COEFFICIENT QUANTIZATION EFFECTS ON POLE LOCATIONS FOR STATE-MODEL DIGITAL FILTERS

By

JAMES DALE LEDBETTER

Bachelor of Science
Oklahoma State University
Stillwater, Oklahoma
1968

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1969

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
May, 1979

COEFFICIENT QUANTIZATION EFFECTS ON POLE LOCATIONS FOR STATE-MODEL DIGITAL FILTERS

Thesis Approved:

Rao Yarlagadda
Thesis Adviser

Donald L. Grace

C. M. Bacon

Craig S. Sims

Norman N. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Dr. Rao Yarlagadda, my thesis adviser and chairman of my doctoral committee, for the continued support and encouragement he has given me throughout my program, which has contributed significantly to the completion of this dissertation.

The interest and comments offered by Dr. Charles Bacon, Dr. Craig Sims, and Dr. Donald Grace, members of my doctoral committee, are also greatly appreciated.

I would like to thank the United States Air Force for giving me the opportunity to finish my program through its AFIT Civilian Institution program. The support and consideration given me by my supervisors at Frank J. Seiler Research Laboratory is also greatly appreciated.

I would like to express my special gratitude to my wife, Colene, for her understanding, encouragement, and sacrifices. Finally, I would like to acknowledge the continued source of motivation offered by my son, James, for whom this effort was begun.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$H(z)$	Digital filter transfer function
$f(z)$	General pole polynomial
K	New system matrix
A_s	Cross-coupled system matrix
A_c	Companion matrix
A_o	One's canonical system matrix
k_1	Variable element in K
k_2	Variable element in K
α	Variable element in A_s
β	Variable element in A_s
d_1	Variable element in A_c
d_2	Variable element in A_c
x	Variable element in A_o
y	Variable element in A_o
ρ	Magnitude of complex pole
Θ	Angle of complex pole
λ	System matrix eigenvalue, $\lambda=\rho e^{\pm j\Theta}$
$\lvert S(\lambda_k,F)\rvert$	Minimum worst case eigenvalue sensitivity magnitude of matrix F
$\lvert S_1(\lambda_k,F)\rvert$	Sensitivity magnitude (not necessarily worst case)
A'	Sensitivity matrix of A_c

$\hat{A}$	General sensitivity matrix designation for matrices other than A_c
$\Delta\rho$	Radial change in pole location
$\Delta\Theta$	Angular change in pole location
α	General variational vector
M	Matrix relating changes in pole polynomial coefficients to changes in system matrix elements
D	Diagonal matrix of coefficient values
S_ρ	Radial sensitivity vector
S_Θ	Angular sensitivity vector
N_a	Wordlength of matrix element a
σ_0^2	Roundoff noise variance
E_0	Quantization increment

CHAPTER I

INTRODUCTION

1.1 Statement of the Problem

The design of digital filters involves three basic steps: (1) the determination of the filter specifications; (2) the approximation of these specifications using a discrete-time system; and (3) the realization of the filter. Although these three steps are not completely independent, this thesis is focused primarily on the third step. The realization of the system as a computer program or in hardware requires that a digital network or structure be chosen. There are many considerations and tradeoffs involved in choosing a structure among which are hardware requirements and/or specifications. A distinction in terminology is being made between requirements and specifications. If the filter is being implemented on a general-purpose computer, the designer will have to work with the existing specification of that computer among which will be included the memory wordlength. If, on the other hand, the filter is being implemented using a special purpose hardware, the designer may or may not have more freedom in establishing hardware requirements, thereby setting hardware specifications, necessary to meet the filter specifications. These hardware requirements will include the accuracy requirements on the A/D and the length of registers in the system. In most hardware realizations, of course, it is economically

desirable to minimize the length of the registers that must be provided to store the filter parameters.

Regardless of the method of implementation, these hardware requirements and/or specifications have an impact upon the accuracy with which the input and the system parameters can be realized. This impact manifests itself in the form of input, filter coefficient, and multiplication quantization errors. The effects of these three sources of error in digital filters has been investigated extensively in the literature [1] [2]. In this thesis, only recursive filters are considered.

For a given system transfer function

$$H(z) = \frac{\sum_{i=o}^{N} a_i z^{-i}}{1 + \sum_{i=1}^{N} b_i z^{-i}} \qquad (1\text{-}1)$$

there is an infinite variety of network realizations that realize the system function when network parameters are realized with infinite precision. It is to be expected that some of these structures will be less sensitive than others to quantization of the parameters; i.e., the system function of the realization will be closer in some sense to the desired system function. State-space techniques provide a convenient method for generating various input/output equivalent structures. Recursive digital filters can be described by the state equations, which are amenable to the incorporation of the effects of possible structure transformation and state amplitude scaling so that an analytical study of the interaction of the filter structure and the quantization errors is made possible. This thesis investigates digital filter coefficient quantization

effects on digital filters that are described by and realized through state equations. The effects are analyzed through the changes in filter pole locations due to the coefficient quantization.

Given a digital transfer function in the form of (1-1), it is well known that there always exists a state model of the form

$$x(n+1) = Ax(n) + Bu(n) \tag{1-2}$$

$$y(n) = Cx(n) + Du(n) \tag{1-3}$$

such that $H(z) = C(zI-A)^{-1}B + D$ and where $x(n)$ is an N-dimensional vector describing the state of the system at time $t = nT$, $u(n)$ is a scalar input, $y(n)$ is the scalar output, and A, B, C, and D are, respectively, NxN, Nx1, 1xN, and 1x1 real, constant matrices. There are an infinite number of state models, all of which will yield the same input-output relationship between $u(n)$ and $y(n)$. Define

$$\bar{x}(n) = T^{-1}x(n) \tag{1-4}$$

where T is a nonsingular matrix of order N, and x is an Nx1 vector. Then

$$\bar{x}(n+1) = \bar{A}\bar{x}(n) + \bar{B}u(n) \tag{1-5}$$

$$y(n) = \bar{C}\bar{x}(n) + Du(n) \tag{1-6}$$

where

$$\bar{A} = T^{-1}AT \tag{1-7}$$

$$\bar{B} = T^{-1}B \tag{1-8}$$

$$\bar{C} = CT \tag{1-9}$$

The state model given by (1-5) and (1-6) realizes (1-1) like that given by (1-2) and (1-3) but may differ greatly in the effects that coefficient

quantization may have on the pole locations of the filter. The pole locations of these filters are determined by the elements of the system matrix A and $\bar{A}$. When realized exactly, the poles of A and $\bar{A}$ are the same. Under the effects of quantization, however, the poles of A and $\bar{A}$ will differ. This thesis compares the properties of various system matrices where elements are subject to variation. A new system matrix suitable for use in digital filter applications is introduced.

The second order filter has been recognized as a basic building block for higher order filters due to its noise characteristics and its suitability for multiplexing [3], therefore only the second order case is considered in this thesis. First- and second-order filters are normally combined in parallel or cascade forms to implement higher order filters [4] [5].

While an analysis of various equivalent state-model formulations for digital filters based on changes in pole location due to coefficient quantization provides a useful basis for the comparison of equivalent system matrices, a criterion more useful for determining hardware requirements for the implementation of a filter might be the number of bits required for each coefficient in order to ensure acceptable performance. Closely associated with the required coefficient wordlength is the location and density of the discrete pole grids which can be realized with a given number of bits. Avenhaus [6] used the density of allowable pole locations in the z-plane as a measure for assessing various filter structures. In this thesis, the second-order system matrices under consideration are compared for wordlength requirements for a given variation in system matrix elements and for realizable pole grids within the unit circle of the z-plane.

For final comparisons, the matrices will be compared regarding their ability to sustain overflow limit cycles and their roundoff noise properties will be discussed. Mills et al. [7] has developed sufficient conditions for the absence of overflow oscillations in second order filters using two's complement arithmetic. Jackson [34] has shown that the roundoff noise of a filter depends on the form of the realization.

Throughout this thesis, only fixed-point arithmetic will be considered. Digital filters are usually realized through implementation on a minicomputer or by the construction of special purpose hardware using fixed point arithmetic. This allows for simplicity in the design [3] and, correspondingly, reduces cost.

1.2 Technical Approach

The approach used in this thesis to evaluate the effects of coefficient quantization upon the pole locations of state-model digital filters is to analyze the eigenvalue sensitivity of the system matrix due to variations in the system matrix elements. Since the eigenvalues of the system matrix of the state-model are the pole locations of the digital filter, the choice of the state-variables for realizing the filter is important [8] [9].

A new system matrix suitable for use in digital filters is presented. The parameter space of the matrix elements is mapped into the unit circle of the z-plane, showing that the second order matrix can realize all real or complex conjugate pole locations within the unit circle. This matrix is compared to other second order matrices that have been presented in the literature [8] [9] by using sensitivity analysis techniques. Expressions for the magnitude, and the corresponding radial and angular

components, of the change in complex conjugate pole locations for each system matrix are developed for both absolute and normalized simultaneous variations in the system matrix elements. Minimum pole sensitivity regions within the unit circle of the z-plane are shown for each system matrix.

Using a sensitivity definition introduced by Singer [9], a sensitivity matrix is derived for each system matrix. Each system matrix can be obtained from the others by the transformation given in (1-7). It is shown that a similar relation exists for the sensitivity matrices.

By using the sensitivity expressions for radial and angular movement, the allowable variation in each element of the matrix is determined for given pole movements in the radial and angular directions. This allowable element variation is then used to determine the coefficient wordlength necessary to constrain the pole movements within the allowable limits.

Since the system matrix elements of a digital filter are implemented with fixed wordlength binary registers, the coefficients of the resultant characteristic polynomial of the system matrix can assume only discrete values and therefore only a fixed set of discrete pole locations within the unit circle can be realized. For a given wordlength, the realizable pole grid of each system matrix is determined by the set of characteristic polynomials consisting of valid combinations of discrete coefficient values. A valid polynomial is one that has roots inside the unit circle.

Mills et al. [7] developed sufficient conditions for the absence of overflow oscillations in second order filters using two's complement arithmetic. This sufficient criteria will be applied to the system

matrices under consideration. The roundoff noise properties of the matrices will be obtained using a method presented by Gold and Rader [25].

1.3 Review of the Literature

The concept of sensitivity in its most basic form is almost as old as the concept of feedback. One of the basic reasons for introducing feedback was to reduce the effect of parameter changes upon system performance. Therefore, it is quite natural that the basic concepts of sensitivity appeared in the fundamental work of Bode [10] which constituted the beginning of the modern theory of control systems. However, the series of ideas and methods which were developed for solving problems connected with parameter variations were contributed by various disciplines. This resulted in a generality which makes it possible to treat these ideas and methods as a fundamental theory--the theory of sensitivity analysis.

Bode [10] introduced the idea of single element sensitivity. The present day definition of this element sensitivity is given by

$$S_{\alpha}^{T(j\omega)} = \frac{d(\ln T(j\omega))}{d(\ln \alpha)}$$

where $T(j\omega)$ is the continuous system transfer function and α is a system parameter.

It has been shown that a continuous or discrete system defined by a transfer function can be expressed in state model format with a companion matrix as the system matrix [11] [12]. Wilkie and Perkins [13] considered the problem of generating the sensitivity functions of all states of a companion matrix state model with respect to any number of parameters for a continuous linear, time-invariant, single-input, controllable system.

As a consequence of an increasing application of the pole-zero approach to the problems of networks and control systems synthesis, the idea of pole-zero sensitivity was developed. Kokotovic and Rutman [14] present sensitivity coefficients for the movements of poles and zeroes as a function of small relative and absolute variations in system parameters.

A fundamental problem within the area of pole-zero sensitivity is the effect of polynomial coefficient variation upon the roots of the polynomial. Maley [15] considered this problem for single parameter variation. Reddy [16] [17] and Morgan [18] considered the more general problem of the eigenvalue sensitivity of a multivariable system expressed in state-model format whose system matrix elements are functions of the system parameters.

Procedures are presented in Huelsman [19], Daryanani [20] and Mitra [21] for the determination of transfer function sensitivity, pole-zero sensitivity, and characteristic polynomial coefficient sensitivity. These presentations are in regard to the analysis and synthesis of continuous, linear, active networks.

Horowitz [28] considered the sensitivity analysis of sampled-data systems by using transformations which made it possible to study the properties of the sensitivity function in terms of continuous system frequency concepts. He found that it is impossible to secure unlimited sensitivity reduction and that a compromise between the values of the sampling period, the system response, and the system sensitivity is necessary.

The use of state-space techniques in systems analysis is very attractive as they lend themselves very well to computer simulation and

sensitivity analysis. Kerlin [23] used state-variable formulation to develop expressions for transfer function and pole sensitivities for the analysis of large systems on digital or analog computers.

Associated with state-space techniques is the selection of the state variables. Mantey [8] has investigated the relationship between eigenvalue sensitivity and state variable selection and found that the eigenvalue sensitivity depends strongly on the choice of the state variables. His search was limited to matrices requiring a minimum number of arithmetic operations. He mentions that no orderly procedure has been devised for the selection of a system matrix to insure minimum eigenvalue sensitivity. Singer [9] studied in further detail the problem of minimizing eigenvalue sensitivity through the selection of state variables. He considered only second order matrices, the number of arithmetic operations not being minimal.

An interesting area for the application of sensitivity analysis is the field of problems concerning roundoff noise in digital filters. It has been shown [4] [24] [25] that equivalent input/output digital filter realizations exhibit different output noise characteristics. Fettweis [26] has shown that there exists a connection between the generation of roundoff noise by a multiplier and the effect that the coefficient word-length limitation of this multiplier has upon the response characteristics of a filter. Bonzanigo [27] has shown that low coefficient sensitivities do not guarantee low roundoff noise output since other factors, such as pairing and ordering of sections in the cascade form, affect the noise output but not the coefficient sensitivities of the filter. Jackson [28] has utilized the sensitivities of a digital filter transfer

function with respect to its coefficients to derive lower bounds on the roundoff noise output.

The effect of coefficient quantization upon the response of a digital filter can be analyzed by calculating the movements of the poles and zeroes of the transfer function. Mitra and Sherwood [29] have presented a technique for estimating pole-zero displacements and for determining coefficient wordlength which insure that the pole-zero movements will stay within prescribed bounds. Gold and Rader [30] and Avenhaus [6] have proposed structures with less pole sensitivity to parameter quantization.

1.4 Quantization Errors in Digital Filters

In this section background information is presented concerning the three sources of errors in digital filters. These sources are: input quantization, product quantization, and coefficient quantization. The procedure used follows that employed by Hwang [31].

Given a digital filter expressed in state-space format as shown in (1-2) and (1-3), the effects of input, product, and coefficient quantizations result in the actual filter implemented by a finite wordlength machine being given by

$$\tilde{x}(n+1) = [(A+\Delta A)\tilde{x}(n)]_r + [(B+\Delta B)\tilde{u}(n)]_r \qquad (1\text{-}11)$$

$$\tilde{y}(n) = [(C+\Delta C)\tilde{x}(n)]_r + [(D+\Delta D)\tilde{u}(n)]_r \qquad (1\text{-}12)$$

or

$$\tilde{x}(n+1) = \tilde{A}\tilde{x}(n) + \tilde{B}\tilde{u}(n) + \alpha(n) + \beta(n) \qquad (1\text{-}13)$$

$$\tilde{y}(n) = \tilde{C}\tilde{x}(n) + \tilde{D}\tilde{u}(n) + \gamma(n) + \delta(n) \qquad (1\text{-}14)$$

where $[\ ]_r$ indicates rounding; $\tilde{u}(n)$, $\tilde{x}(n)$, and $\tilde{y}(n)$ are, respectively,

the actual input, states, and output; $\tilde{A} = A+\Delta A$, $\tilde{B} = B+\Delta B$, $\tilde{C} = C+\Delta C$, $\tilde{D} = D+\Delta D$; and $\alpha(n)$, $\beta(n)$, $\gamma(n)$, and $\delta(n)$ are respectively, N-, N-, 1-, and 1- dimensional error vectors generated due to product quantizations in the $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$ matrices.

Subtracting (1-2) from (1-13), and (1-3) from (1-14)

$$\Delta x(n+1) = \tilde{A}\Delta x(n) + \tilde{B}\Delta u(n) + \Delta Ax(n) + \Delta Bu(n) + \alpha(n) + \beta(n) \quad (1\text{-}15)$$

$$\Delta y(n) = \tilde{C}\Delta x(n) + \tilde{D}\Delta u(n) + \Delta Cx(n) + \Delta Du(n) + \gamma(n) + \delta(n) \quad (1\text{-}16)$$

where $\Delta x(n)$ is the state-error vector, and $\Delta y(n)$ is the output error or noise.

Using the standard method for solving linear, time-invariant vector matrix difference equations [32], the solutions to (1-15) and (1-16) are

$$\Delta x(n) = \tilde{A}^n \Delta x(o) + \sum_{j=o}^{n-1} \tilde{A}^{(n-j-1)}[\tilde{B}\Delta u(j) + \Delta Ax(j) + \Delta Bu(j) + \alpha(j) + \beta(j)] \quad (1\text{-}17)$$

and, assuming that $\Delta x(o) = o$,

$$\Delta y(n) = \Delta y_1(n) + \Delta y_2(n) + \Delta y_3(n) \quad (1\text{-}18)$$

where

$$\Delta y_1(n) = \tilde{C} \sum_{j=o}^{n-1} \tilde{A}^{(n-j-1)}\tilde{B}\Delta u(j) + \tilde{D}\Delta u(n) \quad (1\text{-}19)$$

$$\Delta y_2(n) = \tilde{C} \sum_{j=o}^{n-1} \tilde{A}^{(n-j-1)}[\alpha(j) + \beta(j)] + \gamma(n) + \delta(n) \quad (1\text{-}20)$$

$$\Delta y_3(n) = \tilde{C} \sum_{j=o}^{n-1} \tilde{A}^{(n-j-1)}[\Delta Ax(j) + \Delta Bu(j)] + \Delta Cx(n) + \Delta Du(n) \quad (1\text{-}21)$$

are the errors due to the input, product, and coefficient quantizations, respectively.

One important point of emphasis is necessary at this point. Although this development allows the three types of errors to be expressed distinctly in (1-19), (1-20) and (1-21), it should be kept in mind that the effect of coefficient quantizations, shown in (1-21), couples into those of input quantization in (1-19) and product quantization in (1-20) since $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$ are defined by $\tilde{A} = A + \Delta A$, etc.

From (1-19) - (1-21) it is seen that $\Delta y_3(n)$ is directly proportional to the magnitude of the input, while $\Delta y_1(n)$ and $\Delta y_2(n)$ are independent of it. This means that the signal-to-noise ratio for the error due to coefficient quantizations is fixed for a given network, while those for the input and product quantizations can be improved by increasing the input level.

The state-model given by (1-5) and (1-6) is equivalent to that given by (1-2) and (1-3) and is related by the transformation given in (1-4). The effect of such a transformation on the input quantization error given in (1-19) is now analyzed by substituting (1-7), (1-8), and (1-9) into (1-19). This results in

$$\Delta y_1(n) = \tilde{C} \sum_{j=o}^{n-1} \tilde{A}^{(n-j-1)} \tilde{B} \Delta u(j) + \tilde{D} \Delta u(n) \tag{1-22}$$

$$\Delta y_1(n) = \tilde{C}\, T \sum_{j-o}^{n-1} (T^{-1} \tilde{A}\, T)^{(n-j-1)} T^{-1}\, \tilde{B} \Delta u(j) + \tilde{D} \Delta u(n) \tag{1-23}$$

Now $$(T^{-1} \tilde{A}\, T)^{(n-j-1)} = T^{-1} \tilde{A}^{(n-j-1)} T \tag{1-24}$$

So $$\Delta y_1(n) = \tilde{C} \sum_{j=o}^{n=1} \tilde{A}^{(n-j-1)}\, \tilde{B} \Delta u(j) + \tilde{D} \Delta u(n) \tag{1-25}$$

Therefore, up to first order effects of the quantization level, $\Delta y_1(n)$ is invariant under structure transformation and/or amplitude scalings. A similar examination of $\Delta y_2(n)$ and $\Delta y_3(n)$ shows that they are highly dependent upon transformation effects. Thus, searching for a better network realization entails minimizing the effects of coefficient and product quantizations.

The error due to input quantization, as shown in (1-19) is often referred to as "quantization noise". It is inherent in any A/D conversion process and has been studied in great depth [33] [34].

The error due to product quantizations, as shown in (1-20), is similar to the input quantization error in that it also involves quantization of the data. However, this form of error is different in that the data is already in digital form and the quantization, in the form of either rounding or truncation, takes place within the filter, not just at the input. Generally, this type of error is referred to generically as "roundoff noise", and it is an important design consideration in digital filters which has received extensive research [2] [31] [34].

The last source of error to be considered is that of coefficient quantization as shown in (1-21). The effect of coefficient quantization on the performance of a filter has been of much concern and a number of different approaches to this problem have been suggested [1] [4] [30].

In general, the effect of coefficient quantization is highly dependent on the structure used to implement the system. Oppenheim and Schafer [35] emphasize that the present understanding of the relationship between network structure and coefficient sensitivity is very meager. Although Jackson [34] considers the error due to coefficient quantization as straightforward to analyze, he comments that the

inclusion of coefficient quantization in the initial filter synthesis procedure in order to minimize the resulting filter complexity is a complex problem. No systematic method has yet been developed for determining the best realization given constraints on the number of multipliers, word length, and the number of delays. The only recourse is a comparative search for the best of a set of possible structures.

1.5 Organization of the Thesis

Chapter II presents a new second order system matrix suitable for digital filter applications. The parameter space of the matrix elements is examined and the existence of the third order case, along with a method of solution, is discussed.

Chapter III presents the eigenvalue sensitivity analysis of the new system matrix. Other second order system matrices discussed in the literature are also analyzed for comparison with the new matrix. Both absolute and normalized element variations are assumed. Expressions for the magnitude, and the corresponding radial and angular components, of the eigenvalue displacement due to variation in the matrix elements are presented. Minimum pole sensitivity regions within the unit circle of the z-plane are shown for each system matrix. A relationship between sensitivity matrices of equivalent system matrices is shown to exist. A technique for determining the eigenvalue sensitivity of a system matrix from the sensitivity of an equivalent companion matrix is presented.

Chapter IV presents a comparison of the system matrices based on wordlength requirements necessary to insure that the pole-zero movements will remain within prescribed bounds. For a given coefficient wordlength,

the realizable pole locations within the unit circle of the z-plane are presented for each matrix. The tendency of each matrix to sustain overflow oscillations is also examined. As a final comparison, the roundoff noise properties of each matrix is discussed.

Chapter V presents a summary and suggestions for further study.

CHAPTER II

A TRIDIAGONAL SYSTEM MATRIX FOR DIGITAL FILTERS

2.1 Introduction

In analog filter design, LC ladder structures are noted for the relative insensitivity of their frequency response to the element values. Fettweis [36] has conjectured that digital filter structures modeled after them would have the same coefficient sensitivity properties and could be implemented with shorter coefficient wordlengths. Crochiere [37] investigated this conjecture and found that, in many cases, digital ladder structures can be implemented with shorter wordlengths than conventional cascade structures. Fettweis [36] and Crochiere [37] presented methods for designing digital ladder structures that utilized transformations and digitization methods on an existing analog ladder structure.

Marshall [38] has shown that tridiagonal matrices are related to ladder networks. Yarlagadda [39] has shown that a tridiagonal representation of a system can be obtained directly. As applied to digital filters, the use of a tridiagonal system matrix in a state-model representation allows a digital ladder structure to be obtained directly without using digitization of an analog ladder structure.

In this chapter a class of tridiagonal matrices is investigated for use as system matrices in state-model digital filters. The structure of this class of tridiagonal matrices was chosen in an effort to combine

desirable properties of two other classes of matrices used in state-model representation of systems. These classes of matrices are the diagonal matrix which exhibits minimum eigenvalue sensitivity of 1 for real eigenvalues [9]. Rader and Gold [30] has given a second order coupled-loop structure which has an antisymmetric system matrix and exhibits a constant eigenvalue sensitivity of $\sqrt{2}$ for all eigenvalues inside the unit circle of the z-plane [9].

Considering these two aspects, one logical tridiagonal system matrix is

$$K = \begin{bmatrix} k_1 & 1 & 0 & 0 & 0 \\ -1 & k_2 & 1 & 0 & 0 \\ 0 & -1 & . & . & 0 \\ 0 & 0 & . & . & 1 \\ 0 & 0 & 0 & -1 & k_n \end{bmatrix} \qquad ,(2\text{-}1)$$

where it is clear that the only elements subject to variation are the diagonal entries, as in the diagonal matrix. The entries for the two sub-diagonals are chosen to be antisymmetric, i.e.,

$$k_{i,i+1} = -k_{i+1,i}, \quad i = 1,2, \ldots, n-1 \qquad ,(2\text{-}2)$$

as in the coupled-loop structure. The main diagonal elements are required to be real. For n=3, the matrix K in (2-1) is given by

$$K = \begin{bmatrix} k_1 & 1 & 0 \\ -1 & k_2 & 1 \\ 0 & -1 & k_3 \end{bmatrix} \qquad .(2\text{-}3)$$

However, since second order structures are the basic building blocks for

higher order filters [5], the existence of higher order forms of the matrix K are examined only for completeness. There is a general tridiagonal matrix, known as the Schwarz matrix [32], which exists for any n-th order system. However, when applied to digital filters as a second order building block, the matrix becomes identical to the companion matrix.

