

ELECTRICAL

N69. 20240

NASA. CR 98259

COEFFICIENT QUANTIZATION IN
HYBRID CONTROL SYSTEMS

PREPARED BY

SAMPLED-DATA CONTROL SYSTEMS GROUP

AUBURN UNIVERSITY

C. L. Phillips, Project Leader

Fourteenth Technical Report

28 September, 1968 to 28 December, 1968

**CASE FILE
COPY**

CONTRACT NAS8-11274
GEORGE C. MARSHALL SPACE FLIGHT CENTER
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
HUNTSVILLE, ALABAMA

ENGINEERING EXPERIMENT STATION

AUBURN UNIVERSITY

AUBURN, ALABAMA

E
N
G
I
N
E
E
R
I
N
G

COEFFICIENT QUANTIZATION IN
HYBRID CONTROL SYSTEMS

PREPARED BY

SAMPLED-DATA CONTROL SYSTEMS GROUP

AUBURN UNIVERSITY

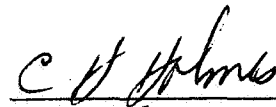
C. L. Phillips, Project Leader

Fourteenth Technical Report

28 September, 1968 to 28 December, 1968

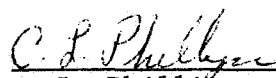
CONTRACT NAS8-11274
GEORGE C. MARSHALL SPACE FLIGHT CENTER
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
HUNTSVILLE, ALABAMA

Approved by:



C. H. Holmes
Head Professor
Electrical Engineering

Submitted by:



C. L. Phillips
Professor
Electrical Engineering

FOREWORD

This document is a technical summary of the progress made since September 28, 1968, by the Auburn University Electrical Engineering Department toward fulfillment of phase B of Contract No. NAS8-11274. This contract was awarded to Engineering Experiment Station, Auburn, Alabama, May 28, 1964, and was extended September 28, 1966 by the George C. Marshall Space Flight Center, National Aeronautics and Space Administration, Huntsville, Alabama.

SUMMARY

The performance of a hybrid control system containing a finite word-length digital subsystem deviates from that obtainable if infinite word-length capabilities are assumed. This deviation arises for two reasons: (1) the system variables processed by the digital hardware are quantized, and (2) the nominal coefficients of the pulse transfer function $D(z)$ to be realized by the digital system are quantized. In this report, quantization errors of the second category are considered. Specifically, a technique is presented for precisely correcting the errors which arise when the nominal coefficients of $D(z)$ are approximated by digital words of finite length. The correction technique is developed initially for the case of a digital filter with one approximated coefficient and is then extended to closed-loop hybrid systems with several approximated coefficients.

The case wherein the correction technique is not feasible is also considered, and a method is described for optimizing the selection of quantized coefficients relative to a given performance index. The performance index employed is based on the use of frequency-domain design specifications.

Finally, an algorithm is presented for generating envelopes on the frequency-response characteristics of continuous- and

discrete-time systems subject to parameter anomalies.

LIST OF PERSONNEL

The following named staff members of Auburn University have actively participated on this project:

- C. L. Phillips--Professor of Electrical Engineering
- R. K. Cavin III--Graduate Assistant in Electrical Engineering
- D. L. Chenoweth--Graduate Assistant in Electrical Engineering
- J. C. Johnson--Graduate Assistant in Electrical Engineering

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF SYMBOLS	xi
I. INTRODUCTION	1
II. CORRECTION OF COEFFICIENT QUANTIZATION ERRORS	4
Digital Filter With One Corrected Coefficient	
Correction of a single numerator coefficient of $D(z)$	
Correction of a single denominator coefficient of $D(z)$	
Coefficient Correction in Hybrid Control Systems	
Example	
Digital Filter With Several Corrected Coefficients	
Correction of several numerator coefficients of $D(z)$	
Correction of several denominator coefficients of $D(z)$	
Hybrid Feedback Control Systems With Several Corrected Coefficients	
Some Practical Considerations	
A Note on Other Realizations of $D(z)$	
III. SELECTION OF OPTIMAL QUANTIZED COEFFICIENTS.	46
Performance Index	
Definitions	
Minimization of $J(\beta)$	
Modification of steepest descent method	
Second stage	
Example	

IV. AN ALGORITHM FOR COMPUTING FREQUENCY-RESPONSE BOUNDS FOR SYSTEMS SUBJECT TO PARAMETER ANOMALIES	63
Definitions and Theorems	
Development of the Algorithm for Toleranced Parameters in Continuous-Time Systems	
Change of variables	
Absolute bounds on U and V	
Bounds on Magnitude and Phase	
Simplifications	
Example	
Extention to Digital Systems With Coefficient Anomalies	
V. CONCLUSIONS	85
REFERENCES	88
APPENDIX	90

LIST OF TABLES

1. Minimization of performance index	61
--	----

LIST OF FIGURES

1.	Generalized block diagram of digital filter	6
2.	Schematic implementation of numerator coefficient correction system via time-shared digital filter.	10
3.	Schematic implementation of numerator coefficient correction system via duplicate digital filters	11
4.	Schematic implementation of denominator coefficient correction system via time-shared digital filter	15
5.	Schematic implementation of denominator coefficient correction system via duplicate digital filters	16
6.	Schematic diagram of generalized hybrid system	18
7.	Discrete-time model of modified hybrid system illustrating independence of $y(k)$ and Δ_i	21
8.	Composite corrected hybrid system illustrating schematically the correction system implementation for a single denominator coefficient	24
9.	Comparison of corrected and noncorrected response	26
10.	Schematic implementation of coefficient correction system for two numerator coefficients of $D(z)$	31
11.	Discrete-time model of coefficient correction system for two denominator coefficients of $D(z)$	38
12.	Schematic representation of simplified denominator coefficient correction system implementation	43
13.	Discrete-time model of simplified correction system for two numerator coefficients of $D(z)$	44
14.	Example	55
15.	Nyquist diagram of nominal system	58

16.	Nyquist diagram of system with quantized coefficients	60
17.	Typical rectangular region in the U-V plane.	72
18.	Special cases where magnitude or phase bounds are not obtained at a vertex of the rectangular region	73
19.	Envelope on the gain characteristic of $G(j\omega, \bar{\mathbf{a}})$ for $0.9 \leq \bar{a}_1 \leq 1.1$, $1.8 \leq \bar{a}_2 \leq 2.2$, and $0.9 \leq \bar{a}_3 \leq 1.1$	81
20.	Envelope on the phase characteristic of $G(j\omega, \bar{\mathbf{a}})$ for $0.9 \leq \bar{a}_1 \leq 1.1$, $1.8 \leq \bar{a}_2 \leq 2.2$, and $0.9 \leq \bar{a}_3 \leq 1.1$	82

LIST OF SYMBOLS

$\bar{a}, \bar{b}, \bar{\alpha}, \bar{\beta}$	anomalous parameter vectors
$\underline{a}, \underline{b}, \underline{\alpha}, \underline{\beta}$	nominal parameter vectors
A/D	analog-to-digital
D/A	digital-to-analog
$D(z)$	transfer function of digital element
$\bar{e}_o(k), \bar{x}(k), \bar{y}(k)$	perturbed system variables at the kth sampling instant
$e_o(k), x(k), y(k)$	nominal system variables at the kth sampling instant
$e_o'(t), x'(t)$	continuous-time variables reconstructed from $e_o(k)$, and $x(k)$
$e_{oi}^{(n)}(k)$	symbol denoting $\partial e_o^n(k) / \partial \beta_i^n$
$e_{oij}^{(m)(n)}(k)$	symbol denoting $\partial^{m+n} e_o(k) / \partial \beta_i^m \partial \beta_j^n$
$E_o(z)$	the z-transform of the discrete-time variable $e_o(k)$
E_{2N+1}^β	symbol denoting 2N+1-dimensional Euclidean space with coordinant axes $\beta_1, \beta_2, \dots, \beta_{2N+1}$
$G(s)$	continuous-time system transfer function
H	a matrix
h_i	granularity associated with the coefficient word-length of β_i
$J(\underline{\beta})$	performance index used in the selection of optimal quantized coefficients
Q	a point set in E_{2N+1}^β

Q_i	set of anomalies of β_i
s	complex variable associated with the Laplace transform
$\underline{s}$	vector in the direction of descent of $J(\underline{\beta})$
$S(k, \Delta_i)$	correction coefficient associated with $\bar{\beta}_i$
T	system sampling period
T_U, T_V	point sets containing candidates for extremizing U and V
U	real component of a complex number
V	imaginary component of a complex number
$y^n(t)$	symbol denoting $d^n y(t)/dt$
$\underline{\beta}^o$	starting point in E_{2N+1}^β for minimizing $J(\underline{\beta})$
Δ	symbol denoting $\det[H]$
Δ_i	scalar representing the difference between β_i and $\bar{\beta}_i$
ϵ	means "is a member of"
$\subset$	means "is a subset of"

I. INTRODUCTION

As a result of the rapid advances in digital computer technology in recent years and the increasing versatility and commercial availability of digital components, the use of digital computers as components in otherwise continuous-time control systems is becoming increasingly prevalent. Systems which contain both analog and digital components are normally referred to as "hybrid systems," and a large body of knowledge, based primarily on z-transform calculus, has evolved for the analysis and synthesis of systems of this class.

One of the major sources of error in a hybrid system is the presence of quantization effects within the digital elements of the system. There are two categories of quantization errors which affect the system performance. These categories are: (1) amplitude quantization of system variables, and (2) quantization of coefficients of the difference equation to be realized by the digital device. This dissertation considers quantization errors of the second classification.

Coefficient quantization in hybrid systems is present in physically realizable digital devices because the coefficients must be represented with binary words of finite length. For example, consider the case wherein the difference equation

$$e_o(k) = \beta_1 e_i(k) + \beta_2 e_o(k-1) \quad (I-1)$$

is to be realized by the digital portion of a hybrid system, where $e_o(k)$ and $e_i(k)$ denote the output and input, respectively, of the digital subsystem at the k th sampling instant. Assume that the nominal coefficients β_1 and β_2 are to be implemented with binary coded words having one sign bit and 10 magnitude bits to the right of the binary point. Consequently, the least significant magnitude bit which may be exercised in realizing β_1 and β_2 has a decimal weight of 2^{-10} . Therefore, the set B of permissible coefficient quantization levels of the device is

$$B = [\pm n2^{-10}; n = 0, 1, 2, \dots, 1023] , \quad (I-2)$$

and in general, β_1 and β_2 must be approximated by a suitable selection from the members of B . The coefficient quantization errors of course depend upon the members of B that are chosen to approximate β_1 and β_2 and upon the granularity of the set B , which in this case is 2^{-10} .

The above example was based on the assumption that the coefficients are uniformly binary coded; however, the quantization phenomenon results regardless of the coding scheme employed. Different coding schemes simply change the contents of B .

It is sometimes the case, due to the presence of rigid design specifications, that one or more digital coefficients must be realized with a much higher degree of accuracy than is permitted by the digital system word-lengths. A technique for correcting coefficient quantization errors in these cases and for attaining effectively infinite word-length

resolution of digital coefficients is presented in Chapter II.

The developments of Chapter III are based on the premise that the set of permissible digital coefficient levels are adequate for approximating the nominal coefficients, and a technique is presented for optimizing the selection of quantized coefficients. The performance index employed is based upon frequency-domain specifications. Finally, in Chapter IV, an algorithm is presented for generating frequency-response bounds of continuous-time systems subject to parameter tolerances. The technique is then extended to digital systems where coefficient anomalies are investigated. The contents of Chapter V are the conclusions resulting from the work in Chapters II, III, and IV.

II. CORRECTION OF COEFFICIENT QUANTIZATION ERRORS

In this chapter a technique is developed for eliminating the output error arising from the quantization of system parameters in linear, time-invariant, discrete-time systems. A generalized hybrid feedback control system is considered and the technique is employed to correct for the effects of coefficient quantization within the digital elements of the system.

The correction method is developed first for the case of coefficient quantization in digital filters with a single quantized coefficient; it is then extended to closed-loop hybrid systems containing digital elements with more than one quantized coefficient.

A. Digital Filter With One Corrected Coefficient

The input-output characteristics of an Nth-order digital filter may be represented by an Nth-order transfer function in z , which will be denoted by $D(z)$ [1]. Numerous techniques are available for physically implementing this transfer function [2,3]. Kaiser [4] has shown that in order to minimize the influence of coefficient quantization upon the pole locations of $D(z)$, any digital filter can, and indeed should, be realized by a combination of first- and second-order systems. Furthermore, it has been demonstrated that certain types of implementation schemes, for a given $D(z)$, are more attractive than others from the viewpoint of minimization of errors which arise due to the quantization of

system variables [5].

However, rather than confining the following discussion to a few specialized techniques for implementing $D(z)$, a more general approach will be taken. A generalized representation of $D(z)$ will be considered, and the correction techniques developed for this form will be shown to be applicable to any of the various other realization methods of $D(z)$.

The assumed form of $D(z)$ is

$$D(z) = \frac{E_o(z)}{E_i(z)} = \frac{\beta_1 z^N + \beta_2 z^{N-1} + \dots + \beta_{N+1}}{z^N - \beta_{N+2} z^{N-1} - \dots - \beta_{2N+1}}, \quad (\text{II-1})$$

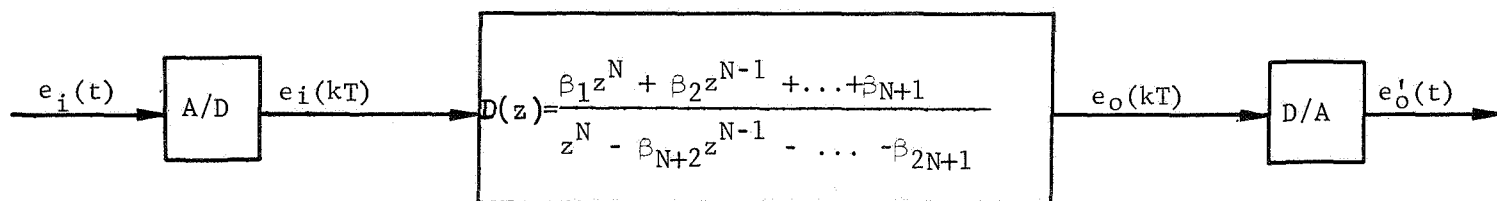
where β_i , $i=1, 2, \dots, 2N+1$, are the nominal, i.e., nonquantized, coefficients to be realized by the digital filter. A generalized block diagram of the digital filter and the associated analog-to-digital (A/D) and digital-to-analog (D/A) interfaces are depicted in Figure 1. Note that the variable $e'_o(t)$ defines the reconstructed continuous-time output of the D/A converter, which in practice is usually a zero-order data-hold.

