g \in G$, we have $\bar{R}(g\omega) = M(\rho(g\omega^r)) = M(\rho(g)g(\rho\omega^r)) = \rho(g)M(g(\rho\omega^r)) = \rho(g)M(\rho\omega^r) = \rho(g)\bar{R}(\omega)$.

That is, $\bar{R}$ is a relative invariant and has the given representation ρ as its modulus. We have,

therefore, proved the following remarkable result:

Theorem 5. If B(G) admits an invariant mean M_0, ρ is any unitary representation of G in GL(V), and a non-trivial bounded measurement function $R: \Omega \to V$ exists, then there exists a non-trival relative invariant $\bar{R}: \Omega \to V$ with modulus ρ. $\bar{R}$ is given implicitly by

$$\bar{R}(\omega) = M(\rho\omega^r), \qquad (15)$$

where M is the lift of M_0 to B(G,V).

The appearance of the words non-trival in the above requires slight explanation. We can clearly define $\bar{M}: B(G,V) \to V$ by $\bar{M}(f) = M(\rho f)$ and deduce that $\bar{M}(gf) = \rho(g)\bar{M}(f)$. The fact that $M \not\equiv 0$ gives $\bar{M} \not\equiv 0$. Since $R(\omega) = M(\omega^r)$, we see the sense in which $\bar{R}$ is non-trivial, i.e., it is the restriction of $\bar{M}$ to the functions $\Omega^r = \{\omega^r | \omega \in \Omega\}$. Nevertheless, it could happen that each $\rho\omega^r$ is annihilated by M so that $\bar{R} \equiv 0$ even though $R \not\equiv 0$. This is unlikely and can be ignored if, for instance, we have some $\rho\omega^r > 0$, since for such $\omega \in \Omega$ we see that $\bar{R}(\omega) > 0$.

6. Summary and Suggestions for Further Research

We have shown that every set of patterns subject to a transformation group is representable as functions defined on the group and that such representations are implicit in the measurement process. It has also been shown that every relative invariant is equivalent to a weighted average of a measurement on the patterns taken over the relevant group of transformations. Moreover, the existence of suitably many relative invariants have been demonstrated in any situation in which measurements are bounded and the group admits an invariant mean.

Several avenues for further research may be suggested. Application of group theory to template matching is a possibility which should be explored. The necessary computational methods are by no means trivial, even in the simplest of cases (e.g., compact groups, one-parameter groups, finite groups). Also, we have totally ignored noise related questions. The impact of noise models on the use of invariants and relative invariants should be investigated rigorously. Experimental results reported to date, for example in [1], [5] and [18], indicate that noise perturbations may be small in comparison with the deterministic factors involved in groups of transformations. However, definitive results are very meager.

Another area for possible investigation involves the use of group theoretic methods in search of imbedded subpatterns. Since this evidently involves local features, it follows that invariance holds little promise as a tool. Indeed, by Theorem 4, invariant features are necessarily global in nature. Some hope remains, however, in the case in which features may be represented as analytic functions, since global information may be obtained by local measurement due to the existence of power series expansions.

Finally, since a large number of groups appearing in applications admit continuous parameters, the use of control theory in pattern matching is suggested. Problems which involve patterns in continuous motion can be modelled in such an environment and a group theoretic approach should be quite fruitful in such cases.

References

[1] F. Alt, Digital Recognition by Moments; in Optical Character Recognition, Washington, D.C., Spartan, 1962, pp. 152-179.

[2] E. Cassirer, The Concept of Group and the Theory of Perception, Philosophy and Phenomenological Research, Vol. V, 1944, pp. 1-35.

[3] P. M. Cohn, Lie Groups, Cambridge University Press, 1957.

[4] H. Dirilten, Ph.D. Dissertation, Texas Tech University, 1974.

[5] H. Dirilten, and T. G. Newman, Pattern Matching Under Affine Transformations, IEEE Trans. Comp., Vol. C-26, pp. 314-317, 1977.

[6] R. Duda, and P. Hart, Pattern Classification and Scene Analysis, New York, John Wiley and Sons, 1973.

[7] J. C. Dunn, Continuous Group Averaging and Pattern Classification Problems, SIAM J. Comput., Vol. 2, 1973, pp. 253-272.