For the second order case, the parameter space of k_1, k_2 is examined regarding its mapping into the unit circle of the z-plane. The parameter dynamics, or the relationship of k_1 and k_2 as a function of pole locations in the unit circle, is also examined. A second order state-model is also presented.

2.2 Second Order System Matrix

For any state-model representation of digital filter specifications given by the transfer function

$$H(z) = \frac{N(z)}{P(z)} \qquad ,(2\text{-}4)$$

the characteristic polynomial of the system matrix A, given by

$$f(z) = |zI-A| \qquad ,(2\text{-}5)$$

must satisfy

$$P(z) = |zI-A| \qquad .(2\text{-}6)$$

For the second order case let X(z) in (2-4) be given by $X(z) = z^2 + az + b$ where a and b are real. Then the second order form of (2-1), given by

$$K = \begin{bmatrix} k_1 & 1 \\ -1 & k_2 \end{bmatrix} \quad ,(2\text{-}7)$$

must satisfy

$$|zI - K| = z^2 + az + b \quad (2\text{-}8)$$

or

$$z^2-(k_1+k_2)z+k_1k_2+1 = z^2+az+b \quad .(2\text{-}9)$$

Equating coefficients in (2-9) and solving for k_1 and k_2 yields

$$k_1,k_2 = \frac{-a\pm \sqrt{a^2-4(b-1)}}{2} \quad ,(2\text{-}10)$$

where the +(-) sign corresponds to $k_1(k_2)$.

It is clear that the coefficients a and b in (2-8) must satisfy

$$a = -(z_1+z_2) \quad (2\text{-}11)$$

$$b = z_1z_2 \quad (2\text{-}12)$$

where z_1 and z_2 are the zeroes of the polynomial X(z). For digital filter applications, the zeroes of P(z) are the poles of the transfer function H(z) and, for a stable filter, must lie inside the unit circle in the z-plane, i.e., $|z_i|<1$, $i=1,2$.

Therefore b, as given in (2-12), has a magnitude less than one, and this assures that the values of k_1 and k_2 in (2-10) are real. This is the reason that a real matrix K, in (2-7), exists for all stable digital filter operations.

For filters requiring real pole locations, equations (2-10)-(2-12) produce the necessary values of k_1 and k_2.

For filters with complex poles, the restriction that a and b are

real requires the poles to be complex conjugates given, in polar coordinates, by

$$z_1 = \rho\cos\Theta + j\rho\sin\Theta \tag{2-13}$$

$$z_2 = \rho\cos\Theta + j\rho\sin\Theta \tag{2-14}$$

Then (2-11) and (2-12) yield

$$a = -2\rho\cos\Theta \tag{2-15}$$

$$b = \rho^2 \tag{2-16}$$

which, when substituted into (2-10), results in

$$k_1 = \rho\cos\Theta + \sqrt{1-\rho^2\sin^2\Theta} \tag{2-17}$$

$$k_2 = \rho\cos\Theta - \sqrt{1-\rho^2\sin^2\Theta} \tag{2-18}$$

Examination of the possible values of k_1 and k_2 for stable pole locations inside the unit circle results in

$$0<k_1<2 \tag{2-19}$$

$$-2<k_2<0 \tag{2-20}$$

This completes the demonstration of the existence of the second order tridiagonal system matrix for all root locations inside the unit circle. The second order matrix in (2-7) has not been given before and has considerable potential in digital filter synthesis. A comparison of this matrix with other second order matrices is presented in Chapter 3. Equations defining the parameters k_1 and k_2 are given along with the range of parameter values. In the next section the mapping of the parameter space k_1, k_2 into the unit circle of the z-plane is examined.

2.2.1 Parameter Space

A mapping of acceptable values of k_1 and k_2 that permit the realizations of real or complex pole locations inside the unit circle of the z-plane is shown in Figure 1 and Figure 2. The labeled points in Figure 1 are mapped into the corresponding points in Figure 2. Points within bounded regions in Figure 1, such as ABEA, map into pole locations inside the corresponding region of Figure 2. Points in the parameter space along AB are mapped into complex conjugate pole locations on the right half of the unit circle while pole locations on the left half of the unit circle are obtained from a mapping of the parameter space along BC.

The left hand side of (2-9) can be solved for the pole locations z_1 and z_2 in terms of the parameters k_1 and k_2. These solutions are

$$z_1 = \frac{k_1+k_2+\sqrt{(k_1-k_2)^2-4}}{2} \qquad (2\text{-}21)$$

$$z_2 = \frac{k_1+k_2-\sqrt{(k_1-k_2)^2-4}}{2} \qquad (2\text{-}22)$$

Examination of these solutions yields the following information about regions in the parameter space:

If $k_1-k_2>2$ - the poles are real and distinct.

If $k_1-k_2=2$ - the poles are real multiples.

If $k_1-k_2>2$ - the poles are complex conjugates.

Line segment AEC in Figure 1 corresponds to $k_1=2+k_2$. From the above it is clear that the region to the left of AEC, bounded by ABCEA but not including the boundary, maps into complex conjugate pole locations inside the unit circle. Points inside the parameter space bounded by ABEA

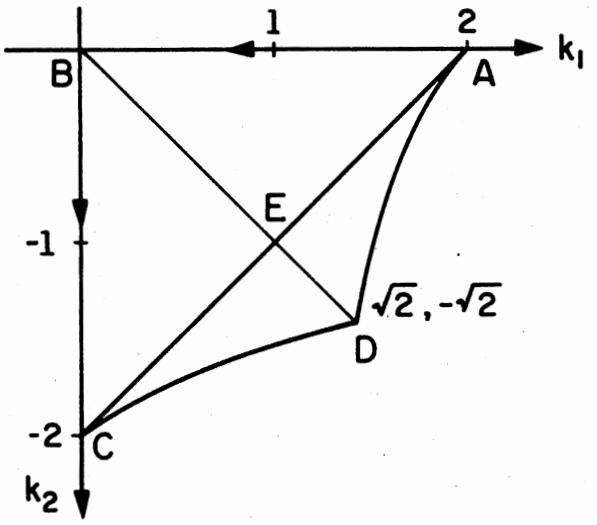


Figure 1. Parameter Space for Valid Poles

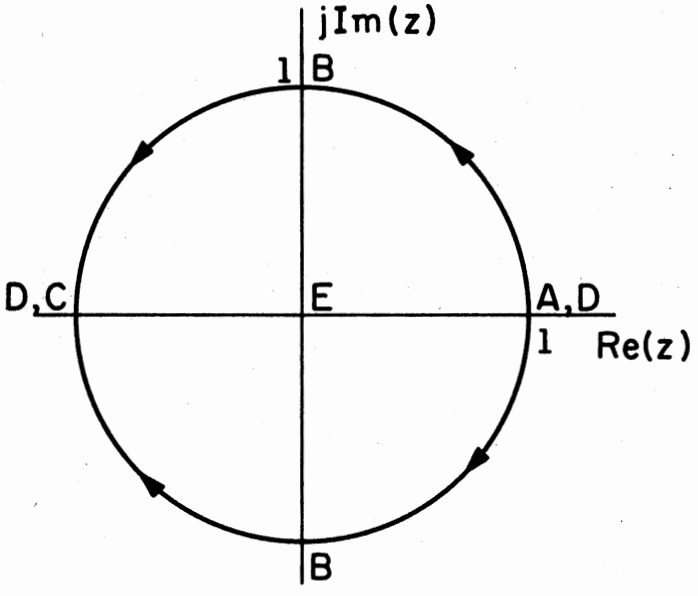


Figure 2. z-Plane Map of Parameter Space

realize complex conjugate poles with positive Re(z) while points inside BCEB realize complex conjugate poles with negative Re(z).

The region to the right of AEC, bounded by AECDA including the boundary, maps into pole locations on the real axis. The boundary ADC corresponds to the case when one or both of the real poles lie on the unit circle. Points on AD in Figure 1 have one pole at 1 while the other can be at any point on the real axis. Points on CD have one pole at -1 while the other can be at any point on the real axis. The equations defining the boundary segments AD and CD are

$$AD \triangleq k_1k_2-k_1-k_2+2 = 0 \qquad \sqrt{2}\leq k_1 \leq 2 \tag{2-23}$$

$$CD \triangleq k_1k_2+k_1+k_2+2 = 0 \qquad 0\leq k_1 \leq \sqrt{2} \tag{2-24}$$

The boundary ABCDA of the parameter space always realizes at least one pole on the unit circle of the z-plane. Since most stable digital filters require poles inside the unit circle, acceptable values for the parameters will not include the boundary.

2.2.2 Parameter Dynamics

The parameter dynamics of the matrix addresses the relationship of k_1 and k_2 as a function of pole location in the unit circle. Such an analysis gives information regarding regions of pole location that require little relative change in magnitude for one parameter as opposed to the other. This information will be used in later chapters.

Using (2-17) and (2-18), Figure 3 shows plots of k_1 and k_2 for poles $\lambda = \rho e^{+j\Theta}$ for different values of magnitude ρ and angle Θ. Also shown is a plot of the curve $2\cos\Theta$ which is the function approached by the sum

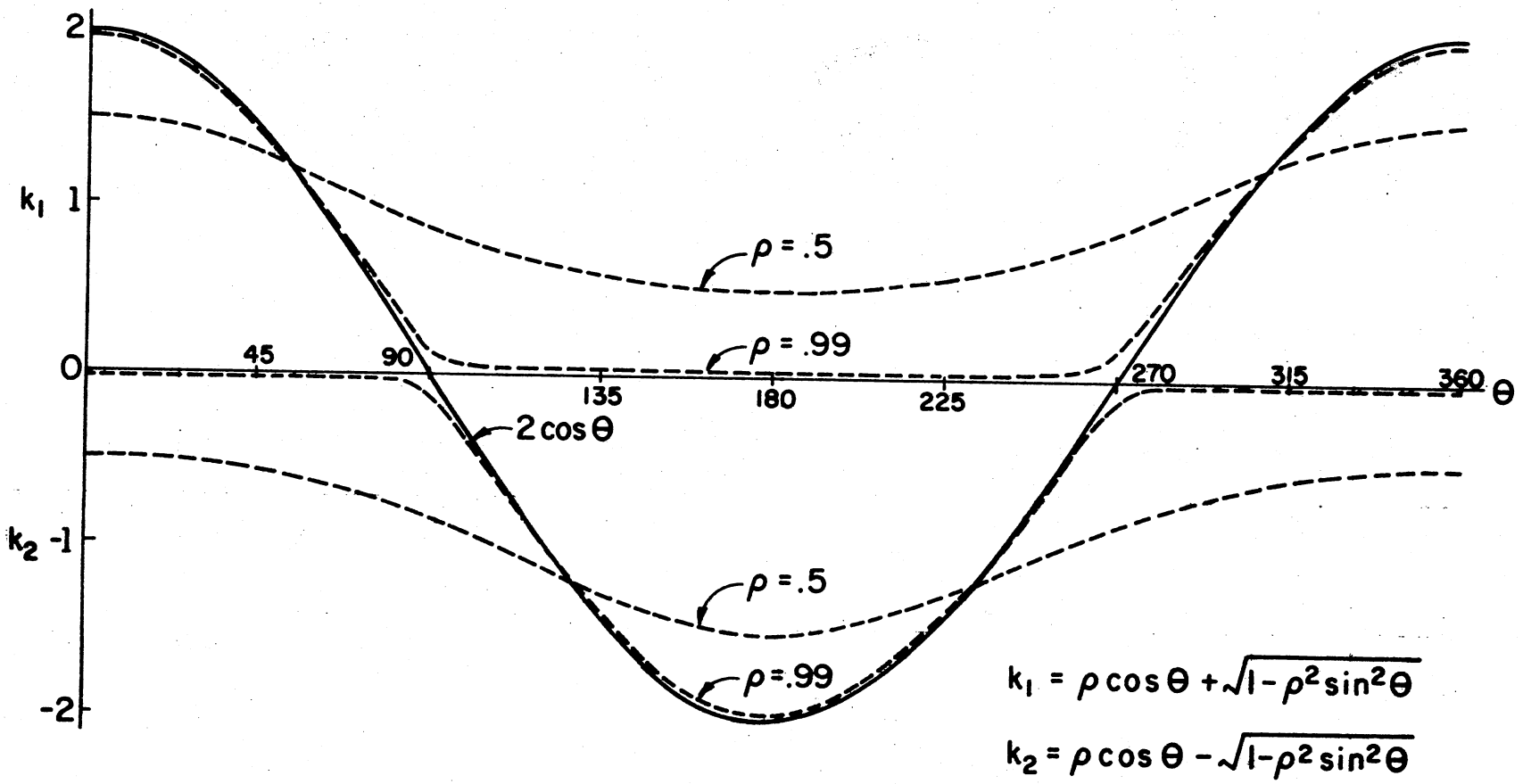


Figure 3. Matrix Parameter Dynamics

of k_1 and k_2 as ρ approaches 1. As shown in Figure 3, for pole locations close to the unit circle the parameters have definite regions of angular location Θ where the location of the pole is primarily determined by only one of the parameters. The limiting case of $\rho=1$ exhibits this feature very well. For angular locations in the first and fourth quadrant of the unit circle ($0\leq\Theta\leq 90$, $270\leq\Theta\leq 360$) the parameter $k_2=0$ while the parameter k_1 determines the pole location. In the second and third quadrants ($90\leq\Theta\leq 270$) the parameter $k_1=0$ while k_2 determines the pole location. As ρ decreases the angular regions primarily dominated by one parameter decrease in size. An interesting result of these parameter characteristics is that the effects of variations in the parameters will be primarily due to variations in the dominant parameter. As ρ approaches 1, the dominant parameter primarily affects the angular location of the pole since the increasing magnitude ρ is a result of the other parameter becoming smaller and more constant in magnitude. Therefore, for poles close to the unit circle, the angular location Θ is more sensitive than the magnitude ρ to changes in the parameters. This intuitive concept is explained analytically in Chapter III.

2.3 Second Order State Model

A state-model utilizing the new second order matrix as a system matrix is now presented. For the implementation of the state-model, a useful characteristic of the second order matrix is applied. As was noted in (2-19) and (2-20), the range of values of k_1 and k_2 for stable pole locations satisfy $0<k_1<2$ and $-2<k_2<0$. For these ranges, k_1 and k_2 can be written as $k_1=x+1$ and $k_2=y-1$, where $|x|<1$ and $|y|<1$. Now K in (2-7) can be written as

$$K = \begin{bmatrix} k_1 & 1 \\ -1 & k_2 \end{bmatrix} = \begin{bmatrix} x & o \\ 0 & y \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} = Q + I \quad ,(2\text{-}25)$$

where I is an invariant matrix consisting of ±1 elements. By using K in this form, the state-model closely represents a simplified implementation of the filter since the matrix I corresponds to hardwired connections and no contingency plan is required to check whether the coefficient magnitudes are less than one [2]. This is a significant advantage over the companion form realization.

A state model for the realization of a digital transfer function

$$H(z) = a_1 \frac{z^2+b_1z+c_1}{z^2+az+b} \; ; \quad V(z)=H(z)U(z) \qquad (2\text{-}26)$$

is given by

$$\begin{bmatrix} X_1(n+1) \\ X_2(n+1) \end{bmatrix} = \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \begin{bmatrix} X_1(n) \\ X_2(n) \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} X_1(n) \\ X_2(n) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(n)$$

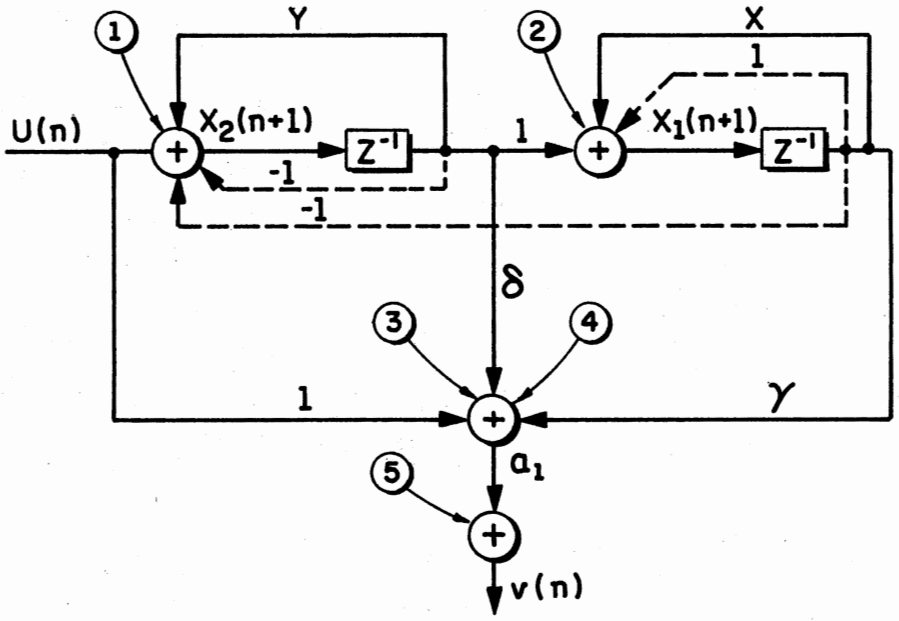
$$W(n) = [\gamma \quad \delta] \begin{bmatrix} X_1(n) \\ X_2(n) \end{bmatrix} + u(n) \; ; \quad v(n) = a_1 W(n)$$

where $\delta = (b_1-a)$, $\gamma = (c_1-b) + \delta k_1$.

The implementation of the model is shown in Figure 4.

2.4 Tridiagonal Realizations Greater Than Second Order

The general existence of the n-th order matrix given in (2-1) that



(i) DENOTES ith EFFECTIVE NOISE INPUT

---- DENOTES HARD-WIRED CONNECTIONS

Figure 4. K-Matrix Realization of Second Order Filter

can realize any given set of n poles is not known for $n \geq 3$. For $n>2$ it becomes necessary to solve an increasingly difficult set of nonlinear equations in the parameters k_i to obtain the matrix. This problem is illustrated for the third order case. Conditions under which it is relatively easy to solve for the parameters are presented along with a method of solution. It is not clear that a solution always exists under these conditions.

In order for a third order tridiagonal matrix given in (2-3) to realize a given pole polynomial

$$f(z) = z^3-az^2+bz-c = 0 \tag{2-27}$$

the determinant $|zI-K|$ must satisfy

$$|zI-K| = f(z) \tag{2-28}$$

or,

$$z^3-(k_1+k_2+k_3)z^2+(k_1k_2+k_1k_3+k_2k_3+2)z-k_1k_2k_3-k_1-k_3=z^3-az^2+bz-c \tag{2-29}$$

Then

$$k_1+k_2+k_3 = a \tag{2-30}$$

$$k_1k_2+k_1k_3+k_2k_3 = b-2 \tag{2-31}$$

$$k_1k_2k_3+k_1+k_3 = c \tag{2-32}$$

are the set of nonlinear equations to be solved for real values of k_1, k_2, and k_3.

By the elimination process or by use of Bezout's determinant [40] equations (2-30)-(2-32) can be reduced to a sixth order polynomial in one parameter given by

$$k_1^6-2ak_1^5+(a^2+2b-3)k_1^4-(2c+2ab-4a)k_1^3+(2ac-a^2-7b+2b^2+7)k_1^2$$

$$-(ab-2a+3c-2bc)k_1+b-2+c^2-ac = 0 \qquad .(2\text{-}33)$$

If a real solution to (2-33) exists that satisfies (2-30)-(2-32) the matrix can be realized. Further discussion on this is omitted as it is fairly routine. In the following, a simpler method of solution is discussed.

2.4.1 Proposed Technique for Third Order Matrices

Examination of (2-30)-(2-32) points out a condition under which solutions might be obtained relatively easily. If the coefficients a and c in the pole polynomial given by (2-27) were equal then (2-29)-(2-31) could be satisfied by $k_2=0$ or $k_1k_3=1$.

Case I: For a=c and $k_2=0$, (2-30)-(2-32) reduce to

$$k_1+k_3 = a \qquad (2\text{-}34)$$

$$k_1k_3 = b-2 \qquad ,(2\text{-}35)$$

for which the solutions for k_1 and k_3 are

$$k_3 = \frac{a\pm\sqrt{a^2+4(2-b)}}{2} \qquad (2\text{-}36)$$

$$k_1 = \frac{a\mp\sqrt{a^2+4(2-b)}}{2} \qquad .(2\text{-}37)$$

Case II: For a=c and $k_1k_3=1$, (2-30)-(2-32) reduce to

$$k_1+k_2+k_3 = a \qquad (2\text{-}38)$$

$$k_1k_2+k_2k_3 = b-3 \qquad (2\text{-}39)$$

$$k_3 = \frac{1}{k_1} \quad ,(2\text{-}40)$$

for which the solutions for k_1, k_2, k_3 are

$$k_1 = \frac{1}{k_3} \quad (2\text{-}41)$$

$$K_3 = \frac{a-k_2 \pm \sqrt{(a-k_2)^2-4}}{2} \quad (2\text{-}42)$$

$$k_2 = \frac{a \pm \sqrt{a^2+4(3-b)}}{2} \quad .(2\text{-}43)$$

For most f(z) given in (2-27) the coefficients a and c will not be equal. However, by employing the change of variable

$$z = \lambda+\varepsilon \quad (2\text{-}44)$$

in (2-27), the resultant $f(\lambda)$, given by

$$f(\lambda) = \lambda^3+\alpha\lambda^2+\beta\lambda+\gamma = 0 \quad ,(2\text{-}45)$$

has coefficients α, β, and γ that are polynomials in ε. It is then possible to find a real ε (since γ is a third order polynomial in ε) such that $\alpha = \gamma$. The previously discussed Case I or Case II can then be applied to obtain possible solutions for the set of equations

$$k_1'+k_2'+k_3' = \alpha \quad (2\text{-}46)$$

$$k_1'k_2'+k_1'k_3'+k_2'k_3' = \beta-2 \quad (2\text{-}47)$$

$$k_1'k_2'k_3'+k_1'+k_3' = \alpha \quad .(2\text{-}48)$$

If a solution exists, then matrix elements k_1, k_2, and k_3 for realization of f(z) in (2-27) are obtained by

$$k_i = k_i' + \varepsilon \qquad i = 1,2,3 \qquad .(2\text{-}49)$$

2.4.2 Example

Given $f(z) = z^3+z^2+0.5z = 0$

find k_1, k_2, k_3 such that

$$\begin{vmatrix} z-k_1 & -1 & 0 \\ 1 & z-k_2 & -1 \\ 0 & 1 & z-k_3 \end{vmatrix} = f(z) \qquad .(2\text{-}50)$$

Applying $z=\lambda+\varepsilon$ to $f(z)$ results in

$$f(\lambda) = \lambda^3+(3\varepsilon+1)\lambda^2+(3\varepsilon^2+2\varepsilon+.5)\lambda+\varepsilon^3+\varepsilon^2+.5\varepsilon = 0 \qquad .(2\text{-}51)$$

Solving for ε such that

$$\varepsilon^3+\varepsilon^2+.5\varepsilon = 3\varepsilon+1$$

results in $\varepsilon = 1.366$. Then $f(\lambda)$ in (2-51) is given by

$$f(\lambda) = \lambda^3+5.098\lambda^2+8.829\lambda+5.098 \qquad ,(2\text{-}52)$$

and the set of equations to be solved are:

$$k_1'+k_2'+k_3' = -5.098 \qquad (2\text{-}53)$$

$$k_1'k_2'+k_1'k_3'+k_2'k_3' = 6.829 \qquad (2\text{-}54)$$

$$k_1'k_2'k_3'+k_1'+k_3' = -5.098 \qquad .(2\text{-}55)$$

Applying Case I for $k_2' = 0$, (2-36) yields

$$k_3' = 2.544 \pm j.576$$

which is complex so this solution does not apply.

Applying Case II for $k_1'k_3' = 1$, equations (2-41)-(2-43) yield $k_1' = -3.038$, $k_2' = -1.732$, $k_3' = -0.329$. Applying (2-49) for the realization of f(z) results in $k_1 = -1.672$, $k_2 = -0.366$, $k_3 = 1.037$. Therefore

$$\begin{vmatrix} z+1.672 & -1 & 0 \\ 1 & z+0.366 & -1 \\ 0 & 1 & z-1.037 \end{vmatrix} = z^3+z^2+0.5z \quad .$$

As was mentioned previously, the general existence of third and higher order matrices for all pole locations is not known. The proposed methods of solving the nonlinear equations (2-30)-(2-32), under the condition that the coefficients a and c in (2-27) are equal, has not been proved to apply for all pole locations. In the example presented, Case I did not provide a solution but Case II did.

In digital filter applications, third and higher order filters are generally realized through first and second order sections in cascade or parallel combinations. This treatment of the third and higher ordered matrices is included only for completeness.

In passing, it should be pointed out that two other third order tridiagonal matrices with ones on the sub-diagonals can be considered. The first matrix is

$$K' = \begin{bmatrix} k_1 & 1 & 0 \\ 1 & k_2 & 1 \\ 0 & 1 & k_3 \end{bmatrix} \qquad (2\text{-}56)$$

Since K' is symmetric, it can realize only real poles and, therefore, is not of much interest. The second matrix

$$K'' = \begin{bmatrix} k_1 & 1 & 0 \\ 1 & k_2 & 1 \\ 0 & -1 & k_3 \end{bmatrix} \tag{2-57}$$

is not of much interest either as it cannot realize most pole locations in the unit circle. This can be proved by using root locus arguments [41].