The transfer function of (II-1) may be described by an N th-order difference equation of the form

$$\begin{aligned} e_o(kT) = & \beta_1 e_i(kT) + \beta_2 e_i[(k-1)T] + \dots + \beta_{N+1} e_i[(k-N)T] \\ & + \beta_{N+2} e_o[(k-1)T] + \dots + \beta_{2N+1} e_o[(k-N)T], \end{aligned} \quad (\text{II-2})$$

where T is the system sampling period. However, in the interest of simplifying notation in the subsequent discussion, T will hereafter be omitted.

On the basis of (II-2), the correction technique will now be developed.



9

Fig. 1--Generalized block diagram of digital filter.

1. Correction of a single numerator coefficient of $D(z)$

The solution of (II-2) represents the nominal response of the digital filter to initial conditions and to the input $e_i(k)$. Consider now the resulting response if one of the $D(z)$ numerator coefficients, for example β_i , $1 \leq i \leq N+1$, is perturbed due to the effects of quantization. In other words, assume that each of the coefficients, with the exception of β_i , lies on one of the admissible quantization levels of the digital filter. Let $\bar{\beta}_i$ denote the quantized value of β_i such that

$$\beta_i = \bar{\beta}_i + \Delta_i, \quad (\text{II-3})$$

where Δ_i defines the value of the perturbation of β_i . Thus, the filter response with β_i quantized is the solution of the following difference equation:

$$\begin{aligned} \bar{e}_o(k) = & \beta_1 e_i(k) + \beta_2 e_i(k-1) + \dots + \bar{\beta}_i e_i(k-i+1) + \dots \\ & + \beta_{N+1} e_i(k-N) + \beta_{N+2} \bar{e}_o(k-1) + \dots + \beta_{2N+1} \bar{e}_o(k-N), \end{aligned} \quad (\text{II-4})$$

where $\bar{e}_o(k)$ is used to denote the response of the filter with coefficient quantization effects included.

The solution $e_o(k)$ of (II-2) at any sampling instant is a function of the coefficient of β_i . Thus, it is convenient to relate $e_o(k)$ and $\bar{e}_o(k)$ through a Taylor series expansion in one variable of $e_o(k)$ about the perturbed response $\bar{e}_o(k)$.

In generalized terms, the series takes the form

$$e_o(k) = \bar{e}_o(k) + \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_o^{(n)}(k) \Big|_{\beta_i = \bar{\beta}_i}, \quad (\text{II-5})$$

where $e_o^{(n)}(k)$ denotes the n th derivative of $e_o(k)$ with respect to β_i . At this point it is advantageous to introduce an auxiliary variable $S(k, \Delta_i)$, which is defined by

$$S(k, \Delta_i) = \frac{1}{\Delta_i} \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_o^{(n)}(k) \Big|_{\beta_i = \bar{\beta}_i} . \quad (\text{II-6})$$

Thus (II-5) can be expressed as

$$e_o(k) = \bar{e}_o(k) + \Delta_i S(k, \Delta_i) . \quad (\text{II-7})$$

Due to its usage in (II-7) as a corrective term relating the nominal and perturbed responses, the variable $S(k, \Delta_i)$ will henceforth be referred to as a "correction coefficient."

Each of the terms comprising $S(k, \Delta_i)$ may be directly evaluated from (II-2); i.e.,

$$\begin{aligned} \frac{1}{1!} e_o^{(1)}(k) \Big|_{\beta_i = \bar{\beta}_i} &= e_i(k-i+1) + \beta_{N+2} \bar{e}_o^{(1)}(k-1) + \dots \\ &\quad + \beta_{2N+1} \bar{e}_o^{(1)}(k-N) , \\ \frac{\Delta_i}{2!} e_o^{(2)}(k) \Big|_{\beta_i = \bar{\beta}_i} &= \frac{\Delta_i}{2!} \left\{ \beta_{N+2} \bar{e}_o^{(2)}(k-1) + \dots + \beta_{2N+1} \bar{e}_o^{(2)}(k-N) \right\} , \\ \frac{\Delta_i^2}{3!} e_o^{(3)}(k) \Big|_{\beta_i = \bar{\beta}_i} &= \frac{\Delta_i^2}{3!} \left\{ \beta_{N+2} \bar{e}_o^{(3)}(k-1) + \dots + \beta_{2N+1} \bar{e}_o^{(3)}(k-N) \right\} , \end{aligned} \quad (\text{II-8})$$

and so on. It should be noted that implicit in (II-8) is the assumption that the nominal filter coefficients are independent of each other ($d\beta_i/d\beta_j = 0$; $i \neq j$) and the input $e_i(k)$ is independent of the filter coefficients ($de_i(k)/d\beta_i = 0$). However, this does not impose severe restrictions upon the application of the correction technique, since

dependencies of this type normally do not exist.

The combination of (II-6) and (II-8) yields

$$S(k, \Delta_i) = \beta_{N+2}S(k-1, \Delta_i) + \dots + \beta_{2N+1}S(k-N, \Delta_i) + e_i(k-i+1). \quad (\text{II-9})$$

Therefore, the correction coefficient $S(k, \Delta_i)$ is actually the solution of what might be called an auxiliary difference equation as given by (II-9). Furthermore, the system that this equation describes responds to $e_i(k-i+1)$ from its zero initial state, since the choice of initial conditions of the digital filter is completely independent of the filter coefficients. This auxiliary system, defined by (II-9), will be referred to as the "correction system" in the sequel.

From the viewpoint of physical implementation of the correction system, (II-9) has two significant properties: (1) each of its coefficients are realizable by the digital filter, and (2) its order and its characteristic equation are identical to those of $D(z)$. Hence, the correction system may be implemented using (II-9) by either of the following configurations: (1) by the digital filter itself if time-share operation of the digital filter is feasible, as illustrated in Figure 2, and (2) by two duplicate digital filters operating in parallel as shown in Figure 3. In each of these two configurations the correction coefficient $S(k, \Delta_i)$ is generated as a digital variable and is then multiplied by Δ_i following the A/D conversion process. The resulting correction term $\Delta_i S(k, \Delta_i)$ is then added to the perturbed response $\bar{e}_0(k)$, which completes the implementation of (II-7) and produces the nominal response $e_0(k)$.

2. Correction of a single denominator coefficient of $D(z)$

Development of the correction technique for denominator coefficients

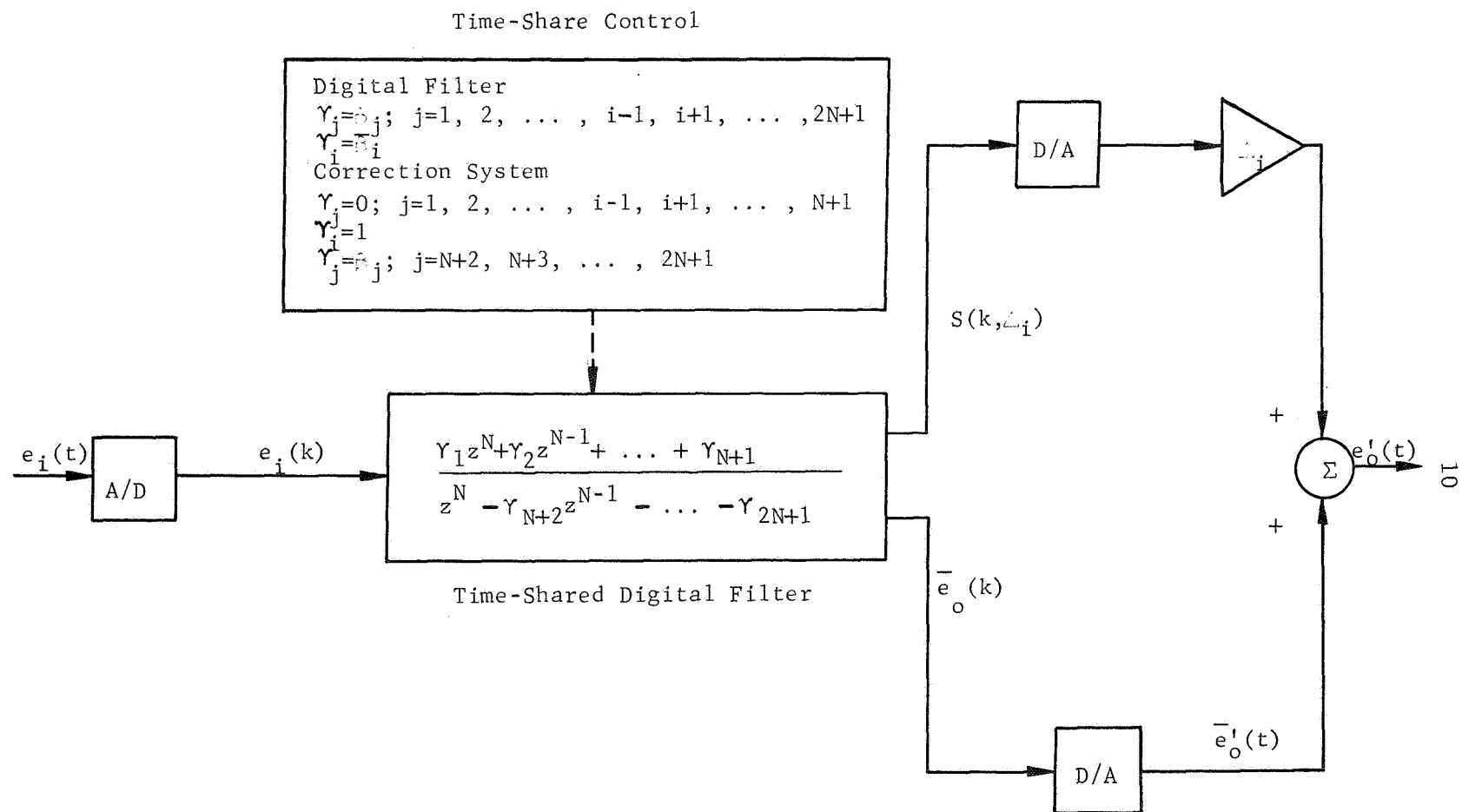


Fig. 2--Schematic implementation of numerator coefficient correction system via time-shared digital filter.

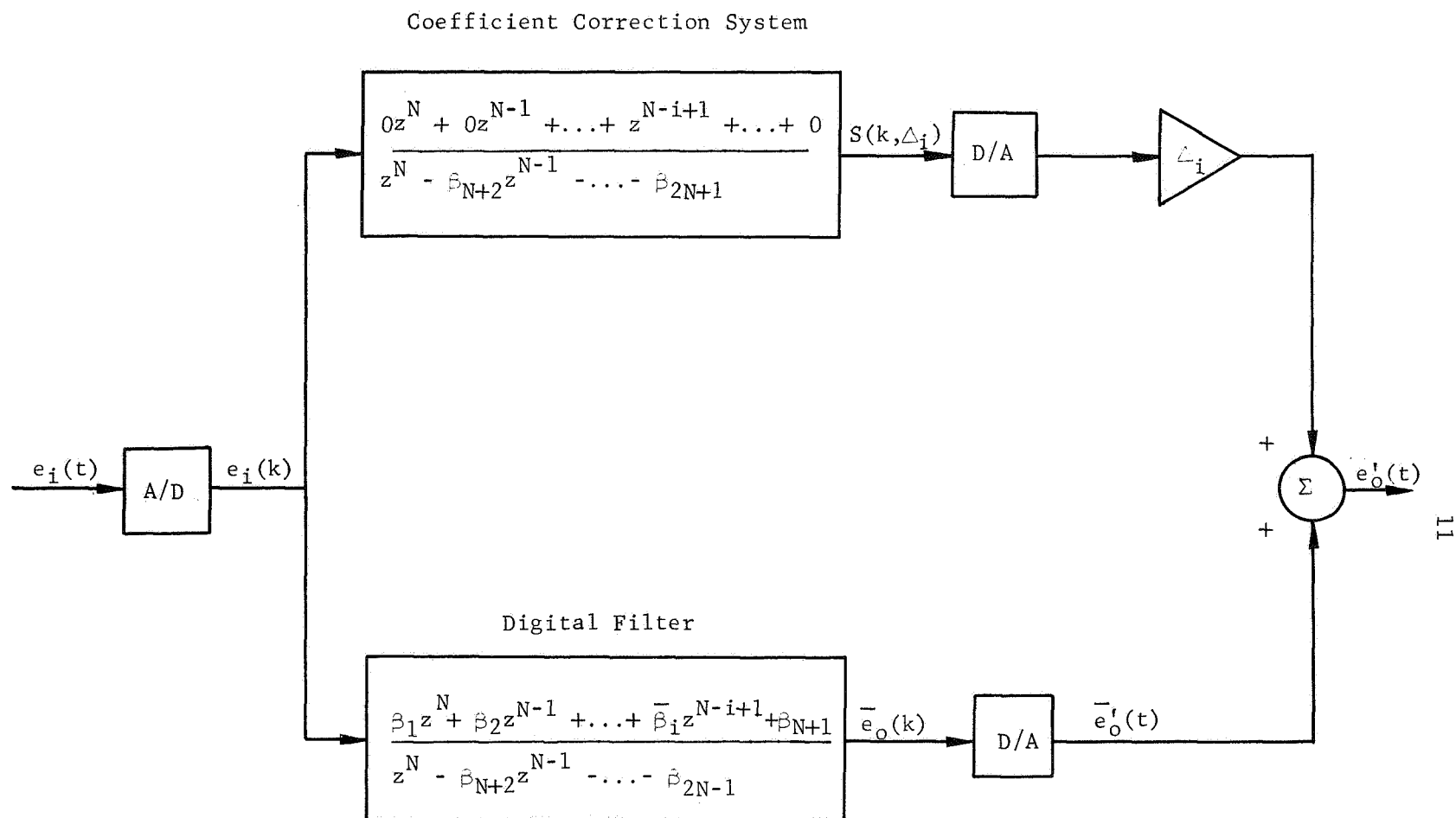


Fig. 3--Schematic implementation of numerator coefficient correction system via duplicate digital filters.

will now proceed and is based largely on arguments and assumptions advanced in the previous section.