[8] J. C. Dunn, Group Averaged Linear Transforms that Detect Corners and Edges, IEEE Trans. Comp., Vol. C-24, 1975, pp. 1191-1201.

[9] E. Hewitt, and K. Ross, Abstract Harmonic Analysis I, New York, Academic Press Inc., 1963.

[10] W. C. Hoffman, The Lie Algebra of Visual Perception, J. Math Psych., Vol. 3, 1966, pp. 65-98.

[11] W. C. Hoffman, The Neuron as a Lie Group Germ and a Lie Product, Quart. Appl. Math., Vol. 25, 1968, pp. 423-440.

[12] M. K. Hu, Visual Recognition by Moment Invariants, IRE Trans. Inform. Thy., Vol. IT-8, 1952, pp. 179-187.

[13] R. B. McGee, Automatic Recognition of Complex Three-Dimensional Objects from Optical Images, Report AFOSR-TR-0090 under contract AFOSR-71-2048, National Technical Information Service, Oct. 1973.

[14] L. Nachbin, The Haar Integral, Princeton, N.J., Van Nostrant Inc., 1965.

[15] G. Nagy, State of the Art in Pattern Recognition, Proc. IEEE, Vol. 26, 1968, pp. 836-862.

[16] W. Pitts and W. S. McCulloch, How we know Universals - The Perception of Auditory and Visual Forms, Bull. Math. Biophysics, Vol. 9, 1947, pp. 127-147.

[17] J. M. Richardson, Pattern Recognition and Group Theory, in Frontiers of Pattern Recognition, New York, Academic Press, 1972, pp. 453-477.

[18] A. D. Van der Lugt, Signal Detection by Complex Spatial Filtering, IEEE Trans. Info. Thy., Vol. IT-10, 1964.

[19] H. Weyl, The Classical Groups, Princeton University Press, Princeton, N.J., 1946.

[20] E. Wong, and J. A. Steppe, Invariant Recognition of Geometric Shapes, in Methodologies in Pattern Recognition, New York.

9. Abstract of "An Inverse Problem Related to Video Tracking", by T.G. Newman

Consider a time varying two-dimensional image in which objects are in motion along trajectories arising from horizontal and vertical translations, magnification and rotation. For such images a first order linear P.D.E. holds, provided the images are modeled as functions $F(t,x,y)$ of time t and the spatial variables x and y. In this equation the (unknown) parameters which determine the motion also appear linearly. Evaluation on a grid produces a system of linear equations which may be solved for the trajectory parameters.

In practice, the evaluation proceeds by numerical approximation of the required partial derivatives. In view of the ill-posed nature of numerical differentiation, inherent noise and sampling truncation present great difficulties.

Although no elegant solutions are at hand, examples are given to show the effect of somewhat naive methods of solution on real data.

10. Abstract of "Lie Groups and Lie Algebras in Video Tracking", by T.G. Newman

Motion of objects in time-varying images can sometimes be described by the action of a group of transformations on the image plane, regarded as a manifold. Moreover, the transformation groups occurring in applications can generally be described analytically in terms of a finite number of parameters; that is to say, they are Lie groups. In this situation we show that that data satisfies a linear partial differential equation in which the parameters of motion appear as linear coefficients. More or less standard numerical methods permit these parameters to be determined.

The parameters of motion determined as indicated above may be regarded as a velocity profile. This profile has the useful property of being spatially constant for each moving object in the image. In principle, at least, this permits detection and tracking of various objects having different trajectories.

Following development of the appropriate theory, the paper concludes by presenting the results of applying the technique to a number of real images in the form of digitized video.

11. Abstract of "Results in Differential Geometry with Application to Video Tracking", by G.A. Fredricks and T.G. Newman

We may begin many investigations in pattern recognition by assuming that a pattern is represented by a map on a smooth manifold and that the action of a transformation group on the set of patterns is represented by translation on the associated maps. Such an approach has recently been found to be of value in image processing as well. In this paper we present some theoretical results concerning the interplay between various vector fields which arise from the action of a Lie group on a smooth manifold. We further indicate how these results may be interpreted in the analysis of video data, permitting a new approach to target tracking.