2.5 Summary

A second order tridiagonal matrix suitable for realizing all stable pole locations for state-model digital filter applications is presented. The mapping of the matrix element parameter space into the unit circle of the z-plane is discussed. The dynamics of the parameter interaction as a function of filter pole location is presented. A state-model using the second order tridiagonal matrix as the system matrix is presented. Third and higher order realizations of all pole locations using this class of tridiagonal matrices is not known. An example for the existence of a third order matrix is presented.

CHAPTER III

SYSTEM MATRIX EIGENVALUE SENSITIVITY

3.1 Introduction

In this chapter the second order form of the class of tridiagonal matrices introduced in Chapter 2 is compared with other second order matrices with similar characteristics that are capable of realizing all pole locations inside the unit circle of the z-plane. The eigenvalue sensitivity of the matrices is used as the method of evaluation. Eigenvalue sensitivity is defined as the change in the eigenvalue locations of a matrix due to changes in the elements of the matrix. This sensitivity measure can be applied to the coefficient accuracy problem of state-model digital filters since the constraint of finite wordlength is a cause of changes in the system matrix elements. The research of Mantey [8] and Singer [9] involved searches for minimally sensitive system matrices for equivalent input/output state-models. As pointed out by Mantey [8] and Oppenheim and Schafer [35], no systematic method has yet been devised for determining the best realization in terms of insensitivity to the effects introduced by the constraints of finite wordlength and the number of multipliers. The only recourse is a search for the best of a set of possible realizations. In this thesis, a simple procedure for selecting a second order matrix with two variational elements is proposed which uses the concept of sensitivity analysis.

The emphasis, in this thesis, will be on system matrices with complex conjugate eigenvalues. Singer [9] has shown that for real eigenvalues the diagonal matrix exhibits a minimal sensitivity of one. In some instances when real eigenvalues have large separations between them they may be associated together in a companion matrix form to achieve a sensitivity, for the smaller eigenvalue, that is less than one. With this exception, real eigenvalues are best realized through diagonal system matrices.

The matrices to be considered in this comparison are:

$$A_s = \begin{bmatrix} \alpha & \pm\beta \\ \mp\beta & \alpha \end{bmatrix}, \quad A_c = \begin{bmatrix} 0 & 1 \\ d_1 & d_2 \end{bmatrix}, \quad A_o = \begin{bmatrix} 1 & 1 \\ x & y \end{bmatrix}, \quad K = \begin{bmatrix} k_1 & 1 \\ -1 & k_2 \end{bmatrix}. \tag{3-1}$$

A characteristic common to these matrices, with the exception of A_s, is that they have only two elements subject to variation. Any zero or unity elements which are in a fixed position of a matrix and do not occur as a result of particular eigenvalue locations are considered to be free of variation. The zero elements require no additions or multiplications to be performed. Since their presence is an indication of an absence of operations, no storage or computer error is associated with them. Similarly, no multiplications are performed when an element of the system matrix is unity. It represents a direct hardwire connection on a special purpose realization and a simple addition (subtraction for a minus sign) in a computer.

The variational elements in the companion matrix A_c are the coefficients of a given pole polynomial. Since these coefficients are functions of the variational elements in the other system matrices of (3-1), it is shown that the eigenvalue sensitivity of the matrices A_s, A_o, and K can

be derived from the eigenvalue sensitivity of the companion matrix A_c. Using this technique, expressions for the magnitude, and the corresponding radial and angular components, of the eigenvalue displacement due to simultaneous variation in the matrix elements are presented. Minimum pole sensitivity regions within the unit circle of the z-plane are shown for each system matrix.

Although second order system matrices are of primary emphasis, the eigenvalue sensitivity technique presented is applicable to a general n-th order system. A qualitative discussion of the extension to an n-th order system is presented later.

3.2 Matrix Eigenvalue Sensitivity Relationship

The following is from Huelsman [19] and is concerned with the variations in the simple roots of a polynomial due to variations in the polynomial coefficients. For second order polynomials with complex conjugate roots, to which this analysis is primarily applied in this thesis, the case of multiple roots does not arise. For the derivation which follows for a general n-th order polynomial, it should be remembered that the analysis applies only to simple roots.

Let f(z) be given by

$$f(z) = \prod_{k=1}^{n} (z-z_k) = \sum_{i=0}^{n} -d_{i+1} z^i \qquad (3\text{-}2)$$

where, without loss of generality, $d_{n+1}=1$. The change in the simple root z_k due to small changes in the coefficients d_j is given by

$$\Delta z_k = \sum_{j=1}^{n} \frac{\partial z_k}{\partial d_j} \Delta d_j \qquad ,(3\text{-}3)$$

where $\frac{\partial z_k}{\partial d_j}$ is obtained from

$$\frac{\partial z_k}{\partial d_j} = \frac{z_k^{j-1}}{Q_k(z_k)} \quad ,(3\text{-}4)$$

and $$Q_k(z_k) = \left.\frac{f(z)}{(z-z_k)}\right|_{z=z_k} \quad .(3\text{-}5)$$

By defining

$$q_{kj} \triangleq \frac{z_k^{j-1}}{Q_k(z_k)} \quad ,(3\text{-}6)$$

equation (3-3) can be written as

$$\vec{\Delta z} = Q\vec{\Delta d} \quad ,(3\text{-}7)$$

where $$\vec{\Delta z} = \text{col}(\Delta z_1, \ldots, \Delta z_n) \quad ,(3\text{-}8)$$

$$\vec{\Delta d} = \text{col}\ (\Delta d_1, \ldots, \Delta d_n) \quad ,(3\text{-}9)$$

and Q is an nxn matrix.

Equation (3-7) relates the polynomial root variations to the polynomial coefficient variations.

The variational elements of a companion matrix are simply the negative of the pole polynomial coefficients. Since the eigenvalues λ_i of the matrix are equal to the roots of the polynomial, (3-7) gives the variation of the matrix eigenvalues λ_i as a result of variations in the matrix elements. For the second order matrix A_c in (3-1) the characteristic polynomial is

$$f(z) = z^2 - d_2 z - d_1 \quad .(3\text{-}10)$$

Application of (3-6) and (3-7) for the matrix eigenvalues λ_1 and λ_2

$$\begin{bmatrix} \Delta\lambda_1 \\ \Delta\lambda_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\lambda_1-\lambda_2} & \frac{\lambda_1}{\lambda_1-\lambda_2} \\ \frac{1}{\lambda_2-\lambda_1} & \frac{\lambda_2}{\lambda_2-\lambda_1} \end{bmatrix} \begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} \quad .(3\text{-}11)$$

For complex eigenvalues given by

$$\lambda_1 = \rho\cos\theta + j\rho\sin\theta \quad (3\text{-}12)$$

$$\lambda_2 = \rho\cos\theta + j\rho\sin\theta \quad ,(3\text{-}13)$$

equation (3-11) yields for $\Delta\lambda_1$

$$\Delta\lambda_1 = \frac{\Delta d_1 + \rho\cos\theta\Delta d_2 + j\rho\sin\theta\Delta d_2}{j2\rho\sin\theta} \quad .(3\text{-}14)$$

It is clear from (3-11) that the variation $\Delta\lambda_2$ is simply the complex conjugate of $\Delta\lambda_1$.

Singer [9] defined the magnitude of the eigenvalue sensitivity of a square matrix F at an eigenvalue λ_k as

$$|S(\lambda_k,F)| = \lim_{\substack{\Delta\to 0 \\ |\Delta f_{ij}|\le\Delta}} \frac{1}{\Delta} \frac{\sup \sqrt{\alpha^T \hat{A}\alpha}}{\prod_{\substack{i=1 \\ i\neq k}}^{n} |\lambda_k-\lambda_i|} \quad ,(3\text{-}15)$$

where α is a vector composed of the matrix element variations Δf_{ij} and $\hat{A}$ is a square, positive definite, symmetric matrix. The limit as $\Delta\to 0$ operation results in an expression for the magnitude of the sensitivity that allows convenient geometrical comparisons of matrix sensitivity. What it effectively does is to normalize the magnitude of the element variations to one since

$$\lim_{\substack{\Delta\to 0 \\ |\Delta f_{ij}|\leq\Delta}} \frac{\Delta f_{ij}}{\Delta} = \pm 1 \qquad ,(3\text{-}16)$$

where the sign is determined by the direction of the element variation. Therefore the vector $\frac{\alpha}{\Delta}$ consists of elements ± 1 when the limit operation is performed in (3-15).

$$|\Delta\lambda_1| = \frac{\sqrt{\Delta^2 d_1 + 2\rho\cos\theta\Delta d_1\Delta d_2 + \Delta^2 d_2\rho^2}}{2\rho|\sin\theta|} \qquad .(3\text{-}17)$$

Employing the limit operation of (3-15) and taking the supremum of (3-17) yields the same result as derived by Singer [9] for the eigenvalue sensitivity of the companion matrix A_c. That result is

$$|S(\lambda_k, A_c)| = \frac{\sqrt{1+2\rho|\cos\theta|+\rho^2}}{2\rho|\sin\theta|} \qquad .(3\text{-}18)$$

From (3-15) and (3-17) it is clear that for the companion matrix A_c, $\alpha^T\hat{A}\alpha$ is given by

$$\alpha^T\hat{A}\alpha = [\Delta d_1 \Delta d_2]\begin{bmatrix} 1 & \rho\cos\theta \\ \rho\cos\theta & \rho^2 \end{bmatrix}\begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} \qquad ,(3\text{-}19)$$

or $$\alpha^T\hat{A}\alpha = \Delta d^T A' \Delta d \qquad .(3\text{-}20)$$

where A' is defined as the sensitivity matrix for the companion matrix A_c. For a given system matrix F, there will be a corresponding variational vector α and a sensitivity matrix $\hat{A}$.

The coefficients of a given pole polynomial are functions of the variational elements of a system matrix used to realize those poles. Since the variational elements of a companion matrix are given by the

polynomial coefficients, it is clear that those elements can be expressed as functions of the elements of any other system matrix used to realize the same poles. As a result of this relationship, a product of this thesis is a technique for the derivation of $|S(\lambda_k,F)|$ for any matrix F based on $|S(\lambda_k,A_c)|$. The method is presented for second order matrices without the loss of generality.

For a second order companion matrix A_c, as given in (3-1), and any second order matrix F with matrix elements f_{ij}, i,j=1,2, the functional relationship of d_1 and d_2 to the elements f_{ij} can be given by

$$d_1 = g(f_{ij}) \tag{3-21}$$

$$d_2 = h(f_{ij}) \tag{3-22}$$

For small variations in f_{ij}, the resultant changes in d_1 and d_2 are given by

$$\Delta d_1 = \sum_{i,j} \frac{\partial g}{\partial f_{ij}} \Delta f_{ij} \tag{3-23}$$

$$\Delta d_2 = \sum_{i,j} \frac{\partial h}{\partial f_{ij}} \Delta f_{ij} \tag{3-24}$$

Equations (3-23) and (3-24) can be written as

$$\begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} = \begin{bmatrix} \frac{\partial g}{\partial f_{11}} & \frac{\partial g}{\partial f_{12}} & \frac{\partial g}{\partial f_{21}} & \frac{\partial g}{\partial f_{22}} \\ \frac{\partial h}{\partial f_{11}} & \frac{\partial h}{\partial f_{12}} & \frac{\partial h}{\partial f_{21}} & \frac{\partial h}{\partial f_{22}} \end{bmatrix} \begin{bmatrix} \Delta f_{11} \\ \Delta f_{12} \\ \Delta f_{21} \\ \Delta f_{22} \end{bmatrix} \tag{3-25}$$

or $$\Delta d = M \cdot \Delta f \tag{3-26}$$

As shown in (3-23) and (3-24), if a matrix element is invariant it does

not appear in (3-25). For example, for two elements subject to variation, M is 2x2 and Δf is 2x1 in size.

Substituting (3-26) into (3-20) results in

$$\alpha^T \hat{A} \alpha = \Delta f^T M^T A' M \Delta f \qquad ,(3\text{-}27)$$

where $\hat{A} = M^T A' M$ is the sensitivity matrix of F derived from the sensitivity matrix A' of the companion matrix A_c and Δf is the corresponding variational vector α.

Equation (3-27) is the relationship between matrix eigenvalue sensitivities that was being sought. Given the sensitivity matrix A' for the companion matrix, the eigenvalue sensitivity, as defined in (3-15), of any other system matrix of the same order can be obtained through (3-25), (3-26), and (3-27). For the second order matrix A_c, A' as defined in (3-20) is given in (3-19) as

$$A' = \begin{bmatrix} 1 & \rho\cos\Theta \\ \rho\cos\Theta & \rho^2 \end{bmatrix} \qquad .(3\text{-}28)$$

Singer [9] presents A' for the general n-th order companion matrix.

As an illustration of the described method for determining matrix eigenvalue sensitivity, the sensitivity of the matrix A_s in (3-1) will be determined. Singer [9] has shown that the sensitivity given in (3-15) for this matrix is $|S(\lambda_k, A_s)| = \sqrt{2}$. It is a good example for illustrating the above procedure because all four elements are subject to variation.

3.2.1 Example

For the matrix

$$A_s = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \quad ,(3\text{-}29)$$

equations (3-21) and (3-22) are given by

$$d_1 = -\alpha^2-\beta^2 = f_{11}f_{22}-f_{12}f_{21} \quad (3\text{-}30)$$

$$d_2 = 2\alpha = f_{11}+f_{22} \quad ,(3\text{-}31)$$

where $\alpha=\rho\cos\Theta$ and $\beta=\rho\sin\Theta$ for eigenvalues

$$\lambda_1,\lambda_2 = \rho\cos\Theta\pm j\rho\sin\Theta \quad .(3\text{-}32)$$

Applying (3-25) results in (3-26) being given by

$$\begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} = \begin{bmatrix} -\alpha & \beta & -\beta & -\alpha \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta f_{11} \\ \Delta f_{12} \\ \Delta f_{21} \\ \Delta f_{22} \end{bmatrix} \quad .(3\text{-}33)$$

Even though there are four variational elements in this matrix, the variation of the elements is not independent. A given variation in f_{11} will also exist in f_{22}. Also, $\Delta f_{12} = -\Delta f_{21}$ due to the antisymmetric nature of the matrix. Due to these relationships (3-33) can be conveniently condensed, as is also evident from (3-30) and (3-31), into

$$\begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} = \begin{bmatrix} -2\alpha & -2\beta \\ 2 & 0 \end{bmatrix} \begin{bmatrix} \Delta\alpha \\ \Delta\beta \end{bmatrix} \quad .(3\text{-}34)$$

Equation (3-34) as (3-33) can be used in (3-27) to determine the sensitivity matrix $\hat{A}$ for the system matrix A_s. For convenience (3-34) will

be used. The fact that A_s can be analyzed in terms of two variational elements, although all four elements are subject to variation, is the reason it was included in (3-1).

Substituting (3-34) and (3-28) into (3-27) results in the sensitivity matrix $\hat{A}$ for the matrix A_s being given by

$$\hat{A}(A_s) = \begin{bmatrix} 4\rho^2\sin^2\Theta & 0 \\ 0 & 4\rho^2\sin^2\Theta \end{bmatrix} \qquad .(3\text{-}35)$$

Using (3-35) in (3-15) with $\alpha^T=[\Delta\alpha\ \Delta\beta]$ results in

$$|S(\lambda_k,A_s)| = \lim_{\substack{\Delta\to 0 \\ |\Delta f_{ij}|\leq\Delta}} \frac{1}{\Delta} \frac{\sup\sqrt{4\rho^2\sin^2\Theta(\Delta^2\alpha+\Delta^2\beta)}}{2\rho|\sin\Theta|} \qquad .(3\text{-}36)$$

This yields

$$|S(\lambda_k,A_s)| = \sup \frac{\sqrt{8\rho^2|\sin^2\Theta|}}{2\rho|\sin\Theta|} = \sqrt{2} \qquad .(3\text{-}37)$$

3.3 Eigenvalue Sensitivity Expressions for Second Order System Matrices

In the previous section a method was presented for obtaining the eigenvalue sensitivity of a matrix from the sensitivity expression for a companion matrix. For illustration of the method, an example was presented for the matrix A_s in (3-1). In this section, eigenvalue sensitivity expressions for the remaining system matrices in (3-1) are developed using the same method. Since information on the variation of the matrix elements may be given as absolute variations or as tolerances, expressions are developed for both absolute and normalized element variations.

3.3.1 Absolute Element Variations

Absolute element variations refer to the information for the variational vector α in (3-15) being given in terms of a magnitude and a direction either positive or negative. When the limit operation is performed, the element variations are normalized to plus or minus one depending on the direction assumed for each element variation. This allows the comparison of the eigenvalue sensitivity of various system matrices to be based on equal magnitude of variation in the matrix elements. As shown in the example in 3.2.1, the supremum is determined by the assumed directions, plus or minus, the variations take. Without employing the limit operation, (3-15) gives the sensitivity of an eigenvalue λ_k due to given element variations Δf_{ij} in the vector α. It should be remembered that (3-15) is derived under the assumption of small element variations.

In section 3.2 expressions for the eigenvalue sensitivity of the matrices A_c and A_s were presented. These expressions, repeated here for convenience, are:

$$|S(\lambda_k,A_c)| = \frac{\sup \sqrt{\Delta^2 d_1+2\rho\cos\Theta\Delta d_1\Delta d_2+\Delta^2 d_2\rho^2}}{2\rho|\sin\Theta|} \qquad (3\text{-}38)$$

with normalized, equal variations resulting in

$$|S(\lambda_k,A_c)| = \frac{\sqrt{1+2\rho|\cos\Theta|+\rho^2}}{2\rho|\sin\Theta|} \qquad ,(3\text{-}39)$$

and

$$|S(\lambda_k,A_s)| = \sqrt{2} \qquad .(3\text{-}40)$$

Sensitivity equations for the remaining matrices, A_o and K, in (3-1) are now developed.

Matrix A_o

For the matrix

$$A_o = \begin{bmatrix} 1 & 1 \\ x & y \end{bmatrix} \tag{3-41}$$

equations (3-21) and (3-22) are given by

$$d_1 = x-y \tag{3-42}$$

$$d_2 = 1+y \tag{3-43}$$

For this matrix only two elements are subject to variation. Therefore (3-25) is given by

$$\begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{3-44}$$

Substituting (3-44) and (3-28) into (3-27) results in the sensitivity matrix $\hat{A}$ for the matrix A_o being given by

$$\hat{A}(A_o) = \begin{bmatrix} 1 & \rho\cos\Theta-1 \\ \rho\cos\Theta-1 & \rho^2-2\rho\cos\Theta+1 \end{bmatrix} \tag{3-45}$$

Using (3-45) in (3-15) with $\alpha^T=[\Delta x \quad \Delta y]$ results in

$$|S(\lambda_k,A_o| = \lim_{\substack{\Delta\to 0 \\ |\Delta f_{ij}|\leq\Delta}} \frac{1}{\Delta} \frac{\sup\sqrt{\Delta^2 x-2(1-\rho\cos\Theta)\Delta x\Delta y+(1-2\rho\cos\Theta+\rho^2)\Delta^2 y}}{2\rho|\sin\Theta|} \tag{3-46}$$

or

$$|S(\lambda_k,A_o)| = \frac{\sqrt{4-4\rho\cos\Theta+\rho^2}}{2\rho|\sin\Theta|} \tag{3-47}$$

where the supremum is obtained by setting $\Delta x=-\Delta y$.

Matrix K:

For the matrix

$$K = \begin{bmatrix} k_1 & 1 \\ -1 & k_2 \end{bmatrix} \tag{3-48}$$

equations (3-21) and (3-22) are given by

$$d_1 = -k_1k_2-1 \tag{3-49}$$

$$d_2 = k_1+k_2 \tag{3-50}$$

For this matrix only two elements are subject to variation. Therefore (3-25) is given by

$$\begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} = \begin{bmatrix} -k_2 & -k_1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \Delta k_1 \\ \Delta k_2 \end{bmatrix} \tag{3-51}$$

Substituting (3-51) and (3-28) into (3-27), and using $k_1,k_2=\rho\cos\Theta \pm \sqrt{1-\rho^2\sin^2\Theta}$, results in the sensitivity matrix $\hat{A}$ for the matrix K being given by

$$\hat{A}(K) = \begin{bmatrix} 1 & 2\rho^2\sin^2\Theta-1 \\ 2\rho^2\sin^2\Theta-1 & 1 \end{bmatrix} \tag{3-52}$$

Using (3-52) in (3-15) with $\alpha^T = [\Delta k_1 \ \Delta k_2]$ results in

$$|S(\lambda_k,K)| = \lim_{\substack{\Delta\to 0 \\ |\Delta f_{ij}|\leq\Delta}} \frac{1}{\Delta} \frac{\sup\sqrt{\Delta^2k_1+2(2\rho^2\sin^2\Theta-1)\Delta k_1\Delta k_2+\Delta^2k_2}}{2\rho|\sin\Theta|} \tag{3-53}$$

or

$$|S(\lambda_k,K)| = \begin{bmatrix} 1 & \text{if } 2\rho^2\sin^2\theta \geq 1 \\ \sqrt{\dfrac{1}{\rho^2\sin^2\theta} - 1} & \text{if } 2\rho^2\sin^2\theta < 1 \end{bmatrix} \quad .(3\text{-}54)$$

The sensitivity expressions derived in this section have been presented by Singer [9] for all the system matrices except the matrix K introduced in this thesis. The method of derivation, however, is new and follows the manner presented in Section 3.2. All of the expressions thus far have given the supremum, or maximum, sensitivity that can be expected for a given eigenvalue. While it is beneficial to compare the worst case sensitivity of matrices, it is also interesting to compare the sensitivity obtained without taking the supremum. In some cases this might be the actual sensitivity experienced, depending on the direction in which each element varies, and could affect the choice of a system matrix to realize a given set of complex poles. For all the matrices considered, this sensitivity is obtained when the variations of the elements are all in the same direction. Under this condition, the eigenvalue sensitivity of A_c, A_s, A_o, and K are obtained from (3-38), (3-40), (3-46), and (3-53), respectively, and are given as

$$|S_1(\lambda_k,A_s)| = \sqrt{2} \quad , \quad |S_1(\lambda_k,A_c)| = \frac{\sqrt{1+2\rho\cos\theta+\rho^2}}{2\rho|\sin\theta|} \quad ,(3\text{-}55)$$

$$|S_1(\lambda_k,K)| = 1 \quad , \quad |S_1(\lambda_k,A_o)| = \frac{1}{2|\sin\theta|} \quad ,(3\text{-}56)$$

where the subscript 1 is used to differentiate these sensitivities from those obtained using the supremum.

The sensitivity of the matrix K is shown to be one for parameter variations in the same direction. This is an important product of the introduction of the matrix. Using this matrix, complex poles can achieve

a constant sensitivity, for all pole locations, equal to the sensitivity of real poles realized by a diagonal matrix. In Chapter II the rationale behind choosing the structure of K included the desire to have the variational elements on the main diagonal since diagonal matrices realized real poles with a sensitivity of one. It is evident that the sensitivity properties of the diagonal matrix have been carried over to complex poles due to the structure of K.

The results obtained in this section are summarized in Table I. These results will be used in section 3.3.3 in a comparison of the eigenvalue sensitivities of the matrices.

3.3.2 Normalized Variations

The sensitivity expressions derived in the last section allow the computation of the eigenvalue sensitivity in terms of the absolute variations of matrix elements. In many cases, however, information on the variational elements is given in terms of tolerances rather than absolute variations. In these cases, the vector α in (3-15) should be expressed as normalized values $\Delta f_{ij}/f_{ij}$. Since $\alpha^T = [\Delta f_1 \ldots \Delta f_m]$, where m is the number of elements subject to variation, and the normalized variation $\Delta f_{iN} = \Delta f_i/f_i$ (i=1, ..., m) then

$$\alpha^T = \alpha_N^T D \tag{3-57}$$

where $\alpha_N^T = [\Delta f_{1N} \ldots \Delta f_{mN}]$ and D is a diagonal matrix with main diagonal elements $f_1, \ldots, f_m$.

Substituting (3-57) into (3-15) without the supremum or limit operation, the matrix eigenvalue sensitivity in terms of normalized variational elements is given by

TABLE I

SYSTEM MATRIX EIGENVALUE SENSITIVITIES

	A_s	A_c	A_o	K
$\lvert S(\lambda_k,F)\rvert$	$\sqrt{2}$	$\dfrac{\sqrt{1+2\rho\lvert\cos\theta\rvert+\rho^2}}{2\rho\lvert\sin\theta\rvert}$	$\dfrac{\sqrt{4-4\rho\cos\theta+\rho^2}}{2\rho\lvert\sin\theta\rvert}$	$1,\ (2\rho^2\sin^2\theta\geq 1)$ $\sqrt{\dfrac{1}{\rho^2\sin^2\theta}}-1,\ (2\rho^2\sin^2\theta\leq 1)$
$\lvert S_1(\lambda_k,F)\rvert$	$\sqrt{2}$	$\dfrac{\sqrt{1+2\rho\cos\theta+\rho^2}}{2\rho\lvert\sin\theta\rvert}$	$\dfrac{1}{2\lvert\sin\theta\rvert}$	1

$$|S(\lambda_k,F)| = \frac{\sqrt{\alpha_N^T D\hat{A}D\alpha_N}}{\prod_{\substack{i=1\\ i\neq k}}^{n} |\lambda_k-\lambda_i|} \qquad .(3\text{-}58)$$

Eigenvalue sensitivity expressions utilizing the tolerance description of element variations have not been previously presented for the matrices under consideration. Using (3-58) these expressions are now developed.