Consider the perturbed filter response $\bar{e}_o(k)$ which might result due to the quantization of one denominator coefficient, for example β_i , where $N+2 \leq i \leq 2N+1$. The difference equation describing the perturbed filter response is

$$\begin{aligned} \bar{e}_o(k) = & \beta_1 e_i(k) + \beta_2 e_i(k-1) + \dots + \beta_{N+1} e_i(k-N) + \beta_{N+2} \bar{e}_o(k-1) \\ & + \dots + \bar{\beta}_i \bar{e}_o(k-i+N+1) + \dots + \beta_{2N+1} \bar{e}_o(k-N), \end{aligned} \quad (\text{II-10})$$

The nominal response $e_o(k)$ may be expressed in a Taylor series expansion in the variable β_i about the perturbed response $\bar{e}_o(k)$ as follows:

$$e_o(k) = \bar{e}_o(k) + \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_o^{(n)}(k) \Big|_{\beta_i = \bar{\beta}_i}, \quad (\text{II-11})$$

where $e_o^{(n)}(k)$ denotes the n th derivative of $e_o(k)$ with respect to β_i .

Once again, it is convenient to define an auxiliary variable $S(k, \Delta_i)$, which is made up of the derivative terms in the Taylor series expansion; i.e.,

$$S(k, \Delta_i) = \frac{1}{\Delta_i} \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_o^{(n)}(k) \Big|_{\beta_i = \bar{\beta}_i} \quad (\text{II-12})$$

Thus,

$$e_o(k) = \bar{e}_o(k) + \Delta_i S(k, \Delta_i). \quad (\text{II-13})$$

Each of the terms comprising $S(k, \Delta_i)$ may be generated directly from (II-2). It is assumed as before that the coefficients are independent

and that the input $e_i(k)$ is independent of the coefficients. Therefore,

$$\begin{aligned}
\left. \frac{1}{1!} e_o^{(1)}(k) \right|_{\beta_i = \bar{\beta}_i} &= \beta_{N+2} \bar{e}_o^{(1)}(k-1) + \beta_{N+3} \bar{e}_o^{(1)}(k-2) + \dots \\
&\quad + \bar{\beta}_i \bar{e}_o^{(1)}(k-i+N+1) + \bar{e}_o^{(1)}(k-i+N+1) + \dots \\
&\quad + \beta_{2N+1} \bar{e}_o^{(1)}(k-N) \\
\left. \frac{\Delta_i}{2!} e_o^{(2)}(k) \right|_{\beta_i = \bar{\beta}_i} &= \frac{\Delta_i}{2!} \left\{ \beta_{N+2} \bar{e}_o^{(2)}(k-1) + \beta_{N+3} \bar{e}_o^{(2)}(k-2) + \dots \right. \\
&\quad + \bar{\beta}_i \bar{e}_o^{(2)}(k-i+N+1) + 2 \bar{e}_o^{(1)}(k-i+N+1) + \dots \\
&\quad \left. + \beta_{2N+1} \bar{e}_o^{(2)}(k-N) \right\}, \\
\left. \frac{\Delta_i^2}{3!} e_o^{(3)}(k) \right|_{\beta_i = \bar{\beta}_i} &= \frac{\Delta_i^2}{3!} \left\{ \beta_{N+2} \bar{e}_o^{(3)}(k-1) + \beta_{N+3} \bar{e}_o^{(3)}(k-2) + \dots \right. \\
&\quad + \bar{\beta}_i \bar{e}_o^{(3)}(k-i+N+1) + 3 \bar{e}_o^{(2)}(k-i+N+1) + \dots \\
&\quad \left. + \beta_{2N+1} \bar{e}_o^{(3)}(k-N) \right\}, \tag{II-14}
\end{aligned}$$

and so on.

Through the combination of (II-12) and (II-14), it can be shown that the correction coefficient $S(k, \Delta_i)$ satisfies the following recursive relationship:

$$\begin{aligned}
S(k, \Delta_i) &= \beta_{N+2} S(k-1, \Delta_i) + \beta_{N+3} S(k-2, \Delta_i) + \dots + \bar{\beta}_i S(k-i+N+1, \Delta_i) \\
&\quad + \dots + \beta_{2N+1} S(k-N, \Delta_i) + \dots + \bar{e}_o(k-i+N+1) \\
&\quad + \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_o^{(n)}(k-i+N+1) \Big|_{\beta_i = \bar{\beta}_i}. \tag{II-15}
\end{aligned}$$

Note, however, that the last two quantities in (II-15) are actually the Taylor series representation of the nominal response $e_o(k-i+N+1)$; consequently,

$$\begin{aligned} S(k, \Delta_i) = & \beta_{N+2} S(k-i, \Delta_i) + \beta_{N+3} S(k-2, \Delta_i) + \dots + \beta_i S(k-i+N+1, \Delta_i) \\ & + \dots + \beta_{2N+1} S(k-N, \Delta_i) + e_o(k-i+N+1) . \end{aligned} \quad (\text{II-16})$$

Thus, the correction coefficient for the case of quantized denominator coefficients of $D(z)$ is the solution of an N th-order difference equation having coefficients that are precisely realizable by the digital filter. Furthermore, the correction system described by this equation responds to the nominal filter output sequence $e_o(k-i+N+1)$ initially from its zero state, since the initial conditions of the digital filter are independent of the coefficients of the filter.

Essentially the same type of correction system implementation schemes as mentioned previously for quantized numerator coefficients may be utilized for correction of quantized denominator coefficients of $D(z)$. One notable similarity of the auxiliary system equations, (II-9) and (II-16), is that in each case the correction system and the digital filter have identical characteristic equations. However, they are dissimilar in that the filter input is also the input to the numerator coefficient correction system; while in the case of quantized denominator coefficients, the nominal filter response $e_o(k-i+N+1)$ is the input to the correction system.

Schematic diagrams of two forms of denominator coefficient correction system implementations are depicted by Figure 4 and Figure 5. In each example the corrected filter response $e_o(k-i+N+1)$ is converted from analog

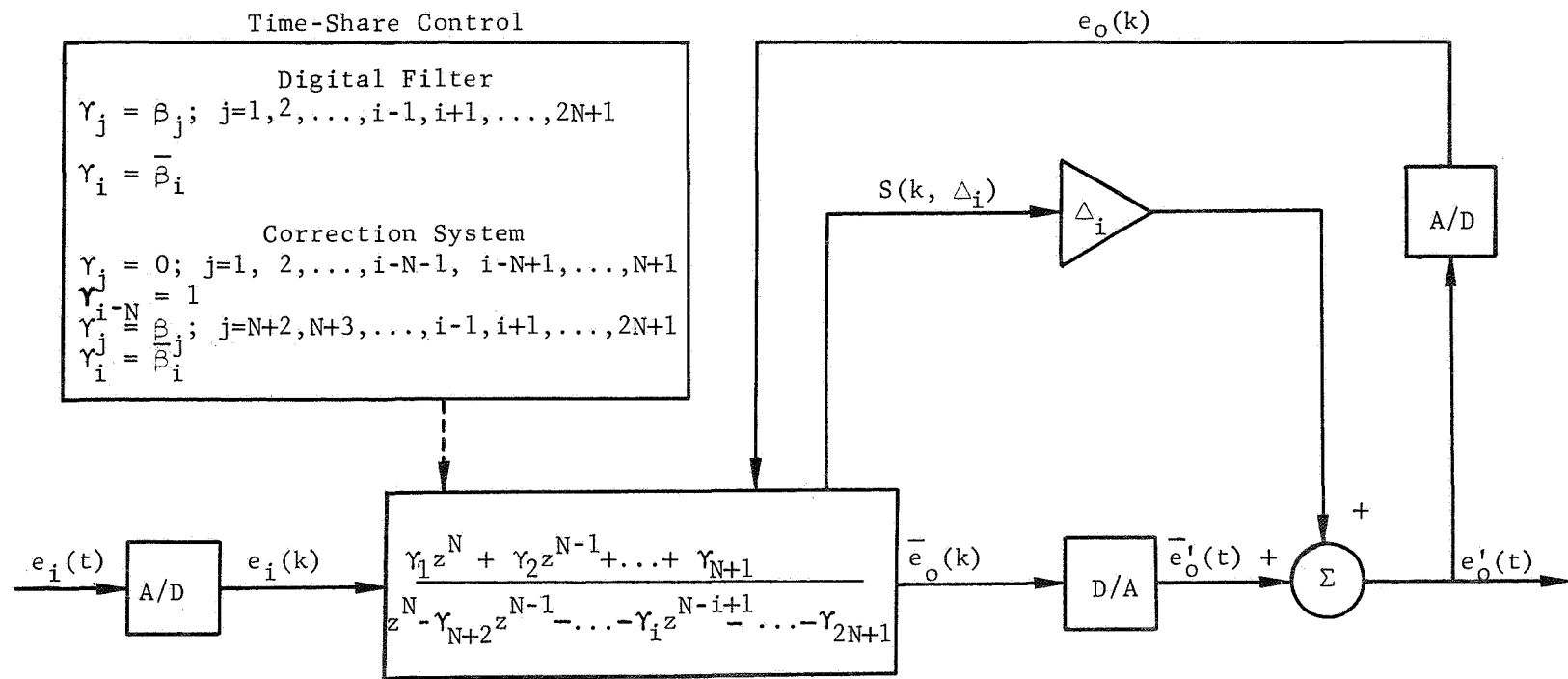


Fig. 4--Schematic implementation of denominator coefficient correction system via time-shared digital filter.

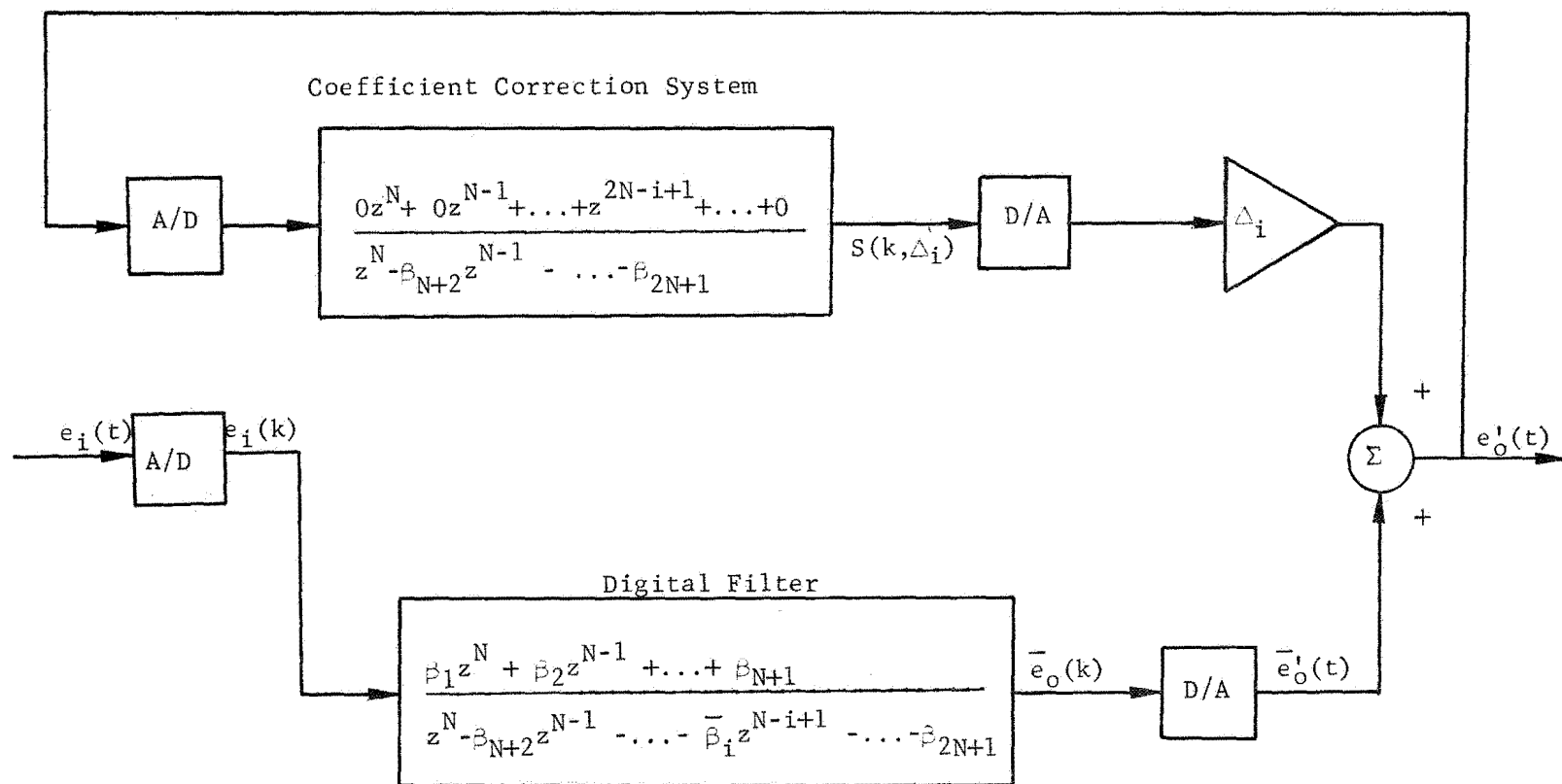


Fig. 5--Schematic implementation of denominator coefficient correction system via duplicate digital filters.

to digital form and is then processed by the auxiliary system, which generates the digitized correction coefficient $S(k, \Delta_i)$. A D/A conversion then takes place and the product $\Delta_i S(k, \Delta_i)$ is formed in analog fashion and added to $\bar{e}_o'(t)$. This completes the implementation of (II-13) and generates the nominal filter response $e_o'(t)$.