12. Abstract of "Lie Theoretic Methods in Video Tracking", by T.G. Newman and D.A. Demus

Consider a 2-dimensional image in which objects are in motion through trajectories describable by translation (both horizontal and vertical), rotation, and magnification. The trajectory of such an object can be completely described by a 4-vector of parameters $\lambda(t) = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ which determine the velocities with respect to the four possible motions. If the data at time t and position x in the view plane is written as F(t,x), then we can show that

$$\frac{\partial F}{\partial t} = \sum_{i=1}^{4} \lambda_i(t) X_i F,$$

where X_1, X_2, X_3 and X_4 are certain (known) differential operators associated with the group of motions.

The derivatives appearing above may be evaluated numerically at various points in a given time slice to produce a system of linear equations which may be solved for the motion parameters. Evaluation at points within a moving rigid body leads to a vector of motion parameters unique to that particular body. In principle, at least, this technique permits application of tracking as well as segmentation of images based on relative motion of various objects.

The paper concludes by presenting the results of having implemented the above method on digitized video images.

Grants and Contracts Administered by JSEP Personnel

A. Funded

Chao, K.S., NSF Grant ENG-77-22991, "Continuation Methods in Computer-Aided Design", $37,421, 2 yrs.

Hunt, L.R., NSF Grant MCS 76-05267-A01, "Uniqueness and Hypoellipticity for Partial Differential Equations", $35,301, 2 yrs., (with M.J. Strauss).

Krile, T.F., SORF Grant, "Electro-Optical Interface", $7,650, 1 yr.

Newman, T.G., ONR Contract (administered as an add-on to contract 76-C-1136), "Pattern Recognition", $15,000, 1 yr.

Saeks, R., ONR Contract 76-C-1136, "Joint Services Electronics Program", $561,000, 3 1/2 yrs (with the JSEP staff).

Saeks, R., NSF Grant ENG-79-11315, "Frequency Domain-Like Methods for the Analysis and Design of Nonlinear and Time-Varying Systems", $79,240, 3 yrs.

Saeks, R., NSF Grant ENG-78-24414, "Large Scale Systems - The Commonality and Diversity", $20,512, 1 yr.

Saeks, R., ONR Contract 79-C-0170, "Fault Diagnosis", $15,000, 1 yr.

Walkup, J.F., AFOSR Grant 75-2855 and 79-0076, "Space-Variant Optical Systems", $375,464, 5 1/4 yrs.

Total Annual Funding $391,436

B. Proposed or in Preparation

Murray, J., Proposal to AFOSR, "The Application of Crossed Products to the Stability and Design of Time-Varying Systems", $22,091, 1 yr.

Murray, J., Proposal to NSF, "The Design of Two-Dimensional Recursive Digital Filters", (in preparation).

Newman, J., Proposal to ARO, "Lie Algebras and Lie Groups in Video Tracking", $20,000, 1 yr.

Walkup, J., Proposal to SPIE, "Undergraduate Optics Laboratory Equipment", $7,500, 1 yr.

Walkup, J., Proposal to NSF, "Detection and Estimation in Signal Dependent Noise", (in preparation).

Grants and Contracts in Electrical Engineering

A. Systems Research

Saeks, R., et al., Joint Services Electronics Program, ONR/AFOSR/ARO, $182,500, 1 yr.

Krile, T.F., "Electro-Optical Interface", State of Texas, $7,650, 1 yr.

Chao, K.S., "Continuation Methods in Computer-Aided Design", NSF, $37,421, 2 yrs.

Saeks, R., "Fault Diagnosis", ONR (NAVMAT), $15,000, 1 yr.

Saeks, R., "Frequency Domain-Like Methods for the Analysis and Design of Nonlinear and Time-Varying Systems", NSF, $79,240, 3 yrs.

Walkup, J.F., "Space-Variant Optical Systems", AFOSR, $85,000, 1 yr.

Total Annual Funding in Systems Research, $335,274.

B. Electro-Physics Research

Williams, P.F., "Studies of Wet Photo-Voltaic Solar Cells", State of Texas, $10,000, 1 yr.

Portnoy, W., "Gallium Arsenide Solar Cells", State of Texas, $10,000, 1 yr.