Matrix A_s:

For A_s, (3-57) is given by

$$[\Delta\alpha \quad \Delta\beta] = [\Delta\alpha_N \quad \Delta\beta_N] \begin{bmatrix} \rho\cos\Theta & 0 \\ 0 & \rho\sin\Theta \end{bmatrix} \qquad ,(3\text{-}59)$$

where $\alpha=\rho\cos\Theta$, $\beta=\rho\sin\Theta$.

Using (3-59) as $\alpha_N^T D$ and $\hat{A}$ given in (3-35), (3-58) gives the sensitivity in terms of element tolerances as

$$|S(\lambda_k,A_s)| = \sqrt{\rho^2\cos^2\Theta\Delta^2\alpha_N+\rho^2\sin^2\Theta\Delta^2\beta_N} \qquad .(3\text{-}60)$$

Matrix A_c:

For A_c, (3-57) is given by

$$[\Delta d_1 \quad \Delta d_2] = [\Delta d_{1N} \quad \Delta d_{2N}] \begin{bmatrix} -\rho^2 & 0 \\ 0 & 2\rho\cos\Theta \end{bmatrix} \qquad ,(3\text{-}61)$$

where $d_1=-\rho^2$, $d_2=2\rho\cos\Theta$.

Using (3-61) as $\alpha_N^T D$ and A given in (3-28), (3-58) gives the sensitivity in terms of element tolerances as

$$|S(\lambda_k,A_c)| = \frac{\sqrt{\rho^4\Delta^2 d_{1N}-4\rho^4\cos^2\Theta\Delta d_{1N}\Delta d_{2N}+4\rho^4\cos^2\Theta\Delta^2 d_{2N}}}{2|\rho\sin\Theta|} \qquad .(3\text{-}62)$$

Matrix A_o:

For A_o, (3-57) is given by

$$[\Delta x \quad \Delta y] = [\Delta x_N \quad \Delta y_N] \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \qquad ,(3\text{-}63)$$

where $x = 2\rho\cos\Theta - 1 - \rho^2$ and $y = 2\rho\cos\Theta - 1$.

Using (3-63) as $\alpha_N^T D$ and A given in (3-45), (3-58) gives the sensitivity in terms of element tolerances as

$$|S(\lambda_k, A_o)| = \frac{\sqrt{x^2\Delta^2 x_N + 2xy(\rho\cos\Theta - 1)\Delta x_N \Delta y_N - xy^2\Delta^2 y_N}}{2\rho|\sin\Theta|} \qquad .(3\text{-}64)$$

Matrix K:

For K, (3-57) is given by

$$[k_1 \quad k_2] = [k_{1N} \quad k_{2N}] \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \qquad ,(3\text{-}65)$$

where $k_1 = \rho\cos\Theta + \sqrt{1-\rho^2\sin^2\Theta}$, $k_2 = \rho\cos\Theta - \sqrt{1-\rho^2\sin^2\Theta}$.

Using (3-65) as $\alpha_N^T D$ and A given in (3-52), (3-58) gives the sensitivity in terms of element tolerances as

$$|S(\lambda_k, K)| = \frac{\sqrt{k_1^2\Delta^2 k_{1N} + k_1 k_2(2\rho^2\sin\Theta - 1)\Delta k_{1N}\Delta k_{2N} + k_2^2\Delta^2 k_{2N}}}{2\rho|\sin\Theta|} \qquad .(3\text{-}66)$$

3.3.3 Matrix Eigenvalue Sensitivity Comparison

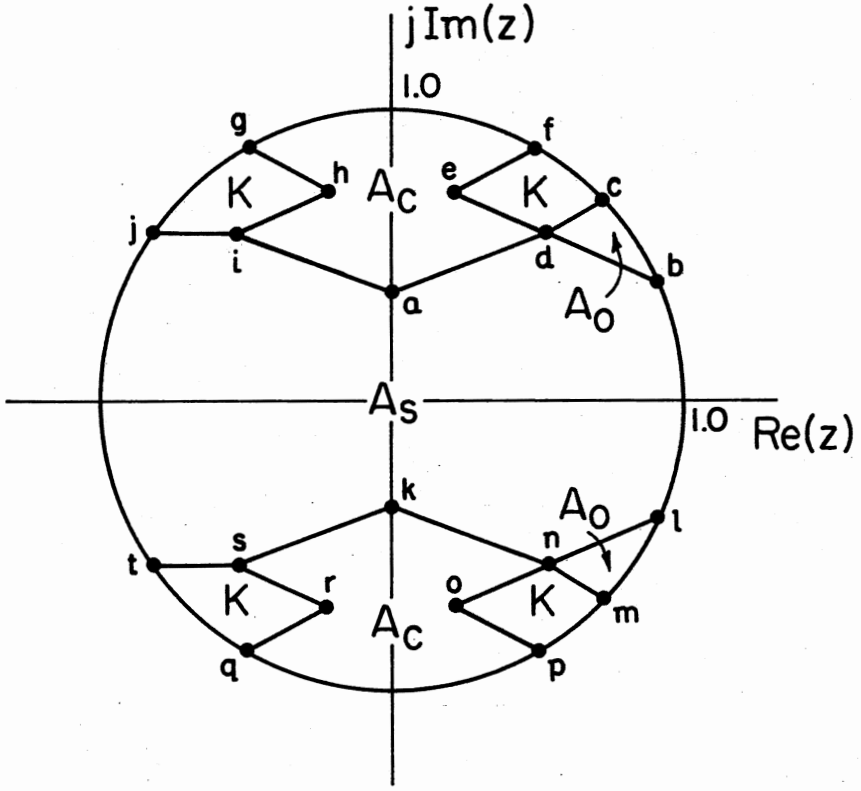
Singer's definition of eigenvalue sensitivity, given in (3-15), is convenient in that it allows the sensitivity characteristics of different system matrices to be compared and summarized in a geometrical manner. Figure 5 and Figure 6 summarize the eigenvalue sensitivity expressions listed in Table I. They indicate the proper choice of a system matrix in order to provide minimum sensitivity for complex eigenvalues λ_k. In

COORDINATES

k - (0.0, -0.378)
l - (0.911, -0.411)
m - (0.725, -0.689)
n - (0.528, -0.577)

o - (0.225, -0.707)
p - (0.500, -0.866)
q - (-0.500, -0.866)
r - (-0.225, -0.707)

s - (-0.528, -0.577)
t - (-0.816, -0.577)



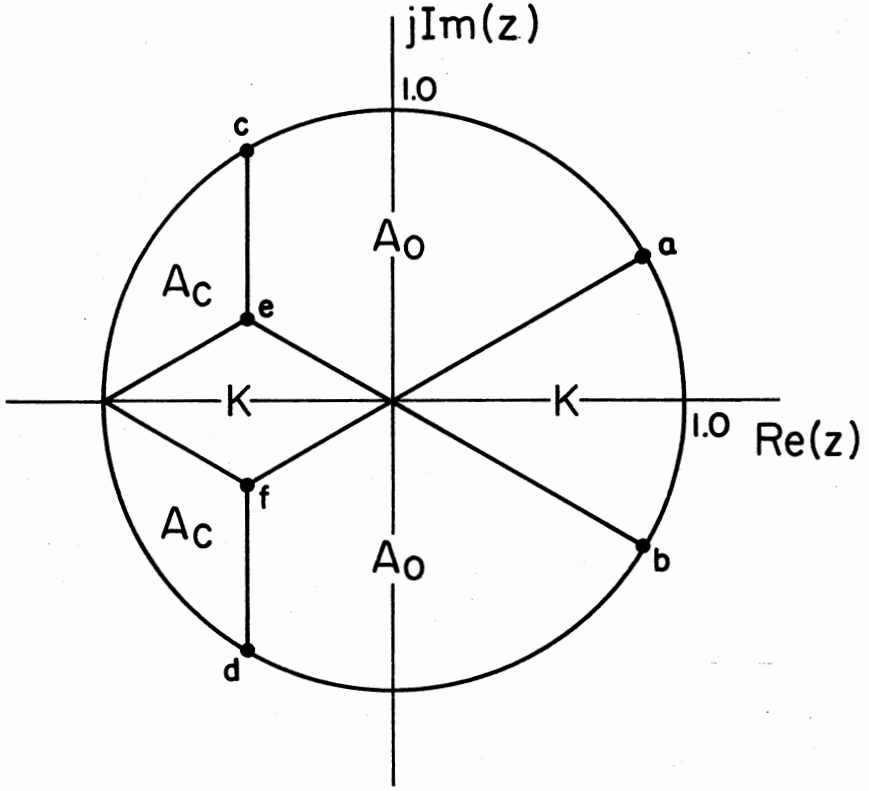
COORDINATES

a - (0.0, 0.378)
b - (0.911, 0.411)
c - (0.725, 0.689)
d - (0.528, 0.577)

e - (0.225, 0.707)
f - (0.500, 0.866)
g - (-0.500, 0.866)
h - (-0.225, 0.707)

i - (-0.528, 0.577)
j - (-0.816, 0.577)

Figure 5. Minimum $|S(\lambda_k, F)|$ Regions



COORDINATES

a - (0.866, 0.500)
b - (0.866, -0.500)
c - (-0.500, 0.866)
d - (-0.500, -0.866)
e - (-0.500, 0.288)
f - (-0.500, -0.288)

Figure 6. Minimum $|S_1(\lambda_k, F)|$ Regions

Figure 5 the areas indicate regions of pole locations in the unit circle for which the matrices A_s, A_c, A_o, and K offer the minimum, worst case sensitivity given by $|S(\lambda_k,F)|$ in Table I. The comparison shown in Figure 6 has not been given before. In this figure the areas indicate regions of pole locations in the unit circle for which the matrices A_c, A_o, and K offer the minimum sensitivity when the supremum is not taken and the element variations are in the same direction, as given by $|S_1(\lambda_k,F)|$ in Table I. The matrix A_s is not shown since the sensitivity of K is always less than that of A_s, as shown in Table I. The striking difference between Figure 5 and Figure 6 points to the importance of the direction in which the matrix elements vary. As explained previously, taking the supremum of the sensitivity is, in effect, placing the requirement that the elements magnitude varies in the same or opposite directions depending on pole location. For example, the supremum of the sensitivity of A_c requires that the element variations be in the same direction for pole angles Θ such that $\cos\Theta$ is positive, and that they be in the opposite direction when $\cos\Theta$ is negative.

In comparison with the other matrices, the newly introduced matrix K does fairly well. In the worst case it offers minimum sensitivity for pole locations close to the unit circle where the stability of a digital filter is of great concern. This is an important factor in digital filter design since optimal filters have poles very close to the unit circle. In the design of digital resonators Gold and Rader [25] point out a common practice of moving pole locations inside the unit circle by an amount $\varepsilon \approx 2^{-20}$ so that the radius is given by $\rho=1-2^{-20}$ and stability problems due to coefficient quantization can be avoided. When the element

variations are in the same direction, K offers minimum sensitivity for pole locations near the real axis.

3.3.4 Example

As an example of the use of Table I and Figure 5, consider a six pole digital filter with pole polynomial given as

$$f(z)=z^6-2.778z^5+4.622z^4-4.812z^3+3.68z^2-1.450z+0.306 \qquad .(3\text{-}67)$$

The poles of this filter are located at $\lambda_{1,2} = 0.3\pm j0.935$, $\lambda_{3,4} = 0.5\pm j0.707$, $\lambda_{5,6} = 0.589\pm j0.276$. Utilizing second order filters as building blocks, determine which matrices offer minimum worst case sensitivity for each complex pole pair.

Placement of the poles on Figure 5 indicates the following choices for realizations: A_c for $\lambda_{1,2}$; K for $\lambda_{3,4}$; and A_s for $\lambda_{5,6}$.

Evaluation of $|S(\lambda_k,F)|$ in Table I for each matrix gives the following results:

$$|S(\lambda_{1,2},A_c)| = 0.85 \quad |S(\lambda_{3,4},A_c)| = 1.17 \quad |S(\lambda_{5,6},A_c)| = 2.92$$

$$|S(\lambda_{1,2},A_s)| = 1.41 \quad |S(\lambda_{3,4},A_s)| = 1.41 \quad |S(\lambda_{5,6},A_s)| = 1.41$$

$$|S(\lambda_{1,2},A_o)| = 1.04 \quad |S(\lambda_{3,4},A_o)| = 1.95 \quad |S(\lambda_{5,6},A_o)| = 2.60$$

$$|S(\lambda_{1,2},K)| = 1.00 \quad |S(\lambda_{3,4},K)| = 1.00 \quad |S(\lambda_{5,6},K)| = 3.48$$

The results of calculating the sensitivities in Table I confirm the original choice of realizing each complex pole pair based on pole location in Figure 5.

3.4 Radial and Angular Sensitivity Expressions for Second Order System Matrices

In many cases, the magnitude of the eigenvalue sensitivity may not give enough information. Although it is a good basis for comparing different system matrices, it does not give any information as to the nature of the variation. For filters with poles very near the unit circle, the radial change in pole location is very important from stability considerations. For filters with stringent resonant or cut-off frequency specifications, the angular change in pole location is of most importance. For filter requirements where it is possible to sacrifice sensitivity qualities in either the radial or the angular direction in order to achieve the best sensitivity in the more critical of the two directions, a different system matrix might be chosen than one chosen based on sensitivity magnitude alone.

In this section radial and angular sensitivity expressions in terms of absolute and normalized element variations are developed for each of the second order matrices under consideration. Rader and Gold [30] and Mitra and Sherwood [29] have presented the expressions for A_c and A_s previously. The expressions for A_o and K have not been presented previously. Mitra and Sherwood [29] present a method for determining radial and angular sensitivities for the poles of an n-th order polynomial that involves the partial fraction expansion of the pole polynomial. The method used here for determining the expressions for A_c follow the more direct approach of Rader and Gold [30]. Following the procedure of section 3.3, the expressions for the matrices A_s, A_o, and K are derived from those of the matrix A_c. The sensitivities of the matrices are

compared by showing regions of the unit circle where they offer minimum angular or radial sensitivity.

Radial and angular variations are the components of the change in eigenvalue location. It is shown that the sensitivity matrix $\hat{A}$ of section 3.3 can be obtained from the radial and angular sensitivity expressions. .

3.4.1 Absolute Element Variations

For the companion matrix A_c in (3-1) to realize the complex poles $\lambda_k=\rho e^{\pm j\Theta}$, the elements d_1 and d_2 are given by

$$d_1 = -\rho^2 \tag{3-68}$$

$$d_2 = 2\rho\cos\Theta \tag{3-69}$$

Assuming small variations,

$$\Delta d_1 = \frac{\partial d_1}{\partial \rho}\Delta\rho \tag{3-70}$$

$$\Delta d_2 = \frac{\partial d_2}{\partial \rho}\Delta\rho + \frac{\partial d_2}{\partial \Theta}\Delta\Theta \tag{3-71}$$

Which results in

$$\Delta d_1 = -2\rho\Delta\rho \tag{3-72}$$

$$\Delta d_2 = 2\cos\Theta\Delta\rho - 2\rho\sin\Theta\Delta\Theta \tag{3-73}$$

Solving (3-72) and (3-73) for $\Delta\rho$ and $\Delta\Theta$ yields,

$$\Delta\rho = \frac{-\Delta d_1}{2\rho} \tag{3-74}$$

$$\Delta\Theta = \frac{-\Delta d_1}{2\rho^2\tan\Theta} - \frac{\Delta d_2}{2\rho\sin\Theta} \tag{3-75}$$

These are the results obtained by Rader and Gold [30]. They can be conveniently presented in matrix format as

Matrix A_c:

$$\begin{bmatrix} \Delta\rho \\ \Delta\theta \end{bmatrix} = \begin{bmatrix} \frac{-1}{2\rho} & 0 \\ \frac{-1}{2\rho^2\tan\theta} & \frac{-1}{2\rho\sin\theta} \end{bmatrix} \begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} \quad .(3\text{-}76)$$

As mentioned previously, the elements d_1 and d_2 of the companion matrix are in one-to-one correspondence with the coefficients of the pole polynomial and are, therefore, functions of matrix elements for which this one-to-one correspondence does not apply. Using this relationship in the same manner as was presented for the derivation of the sensitivity magnitude equations, the radial and angular sensitivity expressions for the matrices A_s, A_o, and K are obtained as follows:

For matrix A_s: Substitution of (3-34) into (3-76) results in

Matrix A_s:

$$\begin{bmatrix} \Delta\rho \\ \Delta\theta \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ \frac{-\sin\theta}{\rho} & \frac{\cos\theta}{\rho} \end{bmatrix} \begin{bmatrix} \Delta\alpha \\ \Delta\beta \end{bmatrix} \quad ,(3\text{-}77)$$

which agrees with the results of Rader and Gold [30].

For matrix A_o: Substitution of (3-44) into (3-76) results in

Matrix A_o:

$$\begin{bmatrix} \Delta\rho \\ \Delta\theta \end{bmatrix} = \begin{bmatrix} \frac{-1}{2\rho} & \frac{1}{2\rho} \\ \frac{-1}{2\rho^2\tan\theta} & \frac{\cos\theta-\rho}{2\rho^2\sin\theta} \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad .(3\text{-}78)$$

For matrix K: Substitution of (3-51) into (3-76) results in

Matrix K:

$$\begin{bmatrix} \Delta\rho \\ \Delta\Theta \end{bmatrix} = \begin{bmatrix} \frac{k_2}{2\rho} & \frac{k_1}{2\rho} \\ \frac{k_2\cos\Theta-\rho}{2\rho^2\sin\Theta} & \frac{k_1\cos\Theta-\rho}{2\rho^2\sin\Theta} \end{bmatrix} \begin{bmatrix} \Delta k_1 \\ \Delta k_2 \end{bmatrix} \quad ,(3\text{-}79)$$

where $k_1, k_2 = \rho\cos\Theta \pm \sqrt{1-\rho^2\sin^2\Theta}$.

The sensitivity expressions in (3-78), to present knowledge, and certainly (3-79) have not been presented before.

3.4.2 Normalized Element Variations

To obtain radial and angular sensitivity expressions utilizing the tolerance of the elements as the element variation information, the same procedure as used in section 3.3.2 is employed. Expressing the variational vector as given in (3-57), the following radial and angular sensitivity equations result:

For matrix A_s: Substitution of the transpose of (3-59) into (3-77) yields

Matrix A_s:

$$\begin{bmatrix} \Delta\rho \\ \Delta\Theta \end{bmatrix} = \begin{bmatrix} \rho\cos^2\Theta & \rho\sin^2\Theta \\ -\cos\Theta\sin\Theta & \cos\Theta\sin\Theta \end{bmatrix} \begin{bmatrix} \Delta\alpha_N \\ \Delta\beta_N \end{bmatrix} \quad .(3\text{-}80)$$

For matrix A_c: Substitution of the transpose of (3-61) into (3-76) results in

Matrix A_c:

$$\begin{bmatrix} \Delta\rho \\ \Delta\theta \end{bmatrix} = \begin{bmatrix} \frac{\rho}{2} & 0 \\ \frac{1}{2\tan\theta} & \frac{-1}{\tan\theta} \end{bmatrix} \begin{bmatrix} \Delta d_{1N} \\ \Delta d_{2N} \end{bmatrix} \quad .(3\text{-}81)$$

For matrix A_o: Substitution of the transpose of (3-63) into (3-78) yields

Matrix A_o:

$$\begin{bmatrix} \Delta\rho \\ \Delta\theta \end{bmatrix} = \begin{bmatrix} \frac{\rho^2-2\rho\cos\theta+1}{2\rho} & \frac{2\rho\cos\theta-1}{2\rho} \\ \frac{\rho^2-2\rho\cos\theta+1}{2\rho^2\tan\theta} & \frac{(2\rho\cos\theta-1)(\cos\theta-\rho)}{2\rho^2\sin\theta} \end{bmatrix} \begin{bmatrix} \Delta x_N \\ \Delta y_N \end{bmatrix} \quad .(3\text{-}82)$$

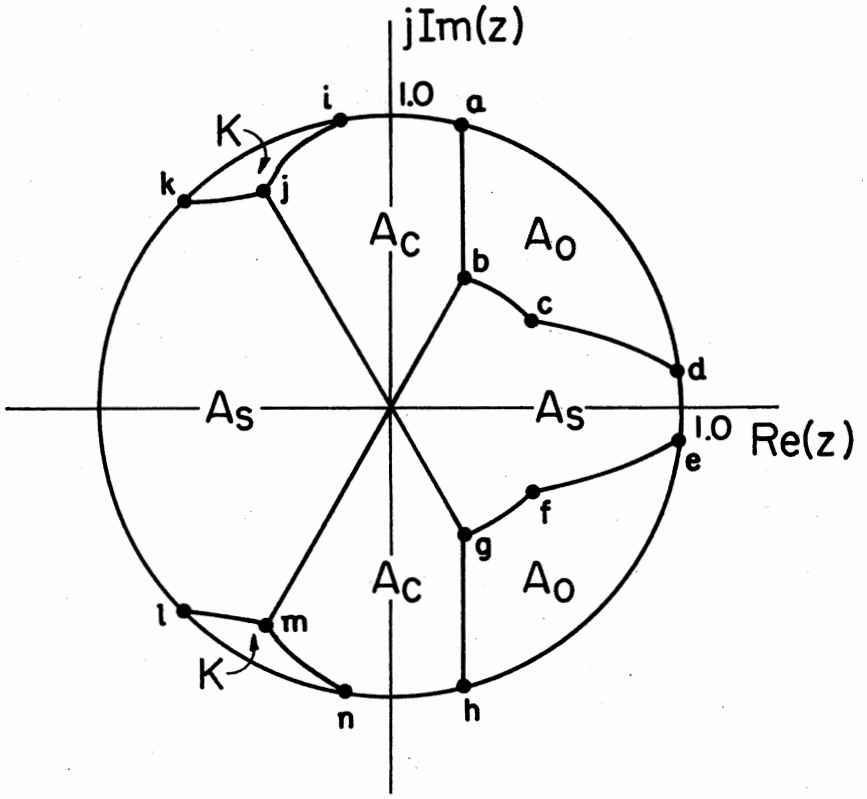
For matrix K: Substitution of the transpose of (3-65) into (3-79) results in

Matrix K:

$$\begin{bmatrix} \Delta\rho \\ \Delta\theta \end{bmatrix} = \begin{bmatrix} \frac{\rho^2-1}{2\rho} & \frac{\rho^2-1}{2\rho} \\ \frac{-\cos\theta-\rho\sqrt{1-\rho^2\sin^2\theta}}{2\rho^2\sin\theta} & \frac{-\cos\theta+\rho\sqrt{1-\rho^2\sin^2\theta}}{2\rho^2\sin^2\theta} \end{bmatrix} \begin{bmatrix} \Delta k_{1N} \\ \Delta k_{2N} \end{bmatrix} \quad .(3\text{-}83)$$

3.4.3 Radial and Angular Sensitivity Comparison

Using the radial and angular sensitivity expressions just presented, the sensitivity characteristics of the second order matrices under consideration can be compared in the same manner as was presented for the total magnitude of the eigenvalue sensitivity in section 3.3.3. Figure 7 and Figure 8 show a summary of the comparison for the minimum, worst case sensitivity for angular and radial pole variations, respectively. In



COORDINATES

a - (0.242, 0.970)
b - (0.250, 0.433)
c - (0.493, 0.296)
d - (0.992, 0.122)
e - (0.992, -0.122)
f - (0.493, -0.296)
g - (0.250, -0.433)
h - (0.242, -0.970)
i - (-0.122, 0.993)
j - (-0.438, 0.758)
k - (-0.707, 0.707)
l - (-0.707, -0.707)
m - (-0.438, -0.758)
n - (-0.122, -0.993)

Figure 7. Minimum Worst Case $\Delta\Theta$ (Normalized Variations)

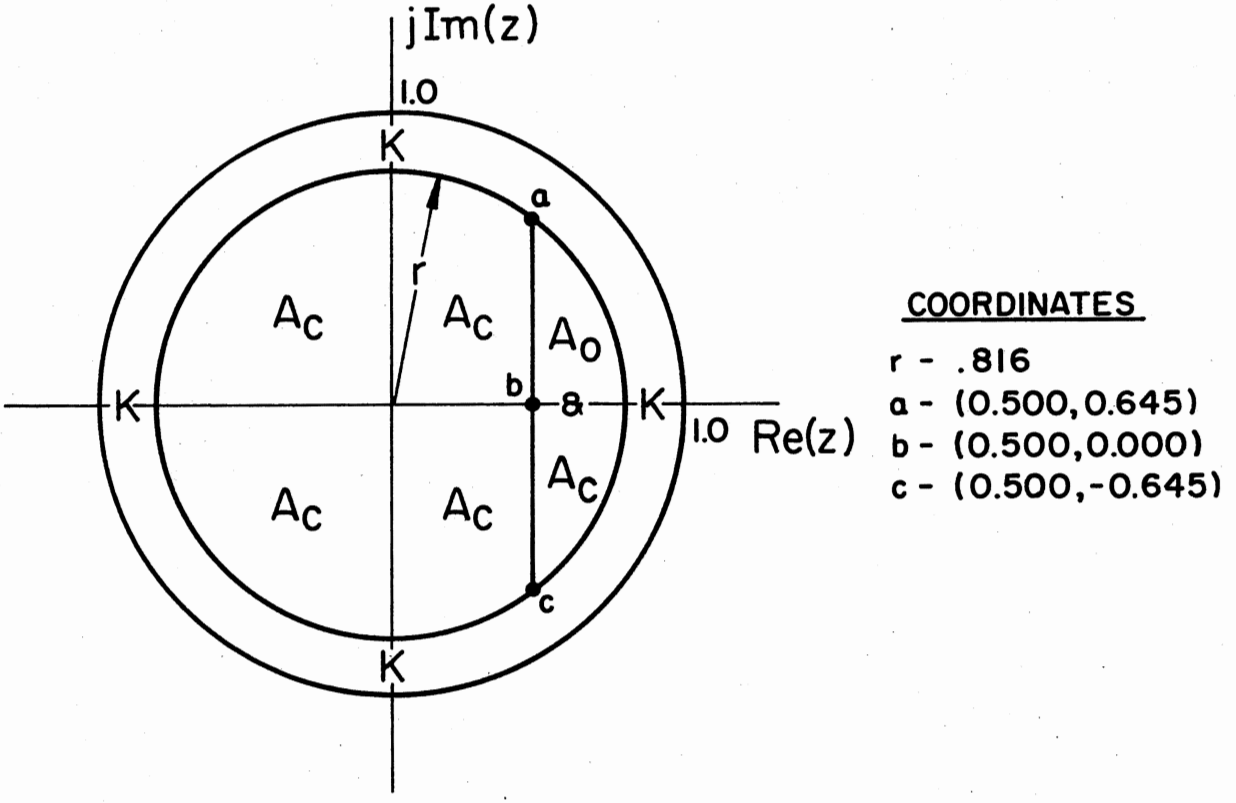


Figure 8. Minimum Worst Case $\Delta\rho$ (Normalized Variations)

these comparisons, the normalized element expressions of (3-80), (3-81), (3-82), and (3-83) are used. The element tolerances have been assumed to be equal in magnitude and the direction has been chosen to achieve the maximum sensitivity. The regions indicate areas of pole locations in which the matrices offer the minimum, worst case radial or angular sensitivity.