B. Coefficient Correction in Hybrid Feedback Control Systems

The foregoing correction technique may be readily extended to systems comprised of closed-loop interconnections of linear, continuous-time, stationary elements and digital elements having quantized coefficients. The schematic diagram illustrated in Figure 6 typifies this class of hybrid systems.

Since the development of the correction technique for hybrid systems closely parallels that previously described for open-loop digital filter configurations, the following discussion will deal, for the purpose of minimizing redundancy, only with the correction of a single denominator coefficient $\bar{\beta}_i$ of $D(z)$, the transfer function of the digital element of the hybrid system. The correction technique for numerator coefficients will be obvious from these results and those of the previous sections. More specifically, the problem being considered here is: "How can the system of Figure 6 be modified to produce the nominal response $y(k)$ even though the denominator coefficient β_i is perturbed?"

The continuous-time elements of the hybrid system may be modelled by the following generalized differential equation:

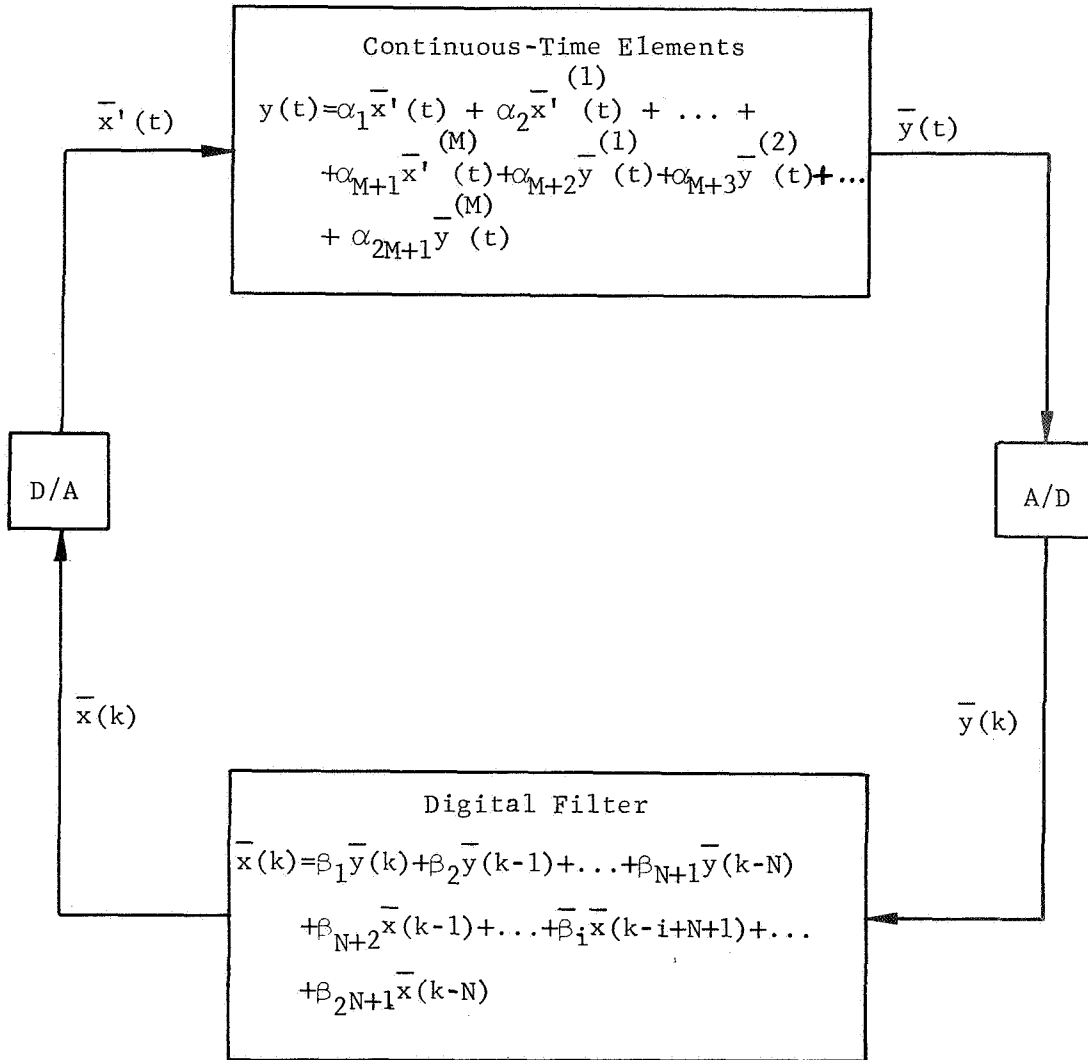


Fig. 6--Schematic diagram of generalized hybrid system.

$$\begin{aligned} \bar{y}(t) = & \alpha_1 \bar{x}'^{(1)}(t) + \alpha_2 \bar{x}'^{(2)}(t) + \dots + \alpha_{M+1} \bar{x}'^{(M)}(t) + \alpha_{M+2} \bar{y}^{(1)}(t) \\ & + \alpha_{M+3} \bar{y}^{(2)}(t) + \dots + \alpha_{2M+1} \bar{y}^{(M)}(t), \end{aligned} \quad (\text{II-17})$$

where $\alpha_i, i=1, 2, \dots, 2M+1$, are determined by the system parameters and M is the order of the system model. In keeping with the notational convention previously established, $\bar{y}(t)$ denotes the perturbed response of the continuous-time elements (resulting from quantized digital filter coefficients) and $\bar{x}'(t)$ represents the perturbed filter response after being reconstructed by the D/A converter, which in most cases is a simple zero-order data-hold. The n th derivative of $\bar{y}(t)$ and $\bar{x}'(t)$ with respect to time are denoted by $\bar{y}^{(n)}(t)$ and $\bar{x}'^{(n)}(t)$, respectively. Since the digital filter accepts inputs and generates outputs every T seconds, where T is the system sampling period, it is necessary to determine the output $\bar{y}(k)$ of the A/D data-hold element every T seconds. For this reason it is desirable to "discretize" the continuous-time portion of the system and formulate a discrete-time model, or difference equation, describing the composite hybrid system. A useful method for accomplishing this is outlined in [6].

The discrete-time model of the continuous-time elements may then be characterized by the following generalized difference equation:

$$\begin{aligned} \bar{y}(k) = & b_1 \bar{x}(k) + b_2 \bar{x}(k-1) + \dots + b_{M+1} \bar{x}(k-M) + b_{M+2} \bar{y}(k-1) \\ & + b_{M+3} \bar{y}(k-2) + \dots + b_{2M+1} \bar{y}(k-M), \end{aligned} \quad (\text{II-18})$$

where $b_i, i=1, 2, \dots, 2M+1$, are constants determined by the system.

The difference equation describing the perturbed response of the

digital filter may be written in general form as

$$\begin{aligned} \bar{x}(k) = & \beta_1 \bar{y}(k) + \beta_2 \bar{y}(k-1) + \dots + \beta_{N+1} \bar{y}(k-N) + \beta_{N+2} \bar{x}(k-1) \\ & + \dots + \beta_i \bar{x}(k-i+N+1) + \dots + \beta_{2N+1} \bar{x}(k-N), \end{aligned} \quad (\text{II-19})$$

where N is the assumed order of the digital filter and β_i carries the same meaning as in the analysis of the open-loop digital filter configuration.

It can be seen in Figure 6 that perturbations of β_i affect not only the output $x(k)$ of the filter but also the input $y(k)$, which is coupled to $x(k)$ through the continuous-time elements. Therefore, the assumption that the filter input is independent of the filter coefficients is a luxury which is no longer available; at least not in the configuration of Figure 6. However, the filter input may be made insensitive to β_i by a simple modification to the structure of the system. Suppose that a correction term $c(k)$ defined by

$$c(k) = x(k) - \bar{x}(k) \quad (\text{II-20})$$

is generated and added to $\bar{x}(k)$, as depicted in Figure 7. This modification in effect results in a system output and digital filter input which are insensitive to perturbations in β_i and behave in a nominal manner regardless of the value of Δ_i . The generation of this correction term will now be considered.

The perturbed filter output $\bar{x}(k)$ and the nominal output $x(k)$ may be related through a Taylor series expansion in the variable β_i as follows:

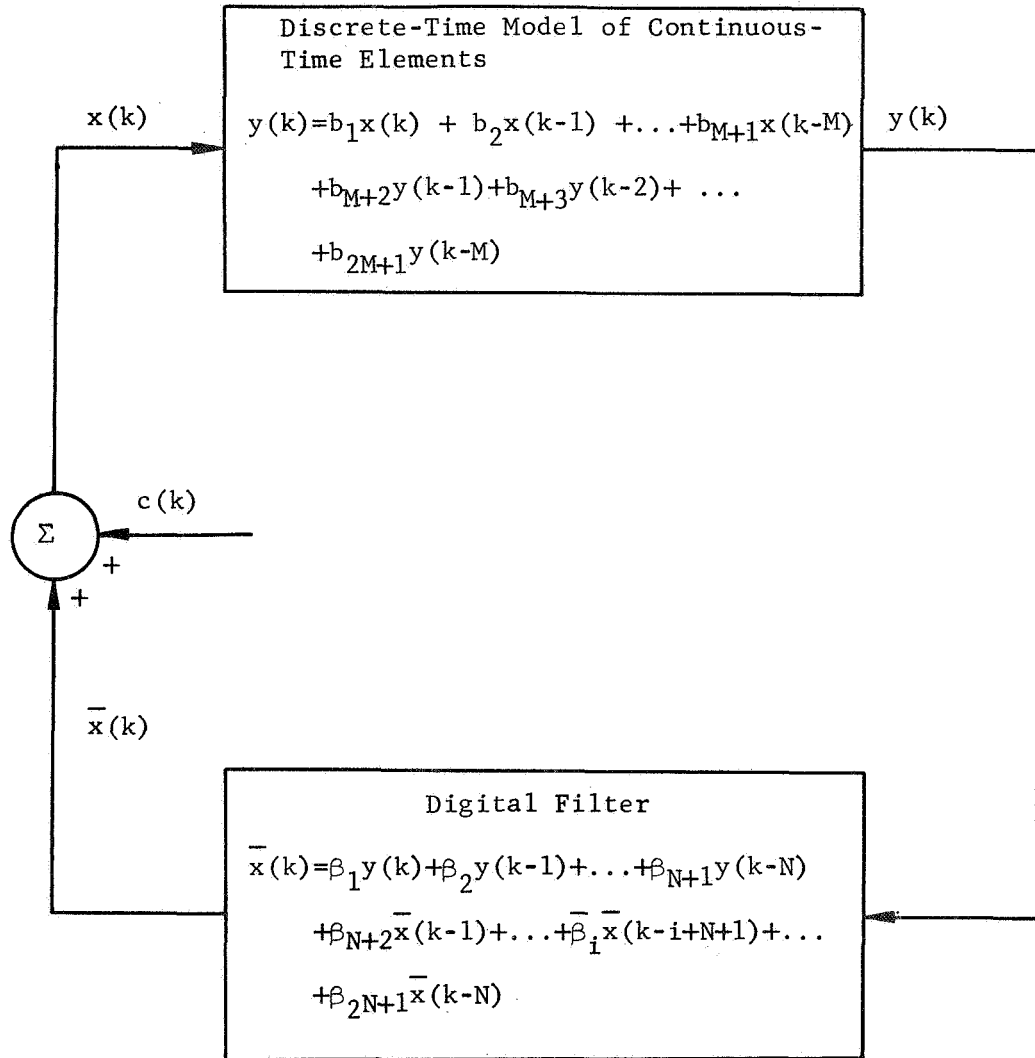


Fig. 7--Discrete-time model of modified hybrid system illustrating independence of $y(k)$ and Δ_i .

$$x(k) = \bar{x}(k) + \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} x^{(n)}(k) \Big|_{\beta_i = \bar{\beta}_i}, \quad (\text{II-21})$$

where the derivative notation conforms to the convention established in previous sections. Once again it is convenient to define an auxiliary variable $S(k, \Delta_i)$ as

$$S(k, \Delta_i) = \frac{1}{\Delta_i} \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} x^{(n)}(k) \Big|_{\beta_i = \bar{\beta}_i}, \quad (\text{II-22})$$

which in turn specifies $c(k)$ as

$$c(k) = \Delta_i S(k, \Delta_i). \quad (\text{II-23})$$

The derivative terms in $S(k, \Delta_i)$ may be written by inspection of the difference equation describing the nominal filter response, which is restated here for convenience:

$$\begin{aligned} x(k) = & \beta_1 y(k) + \dots + \beta_{N+1} y(k-N) + \beta_{N+2} x(k-1) + \dots \\ & + \beta_i x(k-i+N+1) + \dots + \beta_{2N+1} x(k-N). \end{aligned} \quad (\text{II-24})$$

Using the property that $y(k)$ is insensitive to variations of β_i , it can be shown from (II-22) and (II-24) that the recursive relationship defining $S(k, \Delta_i)$ in the corrected hybrid system of Figure 7 is

$$\begin{aligned} S(k, \Delta_i) = & \beta_{N+2} S(k-1, \Delta_i) + \dots + \bar{\beta}_i S(k-i+N+1, \Delta_i) + \dots \\ & + \beta_{2N+1} S(k-N, \Delta_i) + x(k-1+N+1). \end{aligned} \quad (\text{II-25})$$

The correction system defined by (II-25) responds to the nominal filter output sequence $x(k-i+N+1)$ initially from its zero state, since the filter initial conditions and coefficients are selected independently.