Kristiansen, M., "Dense Plasma Heating and Radiation Generations", AFOSR, $99,940, 1 yr.

Gundersen, M., "Laser Research", ERDA (Los Alamos), $69,695, 1 yr.

Kristiansen, M., and M. Hagler, "Theoretical and Experimental Investigations of RF Plasma Heating", NSF, $39,044, 1 yr.

Portnoy, W., "High Temperature Electronic Devices", NRL, $20,436, 1 yr.

Gundersen, M., A Spectroscopic Study of Impurities and Defects in Semiconductors", TI, $10,000, 1 yr.

Gundersen, M., "An Innovative Infrared Detector" NSF $60,000, 2 yrs.

Total Annual Funding in Electro-Physics Research, $289,115

C. Pulsed Power Research

Burkes, T., "High Power Spark Gap Development", SCEEE(AFWL), $63,000, 1 yr.

Burkes, T., "A Critical Assessment and Evaluation of High Power Switches", NSWC, $9,950, 1 yr.

Kunhardt, E., "Breakdown at Overvoltages", NSWC, $54,003, 1 yr.

Burkes, T., "Veredyne Switch Test Program", ERDA (Los Alamos), $8,221, 1 yr.

Kristiansen, M., et al., "Coordinated Research Program in Pulsed Power Physics", AFOSR, $596,128, 1 1/4 yrs.

Kristiansen, M., "Surface Flashover Mechanisms", ERDA (Sandia), $74,000, 1 yr. (Jointly with the University of South Carolina).

Kristiansen, M., "Continuous Discharge Maintained by CO_2 Laser", NSF, $100,000, 3 yrs. (Jointly with the Polish Academy of Sciences).

Total Annual Funding in Pulsed Power Research, $719,409.

D. Power Systems Research

Reichert, J.D., "Crosbyton Solar Power Project", ERDA, $2,866,361, 2 yrs. (Approximately 50% of this contract is spent by Texas Tech with remainder of the funds being subcontracted).

Craig, J., "Power System Studies", TP&L, $8,000, 1 yr.

Total Annual Funding in Power Systems Research $708,000*

E. Funding for Educational Activities

Williams, P.F., "Innovative Undergraduate Laboratory Program in Optical Communications", State of Texas, $8,700, 2 yrs.

Portnoy, W., "Undergraduate Semiconductor Device Fabrication Projects", State of Texas, $6,283, 2 yrs.

Kunhardt, E., "Undergraduate Research Participation", NSF, $19,931, 1 yr.

Portnoy, W., "Undergraduate Semiconductor Device Fabrication Projects", NSF, $13,800, 2 yrs.

Williams, P.F., "Innovative Undergraduate Laboratory Program in Optical Communications", NSF, $8,700, 2 yrs.

Kral, L., "Graduate Fellowship Support", NSF, $1,700, 1 yr.

Kristiansen, M., "Pulsed Power Research Colloquim", AFOSR, $4,841, 1 yr.

Total Annual Funding for Educational Activities $75,213.

* Includes $700,000 spent in-house annually by the Crosbyton Solar Power Project.

F. Support for Conferences, Workships, and Symposia

Kristiansen, M., "Second IEEE International Pulsed Power Conference", AFOSR, $9,968, 1 yr.

Saeks, R., "Workshop on Large-Scale Systems: The Commonality and Diversity", NSF, $20,512, 1 yr.

Portnoy, W., "University/Industry/Government Microelectronics Symposium", NSF, $5,623, 1 yr.

Total Annual Support for Conferences, Workshops and Symposia $36,103.

G. Other

Seacat, R.H., "Research and Development", State of Texas, $18,325, 1 yr.

Total Annual Funding for Other Purposes $18,325.

H. Sources of Funding in Electrical Engineering

Air Force	$ 769,651
Navy*	281,889
Army	-0-
ERDA	851,816
NSF	206,517
Industry	18,000
State of Texas	53,466
Total Annual Grants and Contracts in Elect. Engineering	$2,181,439

* Includes all of JSEP.

Publications by JSEP Personnel*

A. Refereed Journal Articles

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, Vol. CAS-26, pp 589-594, (1979, JSEP).