In Figure 7, a good property of the matrix A_s that has been commented on by Rader and Gold [30] is shown. As the sampling rate of a digital filter is increased, the poles tend to move towards z=1. As shown in Figure 7, under these conditions the matrix A_s offers the minimum angular sensitivity to element variations.

In section 2.2.2 a qualitative analysis of the sensitivity properties of the matrix K for pole locations near the unit circle was made based on the dynamic relationship between the matrix elements. The radial sensitivity expression in (3-83) shows the quantitative confirmation of that analysis. As shown in Figure 8, the matrix K offers the minimum radial sensitivity to element variations, among all the matrices, for pole locations near the unit circle. Most of the pole variation is due to changes in the angular location of the pole. For pole locations near the unit circle, insensitivity in the radial direction has a very good effect on the stability property of the filter due to parameter variations. The introduction of the matrix K has resulted in an inapproved realization for critical pole locations near the unit circle.

3.4.4 Relationship Between Magnitude and Radial, Angular Sensitivity

Since radial and angular variations are components of the change in eigenvalue location, there must be a relationship between the radial and

angular sensitivity expressions and the expression for the magnitude of the eigenvalue sensitivity. In this section, it is shown that the sensitivity matrix $\hat{A}$ of section 3.3 can be obtained from the radial and angular sensitivity expressions.

For complex $\lambda=\rho e^{\pm j\Theta}$ subjected to small changes in ρ and Θ due to matrix element variations, the change in eigenvalue location is given by by

$$\Delta\lambda = \frac{\partial\lambda}{\partial\rho}\Delta\rho + \frac{\partial\lambda}{\partial\Theta}\Delta\Theta \qquad ,(3\text{-}84)$$

which results in

$$\Delta\lambda = \Delta\rho e^{\pm j\Theta} \pm j\rho\Delta\Theta e^{\pm j\Theta} \qquad ,(3\text{-}85)$$

or

$$\Delta\lambda = (\Delta\rho \pm j\rho\Delta\Theta)e^{\pm j\Theta} \qquad .(3\text{-}86)$$

The square of the magnitude of (3-86) is given by

$$|\Delta\lambda|^2 = \Delta^2\rho + \rho^2\Delta^2\Theta \qquad ,(3\text{-}87)$$

which can be written as

$$|\Delta\lambda|^2 = [\Delta\rho \ \ \Delta\Theta]\begin{bmatrix} 1 & 0 \\ 0 & \rho^2 \end{bmatrix}\begin{bmatrix} \Delta\rho \\ \Delta\Theta \end{bmatrix} \qquad .(3\text{-}88)$$

From (3-15) the square of the magnitude of $\Delta\lambda$ can also be written as

$$|\Delta\lambda|^2 = \frac{\alpha^T\hat{A}\alpha}{\prod_{\substack{i=1 \\ i\neq k}}^{n} |\lambda_k-\lambda_i|^2} \qquad .(3\text{-}89)$$

For radial and angular sensitivity expressions given in matrix format as

$$\Delta_{\rho,\Theta} = F\alpha \qquad ,(3\text{-}90)$$

it is clear from (3-88) and (3-89) that

$$\frac{\hat{A}}{\prod_{\substack{i=1 \\ i\neq k}}^{n} |\lambda_k-\lambda_i|^2} = F^T\hat{D}F \qquad ,(3\text{-}91)$$

where

$$\hat{D} = \begin{bmatrix} 1 & 0 \\ 0 & \rho^2 \end{bmatrix} \qquad .(3\text{-}92)$$

Equation (3-91) gives the relationship between the sensitivity matrix $\hat{A}$ and the radial and angular sensitivity expressions.

For an example consider the matrix A_c. From (3-76) and (3-90)

$$F = \begin{bmatrix} \frac{-1}{2\rho} & 0 \\ \frac{-1}{2\rho^2\tan\Theta} & \frac{-1}{2\rho\sin\Theta} \end{bmatrix} \qquad .(3\text{-}93)$$

Substituting (3-93) into (3-91) yields

$$\hat{A} = \begin{bmatrix} 1 & \rho\cos\Theta \\ \rho\cos\Theta & \rho^2 \end{bmatrix} \qquad ,(3\text{-}94)$$

which is the same as the sensitivity matrix A' defined in (3-28) for the companion matrix A_c.

A geometric analysis of a change in eigenvalue location clarifies what is given in (3-87). Figure 9 depicts what is happening when an eigenvalue changes due to small variations $\Delta\rho$ and $\Delta\Theta$. Due to small parameter variations, an eigenvalue $\lambda=\rho e^{\pm j\Theta}$ changes to

$$\lambda+\Delta\lambda=(\rho+\Delta\rho)e^{\pm j(\Theta+\Delta\Theta)} \qquad .(3\text{-}95)$$

In Figure 9 the old location is depicted by the vector OA and the new

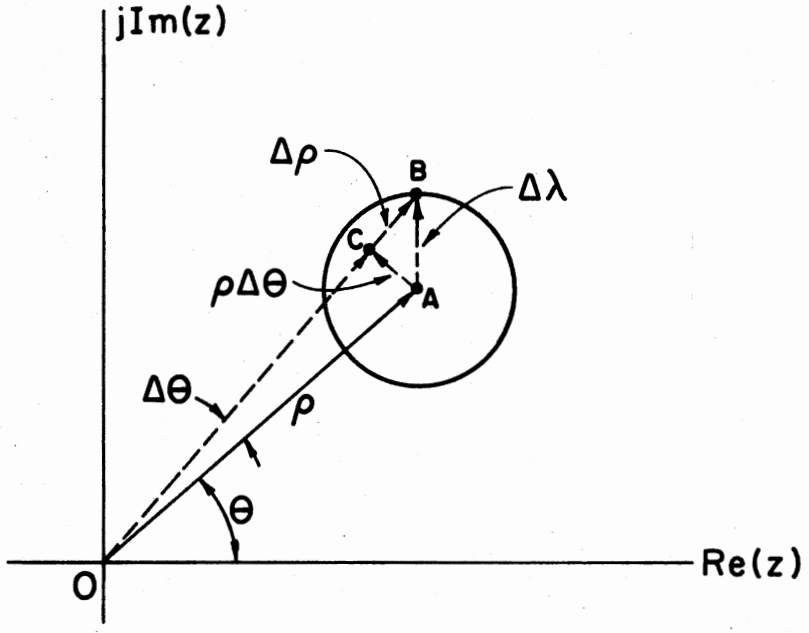


Figure 9. Eigenvalue Change Geometry

location by the vector OB. For small variations it is assumed that the vector AC with magnitude $\rho\Delta\Theta$ is perpendicular to the vector OB. Then

$$|AB|^2 = |AC|^2 + |CB|^2 \tag{3-96}$$

or $|\Delta\lambda|^2=\Delta^2\rho+\rho^2\Delta^2\Theta$, as was given in (3-87).

3.5 General Extension to n-th Order Systems

Although the sensitivity analysis technique presented in section 3.3 is applied, in this thesis, to the analysis and comparison of second order matrices with two variational elements used to realize digital filter poles, the procedure is general and may be applied to larger systems. Any linear analog or digital system that can be represented by a state-model can utilize the technique to determine the effects of parameter variations on the poles and zeroes of the system. For any system specified by a state-model, there exists an equivalent input/output state-model with a companion matrix as the system matrix [11] [12]. Singer [9] has presented the sensitivity matrix A for an n-th order companion matrix. It is given by

$$A' = \begin{bmatrix} 1 & \rho\cos\Theta & \rho^2\cos 2\Theta & \cdots & \rho^{n-1}\cos(n-1)\Theta \\ \rho\cos\Theta & \rho^2 & \rho^3\cos\Theta & \cdots & \rho^n\cos(n-2)\Theta \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \rho^{n-1}\cos(n-1)\Theta & \rho^n\cos(n-2)\Theta & \rho^{n+1}\cos(n-3)\Theta & \cdots & \rho^{2n-2} \end{bmatrix} \tag{3-97}$$

With this matrix, the sensitivity matrix $\hat{A}$ for any other n-th order system matrix can be obtained from (3-27) which gives $\hat{\hat{A}}$ as

$$\hat{A} = M^T A' M \qquad .(3\text{-}98)$$

For any n-th order system, the system matrix must have at least n variable elements in order to realize all possible poles of the system. In general, it is clear that the fewer the number of variable elements, the better the representation in terms of implementation. However, in a completely general case all matrix elements can be variable. Then if m is defined as the number of variational elements, it can assume values in the range $n \leq m \leq n^2$. The matrix M in (3-98) would then have dimension nxm and Δf, the variational vector, would be an mx1 vector. As an example, consider the more general case of a second order matrix with all four elements subject to variation. The dimensions of M and Δf would then be 2x4 and 4x1, respectively. Note that the dimension of the sensitivity matrix is always equal to the number of variational elements. As m changes, that change is reflected in the dimension of the matrix M since it is associated with the variational vector Δf.

Singer's definition of sensitivity as given in (3-15) applies to the n-th order case. As shown previously, when the limit is applied, (3-15) results in

$$|S(\lambda_k, F)| = \frac{\sup \sqrt{\alpha^T \hat{A} \alpha}}{\prod_{\substack{i=1 \\ i \neq k}}^{n} |\lambda_k - \lambda_i|} \qquad ,(3\text{-}99)$$

where the elements of the variational vector, $\alpha^T = [\alpha_1 \alpha_2 \ldots \alpha_m]$, have been normalized to $\alpha_j = \pm 1$. A problem that is encountered early in the application of (3-99) is the determination of sup $\alpha^T \hat{A} \alpha$ where $\alpha_j = \pm 1$. For n=2 this is a simple problem solved by inspection. When the general n-th order problem with m variational elements is considered, where n is

large, the determination of the vector α to maximize $\alpha^T\hat{A}\alpha$ is more complex and requires a systematic procedure.

The problem

$$\max \alpha^T\hat{A}\alpha \qquad (3\text{-}100)$$
$$\alpha_j = \pm 1 \quad j = 1, \ldots, m$$

is a member of a general class of integer nonlinear programming problems (INLP). The general INLP is characterized by [42]

$$\max f(x)$$
$$g_i(x) \leq 0 \quad i = 1, \ldots, p \qquad ,(3\text{-}101)$$
$$x = \text{integer}$$

where $f(x)$ and $g_i(x)$ are real-valued functions. A subset of the problems defined in (3-101) that are more closely related to that of (3-100) are those problems when x takes on binary values 0,1. In fact (3-100) can be translated to a binary problem by defining y such that

$$y_j = \frac{\alpha_j+1}{2} \qquad j = 1,\ldots,m \qquad .(3\text{-}102)$$

Then (3-99) is given by

$$\max[y^T\hat{A}y+\hat{B}y+C]$$
$$y_j = 0,1 \quad j = 1, \ldots, m \qquad .(3\text{-}103)$$

where $\hat{B}$ is a row vector and C is a scalar. Since the function of interest, $\alpha^TA\alpha$, is quadratic in form, a solution to (3-100), translated to (3-103) through (3-102), can be obtained by employing techniques developed for solving quadratic binary programming problems.

Much of the work done in developing solution techniques for these types of problems comes from the disciplines of Operations Research and

Management Science. Under these disciplines the quadratic binary problem may arise in many situations including the classical Traveling Salesman Problem, the Candidates Problem, and Capital-Budgeting Problems [43] [44].

A procedure that has been used to solve (3-103) is known as pseudo-boolean programming [45]. A pseudo-boolean function is a real-valued function of a binary n-vector. Any pseudo-boolean function can be represented by a polynomial to which an enumerative algorithm is applied to eliminate one variable, y_j, at a time until a trivial problem in one variable is solved. The eliminated variables are then obtained from recursive relationships involving the solved variable. Once these binary variables y_j are solved for, the elements α_j in (3-100) are obtained from (3-102) as

$$\alpha_j = 2y_j - 1 \qquad j=1, \ldots, m \qquad ,(3\text{-}104)$$

and the sensitivity of the n-th order matrix F with m variational elements is obtained from (3-100) and (3-99).

To obtain expressions for the magnitude of the eigenvalue sensitivity in terms of the tolerance of the variable elements the same procedure is used as in section 3.3.2. As discussed previously, for an n-th order matrix with m variational elements the vector α has dimension mxl where $n \leq m \leq n^2$. Therefore the matrix D as given in (3-57) has dimension mxm which, when substituted into (3-58) gives the desired expression.

The technique used in section 3.4 of deriving radial and angular sensitivity expressions for second order matrices with two variational elements from the corresponding expressions for a second order companion matrix is also applicable to the general case of an n-th order matrix

with m variational elements. For such a general case, the radial and angular expressions for the n-th order companion matrix may be written as

$$\begin{bmatrix}\Delta\rho\\ \Delta\Theta\end{bmatrix} = \begin{bmatrix}S_\rho\\ S_\Theta\end{bmatrix}\Delta d \qquad ,(3\text{-}105)$$

where Δd is an nxl vector composed of the variations of the companion matrix elements and where S_ρ and S_Θ are lxn radial and angular sensitivity vectors, respectively, that have been presented by Mitra and Sherwood [29]. For the second order case, S_ρ and S_Θ are given in (3-76) as

$$S_\rho = \left[-\frac{1}{2\rho} \quad 0\right] \qquad (3\text{-}106)$$

$$S_\Theta = \left[\frac{-1}{2\rho^2\tan\Theta} \quad \frac{-1}{2\rho\sin\Theta}\right] \qquad .(3\text{-}107)$$

From (3-26) and (3-105) the radial and angular sensitivity expressions for an n-th order matrix F with m variation elements is given as

$$\begin{bmatrix}\Delta\rho\\ \Delta\Theta\end{bmatrix} = \begin{bmatrix}S_\rho\\ S_\Theta\end{bmatrix} M\Delta f \qquad ,(3\text{-}108)$$

where S_ρ and S_Θ are the lxn sensitivity vectors for the companion matrix and M is an nxm matrix relating the mxl variational vector, Δf, of matrix F to the nxl variational vector, Δd, of the companion matrix. Expressions in terms of tolerance information for Δf are obtained in the same manner as was discussed for the sensitivity magnitude expressions.

3.6 Summary

The newly introduced second order system matrix K is compared to other second order system matrices on the basis of the eigenvalue

sensitivity of the matrices. A relationship between the sensitivity matrix A' of the companion matrix A_c and the sensitivity matrices $\hat{A}$ for the other system matrices is shown to exist. Based on this relationship, a method for determining the eigenvalue sensitivity of system matrices from the sensitivity expressions for an equivalent companion matrix is presented. Expressions for the magnitude of the eigenvalue sensitivity and the corresponding radial and angular components of that sensitivity are derived for absolute and normalized matrix element variations. It is shown that the sensitivity matrix of a given system matrix can be obtained from the radial and angular sensitivity expressions for that matrix. Minimum pole sensitivity regions within the unit circle of the z-plane are shown for each matrix. The system matrix K is shown to exhibit very good sensitivity properties for critical pole locations near the unit circle where the stability of a filter subject to element variation is of great concern. Extension to the general case of an n-th order matrix with m variational elements is discussed and the proposed procedure of determining eigenvalue sensitivity of system matrices from the sensitivity expressions of an equivalent companion matrix is shown to apply.

CHAPTER IV

COEFFICIENT WORDLENGTH REQUIREMENTS, REALIZABLE POLE-GRIDS, AND OTHER DATA ON SECOND ORDER SYSTEM MATRICES

4.1 Introduction

There are other properties of a matrix besides eigenvalue sensitivity that can be examined in the process of selecting a system matrix for digital filter applications. One criteria that is useful for determining hardware requirements for the implementation of a filter is the number of bits required for each coefficient in order to insure acceptable performance. Closely associated with the coefficient wordlength is the location and density of the discrete pole grids which can be realized with a given number of bits.

Another effect of finite wordlengths in fixed point recursive digital filters is the introduction of overflow oscillations. The tendency of a filter to sustain overflow oscillations depends on the manner in which it is realized. In state-model realizations it has been shown that certain kinds of system matrices will not sustain oscillations regardless of pole position [7]. This lessens the requirement of using stringent signal scaling, the use of which usually results in higher roundoff noise, in order to avoid the effects of such oscillations.

The realization of the filter also determines its roundoff noise

properties [34]. As explained in Chapter I, roundoff noise results from the quantization of the results of multiplications in the filter and is therefore an unavoidable source of error. Jackson [28] has claimed that realizations with good coefficient sensitivity properties also have good roundoff noise properties.

In this chapter the second order system matrices under consideration are compared with regard to the properties mentioned above. Expressions for determining the coefficient wordlength necessary to keep pole variations within prescribed bounds are developed for each system matrix. For a given set of bounds on pole movements, tabular data representing the required wordlengths necessary for the fractional part of each matrix element is presented for comparison. Realizable pole grids are shown for each of the matrices when the coefficient wordlength is restricted to five bits, including the sign bit and one magnitude bit. From these pole grids, the effects of increasing the size of the wordlength can be evaluated.

A cursory examination of the tendency of each system matrix to sustain overflow limit cycles is made by applying the criteria of Mills et al [7] to each matrix. Using the techniques described by Gold and Rader [25], the mean squared value of the roundoff noise for the realization of a common transfer function by each system matrix is computed and compared.

For each of the properties considered in these comparisons, results have been given in the literature for the matrices A_s and A_c. To present knowledge, none have been presented for A_o and certainly not for the matrix K.

4.2 Coefficient Wordlength Requirements

While an analysis of various state-model formulations for digital filters based on eigenvalue, or pole, sensitivity provides a useful basis for comparing system matrices, the ultimate criteria for the realization of the filter might be the number of bits required for each coefficient in order to insure acceptable performance. Typically a digital filter will be realized in one of two ways. Either the filter will be implemented on a computer or a minicomputer with fixed wordlength, or a special purpose hardware will be built with the possible advantage of using different wordlengths for each coefficient with the intention of obtaining lower cost and/or higher speed. For filters implemented on computers, determining the coefficient wordlengths for a particular realization will indicate whether or not the poles can be realized within a given accuracy. For special purpose hardware implementation, the wordlengths necessary for a given pole location accuracy directly determines the hardware requirements.

The method used for determining coefficient wordlength has been presented by Mitra and Sherwood [29] for the general case of an n-th order matrix with m variational elements. For completeness, a brief discussion of this general case will be presented later. For second order matrices with two variational elements, the procedure is more direct since it allows the matrix element variations to be directly solved for in terms of the maximum allowable pole movements.

The minimum number of bits required for a coefficient γ is determined by the maximum quantization step size $\Delta\gamma$ allowable to insure specified performance, which, in this case, is the realization of a pole

location within a specified distance of the ideal location that could be realized with infinite precision. The radial and angular sensitivity expressions presented in section 3.4 relate complex pole movements to small variations in the matrix elements of the second order matrices of interest. If the radial and angular changes are given as the maximum allowable changes in pole location, then the corresponding element variations constitute the quantization step sizes for those elements. Therefore, solving the sensitivity expressions of section 3.4.1 for the matrix element variations in terms of maximum allowable changes in the radial and angular locations of the pole results in expressions for the quantization step size of each matrix element.

For the companion matrix A_c, solving (3-76) for the parameter changes yields

$$\begin{bmatrix} \Delta d_1 \\ \Delta d_2 \end{bmatrix} = \begin{bmatrix} -2\rho & 0 \\ 2\cos\Theta & -2\rho\sin\Theta \end{bmatrix} \begin{bmatrix} \Delta\rho \\ \Delta\rho \end{bmatrix} \qquad .(4\text{-}1)$$

It is clear from (4-1) that the element variations are a function of pole location and/or the direction in which the pole moves. For a worst case solution, the wordlength should be based on the minimum quantization step size encountered for a given pole location regardless of the direction of pole movement. For the companion matrix, the worst case parameter quantization step sizes are, from (4-1),

A_c:

$$\min|\Delta d_1| = |2\rho\Delta\rho| \qquad (4\text{-}2)$$

$$\min|\Delta d_2| = ||2\cos\Theta\Delta\rho| - |2\rho\sin\Theta\Delta\Theta|| \qquad .(4\text{-}3)$$

For the matrix A_s, solving (3-77) for $\Delta\alpha$ and $\Delta\beta$ yields

$$\begin{bmatrix}\Delta\alpha\\ \Delta\beta\end{bmatrix} = \begin{bmatrix}\cos\theta & -\rho\sin\theta\\ \sin\theta & \rho\cos\theta\end{bmatrix}\begin{bmatrix}\Delta\rho\\ \Delta\theta\end{bmatrix} ,$$

which results in

A_s:

$$\min|\Delta\alpha| = ||\cos\theta\Delta\rho|-|\rho\sin\theta\Delta\theta|| \quad (4\text{-}4)$$

$$\min|\Delta\beta| = ||\sin\theta\Delta\rho|-|\rho\cos\theta\Delta\theta|| \quad .(4\text{-}5)$$

For the matrix A_o, solving (3-78) for Δx and Δy yields

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = \begin{bmatrix}2(\cos\theta-\rho) & -2\rho\sin\theta\\ 2\cos\theta & -2\rho\sin\theta\end{bmatrix}\begin{bmatrix}\Delta\rho\\ \Delta\theta\end{bmatrix} ,$$

which results in

A_o:

$$\min|\Delta x| = ||2(\cos\theta-\rho)\Delta\rho|-|2\rho\sin\theta\Delta\theta|| \quad (4\text{-}6)$$

$$\min|\Delta y = ||2\cos\theta\Delta\rho|-|2\rho\sin\theta\Delta\theta|| \quad .(4\text{-}7)$$

For the matrix K, solving (3-79) for Δk_1 and Δk_2 results in

$$\begin{bmatrix}\Delta k_1\\ \Delta k_2\end{bmatrix} = \frac{1}{k_1-k_2}\begin{bmatrix}2(k_1\cos\theta-\rho) & -2k_2\rho\sin\theta\\ 2(\rho-k_2\cos\theta) & 2k_2\rho\sin\theta\end{bmatrix}\begin{bmatrix}\Delta\rho\\ \Delta\theta\end{bmatrix} ,$$

which results in

K:

$$\min|\Delta k_1| = \frac{1}{k_1-k_2}||2(k_1\cos\theta-\rho)\Delta\rho|-|2k_2\rho\sin\theta\Delta\theta|| \quad (4\text{-}8)$$

$$\min|\Delta k_2| = \frac{1}{k_1-k_2}||2(\rho-k_2\cos\theta)\Delta\rho|-|2k_2\rho\sin\theta\Delta\theta|| \quad .(4\text{-}9)$$

If pole movements due to coefficient quantization are specified to remain within the limits $\Delta\rho=\Delta\rho_{max}$ and $\Delta\Theta=\Delta\Theta_{max}$ radians, then (4-2)-(4-9) give the necessary quantization levels for each matrix element that will insure such a constraint under worst case conditions of matrix element variations. An important point to remember is that since the original sensitivity expressions assumed small parameter variations, the matrix element variations resulting from employing the quantization levels given by (4-2)-(4-9) must be small enough so that second and higher-order factors in the relationship between pole location and parameter variation can be ignored.