It is interesting to note the similarities of the correction coefficient equations in the above case and in the case of the open-loop digital filter configuration of the previous section. The characteristic equations are of the same form; i.e., identical to the corresponding perturbed filter characteristic equation. Furthermore, in both cases, the correction equation coefficients are precisely realizable by the digital filter. These similarities should not be unexpected, however, since the removal of the sensitivity constraints between $y(k)$ and β_i of the hybrid system leads to essentially the same set of conditions as were present in the open-loop digital filter configuration. Therefore, the implementations of the auxiliary correction coefficient equations associated with the open-loop digital filter (see Figure 4 and Figure 5) are also applicable to the hybrid system now under consideration. The composite corrected hybrid system, with the correction equation implementation shown schematically, is depicted by Figure 8.

A simple numerical example will now be considered in order to illustrate the applicability of the coefficient correction technique.

1. Example

Suppose that in Figure 6, the continuous-time elements are describable by the following first-order differential equation:

$$y^{(1)}(t) = 13.8 x(t) - 17.2 y(t) \quad (\text{II-26})$$

Further, suppose that the D/A converter operates as a perfect zero-order data-hold at the rate of 25 Hz. and that the effects of quantization of system variables are absent. Thus, the continuous-time elements, in conjunction with the D/A converter, may be modelled as a discrete-time

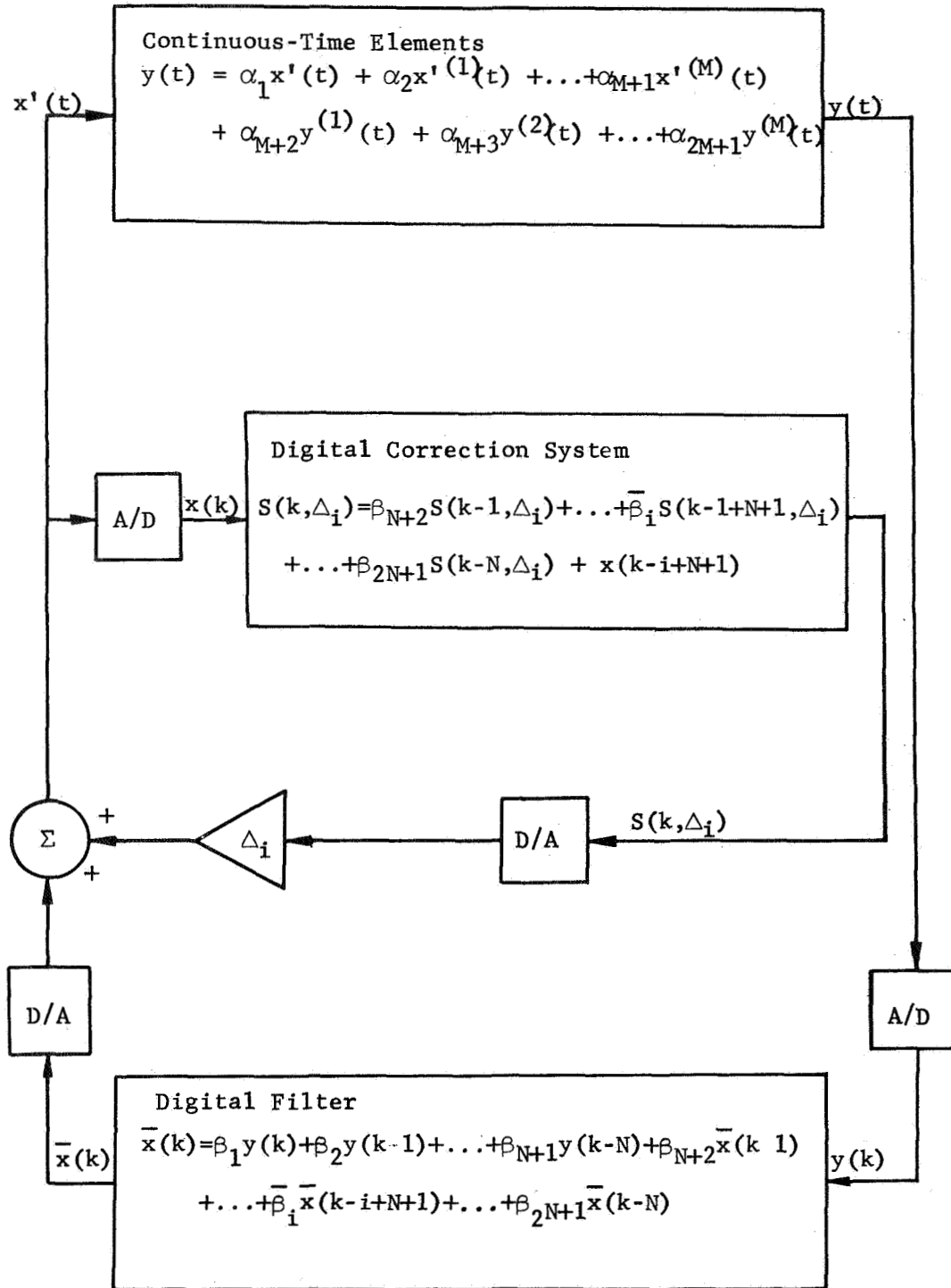


Fig. 8--Composite corrected hybrid system illustrating schematically the correction system implementation for a single denominator coefficient.

system by the following difference equation:

$$y(k) = 0.4 x(k-1) + 0.5 y(k-1) . \quad (\text{II-27})$$

Assume now that the nominal difference equation to be implemented by the digital filter is

$$x(k) = \beta_1 y(k) + \beta_2 x(k-1), \quad \beta_1 = 1.0, \beta_2 = 0.1 , \quad (\text{II-28})$$

and that the set B of realizable coefficients of the filter is given by

$$B = [\pm 0.125n; n = 0, 1, 2, \dots, 15]. \quad (\text{II-29})$$

Obviously, the nominal difference equation cannot be realized precisely, since β_2 , a denominator coefficient of the digital filter transfer function, is not a member of B . Therefore, a reasonable approximation to the nominal difference equation might be

$$\bar{x}(k) = \bar{y}(k) + 0.125\bar{x}(k-1), \quad (\text{II-30})$$

which corresponds to

$$\Delta_2 = \beta_2 - \bar{\beta}_2 = -0.025. \quad (\text{II-31})$$

The correction coefficient difference equation may be obtained directly from (II-30) using the generalized form of the correction equation developed previously in (II-25). Thus, in this example,

$$S(k, \Delta_2) = 0.125 S(k-1, \Delta_2) + x(k-1). \quad (\text{II-32})$$

In order to evaluate the performance of the correction scheme in this example, a digital computer simulation of the discretized nominal and corrected system difference equations was carried out. A program was generated for solving recursively the system equations given by

(II-20), (II-23), (II-27), (II,30), and (II-32).

The results of this simulation are depicted in Figure 9, where the nominal, the corrected, and the noncorrected system responses to initial conditions $x(0) = \bar{x}(0) = 100.0$ are compared. Note that the corrected and the nominal responses behave identically.

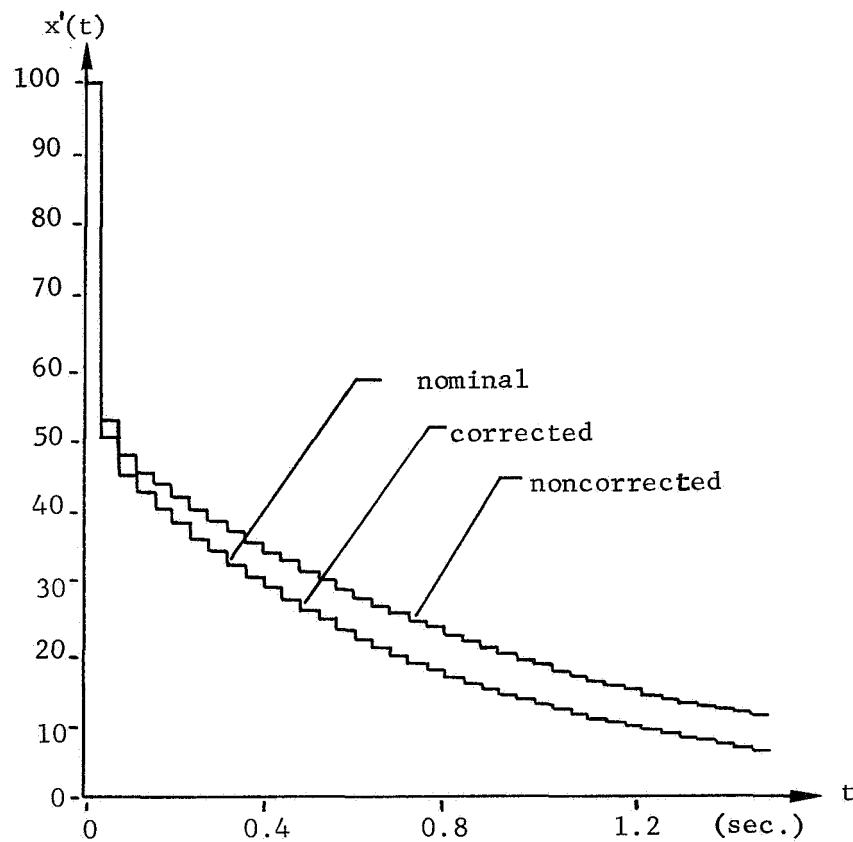


Fig. 9--Comparison of corrected and noncorrected responses.

C. Digital Filter With Several Corrected Coefficients

The remarks thus far have been limited, both in the case of the

digital filter and the closed-loop hybrid system, to the correction of a single coefficient of $D(z)$. However, it is often the case that there exist more than one coefficient of $D(z)$ which are not realizable by the digital element. Thus, it is desirable to extend the correction technique to the case of several perturbed coefficients.

The developemnt will begin with the correction of numerator and denominator coefficients in the open-loop digital filter configuration and then proceed to closed-loop hybrid systems. As might be expected, the Taylor series expansion in several variables will replace the expansion in one variable, which was introduced in the previous sections. Consequently, the related equations, for more than two variables, become extremely unwieldy. Therefore, in order to avoid obscuring the salient features of the technique with mathematical detail, the approach will be to develop the technique first on the basis of two perturbed coefficients and then to generalize to any number of coefficients.

1. Correction of several numerator coefficients of $D(z)$

Consider the digital filter response which results if two numerator coefficients of $D(z)$, for example, β_i and β_j , $i < j$ and $1 \leq i, j \leq N+1$, are perturbed Δ_i and Δ_j , respectively, due to the effects of quantization. From (II-2), the perturbed response may be expressed as

$$\begin{aligned} \bar{e}_o(k) = & \beta_1 e_i(k) + \dots + \bar{\beta}_i e_i(k-i+1) + \dots + \bar{\beta}_j e_i(k-j+1) + \dots \\ & + \beta_{N+1} e_i(k-N) + \beta_{N+2} \bar{e}_o(k-1) + \dots + \beta_{2N+1} \bar{e}_o(k-N). \end{aligned} \quad (\text{II-33})$$

Since $e_o(k)$ is a function of two perturbed coefficients at any sampling instant k , it is reasonable to assume that the correction

coefficient, or coefficients, must evolve from a Taylor series expansion in two variables, β_i and β_j . The nominal response $e_o(k)$ may be expanded about the perturbed response $\bar{e}_o(k)$ in a Taylor series as follows:

$$e_o(k) = \bar{e}_o(k) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(\Delta_i e_{oi}(k) + \Delta_j e_{oj}(k) \right)^{(n)} \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}}, \quad (\text{II-34})$$

where the notation under the summation in the above equation is defined by

$$\begin{aligned} \left(\Delta_i e_{oi}(k) + \Delta_j e_{oj}(k) \right)^{(1)} &= \Delta_i \frac{\partial e_o(k)}{\partial \beta_i} + \Delta_j \frac{\partial e_o(k)}{\partial \beta_j}, \\ \left(\Delta_i e_{oi}(k) + \Delta_j e_{oj}(k) \right)^{(2)} &= \Delta_i^2 \frac{\partial^2 e_o(k)}{\partial \beta_i^2} + 2\Delta_i \Delta_j \frac{\partial^2 e_o(k)}{\partial \beta_i \partial \beta_j} \\ &\quad + \Delta_j^2 \frac{\partial^2 e_o(k)}{\partial \beta_j^2}, \end{aligned} \quad (\text{II-35})$$

and so on. Note that the superscript symbolism in (II-35) now denotes partial derivatives instead of total derivatives as in (II-6) and (II-11).

At this point, two auxiliary variables, analogous to the correction coefficient $S(k, \Delta_i)$ for one perturbed coefficient, will be introduced.

Firstly, let

$$S_1(k, \Delta_i, \Delta_j) = \frac{1}{\Delta_i} \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_{oi}(k) \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}} + \frac{1}{\Delta_i} T_1(k, \Delta_i, \Delta_j) \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}} \quad (\text{II-36})$$

where $T_1(k, \Delta_i)$ is comprised of: (1) all cross-product terms in the summation of (II-34) having the property that the exponent of Δ_i is greater than the exponent of Δ_j , and (2) one half of each of the cross products having equal exponents of Δ_i and Δ_j . That is,

$$\begin{aligned}
T_1(k, \Delta_i, \Delta_j) = & \frac{\Delta_i \Delta_j}{2!} e_{oi j}^{(1)}(k) + \frac{3}{3!} \Delta_i^2 \Delta_j e_{oi j}^{(2)(1)}(k) + \frac{4}{4!} \Delta_i^3 \Delta_j e_{oi j}^{(3)(1)}(k) \\
& + \frac{3}{4!} \Delta_i^2 \Delta_j^2 e_{oi j}^{(2)(2)}(k) + \dots, \quad (II-37)
\end{aligned}$$

where $e_{oi j}^{(m)(n)}$ denotes $\partial^{m+n} e_o(k) / \partial \beta_i^m \partial \beta_j^n$.