2. DeCarlo, R.A., and R. Saeks, "A Root Locus Technique for Interconnected Systems", IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-9, pp. 53-55, (1979, JSEP).

3. DeSantis, R.M., Saeks, R., and L.J. Tung, "Basic Optical Estimation and Control Problems in Hilbert Space", Math, System Theory, Vol. 12, pp. 175-203, (1978, AFOSR-74-2631).

4. Desoer, C.A., Liu, R.-W., Murray, J., and R. Saeks, "Feedback System Design: The Fractional Representation Approach to Analysis and Synthesis", IEEE Trans. on Auto. Cont., (to appear, JSEP).

5. Froehlich, G.K., Walkup, J.F., and R.B. Asher, "Optimal Estimation in Signal-Dependent Noise", Jour. of the Optical Soc. of Am., Vol. 68, pp. 1665-1671, (1978, JSEP).

6. Hunt, L.R., "Controllability of General Nonlinear Systems", Math. Sys. Theory, Vol. 12, pp. 361-370, (1979, JSEP).

7. Hunt, L.R., and J.J. Murray, "q-Plurisubharmonic Functions and a Generalized Dichlet Problem", Michigan Math, Jour., Vol. 25, pp. 299-315, (1978, NSF-MCS-76-05267).

8. Hunt, L.R., M. Kazlow, "A Two-sided H. Lewy Extension Phenomenon", Proc. of the AMS, Vol. 74, pp. 95-99, (1979, NSF-MCS-76-05267).

9. Lu, K.S., and R. Saeks, "Failure Prediction for an On-Line Maintenance System in a Poisson Shock Environment", IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-9, pp. 356-362, (1979, JSEP).

10. Marks, R.J., Walkup, JF., and M.O. Hagler, "Methods of Linear System Characterization Through Response Cataloging", Applied Optics, Vol. 18, pp. 655-659, (1979, AFOSR-79-0076).

11. Marks, R.J., Jones, M.I., Kral, L., and J.F. Walkup, "One-Dimensional Linear Coherent Processing Using a Single Optical Element", Applied Optics, Vol. 18, pp. 2783-2786, (1979, AFOSR-79-0076).

12. Murray, J., "Spectral Factorization and Quarter-Plane Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-25, pp. 586-592, (1978, JSEP and AFOSR-74-2631).

* Includes all publications by JSEP personnel with source of support.

13. Murray, J.J., Hunt, L.R., and M.J. Strauss, "Liouville's Theorem for First-Order Partial Differential Equations", Colloquim Math., (to appear, NSF-MCS-76-05267).

14. Pan, C.T., and K.S.Chao, "Multiple Solutions of Nonlinear Equations: Roots of Polynomials", IEEE Trans. on Circuits and Systems (to appear, NSF-ENG-77-22991).

15. Pan, C.T., and K.S. Chao, "A Computer-Aided Root-Locus Mehtod", IEEE Trans. on Automatic Control, Vol. AC-23, pp 856-860. (1978, JSEP and NSF-ENG-77-22991).

16. Saeks, R., "On the Decentralized Control of Interconnected Dynamical Systems", IEEE Trans. on Auto. Control, Vol. AC-24, pp. 269-271, (1979, JSEP).

17. Saeks, R., "An Approach to Built-in Testing", IEEE Trans. on Aerospace and Electronic Systems, Vol. AES-14, pp. 813-818, (1979, JSEP).

18. Saeks, R., "A Continuation Algorithm for Sparse Matrix Inversion", IEEE Proc., Vol. 67, pp. 682-683, (1979, JSEP).

19. Saeks, R., "Forward to the Special Issue on the Mathematical Foundations of System Theory", IEEE Trans. on Circuits and Systems, Vol. CAS-25, p. 649, (1978, unsupported).

20. Saeks, R., "Review of 'Monotone Operators and Applications in Control and Network Theory' by V. Dolezal", Bull. of AMS, (to appear, unsupported).

21. Sen, N., and R. Saeks, "Fault Diagnosis for Linear Systems Via Multi-frequency Measurements", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 457-465, (1979, JSEP).

22. Tung, L.J., and R. Saeks, "Reproducing Kernel Resolution Space and its Applications II", Jour. of the Franklin Inst., Vol. 306, pp. 425-447, (1978, AFOSR-74-2631).