When the filter is implemented, the coefficients will be stored in finite-length binary form. If n is the number of bits to the right of the binary point, then the coefficient quantization step size $\Delta\gamma$ is given by

$$\Delta\gamma = 2^{-n} \qquad .(4\text{-}10)$$

Solving for n results in

$$n = [-\log_2\Delta\gamma] \qquad ,(4\text{-}11)$$

if coefficient quantization is done by rounding, and

$$n = [-\log_2\Delta\gamma]+1 \qquad ,(4\text{-}12)$$

if coefficient quantization is done by truncation, and where [x] stands for the largest integer in x.

Using equations (4-2)-(4-9) and (4-11), the wordlength required to the right of the binary point in order to constrain movements of complex poles within the limits $\Delta\rho_{max}$=.001 and $\Delta\Theta_{max}$=.001 radian were computed

for each matrix element of the second order matrices of interest. The values of $\Delta\rho_{max}$ and $\Delta\Theta_{max}$ that were used were chosen in order to insure that the matrix element variations were small with respect to the nominal values of the elements for each pole location and also to provide a common basis for comparing the requirements of the matrices. The results are given in Table II–Table XI for various pole locations.

The data presented in the tables show general trends of wordlength requirements as pole locations vary throughout the unit circle of the z-plane. For example, as the radius of the pole location becomes smaller, the matrix K requires an increased wordlength to maintain pole variations within the prescribed limits. This is a characteristic of the other matrices also but not to the same degree as it is for K. Also apparent is the cyclic nature of the wordlength requirements of a given matrix as the pole angle varies on a given radius. This is true of all the matrices except A_0. Recall from Figure 3 in Chapter II that, as ρ approaches one, one or the other of the parameters k_1 and k_2 of the matrix K maintains a constant value close to zero for particular angular regions of pole locations. This cyclic nature is also shown in the tables. The element x of the matrix A_0 is shown to require dramatically increasing wordlength requirements as Θ approaches ninety degrees. The dash entry at that point indicates a very high wordlength requirement. Examination of (4-6) shows this to be true for any equal bounds placed on $\Delta\rho$ and $\Delta\Theta$. Element d_1 of the matrix A_c shows a constant wordlength requirement for a given pole radius. Since d_1 determines the pole radius and not the angle this is not surprising.

The extent to which the wordlength requirements of the elements of a particular matrix change as Θ varies is not indicated in the tables.

TABLE II

WORDLENGTH WHEN ρ=.99

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	10	8	8	16	8	10	9
9	10	10	8	9	11	9	10	9
18	10	10	8	9	10	9	11	9
27	11	11	8	10	10	10	12	9
36	12	12	8	11	10	11	10	9
45	17	17	8	16	10	16	9	9
54	12	12	8	11	10	11	9	9
63	11	11	8	10	10	10	10	8
72	10	10	8	9	10	9	10	8
81	10	10	8	9	11	9	8	7
90	9	9	8	8	--	8	7	7
99	10	10	8	9	11	9	7	8
108	10	10	8	9	10	9	8	10
117	11	11	8	10	9	10	8	10
126	12	12	8	11	9	11	9	9
135	17	17	8	16	8	16	9	9
144	12	12	8	11	8	11	9	10
153	11	11	8	10	8	10	9	12
162	10	10	8	9	8	9	9	11
171	10	10	8	9	8	9	9	10
179	9	10	8	8	7	8	9	10

TABLE III

WORDLENGTH WHEN ρ=.90

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	10	9	8	12	8	10	9
9	10	10	9	9	13	9	10	9
18	10	10	9	9	11	9	11	9
27	11	11	9	10	10	10	13	9
36	11	12	9	10	10	10	10	9
45	13	13	9	12	10	12	9	9
54	12	11	9	11	10	11	9	9
63	11	11	9	10	10	10	10	9
72	10	10	9	9	10	9	14	9
81	10	10	9	9	11	9	10	9
90	10	9	9	9	--	9	9	9
99	10	10	9	9	11	9	9	10
108	10	10	9	9	10	9	9	14
117	11	11	9	10	9	10	9	10
126	12	11	9	11	9	11	9	9
135	13	13	9	12	9	12	9	9
144	11	12	9	10	8	10	9	10
153	11	11	9	10	8	10	9	13
162	10	10	9	9	8	9	9	11
171	10	10	9	9	8	9	9	10
179	9	10	9	8	8	8	9	10

TABLE IV

WORDLENGTH WHEN ρ=.80

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	10	9	8	11	8	9	9
9	10	10	9	9	12	9	10	9
18	10	11	9	9	12	9	11	9
27	10	11	9	9	10	9	13	9
36	11	14	9	10	10	10	11	9
45	12	12	9	11	10	11	10	9
54	14	11	9	13	10	13	10	9
63	11	10	9	10	10	10	10	9
72	11	10	9	10	10	10	11	10
81	10	10	9	9	11	9	12	10
90	10	9	9	9	--	9	10	10
99	10	10	9	9	11	9	10	12
108	11	10	9	10	10	10	10	11
117	11	10	9	10	9	10	9	10
126	14	11	9	13	9	13	9	10
135	12	12	9	11	9	11	9	10
144	11	14	9	10	8	10	9	11
153	10	11	9	9	8	9	9	13
162	10	11	9	9	8	9	9	11
171	10	10	9	9	8	9	9	10
179	9	10	9	8	8	8	9	9

TABLE V

WORDLENGTH WHEN ρ=.70

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	10	9	8	10	8	9	9
9	10	10	9	9	11	9	10	10
18	10	11	9	9	13	9	10	10
27	10	12	9	9	11	9	12	10
36	11	15	9	10	10	10	12	10
45	12	12	9	11	9	11	11	10
54	15	11	9	14	10	14	10	10
63	12	10	9	11	10	11	10	10
72	11	10	9	10	10	10	11	10
81	10	10	9	9	11	9	17	10
90	10	9	9	9	--	9	11	11
99	10	10	9	9	11	9	10	17
108	11	10	9	10	10	10	10	11
117	12	10	9	11	9	11	10	10
126	15	11	9	14	9	14	10	10
135	12	12	9	11	9	11	10	11
144	11	15	9	10	8	10	10	12
153	10	12	9	9	8	9	10	12
162	10	11	9	9	8	9	10	10
171	10	10	9	9	8	9	10	10
179	9	10	9	8	8	8	9	9

TABLE VI

WORDLENGTH WHEN ρ=.60

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	10	9	8	10	8	9	9
9	10	11	9	9	10	9	10	10
18	10	11	9	9	11	9	10	10
27	10	13	9	9	14	9	11	10
36	11	13	9	10	11	10	14	10
45	11	11	9	10	10	10	11	10
54	13	11	9	12	10	12	10	10
63	13	10	9	12	10	12	10	10
72	11	10	9	10	10	10	11	10
81	11	10	9	10	11	10	13	11
90	10	9	9	9	--	9	12	12
99	11	10	9	10	11	10	11	13
108	11	10	9	10	10	10	10	11
117	13	10	9	12	9	12	10	10
126	13	11	9	12	9	12	10	10
135	11	11	9	10	9	10	10	11
144	11	13	9	10	8	10	10	14
153	10	13	9	9	8	9	10	11
162	10	11	9	9	8	9	10	10
171	10	11	9	9	8	9	10	10
179	9	10	9	8	8	8	9	9

TABLE VII

WORDLENGTH WHEN ρ=.50

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	11	9	8	9	8	9	9
9	10	11	9	9	10	9	10	10
18	10	12	9	9	10	9	10	10
27	10	16	9	9	11	9	11	10
36	10	12	9	9	15	9	12	10
45	11	11	9	10	11	10	14	10
54	12	10	9	11	10	11	11	10
63	16	10	9	15	10	15	10	10
72	12	10	9	11	10	11	11	11
81	11	10	9	10	11	10	12	11
90	11	9	9	9	--	9	13	13
99	11	10	9	10	11	10	11	12
108	12	10	9	11	10	11	11	11
117	16	10	9	15	9	15	10	10
126	12	10	9	11	9	11	10	11
135	11	11	9	10	9	10	10	14
144	10	12	9	9	8	9	10	12
153	10	16	9	9	8	9	10	11
162	10	12	9	9	8	9	10	10
171	10	11	9	9	8	9	10	10
179	9	11	9	8	8	8	9	9

TABLE VIII

WORDLENGTH WHEN ρ=.40

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	11	10	8	9	8	9	9
9	10	12	10	9	9	9	10	10
18	10	13	10	9	10	9	10	10
27	10	13	10	9	10	9	10	10
36	10	11	10	9	11	9	11	10
45	11	11	10	10	14	10	12	10
54	11	10	10	10	11	10	13	10
63	13	10	10	12	10	12	11	10
72	13	10	10	12	10	12	11	11
81	12	10	10	11	11	11	12	12
90	11	9	10	10	--	10	14	14
99	12	10	10	11	11	11	12	12
108	13	10	10	12	10	12	11	11
117	13	10	10	12	9	12	10	11
126	11	10	10	10	9	10	10	13
135	11	11	10	10	9	10	10	12
144	10	11	10	9	8	9	10	11
153	10	13	10	9	8	9	10	10
162	10	13	10	9	8	9	10	10
171	10	12	10	9	8	9	10	10
179	9	11	10	8	8	8	9	9

TABLE IX

WORDLENGTH WHEN ρ=.30

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	11	10	8	9	8	9	9
9	10	12	10	9	9	9	10	10
18	10	15	10	9	9	9	10	10
27	10	12	10	9	10	9	10	10
36	10	11	10	9	10	9	11	10
45	10	10	10	9	11	9	11	10
54	11	10	10	10	13	10	13	10
63	12	10	10	11	12	11	13	11
72	15	10	10	14	10	14	11	11
81	12	10	10	11	11	11	12	12
90	11	9	10	10	--	10	16	16
99	12	10	10	11	11	11	12	12
108	15	10	10	14	10	14	11	11
117	12	10	10	11	9	11	11	13
126	11	10	10	10	9	10	10	13
135	10	10	10	9	9	9	10	11
144	10	11	10	9	9	9	10	11
153	10	12	10	9	8	9	10	10
162	10	15	10	9	8	9	10	10
171	10	12	10	9	8	9	10	10
179	9	11	10	8	8	8	9	9

TABLE X

WORDLENGTH WHEN ρ=.20

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	12	11	8	9	8	9	9
9	10	14	11	9	9	9	10	10
18	10	13	11	9	9	9	10	10
27	10	11	11	9	9	9	10	10
36	10	11	11	9	9	9	10	10
45	10	10	11	9	10	9	11	10
54	11	10	11	10	11	10	11	10
63	11	10	11	10	12	10	13	11
72	13	10	11	12	12	12	13	11
81	14	10	11	13	11	13	12	12
90	12	9	11	11	--	11	17	17
99	14	10	11	13	11	13	12	12
108	13	10	11	12	10	12	11	13
117	11	10	11	10	10	10	11	13
126	11	10	11	10	9	10	10	11
135	10	10	11	9	9	9	10	11
144	10	11	11	9	9	9	10	10
153	10	11	11	9	8	9	10	10
162	10	13	11	9	8	9	10	10
171	10	14	11	9	8	9	10	10
179	9	12	11	8	8	8	9	9

TABLE XI

WORDLENGTH WHEN ρ=.10

	A_s		A_c		A_o		K	
Θ	N_α	N_β	N_{d_1}	N_{d_2}	N_x	N_y	N_{k_1}	N_{k_2}
1	9	13	12	8	9	8	9	9
9	10	14	12	9	9	9	10	9
18	10	12	12	9	9	9	10	10
27	10	11	12	9	9	9	10	10
36	10	10	12	9	9	9	10	10
45	10	10	12	9	9	9	10	10
54	10	10	12	9	10	9	11	10
63	11	10	12	10	10	10	11	11
72	12	10	12	11	12	11	13	11
81	14	10	12	13	13	13	14	12
90	13	9	12	12	--	12	20	20
99	14	10	12	13	11	13	12	14
108	12	10	12	11	10	11	11	13
117	11	10	12	10	10	10	11	11
126	10	10	12	9	9	9	10	11
135	10	10	12	9	9	9	10	10
144	10	10	12	9	9	9	10	10
153	10	11	12	9	9	9	10	10
162	10	12	12	9	8	9	10	10
171	10	14	12	9	8	9	9	10
179	9	13	12	8	8	8	9	9

For example, at ρ=.9 and Θ=48 degrees, the matrix A_o requires a twenty bit wordlength for the element y. The data was taken at generally nine degree increments in Θ for a given radius ρ. This wide separation made the amount of data manageable but it has tended to mask some of the dynamics of the changing requirements as a function of Θ. Nevertheless, the tables do show the general characteristics of each matrix as was intended. Specific comparisons of wordlength requirements should be obtained through the use of (4-2)-(4-9) at particular pole locations.

The one case where (4-2)-(4-9) and the data listed in the tables is misleading and does not apply in determining wordlength requirements is when the nominal values of the matrix elements necessary to realize a given pole location can be expressed as an integer power of two. In those cases there are no variations in the matrix elements due to quantization and, therefore, the poles are realized exactly. The set of poles corresponding to these conditions constitute an exactly realizable pole grid for the matrix which will be examined in section 4.3. For the moment, however, it is important to realize that, if the nominal value of the matrix element is an integer power of two, then $\Delta\rho$ and $\Delta\Theta$ are zero. Application of (4-2)-(4-9), with assumed limits on $\Delta\rho$ and $\Delta\Theta$, will result in a wordlength from (4-11) and (4-12) that is misleading. As an example, consider the complex pole pair z_1, z_2=0.5$\pm$j0.5. In polar coordinates, this pole pair is given by ρ=.707, Θ= 45 degrees. Each of the matrices A_s, A_c. and A_o can realize these poles exactly. For A_s the matrix elements are: α=0.5, β=0.5. For A_c the matrix elements are: d_1=-0.5, d_2=1.0. The matrix elements of A_o are given by: x=-0.5, y=0.0. It is clear that, with the exception of d_2 and y, only one bit to the right of the binary point will realize these elements, and therefore the poles,

exactly. The elements d_2 and y don't require any fractional bits. If, however, $\Delta\rho$ and $\Delta\Theta$ are assumed to be .001 in (4-2)-(4-7), the resultant fractional wordlengths are: $N_\alpha=N_\beta=12$; $N_{d_1}=9$, $N_{d_2}=11$; and $N_x=9$, $N_y=11$.

For an example of the utilization of the wordlength equations, the following example is presented.

4.2.1 Example

Determine the necessary wordlengths to the right of the binary point for k_1 and k_2 in order to realize the pole pair given by $\rho=0.9$, $\Theta=\pm 27$ degrees (.471238898 radians) within the limits $\Delta\rho_{max}=.001$, $\Delta\Theta_{max}=.001$ radians.

For these poles, k_1 and k_2 are given by $k_1=1.714623258$ and $k_2=-0.1108115145$, and the resultant pole polynomial is given by

$$z^2-1.603811744z+.81=0 \qquad .(4\text{-}13)$$

Applying (4-8), (4-9), and (4-11) with $\Delta\rho_{max}=\Delta\Theta_{max}=.001$ results in $N_{k_1}=13$, $N_{k_2}=9$.

The values of k_1 and k_2 realized with 13 bits and 9 bits, respectively, with rounding, are given by $k_{1R}=1.714599434$ and $k_{2R}=-0.111328125$. The resulting pole polynomial is given by

$$z^2-1.603271309z+.8091168599=0 \qquad ,(4\text{-}14)$$

which realizes poles given by

$$\rho = .8995092328$$

$$\Theta = .4708296436 \text{ radians}$$

which are clearly within the prescribed limits.

If truncation were the method of quantization, then (4-12) yields $N_{k_1}=14$, $N_{k_2}=10$. The values k_1 and k_2 realized with 14 bits and 10 bits respectively, with truncation, are: $k_{1T}=1.714600838$ and $k_{2T}=-0.1103515625$. The resulting pole polynomial is given by

$$z^2-1.604249276z+.8107911185=0 \qquad ,(4\text{-}15)$$

which realizes poles given by

$$\rho = .900439403$$

$$\Theta = .4716612977 \text{ radians}$$

which are clearly within the prescribed limits.

4.2.2 Coefficient Wordlength Requirement - General Case

The method used to determine the wordlength requirements of the various matrix elements made use of the radial and angular sensitivity expressions. Recall from (3-105) that these expressions for an n-th order matrix with m variational elements can be given as

$$\begin{bmatrix} \Delta\rho \\ \Delta\Theta \end{bmatrix} = \begin{bmatrix} S_\rho \\ S_\Theta \end{bmatrix} \Delta f \qquad ,(4\text{-}16)$$

where Δf is interpreted here as the mxl vector of the element variations and S_ρ and S_Θ are lxm radial and angular sensitivity vectors, respectively. For a second order matrix with two variational elements, S_ρ and S_Θ combine to form a square matrix which can be inverted to solve for the parameter variations directly. For the general case, however, the matrix is not square and the parameter variations cannot be obtained as directly.

Mitra and Sherwood [29] use the following approach to obtain a solution.

For each pole $z_k=\rho_k e^{\pm j\Theta_k}$, an expression given by (4-16) exists and is given by

$$\begin{bmatrix}\Delta\rho_k\\ \Delta\Theta_k\end{bmatrix}=\begin{bmatrix}S_{\rho k}\\ S_{\Theta k}\end{bmatrix}\Delta f \qquad ,(4\text{-}17)$$

where
$$S_{\rho k} = \left[\frac{\partial\rho_k}{\partial f_1} \quad \frac{\partial\rho_k}{\partial f_2} \quad \cdots \quad \frac{\partial\rho_k}{\partial f_m}\right] \qquad ,(4\text{-}18)$$

$$S_{\Theta k} = \left[\frac{\partial\Theta_k}{\partial f_1} \quad \frac{\partial\Theta_k}{\partial f_2} \quad \cdots \quad \frac{\partial\Theta_k}{\partial f_m}\right] \qquad ,(4\text{-}19)$$

and
$$\Delta f = [\Delta f_1 \quad \Delta f_2 \quad \cdots \quad \Delta f_m]^T \qquad .(4\text{-}20)$$

For a given $\Delta\rho_{kmax}$ and $\Delta\Theta_{kmax}$, the allowable pole movement is divided equally among all the variable elements. Then for each variable element f_d, d=1,2,...,m, the corresponding variation is given by

$$\Delta f_{d\rho} = \frac{1}{m}\left|\frac{\Delta\rho_{kmax}}{\frac{\partial\rho_k}{\partial f_d}}\right| \qquad d=1,2,\ldots,m \qquad (4\text{-}21)$$

for radial movements of the pole, and by

$$\Delta f_{d\Theta} = \frac{1}{m}\left|\frac{\Delta\Theta_{kmax}}{\frac{\partial\Theta_k}{\partial f_d}}\right| \qquad d=1,2,\ldots,m \qquad (4\text{-}22)$$

for angular pole movements.

For a given element f_d define

$$\Delta f_d^{\rho} \triangleq \min_k \Delta f_{d\rho} \qquad d=1,2,\ldots,m \tag{4-23}$$

$$\Delta f_d^{\Theta} \triangleq \min_k \Delta f_{d\Theta} \qquad d=1,2,\ldots,m \tag{4-24}$$

If the variational vector Δf in (4-17) was composed of elements Δf_d^{ρ} given in (4-23), then the radial variation of all poles z_k would be within the given limits. Similarly, if the variational vector Δf were composed of elements Δf_d^{Θ} given in (4-24), the angular variation of all poles z_k would be within the limits. Therefore, Δf is composed of a composite of the elements given by (4-23) and (4-24). That composite is formed by letting the elements Δf_d of the vector Δf be given by

$$\Delta f_d = \min(\Delta f_d^{\rho}, \Delta f_d^{\Theta}) \qquad d=1,2,\ldots,m \tag{4-25}$$

It is clear that with Δf composed of elements given by (4-25), the radial and angular variations in pole location will be within the limits given for all poles z_k.

Equal allocation of $\Delta\rho_{max}$ and $\Delta\Theta_{max}$ among all the variable elements, as used by Mitra and Sherwood [29], is one method of obtaining a solution for the element variations. If, however, particular matrix elements are known to be critical in determining the pole location, a more realistic approach would be to apply weights to the contribution of each element in proportion to its criticality.

4.3 Realizable Pole Grids

Since the system matrix elements are realized by binary numbers of finite lengths, there exists only a finite set of possible pole location

in the unit circle of the z-plane. Avenhaus [6] used the density of allowable pole positions in the z-plane as a measure for assessing various filter structures. He showed that the distribution of these poles, by choice of a suitable structure, may be arranged to provide a higher density of realizable locations in areas of the z-plane critical to a particular filter requirement.

A unique example of digital filter implementation requirements that affect the realizable pole locations is given by Schmidt [46]. In this example, the implementation of a high speed digital filter on an LSI chip required that the coefficients be represented in canonical signed digit (CSD) form. Coefficients represented in this manner have the least number of non-zero bits, which allows faster multiplication. By limiting the number of non-zero bits in each CSD coefficient to three, multiplier complexity, and therefore multiplier area required on the chip, was reduced. However, this resulted in the elimination of certain coefficient values and, therefore, certain pole locations that could not be realized.

In this section, the effects of coefficient quantization on the poles of the system matrices of interest are shown in pole grids that depict the actual poles that can be realized inside the unit circle of the z-plane for a given wordlength. Only complex poles are considered. The realizable pole grids of the matrices A_s and A_c have been presented previously in the literature [35] [47] [48]. The pole grids for the matrices A_o and K have not been previously presented. For this comparison, the wordlength of the matrix elements is assumed to be five bits long; consisting of three fractional bits, one magnitude bit, and a sign bit.

The realizable pole grids for the matrices A_s, A_c, A_o, and K, for quantization to three fractional bits, are given in Figure 10, Figure 11, Figure 12, and Figure 13; respectively. For A_s, the poles lie on a grid defined by the intersection of vertical lines (corresponding to quantization of $\alpha=\rho\cos\theta$) and horizontal lines (corresponding to quantization of $\beta=\rho\sin\theta$). The separation between these lines is given by the quantization increment given by $2^{-3}=0.125$. As shown in Figure 10, A_s realizes a pole grid with uniform density throughout the unit circle.

For the matrix A_c, the poles lie on a grid defined by the intersection of concentric circles (corresponding to quantization of $-d_1=\rho^2$) and vertical lines (corresponding to quantization of $d_2=2\rho\cos\theta$). The separation between the vertical lines is 0.0625 since the real part of the pole ($\rho\cos\theta$) is one-half the quantization increment of d_2. As shown in Figure 11, A_c does not realize a uniform density pole grid. The density of realizable poles increases with the magnitude, or radius, of the pole.