The second correction coefficient $S_2(k, \Delta_i, \Delta_j)$ will be defined as

$$S_2(k, \Delta_i, \Delta_j) = \frac{1}{\Delta_j} \sum_{n=1}^{\infty} \frac{\Delta_j^n}{n!} e_{oj}^{(n)}(k) \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}} + \frac{1}{\Delta_j} T_2(k, \Delta_i, \Delta_j) \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}} \quad (II-38)$$

where the quantity $T_2(k, \Delta_i, \Delta_j)$ contains all cross-product terms of (II-34) not included in $T_1(k, \Delta_i, \Delta_j)$. The motivation for defining the above auxiliary variables is simply to permit the representation of the nominal response $e_o(k)$ as the sum of the perturbed response $\bar{e}_o(k)$ plus two additional correction terms, one for each perturbed coefficient; i.e.,

$$e_o(k) = \bar{e}(k) + \Delta_i S_1(k, \Delta_i, \Delta_j) + \Delta_j S_2(k, \Delta_i, \Delta_j). \quad (II-39)$$

It can be shown by substitution of the expression for $e_o(k)$, given by (II-2), into the correction coefficient equations, (II-36) and (II-39), and after considerable rearrangement of resulting terms, that the correction coefficients satisfy the following difference equations:

$$\begin{aligned}
S_1(k, \Delta_i, \Delta_j) = & e_i(k-i+1) + \beta_{N+2} S_1(k-1, \Delta_i, \Delta_j) + \beta_{N+3} S_1(k-2, \Delta_i, \Delta_j) \\
& + \dots + \beta_{2N+1} S_1(k-N, \Delta_i, \Delta_j), \quad (II-40)
\end{aligned}$$

and

$$\begin{aligned}
S_2(k, \Delta_i, \Delta_j) = & e_i(k-j+1) + \beta_{N+2} S_2(k-1, \Delta_i, \Delta_j) + \beta_{N+3} S_2(k-2, \Delta_i, \Delta_j) \\
& + \dots + \beta_{2N+1} S_2(k-N, \Delta_i, \Delta_j). \quad (II-41)
\end{aligned}$$

Once again, all of the partial derivatives of the input $e_i(k)$ with respect to β_i or β_j are equal to zero, since the input and the coefficients are independent of each other. Similarly, it is assumed in the development of (II-40) and (II-41) that each of the filter coefficients are independent of the other coefficients. Furthermore, it is evident that the systems represented by (II-40) and (II-41) respond to the delayed input sequences $e_i(k-i+1)$ and $e_i(k-j+1)$, respectively, from the system zero states, since the selection of the digital filter initial conditions is independent of the filter coefficients.

Note the similarity of the correction coefficient equations for the case of two perturbed numerator coefficients, (II-40) and (II-41), and the case of one perturbed numerator coefficient (II-9). Each of the three correction systems have the same characteristic equation as that of the corresponding digital filter with quantized coefficients. Consequently, it is apparent that the same types of correction system realizations may be employed in digital filters with two perturbed numerator coefficients as for filters with one perturbed coefficient. The additional perturbed coefficient simply adds another correction system in parallel with the digital filter, as shown in Figure 10.

Verification and extension to more than two numerator coefficients of $D(z)$. The validity of the above technique may be easily substantiated by an examination of Figure 10. Since the response of the corrected system is stated to be identical to the nominal digital filter response, the z-transfer function relating the input and the output of the corrected system should be identical to $D(z)$. This, as may be seen in Figure 10, is obviously the case.

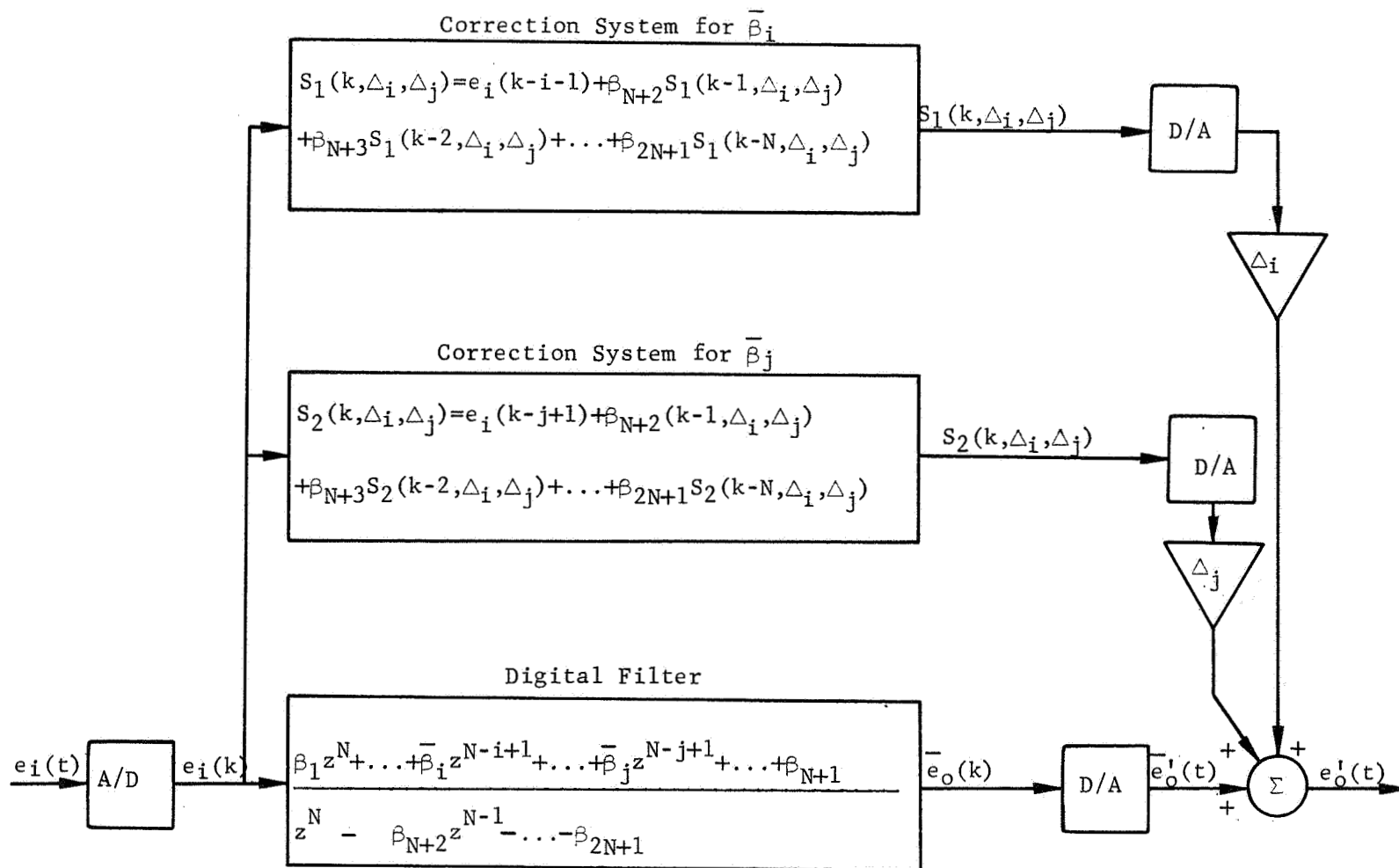


Fig. 10--Schematic implementation of coefficient correction system for two numerator coefficients of $D(z)$.

Furthermore, it is not necessary to resort again to the Taylor series expansion method in more than two variables in order to infer that for N quantized numerator coefficients, there will be N separate auxiliary correction systems in parallel with the digital filter in the corrected system configuration.

The general form of the n th numerator coefficient correction system is evident from Figure 10. The system input is $e_i(k-n+1)$, its characteristic equation is identical to that of the digital filter with quantized coefficients, and the associated analog multiplier is $\Delta_n = \beta_n - \bar{\beta}_n$.

2. Correction of several denominator coefficients of $D(z)$

The application of the correction technique to two denominator coefficients of $D(z)$ is, in principle, equivalent to the development for one perturbed coefficient. Suppose, for instance, that the denominator coefficients, β_i and β_j ; $i < j$ and $N+1 \leq i, j \leq 2N+1$, are perturbed by Δ_i and Δ_j , respectively, as a consequence of the quantization process. Then, the perturbed digital filter response $\bar{e}_o(k)$ may be characterized as

$$\begin{aligned} \bar{e}_o(k) = & \beta_1 e_i(k) + \beta_2 e_i(k-1) + \dots + \beta_{N+1} e_i(k-N) + \beta_{N+2} \bar{e}_o(k-1) \\ & + \dots + \bar{\beta}_i \bar{e}_o(k-i+N+1) + \dots + \bar{\beta}_j \bar{e}_o(k-j+N+1) + \dots \\ & + \beta_{2N+1} \bar{e}_o(k-N) \end{aligned} \quad (\text{II-42})$$

Furthermore, the nominal and the perturbed responses may be related through a Taylor series expansion in β_i and β_j as follows:

$$e_o(k) = \bar{e}_o(k) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(\Delta_i e_{oi}^{(1)}(k) + \Delta_j e_{oj}^{(1)}(k) \right)^{(n)} \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}}, \quad (\text{II-43})$$

where the terms in the above summation carry the same meanings as previously established in (II-35); i.e.,

$$\begin{aligned} \left(\Delta_i e_{oi}^{(1)} + \Delta_j e_{oj}^{(1)} \right)^{(1)} &= \Delta_i \frac{\partial e_o(k)}{\partial \beta_i} + \Delta_j \frac{\partial e_o(k)}{\partial \beta_j} , \\ \left(\Delta_i e_{oi}^{(1)} + \Delta_j e_{oj}^{(1)} \right)^{(2)} &= \Delta_i^2 \frac{\partial^2 e_o(k)}{\partial \beta_i^2} + 2\Delta_i \Delta_j \frac{\partial^2 e_o(k)}{\partial \beta_i \partial \beta_j} \\ &\quad + \Delta_j^2 \frac{\partial^2 e_o(k)}{\partial \beta_j^2} , \end{aligned} \quad (\text{II-35})$$

and so on.

It is advantageous at this point to separate the infinite summation of correction terms which relate $e_o(k)$ and $\bar{e}_o(k)$ in (II-43) into two distinct components, one being associated with Δ_i and the other with Δ_j . In order to accomplish this, two auxiliary correction coefficients, $S_1(k, \Delta_i, \Delta_j)$ and $S_2(k, \Delta_i, \Delta_j)$, will be defined such that

$$e_o(k) = \bar{e}_o(k) + S_1(k, \Delta_i, \Delta_j) + S_2(k, \Delta_i, \Delta_j) \quad (\text{II-44})$$

More explicitly, let the first coefficient be defined as

$$S_1(k, \Delta_i, \Delta_j) = \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_{oi}^{(n)}(k) \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}} + T_1(k, \Delta_i, \Delta_j) \bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}} , \quad (\text{II-45})$$

where the variable $T_1(k, \Delta_i, \Delta_j)$ assumes the same meaning as given previously by (II-37); i.e.,

$$T_1(k, \Delta_i, \Delta_j) = \frac{\Delta_i \Delta_j}{2!} e_{oij}^{(1)(1)}(k) + \frac{3}{3!} \Delta_i^2 \Delta_j e_{oij}^{(2)(1)}(k) + \frac{4}{4!} \Delta_i^3 \Delta_j e_{oij}^{(3)(1)}(k) \\ + \frac{3}{4!} \Delta_i^2 \Delta_j^2 e_{oij}^{(2)(2)}(k) + \dots, \quad (\text{II-37})$$

and so on, (recall that $e_{oij}^{(m)(n)}$ denotes $\partial^{m+n} e_o(k) / \partial \beta_i^m \partial \beta_j^n$). Consequently, the second auxiliary correction coefficient $S_2(k, \Delta_i, \Delta_j)$ may be defined in terms of $T_2(k, \Delta_i, \Delta_j)$ of (II-38), as follows:

$$S_2(k, \Delta_i, \Delta_j) = \sum_{n=1}^{\infty} \frac{\Delta_j^n}{n!} e_{oj}^{(n)}(k) \left| \begin{array}{l} \beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j \end{array} \right. + T_2(k, \Delta_i, \Delta_j) \left| \begin{array}{l} \beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j \end{array} \right. \quad (\text{II-46})$$

Now, on the basis of the correction coefficient definitions, (II-45) and (II-46) stated above, the problem of synthesizing the auxiliary correction system equations may be taken up.

In order to generate these equations, however, a somewhat different approach than was employed in the numerator coefficient correction scheme will be taken. Rather than consider $S_1(k, \Delta_i, \Delta_j)$ and $S_2(k, \Delta_i, \Delta_j)$ separately, it is convenient to deal with the sum $S_1(k, \Delta_i, \Delta_j) + S_2(k, \Delta_i, \Delta_j)$. The method of solution will be, firstly, to substitute the expression for the nominal digital filter response (II-2) into the combined form of $S_1(k, \Delta_i, \Delta_j) + S_2(k, \Delta_i, \Delta_j)$ from (II-45) and (II-46), and secondly, to attempt to restate the resultant form as a recursive relationship in terms of the delayed correction coefficient sequences and the delayed nominal filter output sequence.