23. Tung, L.J., Saeks, R., and R.M. DeSantis, "Wiener-Hopf Filtering in Hilbert Resolution Space", IEEE Trans. on Circuits and Systems, Vol. CAS-25, pp. 702-705, (1978, AFOSR-74-2631).

24. Walkup, J.F., Krile, T.F., Hagler, M.O., and W.D. Redus, "Multiplex Holography with Chirp-Modulated Binary Phase-Coded Reference Beam Masks", Applied Optics, Vol. 18, pp. 52-57, (1979, AFOSR-79-0076).

25. Walkup, J.F., Hagler, M.O., Marks, R.J., and L. Kral, "Scanning Technique for Coherent Processors", Applied Optics, (to appear, AFOSR-79-0076).

26. Walkup, J.F., Novel Techniques for Optical Information Processing: An Introduction", Applied Optics, Vol. 18, pp. 2735-2736, (1979, AFOSR-79-0076).

27. Walkup, J.F., and M.O. Hagler, "Optical Information Processing", IEEE Proceedings, (to appear, AFOSR-79-0076).

28. Walkup, J.F., "Space-Variant Optical Processing", Optical Engineering, (to appear, AFOSR-79-0076).

B. Conference Papers and Abstracts

1. Asher, R.B., and J.F. Walkup, "Introduction to Detection and Estimation Concepts", 1778 Annual Meeting of the Optical Soc. of Amer. San Francisco, Oct. 1978, (abstract only, unsupported).

2. Froehlich, G.K., Walkup, J.F., and R.B. Asher, "Optimal Estimation in Signal-Dependent Film-Grain Noise", Proc. of the 11th Inter. Commission for Optics Conf., Madrid, Sept. 1978, pp. 367-369, (JSEP).

3. Froehlich, G.K., Walkup, J.F., and R.B. Asher, "Estimation in Signal-Dependent Noise", 1978 Annual Meeting of the Optical Soc. of Am., San Francisco, Nob. 1978, (abstract in the Jour. of the OSA, Vol. 68, p. 1385A, JSEP).

4. Hunt, L.R., "Control Theory for Nonlinear Systems", Proc. of the 12th Asilomar Conf. on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1979, pp. 339-343, (JSEP).

5. Hunt, L.R., "Controllability of Nonlinear Systems", Proc. of the 1979 Inter. Symp. on the Mathematics of Networks and Systems, T.H., Delft, July 1979, pp. 466-467, (JSEP).

6. Jones, M.I., Walkup, J.F., and M.O. Hagler, "Multiplex Holography for Space-Variant Optical Computing, Proc. SPIE, Washington, April 1979, pp. 16-21, (AFOSR-79-0076).

7. Karmokolias, C., and R. Saeks, "Optimal Selection of Weighting Matrices in Kalman Regulators", Proc. of the 21st Midwest Symp. on Circuits and Systems, Iowa State Univ., Ames, Ia., Aug. 1978, pp. 71-72, (JSEP).

8. Kral, L., Hagler, M.O., Marks, R.J., and J.F. Walkup, "A Input Scanning Technique for Coherent Processing", 1978 Annual Meeting of the Optical Society of Amer., San Francisco, Oct. 1978, (abstract only, AFOSR-79-0076).

9. Marks, R.J., Walkup, J.F., Hagler, M.O., and L. Kral, "Linear Coherent Processing using an Input Scanning Technique", Proc. of the IEEE Inter. Optical Computing Conference, London, Sept. 1978, (JSEP)

10. Murray, J., "Semidirect Products and the Stability of Time-Varying Systems", Proc. of the Inter. Symp. on the Mathematics of Networks and Systems, Vol. 3, T.H., Delft, July 1979, pp. 121-125, (JSEP).

11. Newman, T.G., and R.M. Anderson, "Treelike Programs from Two Structure Rules", Proc. of the 6th Annual Computer Science Conference, North Texas State Univ., April 1979, (unsupported).

12. Newman, T.G., "A Group Theoretic Approach to Invariance and Pattern Recognition", Proc. of the IEEE Conf. on Pattern Recognition and Image Processing", Chicago, Aug. 1979, pp. 407-412, (JSEP).