The realizable poles of the matrix A_o lie on a grid defined by the intersection of concentric circles (corresponding to quantization of $x-y=\rho^2$) and vertical lines (corresponding to quantization of $y+1=2\rho\cos\theta$). Functions consisting of the sum or subtraction of quantized elements assume discrete values with the same quantization interval as that of the elements. Therefore, the concentric circles defining pole locations for A_o are the same as those for A_c. The vertical lines defining pole locations are also the same as those for A_c, with the exception of pole locations in part of the second and third quadrants. With the assumed wordlength, the largest negative value of the element y, in magnitude, is given by $y=-1.875$. Since $y+1=2\rho\cos\theta$, the largest negative real part

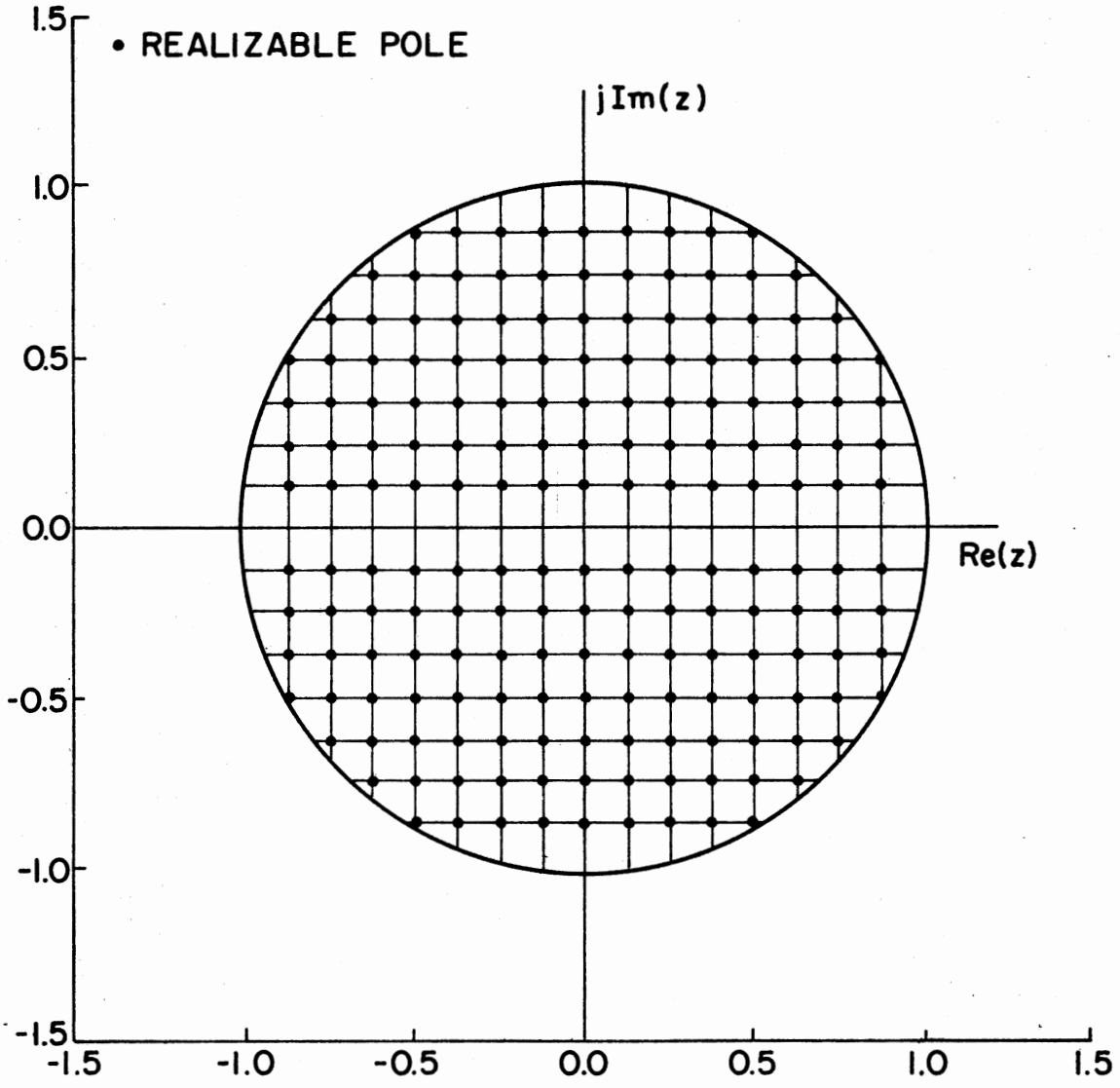


Figure 10. A_s Pole Grid - Three Bit Quantization

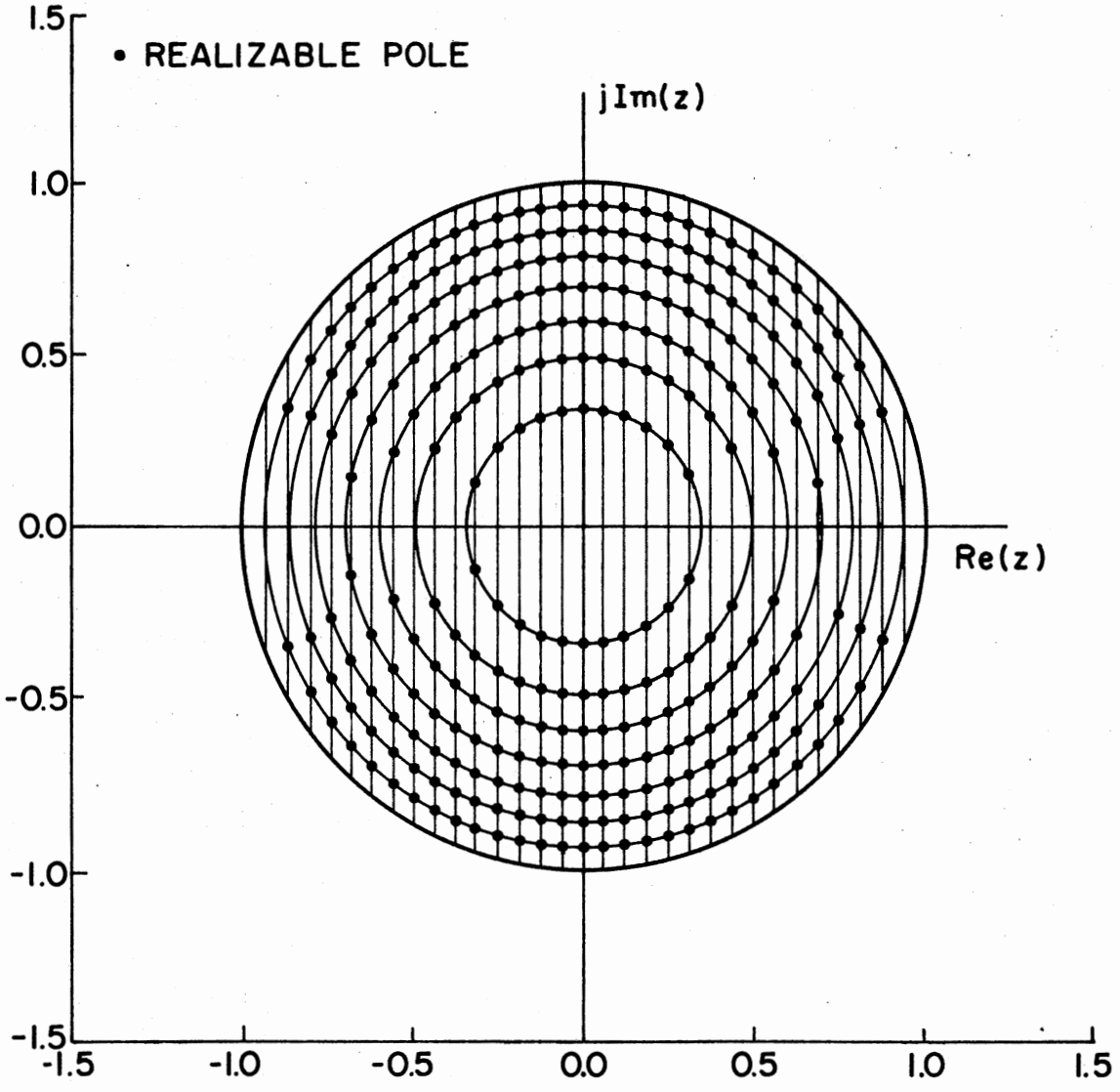


Figure 11. A_c Pole Grid - Three Bit Quantization

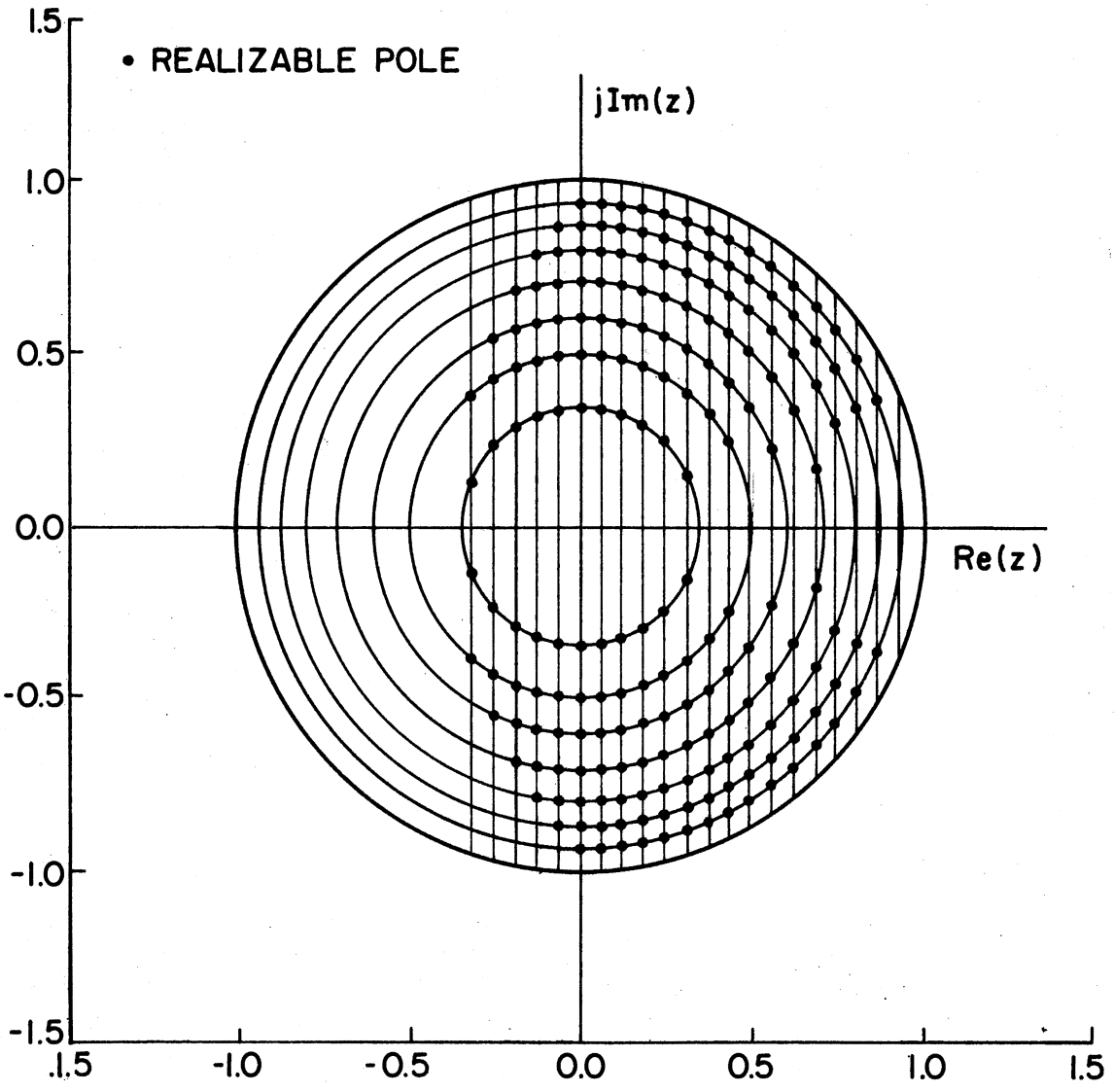


Figure 12. A_o Pole Grid - Three Bit Quantization

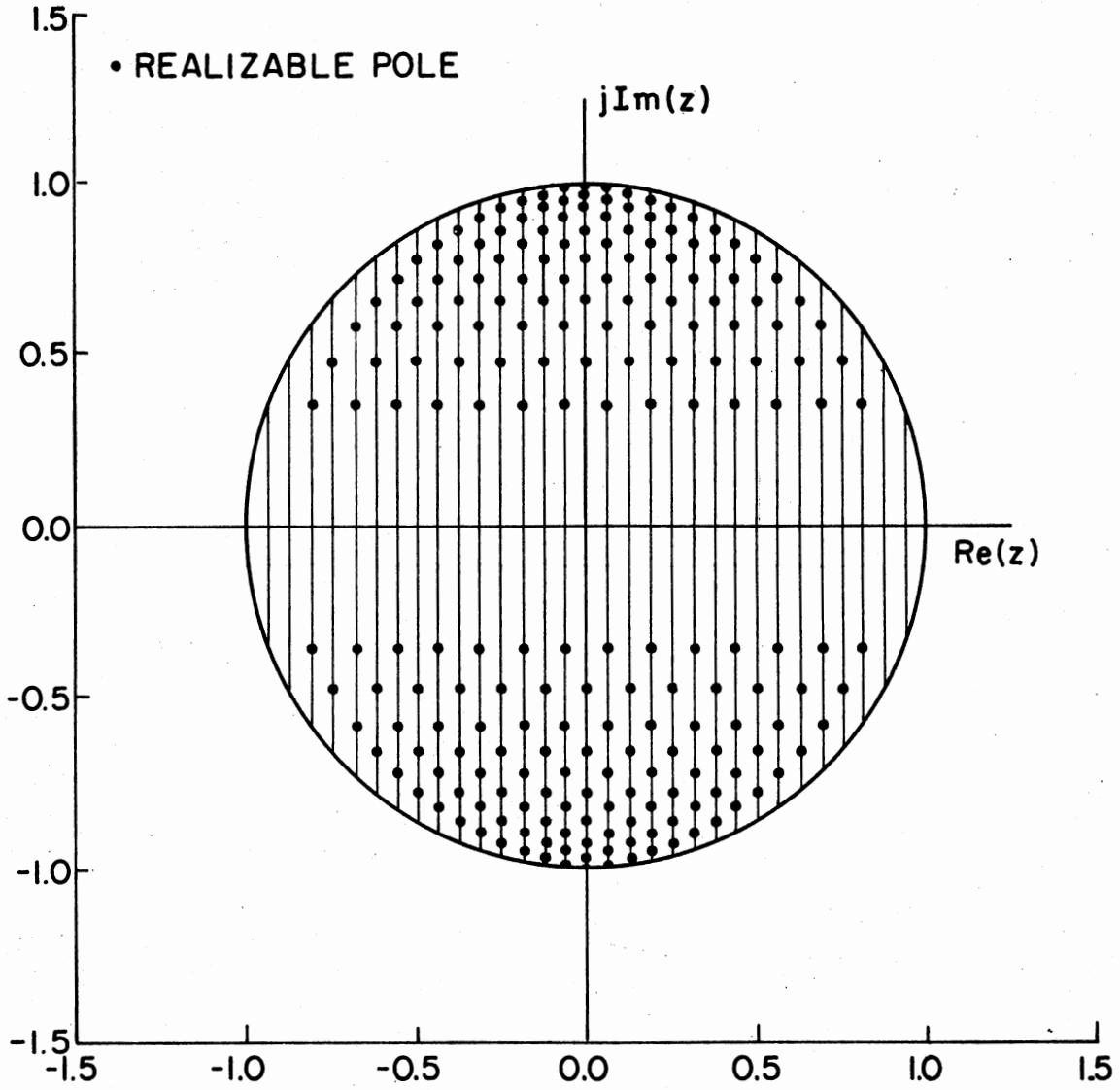


Figure 13. K Pole Grid - Three Bit Quantization

for realizable poles is $\rho\cos = -0.4375$. So, with only one numerical bit to the left of the binary point, A_o cannot realize poles throughout the second and third quadrants, as shown in Figure 12. In general, A_o will require an additional bit in wordlength, as compared to A_c, to realize poles throughout the unit circle. The emphasis, here, is that the range of values required by the element y ($-3<y<1$) in order to realize poles in all areas of the unit circle is not accommodated by the assumed wordlength. However, y is related to the element d_2 of matrix A_c by the expression $y = d_2-1$. If y is implemented in this matter, the range constraint of the assumed wordlength no longer applies and A_o can realize poles throughout the unit circle. The resulting pole grid would then be the same as that for A_c shown in Figure 11.

The realizable poles of the matrix K assume the same quantized locations for the real part of the poles as do the matrices A_c and A_o. However, the magnitudes of the pole locations are now determined by a functional relationship with the product of the quantized elements k_1 and k_2 that is given by $\rho = \sqrt{k_1 k_2 + 1}$. As shown in Figure 13, this drastically changes the nature of the pole grid. The pole grid shows an increasing density as the radius increases towards one, especially at Θ=90 degrees. The matrix K clearly has the highest pole density of all the matrices in this region of the unit circle. With increasing wordlength, this characteristic would cover larger areas of the unit circle. This is another indication of the excellent characteristics of this matrix for pole locations very close to the unit circle.

Although the assumed wordlength in this comparison is five bits, the matrix A_s only requires four bits, including the sign bit, to realize the poles shown in Figure 10. This is because the elements α and β are

less than one. If the matrix K were implemented using the form given in (2-25), the same comments would apply to it. In general, then, the matrices A_s and K require one bit less in wordlength, for a given quantization increment, than does A_c, and two bits less than A_o.

4.4 Overflow Limit Cycle Tendency

Overflow oscillations, or limit cycles, in fixed point recursive digital filters are caused by nonlinearities introduced by finite register lengths. Most digital filters are implemented using 2's complement arithmetic for the addition operation. Register overflow can then occur at the adder and the resulting nonlinearity causes self sustained oscillations of large magnitude that dominate the output of the filter. This has justified the use of highly conservative scaling rules which makes overflows impossible at the expense of increased roundoff noise [34], or the use of saturation arithmetic in order to not sustain the oscillations when overflow occurs [49].

The tendency of a filter to sustain overflow oscillations also depends on the realization. In state-model realizations it has been shown that certain kinds of system matrices will not sustain oscillations regardless of pole position [7]. In this section, a cursory examination of the tendency of each system matrix under consideration to sustain overflow limit cycles is made by applying the criteria of Mills, Mullis, and Roberts [7] to each matrix. Results for the matrices A_s and A_c have been reported previously [7] [49]. The overflow tendencies of A_o and K have not been previously reported.

The criteria given in [7] constitute a sufficient condition for the absence of overflow oscillations for a matrix A and are based upon

finding a diagonal matrix D with positive diagonal elements for which $(D-A^TDA)$ is positive definite. For a second order matrix A, there exists such a D if and only if the elements of A satisfy

$$(1)\quad a_{12}a_{21}\geq 0 \qquad ,(4\text{-}26)$$

or if $a_{12}a_{21}<0$ then

$$(2)\quad |a_{11}-a_{22}| + \det(A)<1 \qquad .(4\text{-}27)$$

The elements of A_s are given by $\alpha=\rho\cos\Theta$ and $\beta=\rho\sin\Theta$. Since $a_{12}=\beta$, $a_{21}= -\beta$, condition (1) in (4-26) is not satisfied. For condition (2) given in (4-27), A_s yields $\alpha^2+\beta^2<1$ which is true for all pole locations inside the unit circle. Therefore A_s will not sustain overflow limit cycles for all stable pole locations.

For the matrix A_c, application of condition (1) in (4-26) yields $d_1\geq 0$ which is not true since $d_1= -\rho^2<0$. Condition (2) in (4-27) yields $|-d_2|-d_1<1$. Since $d_1<0$, this is the same as $|d_1|+|d_2|<1$, which is the result obtained by Ebert, Mazo, and Taylor [49]. Therefore, A_c will not sustain overflow limit cycles when

$$|d_1|+|d_2|<1 \qquad .(4\text{-}28)$$

This means that the ability of A_c to sustain overflow limit cycles depends on the pole location.

For the matrix A_o, the conditions for which poles are realized inside the unit circle are given by

$$1+x-y>0 \qquad (4\text{-}29)$$

$$2+2y-x>0 \qquad (4\text{-}30)$$

$$x<0 \qquad .(4\text{-}31)$$

Applying condition (1) in (4-26) results in $x \geq 0$ which, from (4-31) is not true. Applying condition (2) in (4-27) yields

$$|1-y|+y-x<1 \qquad .(4\text{-}32)$$

If $1-y>0$, then (4-32) yields $x>0$, which, from (4-31), is not true. If $1-y<0$, then (4-29) is not satisfied since $x<0$. Therefore, at no point in the parameter space of A_o are the criteria in (4-26) and (4-27) satisfied. Since the conditions in (4-26) and (4-27) are sufficient and not necessary, this does not mean that there are no conditions under which A_o will not sustain overflow oscillations. It does point out, however, that scaling techniques may be necessary when using this matrix.

Application of condition (1) in (4-26) to the matrix K reveals that it is not satisfied since $a_{12}=1$, $a_{21}= -1$. Applying condition (2) in (4-27) results in

$$|k_1-k_2|+k_1k_2<0 \qquad .(4\text{-}33)$$

Expressing k_1 and k_2 in terms of ρ and Θ, (4-33) becomes

$$2\sqrt{1-\rho^2\sin^2\Theta}+\rho^2<1 \qquad .(4\text{-}34)$$

For no values of ρ and Θ is (4-34) satisfied. Therefore, at no point in the parameter space of the matrix K are the criteria in (4-26) and (4-27) satisfied. As in the case of A_o, scaling techniques may be necessary when using the matrix K in order to avoid overflow oscillations.

4.5 Roundoff Noise Properties

In this section, the roundoff noise that each second order system exhibits when realizing a common transfer function is compared. Fixed

point arithmetic and rounding of products prior to summing is assumed. As discussed in Chapter 1, roundoff noise results from the quantization of the results of multiplications in the filter and has been shown to be a function of the realization. The product of an m bit multiplicand and an n bit coefficient is an m+n bit product. Due to finite register lengths in the filter, the m+n bit product will be rounded to m bits. This quantization introduces errors which can be represented as additive noise sources after each multiplier coefficient. For example, the noise sources due to roundoff quantization in the state-model presented for the second order matrix K are shown in Figure 4.

The technique used to compute the roundoff noise will follow that presented by Gold and Rader [25]. The errors due to roundoff noise are generally assumed to be statistically independent and have a uniform probability density with zero mean (when quantization is performed by rounding). If E_0 is the quantization increment, the mean squared value of each noise source is given by its variance as

$$\sigma^2 = \frac{E_0^2}{12} \qquad .(4\text{-}35)$$

Since the noise sources are statistically independent, the total mean squared value of the output noise of the filter is given by the sum of the output noise due to each noise input. Therefore, the total output noise is given by

$$\sigma_0^2 = \frac{E_0^2}{12} \sum_{i=1}^{n} \frac{1}{2\pi j} \oint H_i(z) H_i\left(\frac{1}{z}\right) z^{-1} dz \qquad ,(4\text{-}36)$$

where n is the number of noise sources, and $H_i(z)$ is the transfer function relating the i-th noise input to the filter output. The integration path

is taken around the unit circle [25].

Gold and Rader [25] offer (4-36) as being easy to apply to find the roundoff noise of filters since evaluation of the integral for linear discrete networks is always possible from the Cauchy residue theorem.

For the purpose of comparison, each second order system matrix of interest will be used to realize a common transfer function given by

$$H(z) = \frac{Gz}{z^2-d_2z-d_1} \qquad ,(4\text{-}37)$$

where G is a constant gain factor and coefficients $d_1 = -\rho^2$, $d_2 = 2\rho\cos\Theta$, realize the pole pair $z_1, z_2 = \rho\cos\Theta \pm j\rho\sin\Theta$. Expressions for the roundoff noise of filters using matrices A_c and A_s to realize (4-37) have been presented in the literature [25] [30]. Roundoff noise expressions for realization of (4-37) using the matrices A_o and K have not been previously presented.

The realization of (4-37) through the use of matrices K, A_s, A_o, and A_c is shown in Figure 14, Figure 15, Figure 16, and Figure 17, respectively. From these figures, the transfer functions necessary in (4-36) can be determined for each of the noise inputs E_i indicated.