Direct substitution of the nominal filter output $e_o(k)$ from (II-2) into the expression for $S_1(k, \Delta_i, \Delta_j) + S_2(k, \Delta_i, \Delta_j)$ results (after a slight rearrangement of terms and with the previously stated assumptions

concerning independence of filter coefficients and the input) in the following expression:

$$\begin{aligned}
S_1(k, \Delta_i, \Delta_j) + S_2(k, \Delta_i, \Delta_j) = & \\
& \left\{ \beta_{N+2} \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_{oi}^{(n)}(k-1) + \beta_{N+2} T_1(k-1, \Delta_i, \Delta_j) + \dots \right. \\
& + \bar{\beta}_i \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_{oi}^{(n)}(k-i+N+1) + \bar{\beta}_i T_1(k-i+N+1, \Delta_i, \Delta_j) + \dots \\
& + \bar{\beta}_j \sum_{n=1}^{\infty} \frac{\Delta_j^n}{n!} e_{oi}^{(n)}(k-i+N+1) + \bar{\beta}_j T_1(k-j+N+1, \Delta_i, \Delta_j) + \dots \\
& + \beta_{2N+1} \sum_{n=0}^{\infty} \frac{\Delta_i^n}{n!} e_{oi}^{(n)}(k-N) + \beta_{2N+1} T_1(k-N, \Delta_i, \Delta_j) \\
& + \beta_{N+2} \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_{oj}^{(n)}(k-1) + \beta_{N+2} T_2(k-1, \Delta_i, \Delta_j) + \dots \\
& + \bar{\beta}_i \sum_{n=1}^{\infty} \frac{\Delta_i^n}{n!} e_{oj}^{(n)}(k-i+N+1) + \bar{\beta}_i T_2(k-i+N+1, \Delta_i, \Delta_j) + \dots \\
& + \bar{\beta}_j \sum_{n=1}^{\infty} \frac{\Delta_j^n}{n!} e_{oj}^{(n)}(k-j+N+1) + \bar{\beta}_j T_2(k-j+N+1, \Delta_i, \Delta_j) + \dots \\
& + \beta_{2N+1} \sum_{n=1}^{\infty} \frac{\Delta_j^n}{n!} e_{oj}^{(n)}(k-N) + \beta_{2N+1} T_2(k-N, \Delta_i, \Delta_j) \\
& + \Delta_i e_o(k-i+N+1) + \Delta_i \sum_{n=1}^{\infty} \frac{1}{n!} \left(\Delta_i e_{oi}^{(1)}(k-i+N+1) + \Delta_j e_{oj}^{(1)}(k-i+N+1) \right)^{(n)} \\
& + \Delta_j e_o(k-j+N+1) + \Delta_j \sum_{n=1}^{\infty} \frac{1}{n!} \left(\Delta_i e_{oi}^{(1)}(k-j+N+1) + \Delta_j e_{oj}^{(1)}(k-j+N+1) \right)^{(n)} \left. \vphantom{\sum_{n=1}^{\infty}} \right\} \Bigg|_{\substack{\beta_i = \bar{\beta}_i \\ \beta_j = \bar{\beta}_j}}
\end{aligned}$$

Notice that the last four quantities in (II-47) are actually the Taylor series representations of the delayed nominal variables $\Delta_i e_o(k-i+N+1)$ and $\Delta_j e_o(k-j+N+1)$. In addition, each of the remaining terms are of the same form as the correction coefficients defined in (II-45) and (II-46). Consequently, it is possible to rewrite (II-47) as follows:

$$\begin{aligned}
S_1(k, \Delta_i, \Delta_j) + S_2(k, \Delta_i, \Delta_j) = & \\
& \beta_{N+2} S_1(k-1, \Delta_i, \Delta_j) + \dots + \bar{\beta}_i S_1(k-i+N+1, \Delta_i, \Delta_j) + \dots \\
& + \bar{\beta}_j S_1(k-j+N+1, \Delta_i, \Delta_j) + \dots + \beta_{2N+1} S_1(k-N, \Delta_i, \Delta_j) \\
& + \beta_{N+2} S_2(k-1, \Delta_i, \Delta_j) + \dots + \bar{\beta}_i S_2(k-i+N+1, \Delta_i, \Delta_j) + \dots \\
& + \bar{\beta}_j S_2(k-j+N+1, \Delta_i, \Delta_j) + \dots + \beta_{2N+1} S_2(k-N, \Delta_i, \Delta_j) \\
& + \Delta_i e_o(k-i+N+1) + \Delta_j e_o(k-j+N+1) . \tag{II-48}
\end{aligned}$$

It is interesting to digress for a moment and note the effect of combining $S_1(k, \Delta_i, \Delta_j)$ and $S_2(k, \Delta_i, \Delta_j)$ into a single expression, as was done in (II-47). The motivation for this step becomes evident in the development of (II-48). As a result of this combination, each of the required partial derivative terms in the Taylor series representations of $e_o(k-i+N+1)$ and $e_o(k-j+N+1)$ are generated, which is desirable since these variables are physically available for use as inputs to the correction system. If $S_1(k, \Delta_i, \Delta_j)$ and $S_2(k, \Delta_i, \Delta_j)$ had been considered separately, the system variables $e_o(k-i+N+1)$ and $e_o(k-j+N+1)$ would not have appeared explicitly.

The difference equation given by (II-48) may now be employed to

realize the correction systems for $\bar{\beta}_i$ and $\bar{\beta}_j$. This equation may be implemented in a number of ways; however, perhaps the most convenient method is to treat (II-48) as the sum of two separate correction system equations, each having zero initial conditions. More specifically,

$$\begin{aligned} S_1(k, \Delta_i, \Delta_j) = & \beta_{N+2} S_1(k-1, \Delta_i, \Delta_j) + \dots + \bar{\beta}_i S_1(k-i+N+1, \Delta_i, \Delta_j) \\ & + \dots + \bar{\beta}_j S_1(k-j+N+1, \Delta_i, \Delta_j) + \dots \\ & + \beta_{2N+1} S_1(k-N, \Delta_i, \Delta_j) + \Delta_i e_o(k-i+N+1), \end{aligned} \quad (\text{II-49})$$

and

$$\begin{aligned} S_2(k, \Delta_i, \Delta_j) = & \beta_{N+2} S_2(k-1, \Delta_i, \Delta_j) + \dots + \bar{\beta}_i S_2(k-i+N+1, \Delta_i, \Delta_j) \\ & + \dots + \bar{\beta}_j S_2(k-j+N+1, \Delta_i, \Delta_j) + \dots \\ & + \beta_{2N+1} S_2(k-N, \Delta_i, \Delta_j) + \Delta_j e_o(k-j+N+1). \end{aligned} \quad (\text{II-50})$$

Therefore, it can be seen that (II-44), (II-49), and (II-50) completely specify the correction system for the open-loop digital filter configuration presently under consideration.

A discrete-time model of the implementation of (II-44), (II-49), and (II-50), illustrating the required D/A and A/D interfaces and the analog multipliers Δ_i and Δ_j is shown in Figure 11. Furthermore, since the correction equations (II-49) and (II-50) are of the same general form as (II-16), which was derived for a single corrected denominator coefficient, the generalized implementation schemes depicted in Figure 4 and Figure 5 may also be employed in correcting two denominator coefficients. The additional perturbed coefficient simply adds one more auxiliary system in parallel with the first coefficient correction

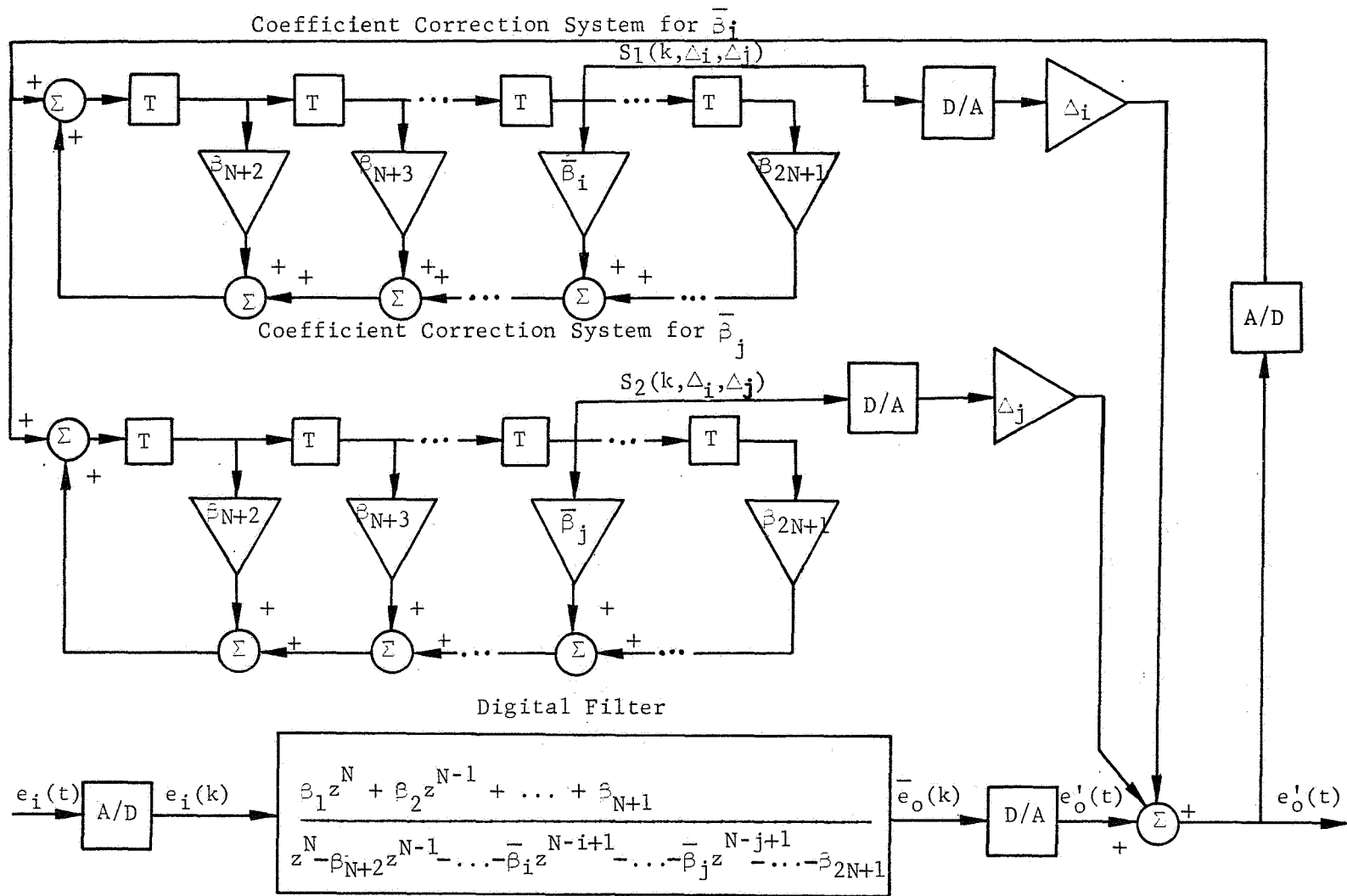


Fig. 11--Discrete-time model of coefficient correction system for two denominator coefficients of $D(z)$.

system (see Figure 11).

Verification and Extension to More than Two Denominator Coefficients of $D(z)$. The validity of the preceding coefficient correction method may be independently verified by inspection of Figure 11. In order for the corrected and the nominal system time responses to be equivalent, it is required that the composite z -transfer functions of the corrected and the nominal configurations be identical. From Figure 11, it can be seen that the corrected system transfer function is in fact identical to $D(z)$; i.e.,

$$\frac{E_o(z)}{E_i(z)} = A(z)B(z) = D(z) \quad (\text{II-51})$$

where $A(z)$ is the digital filter transfer function with quantized coefficients

$$A(z) = \frac{\beta_1 z^N + \beta_2 z^{N-1} + \dots + \beta_{N+1}}{z^N - \dots - \bar{\beta}_i z^{N-i+1} - \dots - \bar{\beta}_j z^{N-j+1} - \dots - \beta_{2N+1}}, \quad (\text{II-52})$$

and $B(z)$ is the overall correction system transfer function,

$$B(z) = \frac{1}{1 - \frac{\Delta_i z^{N-i+1} + \Delta_j z^{N-j+1}}{\beta_{N+2} z^N - \dots - \bar{\beta}_i z^{N-i+1} - \dots - \bar{\beta}_j z^{N-j+1} - \dots - \beta_{2N+1}}} \quad (\text{II-53})$$

Thus, the verification is completed.

Further extension of the above correction technique to any number of perturbed denominator coefficients of $D(z)$ is self-evident from Figure 11 and (II-53). It can be seen that for N perturbed denominator coefficients, there will be N auxiliary correction systems in parallel, each

one having the same characteristic equation as the digital filter with quantized coefficients. Furthermore, the input to the n th correction system will be the delayed nominal output $e_o(k-n+N+1)$ multiplied in analog form by $\Delta_n = \beta_n - \bar{\beta}_n$.

It should be noted at this point that even though the numerator and the denominator coefficient correction methods have been treated separately in the preceding discussion, the techniques may actually be employed simultaneously for any given $D(z)$. This is easily demonstrated by initially correcting the numerator coefficients of $D(z)$ and then by viewing the resulting transfer function as an intermediate function $D'(z)$ with perturbed denominator coefficients. It is a simple matter then to correct the denominator coefficients of $D'(z)$ using the technique presented in this section. The final result is, therefore, an overall system transfer function equivalent to $D(z)$.

It will become apparent in the following section that the above arguments also apply to the hybrid configuration of Figure 6.

D. Hybrid Feedback Control Systems With Several Corrected Coefficients

Consider now the general class of hybrid systems illustrated in Figure 6. However, instead of assuming only a single perturbed denominator or numerator coefficient in the digital element, suppose that several coefficients must be corrected.

It has been demonstrated, both by the use of the Taylor series expansion method and by z -transform analysis, that it is possible to precisely correct for any number of numerator or denominator quantization

errors in the open-loop digital filter configuration; i.e., the nominal and the corrected system input-output characteristics have been proven to be equivalent.

Therefore, since the objective in correcting the hybrid system is to modify the digital feedback element such that its input-output characteristics are nominal, it is apparent that the multiple coefficient correction schemes advanced for the open-loop case are applicable, without modification, to the closed-loop system of Figure 6. The continuous-time elements have no effect on the correction system implementation.

E. Some Practical Considerations

Note that the correction systems discussed to this point have been based on the premise that only digital elements would be used to realize the correction coefficient difference equations. Therefore, each of the proposed implementations have required a separate digital element and an A/D - D/A interface combination for each perturbed coefficient. This obviously increases hardware requirements and introduces additional sources of errors due to quantization of system variables. Therefore, it is advantageous to simplify the correction system implementations whenever possible.

One potential area for economization of hardware requirements is in the A/D conversion equipment necessary in the correction denominator coefficients of $D(z)$. Consider, for instance, the denominator coefficient correction system modelled in Figure 5. This implementation requires one A/D and one D/A conversion for each corrected denominator coefficient.