13. Olivier, P.D., and R. Saeks, "Nonlinear Observers and Fault Analysis", Proc. of the 22nd Midwest Symposium on Circuits and Systems, Univ. of Penn., Philadelphia, June 1979, pp. 535-536, (JSEP).

14. Olivier, P.D., and R. Saeks, "On Large Nonlinear Perturbations of Linear Systems", Proc. of the 12th Asilomar Conf. on Circuits, Systems, and Computers, Pacific Grove, Ca., Nov. 1978, pp. 473-477, (JSEP).

15. Pan, C.T., and K.S. Chao, "Multiple Solutions of a Class of Nonlinear Equations", Proc. of the 1979 IEEE Inter. Symp. on Circuits and Systems, Tokyo, July 1979, pp. 577-580, (JSEP, and NSF-ENG-77-22991).

16. Pan, C.T., and K.S. Chao, "A Continuation Method for Finding the Roots of a Polynomial", Proc. of the 22nd Midwest Symp. on Circuits and Systems, Univ. of Pennsylvania, Philadelphia, June 1979, pp. 428-431, (JSEP and NSF-ENG-77-22991).

17. Saeks, R., and J.J. Murray, "Stability and Homotopy II", Proc. of the 1979 JACC, Denver, June 1979, p. 358, (abstract only, NSF-ENG-79-11315).

18. Saeks, R., "Hilbert Resolution Space: Engineering Concepts", Proc. of the 1979 Inter. Symp. on the Mathematical Theory of Networks and Systems", T.H., Delft, July 1979, p. 213, (abstract only, NSF-ENG-11315).

19. Saeks, R., "CAD Oriented Measures of Testability", Proc. of the Industry/ Joint Services Automatic Test Conference and Workshop, NSIA, San Diego, April 1978, pp. 71-72, (JSEP).

20. Saeks, R., "An Application of Large-Scale Systems Techniques to the Fault Analysis Problem", Proc. of the 21st Midwest Symp. on Circuits and Systems, Iowa State Univ., Ames, Ia., Aug. 1978, p. 314, (abstract only, JSEP).

21. Walkup, J.F., and M.O. Hagler, "Integration of Optics Experiments into an EE Curriculum", Proc. of the 87th Annual Conf. of the ASEE, Baton Rouge, June 1979, (unsupported).

22. Walkup, J.F., Kral, L., and M.O. Hagler, "Correlation Properties of Diffusers for Multiplex Holography", 1979 Annual Meeting of the Optical Soc. of Amer., Rochester, Oct. 1979, JSEP.

C. Preprints

1. Feintuch, A., Saeks, R., and C. Neil, "A New Performance Measure for Stochastic Optimization in Hilbert Space", (submitted for publication, NSF-ENG-79-11315).

2. Fredricks, G., and T.G. Newman, "Results in Differential Geometry with Applications to Video Tracking", (submitted for publication, JSEP).

3. Green, B., "Continuation Algorithms for the Solution of the Eigenvalue Problem", (preliminary draft, JSEP).

4. Hunt, L.R., "Global Controllability in Two Dimensions", (submitted for publication, JSEP).

5. Hunt, L.R., "Controllability of Nonlinear Hypersurface Systems", (submitted for publication, JSEP).

6. Hunt, L.R., "Controllability and Stability", (submitted for publication, JSEP).

7. Hunt, L.R., and M. Kazlow, "A Two-Regular H. Lewy Extension Phenomenon", (submitted for publication, NSF-MCS-76-05267).

8. Karmokolias, C., and R. Saeks, "Suboptimal Control with Optimal Quadratic Regulators", (submitted for publication, JSEP).

9. Karmokolias, C., and R. Saeks, "Suboptimal Design of an Aircraft Landing System", (submitted for publication, JSEP).

10. Newman, T.G., "An Inverse Problem Related to Video Tracking", (submitted for publication, JSEP).

11. Newman, T.G., "Lie Groups and Lie Algebra's in Video Tracking", (submitted for publication, JSEP).

12. Newman, T.G., and D.A. Demus, "Lie Theoretic Methods in Video Tracking", (submitted for publication, JSEP).