To illustrate the process of computing the roundoff noise of a realization through the use of (4-36), consider the K matrix realization in Figure 14. Through standard state-model methods, the transfer functions relating the noise sources E_1 (resulting from multiplication by k_2), E_2 (resulting from multiplication by k_1), and E_3 (resulting from multiplication by G) to the filter output Y(n) are determined to be

$$H_1(z) = \frac{Gz}{z^2-(k_1+k_2)z+k_1k_2+1} \qquad ,(4\text{-}38)$$

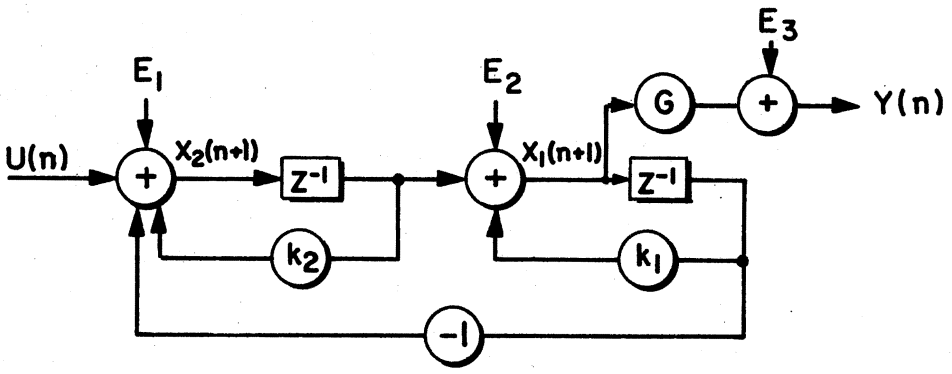


Figure 14. K-Matrix Realization of (4-37)

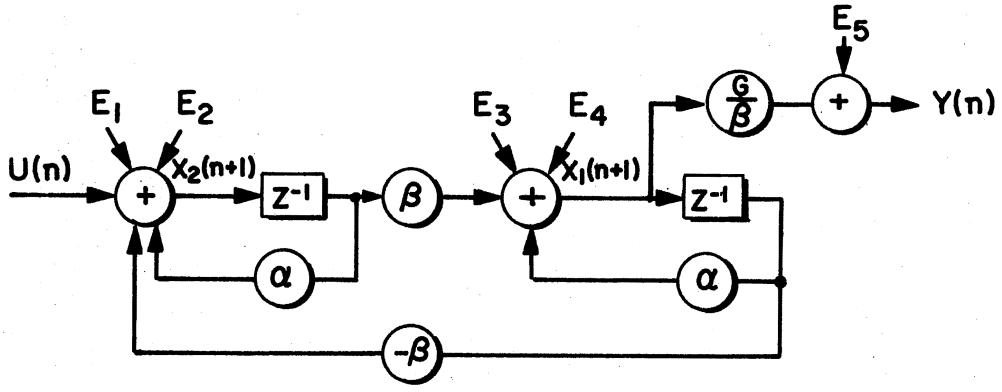


Figure 15. A_s-Matrix Realization of (4-37)

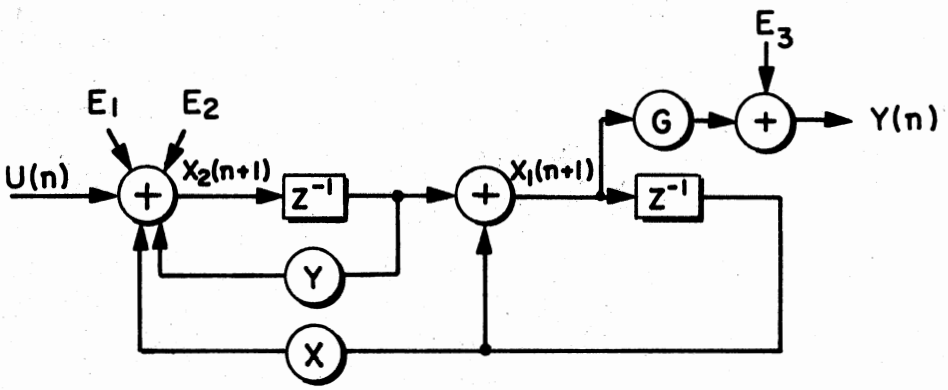


Figure 16. A_o-Matrix Realization of (4-37)

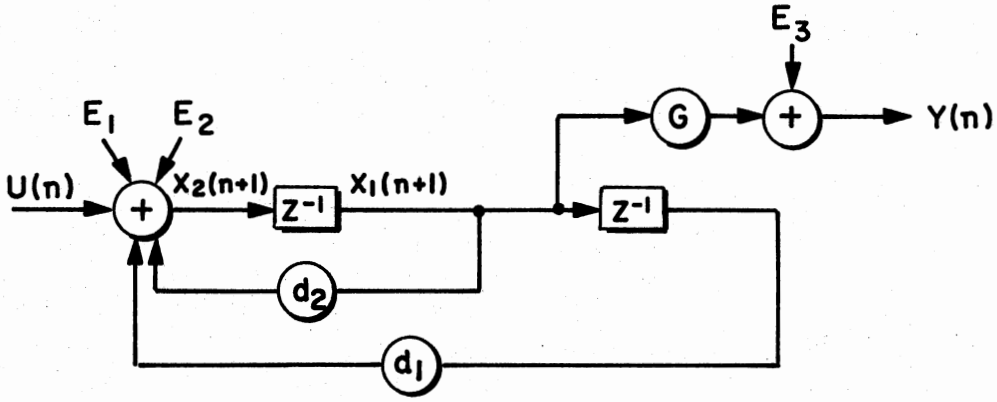


Figure 17. A_c-Matrix Realization of (4-37)

$$H_2(z) = \frac{Gz(z-k_2)}{z^2-(k_1+k_2)z+k_1k_2+1} \quad ,(4\text{-}39)$$

$$H_3(z) = 1 \quad .(4\text{-}40)$$

Then the total output noise is given by

$$\sigma_0^2 = \sigma_1^2+\sigma_2^2+\sigma_3^2 \quad ,(4\text{-}41)$$

where

$$\sigma_1^2 = \frac{E_0^2}{12}\frac{1}{2\pi j}\oint H_1(z)H_1(\tfrac{1}{z})z^{-1}dz \quad ,(4\text{-}42)$$

$$\sigma_2^2 = \frac{E_0^2}{12}\frac{1}{2\pi j}\oint H_2(z)H_2(\tfrac{1}{z})z^{-1}dz \quad ,(4\text{-}43)$$

$$\sigma_3^2 = \frac{E_0^2}{12}\frac{1}{2\pi j}\oint H_3(z)H_3(\tfrac{1}{z})z^{-1}dz \quad .(4\text{-}44)$$

Through the use of the Cauchy residue theorem, with the integration contour being the unit circle, the expressions in (4-42)-(4-44) are determined to be

$$\sigma_1^2 = \frac{E_0^2}{12}\frac{1+\rho^2}{1-\rho^2}\frac{G^2}{\rho^4+1-2\rho^2\cos 2\Theta} \quad ,(4\text{-}45)$$

$$\sigma_2^2 = \frac{E_0^2G^2}{12(1-\rho^2)}\frac{(2\rho\cos\Theta\sqrt{1-\rho^2\sin^2\Theta}(1-\rho^2)+2\rho^2\cos^2\Theta(\rho-1)+2+\rho^2-\rho^4)}{\rho^4+1-2\rho^2\cos 2\Theta} \quad ,(4\text{-}46)$$

$$\sigma_3^2 = \frac{E_0^2}{12} \quad .(4\text{-}47)$$

As the poles of a digital filter approach the unit circle, the roundoff noise increases very rapidly. For comparison purposes it is good to look at the roundoff noise of a realization for poles near the unit circle. Letting $\rho=1-\varepsilon$ in (4-45) and (4-46) results in, as $\varepsilon\to 0$,

$$\sigma_1^2 = \sigma_2^2 = \frac{E_0^2 G^2}{48\varepsilon \sin^2\Theta} \qquad .(4\text{-}48)$$

Substituting (4-48) and (4-47) into (4-41) yields the total roundoff noise of the K matrix realization of (4-37) to be

$$\sigma_0^2 = \frac{E_0^2 G^2}{24\varepsilon \sin^2\Theta} + \frac{E_0^2}{12} \qquad .(4\text{-}49)$$

In a similar manner, the preceding procedure can be used to determine the roundoff noise expressions for the matrices A_s, A_o, and A_c. The results of such an analysis, for pole locations close to the unit circle, are summarized in Table XII.

As shown in Table XII, all of the matrices exhibit the same general roundoff noise properties. The matrix A_s clearly has the highest roundoff noise, with all the other matrices being equal. All of them show that the noise variance is inversely proportional to the distance of the poles from the unit circle. All of them also clearly point out the dependence of the noise on the resonant angle Θ. For very low values of Θ, the noise variance is greatly increased. The newly introduced matrix K compares very well with the other matrices. Although it is no better, it certainly is no worse than the best of the more commonly used matrices.

Since the K matrix is newly introduced, a similar roundoff noise analysis was done for the realization in Figure 4 of the general transfer function given in (2-26). The total roundoff noise variance for Figure 4 is given, for completeness, as

$$\sigma_0^2 = a_1^2(\sigma_1^2+\sigma_2^2 + \frac{E_0^2}{6}) + \frac{E_0^2}{12} \qquad ,(4\text{-}50)$$

TABLE XII

ROUNDOFF NOISE VARIANCE
($\rho=1-\varepsilon$)

	A_s	A_c	A_o	K
σ_0^2	$\frac{E_0^2G^2}{12\varepsilon\sin^2\theta} + \frac{E_0^2}{12}$	$\frac{E_0^2G^2}{24\varepsilon\sin^2\theta} + \frac{E_0^2}{12}$	$\frac{E_0^2G^2}{24\varepsilon\sin^2\theta} + \frac{E_0^2}{12}$	$\frac{E_0^2G^2}{24\varepsilon\sin^2\theta} + \frac{E_0^2}{12}$

where $\sigma_1^2 + \sigma_2^2$ is given by

$$\sigma_1^2 + \sigma_2^2 = \frac{E_0^2}{12} \frac{\frac{-4(\gamma^2+\delta^2)}{k_1 k_2} - (\delta^2 k_1^2 + \gamma^2 k_2^2) + 2\delta\gamma(k_1-k_2)(1+\frac{2}{k_1 k_2})}{3 - (k_1^2 + k_2^2) + (k_1 k_2 + 1)^2} \quad .(4\text{-}51)$$

4.6 Summary

Using the radial and angular sensitivity expressions of section 3.4, expressions for determining the wordlength required to constrain pole movements, due to coefficient quantization, to within prescribed limits are developed for each second order matrix element. Using radial and angular variation limits of 0.001, wordlengths are computed and general trends in wordlength requirements as a function of pole location are noted. With the method used, wordlengths for second order matrices with two variational elements can be obtained directly. Extension of the method to the general n-th order case is discussed.

Realizable pole grids are presented for each system matrix. The matrix elements have a quantization increment of 0.125 as a result of a wordlength consisting of three fractional bits, a magnitude bit, and a sign bit. With this wordlength, the matrix A_o cannot realize poles throughout the entire unit circle unless the element y is implemented indirectly. In general, the matrices A_s and K require one less bit in wordlength than does A_c and two bits less than A_o in order to realize poles throughout the unit circle for a given quantization increment. The matrix K exhibits a very high density of realizable pole locations for poles near the unit circle, especially for pole angles near ninety degrees. With increasing wordlength, this characteristic will be more

pronounced for all poles with magnitudes close to one and will exceed the density offered by the other matrices in this critical pole region.

An examination of the tendency of each of the system matrices to sustain overflow limit cycles is presented. The matrix A_s will not sustain overflow oscillations for any pole location while the tendency of A_c to sustain oscillations is shown to be a function of pole location. The criteria applied in assessing overflow properties did not show A_o and K to be free of overflow oscillations for any pole location. Suitable scaling techniques must be considered when using these matrices.

As a final comparison, the roundoff noise output of a realization, by each system matrix, of a common transfer function is derived. The expressions for each system matrix show the same general properties of increased noise for poles close to the unit circle and/or close to the real axis. The matrix K exhibits roundoff noise properties as good as the best of the more commonly encountered matrices.

CHAPTER V

SUMMARY AND SUGGESTIONS FOR FURTHER STUDY

5.1 Summary

This thesis investigates the effects of coefficient quantization on the pole locations of digital filters realized through state-equations. Since the poles of a digital filter are the eigenvalues of the state-model system matrix, the eigenvalue sensitivity of the system matrix due to variations in the matrix coefficients is used as the method of analysis. A technique for conducting this analysis, based on the sensitivity expressions of a companion matrix, is presented. This technique can be applied to any n-th order linear system, analog or digital, that can be described by state-equations.

If the eigenvalue sensitivity is expressed in terms of its magnitude, it is shown that a sensitivity matrix can be defined for any given system matrix. Furthermore, it is shown that a relationship exists between the sensitivity matrix A' of the companion matrix A_c and the sensitivity matrices $\hat{A}$ of other system matrices. This relationship provides the basis for an easily applied technique for determining the magnitude of the eigenvalue sensitivity of a matrix from the sensitivity magnitude of an equivalent companion matrix.

The same technique applies if the eigenvalue sensitivity of a matrix is expressed in terms of its radial and angular components. It is also

shown that the sensitivity matrix $\hat{A}$ of a given system matrix can be derived from the radial and angular sensitivity expressions for that matrix.

Since second order filters are basic building blocks for higher order filters, second order system matrices are the primary concern of this thesis. A new system matrix suitable for state-model digital filter applications is introduced. This matrix is the second order form of a class of tridiagonal matrices that has the variable elements on the main diagonal while the upper subdiagonal and the lower subdiagonal has invariant elements of positive or negative one. The mapping of the element space of the second order matrix K into the unit circle of the z-plane is discussed and it confirms that the new matrix can realize all stable pole positions inside the unit circle. A state-model, using the matrix K, for the realization of a general second order digital filter transfer function is presented. The general existence of higher order forms of the tridiagonal matrix, for the realization of stable pole locations, is not known. For the third order case, however, a new method is presented for solving the set of nonlinear equations relating the matrix elements and the desired poles. For the specific example presented, where one pole was zero, the matrix is shown to exist. Whether or not it exists for all pole locations inside the unit circle is not known.

The eigenvalue sensitivity of the new matrix K is compared to the sensitivity of other second order matrices. All the matrices are similar in that they have, or can be analyzed as having, only two variational elements. Using the technique presented in this thesis, expressions for the magnitude of the eigenvalue sensitivity, and the corresponding radial

and angular components, are derived for each matrix. The expressions are given in terms of absolute element variations and also in terms of element tolerances. Minimum pole sensitivity regions within the unit circle of the z-plane are shown for each matrix. The new matrix K is shown to exhibit very good sensitivity properties for critical pole locations near the unit circle, where the stability of a filter subject to coefficient variation is of great concern.

The wordlength requirements of each of the second order matrices are compared by deriving expressions for determining the minimum wordlength required to constrain pole movements, due to coefficient quantization, within prescribed limits. For a specific set of limits, wordlengths are computed and general trends in requirements as a function of pole location are noted. With the method used to determine wordlength, the requirements of second order matrices with two variational elements can be obtained directly. Extension of the method to the general n-th order case is discussed.

As another method of comparison, the realizable pole grids obtained when the matrix elements are quantized to three fractional bits are presented for each system matrix. The matrix K exhibits a very high density of realizable pole locations for poles near the unit circle, especially for pole angles near ninety degrees. With increasing wordlength, this characteristic will be more pronounced for all poles with magnitude close to one and will exceed the density offered by the other matrices in that critical pole region.

An examination of the tendency of each of the system matrices to sustain overflow limit cycles is presented. The criteria applied in assessing overflow properties indicates that the matrix K is not immune

to overflow oscillations. Suitable scaling techniques must be considered when using this matrix.

As a final comparison, the roundoff noise output of the realization, by each system matrix, of a common transfer function is derived. The expressions for each matrix show the same general properties of increased noise for poles close to the unit circle and/or close to the real axis. The matrix K exhibits roundoff noise properties as good as the best of the other matrices.

5.2 Suggestions for Further Study

In the following, some extensions to the present effort are suggested. Appropriate references are indicated.

5.2.1 n-th Order Tridiagonal Matrix

The second order class of the general tridiagonal matrix introduced in Chapter 2 has been thoroughly investigated because of the importance of second order sections in digital filter design. It has been shown to exhibit very good properties in this application. Although such a direct application does not exist for higher order matrices, investigation of the properties of the general form of this matrix is warranted from a theoretical nature. The existence of the n-th order form of this matrix for realization of prescribed eigenvalues is not known and needs investigation. A starting point, of course, is the third order matrix which has been shown, in this study, to exist for certain eigenvalues and for which a method of obtaining the matrix elements has been given. Investigation as to whether the proposed solutions always allow real element values to be determined for any combination of eigenvalues is necessary. If the

matrix does not exist for all eigenvalues, for what classes of eigenvalues does it exist?

5.2.2 Quadratic Maximization/Minimization Procedure

The problem of determining the maximum or minimum of $\alpha^T A\alpha$, when the elements α_i of the vector α are $\alpha_i=\pm 1$, is one of determining the signs of the elements such that a maximum or minimum is achieved. For second order matrices this is a simple problem solved by inspection. When the general n-th order problem is considered, where n is large, the determination of α is more complex and requires an efficient, systematic procedure. Although this problem arises in this thesis for the determination of the eigenvalue sensitivity magnitude of an n-th order matrix, McMillan [50] has shown that this quadratic problem also occurs in delta modulation communication problems. Therefore, any contribution to the investigation of this problem could have far ranging effects. As discussed in section 3.5, solutions for this type of problem can possibly be obtained through adaptation of quadratic binary programming procedures [44] [45]. All of the present methods for analyzing this problem are basically enumerative in nature. Although these methods provide a systematic procedure, their enumerative nature is a disadvantage for high order systems. An efficient analytical method for solving this problem is needed.

5.2.3 Eigenvalue Sensitivity Minimization

An interesting class of matrices not considered in this thesis is the second order matrix with all four elements subject to variation.

This type of matrix offers the possibility of being able to minimize the sensitivity of the matrix by proper selection of the values of the matrix elements. If so, it should be realized that the minimization would be achieved at the expense of more numbers of multiplications required per iteration of the filter. Although it is usually desirable to minimize the number of multiplications, it might be more advantageous to employ such a matrix in order to satisfy more important requirements in other areas such as stability. One possible approach to this problem is suggested by a transformation given by Ogata [32] in which the elements a_{12} and a_{22} of a matrix can be expressed in terms of the eigenvalues of the matrix and the elements a_{11}, a_{21}. For eigenvalues $\sigma_1, \sigma_2 = \sigma \pm j\omega$, the result of this transformation is given by

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & \dfrac{(\sigma - a_{11})^2 + \omega^2}{a_{21}} \\ a_{21} & 2\sigma - a_{11} \end{bmatrix} .$$

With the matrix expressed in this manner, the problem now is to find a_{11} and a_{21} such that the eigenvalue sensitivity is minimized. Using either Singer's [9] or Manty's [8] definition of eigenvalue sensitivity, the observation is that the sensitivity of a four variational element matrix can be expressed in terms of σ, ω, which are given, and a_{11}, a_{21}, which are to be determined. At this point it might be possible to employ a minimization procedure, such as steepest descent, to find a_{11}, a_{21} (and, therefore, a_{12}, a_{22}) such that the sensitivity is minimized.

Another approach to the problem of minimizing eigenvalue sensitivity would be to consider the problem of determining a transformation matrix T such that a matrix A is transformed to a minimally sensitive matrix $\bar{A}$,

with the same eigenvalues, by the transformation $\bar{A} = T^{-1}AT$. This is an old problem still awaiting a solution. Recently, Hwang [51] has made some contributions in this area for the problem of roundoff noise minimization. Perhaps some of his results could be applied to the problem of coefficient quantization.

SELECTED BIBLIOGRAPHY

1. Bede Liu, "Effect of finite word length on the accuracy of digital filters - a review," IEEE Trans. Circuit Theory, vol. CT-18 pp. 670-677, 1971.

2. S. A. White, "Recursive-digital-filter accuracy requirements," Real Time Digital Filter and Spectrum Analysis NEC Professional Growth In Electronics Seminar, vol. II, 1970.

3. L. Jackson, J. F. Kaiser, and H. S. McDonald, "An approach to the implementation of digital filters," IEEE Trans. on Audio and Electroacoustics, vol. AU-16, pp. 413-421, 1968.

4. J. F. Kaiser, "Digital filters," in System Analysis by Digital Computers. F. F. Kuo and J. F. Kaiser, Eds., New York: Wiley, pp. 218-235, 1966.

5. C. T. Mullis and R. A. Roberts, "Filter structures which minimize roundoff noise in fixed point digital filters," Proc. IEEE Int. Conf. Accoust., Speech and Signal Processing, pp. 505-508, 1976.

6. E. Avenhaus, "A proposal to find suitable canonical structures for the implementation of digital filters with small coefficient wordlength," Nachrichtentech. Z., vol. 25, pp. 377-382, 1972.

7. W. L. Mills, C. T. Mullis, and R. A. Roberts, "Digital filter realizations without overflow oscillations," IEEE Int. Conf. on Accoust., Speech and Signal Processing, pp. 71-75, 1978.

8. P. E. Mantey, "Eigenvalue sensitivity and state variable selection," IEEE Trans. on Aut. Con., vol. AC-13, pp. 263-269, 1968.

9. R. A. Singer, "Selecting state variables to minimize eigenvalue sensitivity of multivariable systems," Automatica, vol. 5, pp. 85-93, 1969.

10. H. W. Bode, Network Analysis and Feedback Amplifier Design, New York: Van Nostrand, 1945.

11. W. G. Tuel, Jr., "On the transformation to (phase-variable) canonical form," IEEE Trans. on Aut. Con., vol. AC-11, p. 608, 1966.

12. D. S. Rane, "A simplified transformation to (phase-variable) canonical form," IEEE Trans. on Aut. Con., vol. AC-11, p. 608, 1966.

13. D. F. Wilkie and W. R. Perkins, "Essential parameters in sensitivity analysis," Automatica, vol. 5, pp. 191-197, 1969.

14. D. V. Kokotovic and R. S. Rutman, "Sensitivity of automatic control systems (survey)," Automatika: Telemekhanika, vol. 26, pp. 730-750, 1965.

15. C. E. Maley, "The effect of parameters on the roots of an equation system," Computer Journal, vol. 4, pp. 62-63, 1963.

16. D. C. Reddy, "Sensitivity of an eigenvalue of a multivariable control system," Electronics Letters, vol. 2, p. 446, 1966.

17. D. C. Reddy, "Evaluation of the sensitivity coefficient of an eigenvalue," IEEE Trans. on Aut. Con., vol. AC-12, p. 792, 1967.

18. B. S. Morgan, Jr., "Sensitivity analysis and synthesis of multivariable systems," IEEE Trans. on Aut. Con., vol. AC-11, pp. 506-512, 1966.

19. L. P. Huelsman, Theory and Design of Active RC Circuits. New York: McGraw-Hill, pp. 11-58, 1968.

20. G. Daryanani, Principles of Active Network Synthesis and Design. New York: Wiley, pp. 147-174, 1976.

21. S. K. Mitra, Analysis and Synthesis of Linear Active Networks. New York: Wiley, pp. 161-195, 1969.

22. I. M. Horowitz, "The sensitivity problem in sampled-data systems," Trans. IRE, vol. AC-6, pp. 251-259, 1961.

23. T. W. Kerlin, "Sensitivities by the state variable approach," Simulation, pp. 337-345, June 1967.

24. R. M. Golden and J. F. Kaiser, "Design of wideband sampled-data filters," Bell Syst. Tech. J., vol. 43, pt. 2, pp. 1533-1546, 1964.

25. B. Gold and C. M. Rader, Digital Processing of Signals. New York: McGraw-Hill, 1969.

26. A. Fettweis, "On the connection between multiplier word length limitation and roundoff noise in digital filters," IEEE Trans. on Circuit Theory, vol. 19, pp. 486-491, 1972.

27. F. Bonzanigo, "Comment on 'roundoff noise and attenuation sensitivity in digital filters with fixed-point arithmetic'," IEEE Trans. on Circuits and Systems, vol. CAS-21, pp. 809-810, 1974.

28. L. B. Jackson, "Roundoff noise bounds derived from coefficient sensitivities for digital filters," IEEE Trans. on Circuits and Systems, vol. CAS-23, pp. 481-485, 1976.

29. S. K. Mitra and R. J. Sherwood, "Estimation of pole-zero displacements of a digital filter due to coefficient quantization," IEEE Trans. on Circuits and Systems, vol. CAS-21, pp. 116-124, 1974.

30. C. M. Rader and B. Gold, "Effects of parameter quantization on the poles of a digital filter," Proc. IEEE, vol. 55, pp. 686-689, 1967.

31. S. Y. Hwang, "Roundoff noise in state-space digital filtering: a general analysis," IEEE Trans. on Acoust., Speech and Signal Processing, vol. ASSP-24, pp. 256-262, 1976.

32. K. Ogata, State Space Analysis of Control Systems. Englewood Cliffs, N. J.: Prentice-Hall, pp. 294-369, 1967.

33. W. R. Bennett, "Spectra of quantized signals," Bell Syst. Tech. J., vol. 27, pp. 446-472, 1948.

34. L. B. Jackson, An Analysis of Roundoff Noise in Digital Filters, Ph.D. Thesis, Stevens Institute of Technology, Castle Point, Hoboken, New Jersey, 1969.

35. A. V. Oppenheim and R. W. Schafer, Digital Signal Processing. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

36. A Fettweis, "Some principles of designing digital filters imitating classical filter structures," IEEE Trans. on Circuit Theory, vol. CT-18, pp. 314-316, 1971.

37. R. E. Crochiere, "Digital ladder structures and coefficient sensitivity," IEEE Trans. on Audio and Electracoustics, vol. AU-20, pp. 240-246, 1972.

38. T. G. Marshall, "Primitive matrices for doubly terminated ladder networds," Proc. 4th Allerton Conference on Circuit System Theory, University of Illinois, Urbana, pp. 935-943, 1966.

39. R. Yarlagadda, "An application of tridiagonal matrices to network synthesis," SIAM J. Applied Math., vol. 16, pp. 1146-1162, 1968.

40. S. Y. Ku and R. J. Adler, "Computing polynomial resultants: Bezout's determinant vs. Collins reduced P.R.S. algorithm," Communications of the ACM, vol. 12, pp. 23-30, 1969.

41. B. C. Kuo, Automatic Control Systems. Englewood Cliffs, N.J.: Prentice-Hall, pp. 329-388, 1967.

42. R. S. Garfinkel and G. L. Nemhauser, Integer Programming. New York: Wiley, 1972.

43. E. L. Lawler, "The quadratic assignment problem," Management Science, vol. 9, pg. 586, 1963.

44. D. J. Laughhunn, "Quadratic binary programming with application to capital-budgeting problems," Oper. Res., vol. 18, pp. 454-461, 1970.

45. P. L. Manner and S. Rudeanu, Boolean Methods in Operations Research. Berlin: Springer-Verlag, 1968.

46. L. S. Schmidt, "Designing programmable digital filters for LSI implementation," Hewlett-Packard Journal, vol. 29, pp. 15-23, 1978.

47. B. J. Leon and S. C. Bass, "Designers guide to digital filters," Electronic Design News, June 20, 1974.

48. L. R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing. Englewood Cliffs, N.J.: Prentice-Hall, pp. 344-346, 1975.

49. P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow oscillations in digital filters," Bell Syst. Tech. J., vol. 48, pp. 2999-3020, 1969.

50. B. McMillan, "History of a problem," Soc. Indust. Appl. Math. vol. 3, pp. 119-128, 1955.

51. S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," IEEE Trans. on Acoust., Speech and Signal Processing, vol. ASSP-25, pp. 273-281, 1977.

VITA [2]

James Dale Ledbetter

Candidate for the Degree of

Doctor of Philosophy

Thesis: COEFFICIENT QUANTIZATION EFFECTS ON POLE LOCATIONS FOR STATE-MODEL DIGITAL FILTERS

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Miami, Oklahoma, May 14, 1945, the son of Mr. and Mrs. J. E. Ledbetter.

Education: Graduated from Ponca City High School, Ponca City, Oklahoma, in May, 1963; received Associate of Science degree in Engineering from Northern Oklahoma College in May, 1965; received Bachelor of Science degree in Electrical Engineering from Oklahoma State University in May, 1968; received Master of Science in Electrical Engineering from Oklahoma State University in May, 1969; completed requirements for the Doctor of Philosophy degree in Electrical Engineering at Oklahoma State University in May, 1979.

Professional Experience: Engineer in Training, Shell Development Corporation, summer of 1966; Graduate Teaching Assistant, Electrical Engineering, Oklahoma State University, 1968-1969; Graduate Research Assistant, Electrical Engineering, Oklahoma State University, 1970-1973; Electrical Engineer, United States Air Force, 1973.