However, the overall transfer function, $E_o(z)/E_i(z)$, of the corrected digital filter is unchanged if the correction system is removed from its location in Figure 5 and cascaded ahead of the digital filter as shown in Figure 12. By this simple modification, the A/D converter at the input of the filter may also be used as part of the correction system; and consequently, no additional A/D converters are required to correct for denominator coefficient errors.

Another area where hardware requirements might be reduced is in the implementation of multiple-coefficient correction systems, as typified in Figure 10. If special purpose hardware is to be utilized to correct more than one coefficient of $D(z)$, the implementation may be achieved by a single Nth-order correction system, rather than by a separate Nth-order system for each corrected coefficient.

The use of a single Nth-order system for correcting more than one coefficient (in this case two, β_i and β_j) is illustrated in Figure 13. Although this system performs the same function as that of Figure 10, it is evident that the realization which is employed in Figure 13 requires only half as many digital operations as that of Figure 10. Consequently, a considerable saving in equipment may be achieved.

F. A Note on Other Realizations of $D(z)$

From the outset it was assumed that $D(z)$ was to be physically realized by a single Nth-order digital system. This was the motivation for expressing $D(z)$ in the generalized form of (II-1). However, as previously stated, there are in many cases advantages to realizing $D(z)$

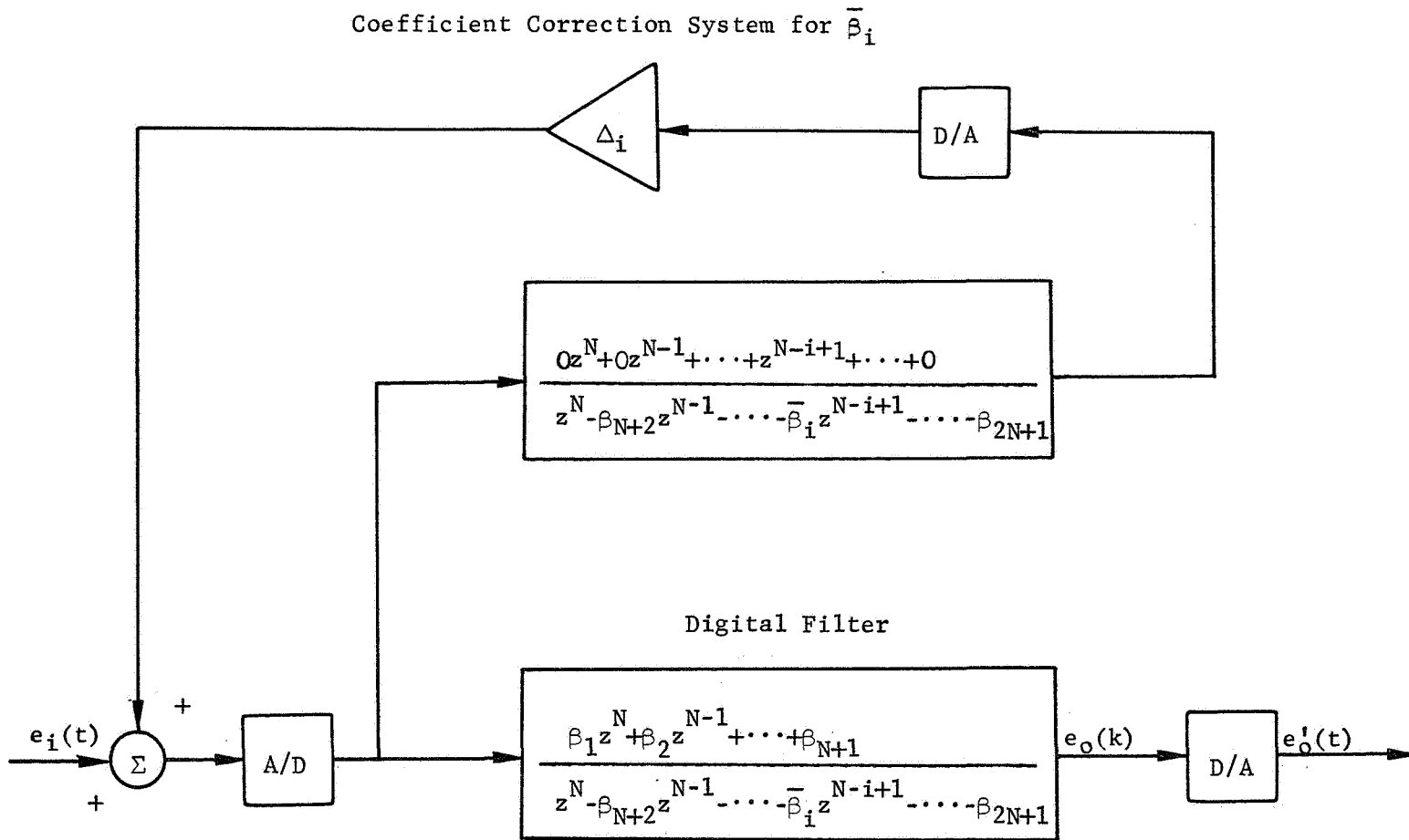


Fig. 12--Schematic representation of simplified denominator coefficient correction system implementation.

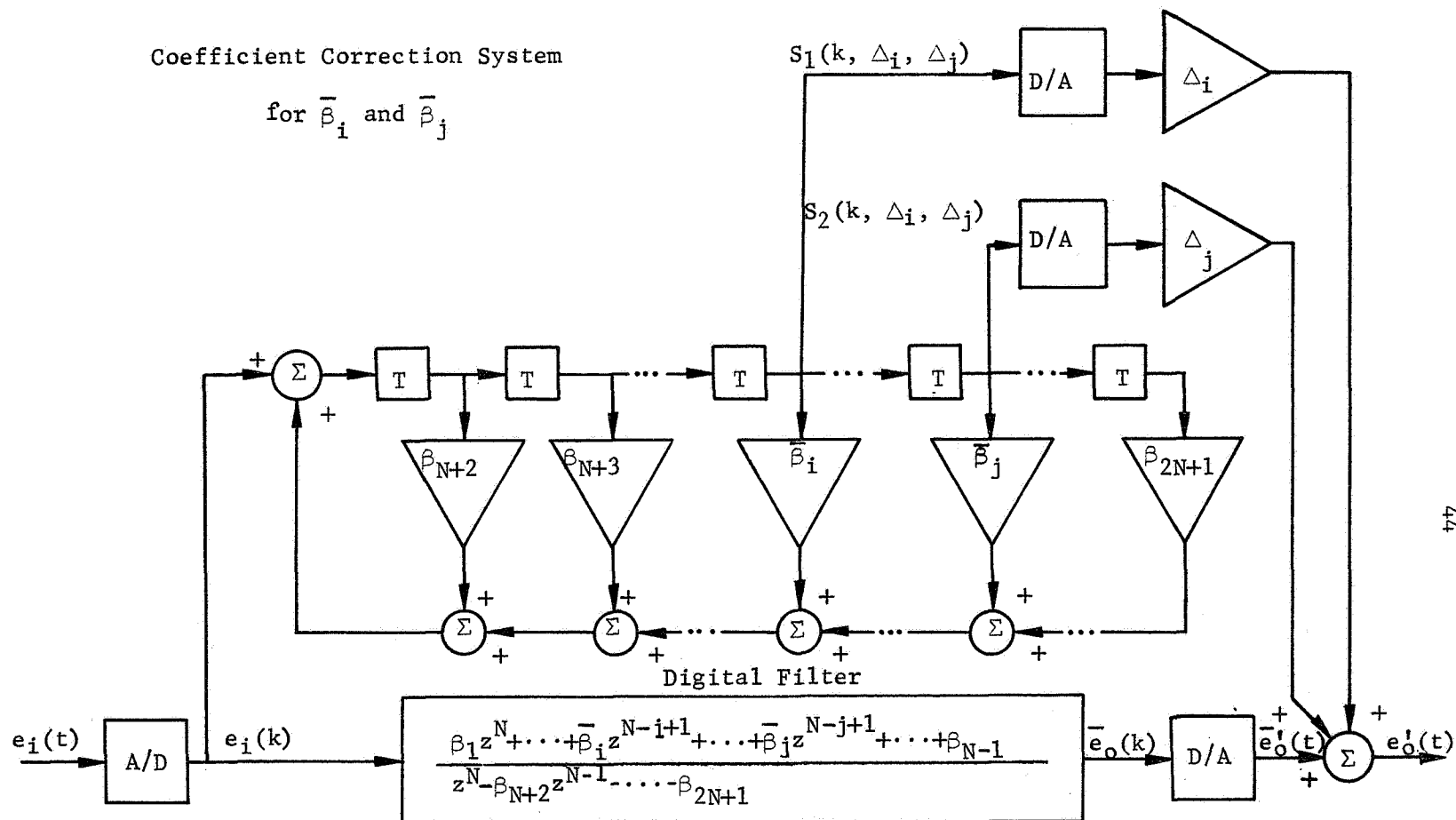


Fig. 13--Discrete-time model of simplified correction system for two numerator coefficients of $D(z)$.

as a combination of first- and second-order parallel or cascaded subsystems, which of course requires that $D(z)$ be expressed in partial fractioned form or in factored form.

If this realization of $D(z)$ is performed, it can be seen that the coefficients become decoupled; that is, the correction of quantized coefficients in one subsystem may be executed independently of the coefficients of the remaining subsystems. It is important to note, however, that each of correction techniques developed for the generalized $D(z)$ of (II-1) is directly applicable to the combination of first- and second-order systems. The only difference is that the correction systems must be implemented for several individual low order systems rather than for a single Nth-order system.

III. SELECTION OF OPTIMAL QUANTIZED COEFFICIENTS

The techniques presented in the preceding chapter for correcting quantized coefficients of $D(z)$ are sometimes impractical to implement, primarily due to the additional hardware which is necessitated. Consequently, the designer is sometimes faced with the problem of approximating the nominal $D(z)$ as closely as possible by a suitable selection of quantized coefficients. This immediately leads to the questions: "What measure of 'closeness' is to be used, and how can the selection of quantized coefficients be systematized?"

In this chapter the above problems will be investigated.

A. Performance Index

Before it is possible to systematize the selection of quantized coefficients of $D(z)$, a performance index must be defined which is, in some sense, indicative of the nearness of the performance of the approximated $D(z)$ to that of the nominal $D(z)$. This performance index must include factors which reflect numerically the design requirements for the overall hybrid system. Therefore, it is apparent that the formulation of the performance index is not independent of the designer's judgement.

Perhaps the most fundamental design requirement which influences the selection of quantized coefficients of $D(z)$ is that of system stability;

i.e., the degradation of system stability due to the approximation of the nominal $D(z)$ coefficients must be held to a minimum. Of course, there may be other design specifications, such as static accuracy, system bandwidth, response time, maximum overshoot, etc., which must also be maintained as nearly as possible to nominal in the selection of quantized coefficients.

Each of the above design specifications is reflected in the shape of the gain-phase plot of $D(z)$ as $z = \exp(sT)$ traverses the unit circle in the z -plane, or as s varies from $-j\pi/T$ to $j\pi/T$ in the s -plane. For instance, if it is required that the static accuracy of the hybrid system be unchanged after the coefficients of $D(z)$ are quantized, it is necessary that $D(1)$ remain unchanged after coefficient quantization takes place. Further, if the relative stability of the hybrid system is to be the same before and after quantization of the coefficients of $D(z)$, it is necessary that the critical phase margins and gain margins of the system be equivalent before and after quantization takes place. In other words, if z_i is a value on the unit circle in the z -plane corresponding to a critical stability margin, it is necessary that $D(z_i)$ be approximately the same before and after quantization of coefficients is effected.

In general, the designer may emphasize any desired combination of design specifications when selecting quantized coefficients of $D(z)$ by simply requiring that the departure from nominal of the gain-phase plot of $D(z)$ be minimized over the ranges of z on the unit circle which are associated with these specifications. There are several ways in which the above quantization policy might be incorporated in a performance

index. However, for the purposes of the developments of this chapter, the following performance index will be employed:

$$J(\underline{\beta}) = \sum_{i=1}^K \left\{ \lambda_i \frac{\operatorname{Re} [D(z_i) - \bar{D}(z_i)]}{\operatorname{Re} [D(z_i)]} \right\}^2 + \sum_{i=1}^K \left\{ \lambda_i \frac{\operatorname{Im} [D(z_i) - \bar{D}(z_i)]}{\operatorname{Im} [D(z_i)]} \right\}^2 \quad (\text{III-1})$$

where z_i , $i = 1, 2, \dots, K$, are the predetermined critical values of z on the unit circle at which the perturbations of the gain-phase plot of $D(z)$ must be minimized. The symbol $\bar{D}(z_i)$ denotes the value of $D(z_i)$ obtained after the nominal coefficients have been replaced by quantized coefficients, and λ_i , $i = 1, 2, \dots, K$, represent weighting constants to be selected by the designer. Consequently, minimizing the performance index given by (III-1) is equivalent to minimizing the sums of the squares of the percentage changes in the real and the imaginary parts of $D(z_i)$, $i = 1, 2, \dots, K$, in a given set of quantized coefficients. No attempt will be made here to show that this particular performance index is superior to any other. Moreover, the principle justification which will be offered for its use is that experience has shown it to be satisfactory in practical problems.

Before proceeding to the development of a numerical method for minimizing the performance index, there are a number of definitions which should be stated.

B. Definitions

Definition III-1: Let E_{2N+1}^{β} denote the space with coordinates

defined by the coefficients of $D(z)$. A "point" or "vector" in E_{2N+1}^β means the number sequence of $2N+1$ terms $[\beta_1, \beta_2, \dots, \beta_{2N+1}]$ and will be denoted by the single underlined letter $\underline{\beta}$.

Definition III-2: Suppose D represents a digital element which is to be used to realize $D(z)$. Let Q denote the set of points in E_{2N+1}^β to which $\underline{\beta}$ belongs if and only if β_i , $i = 1, 2, \dots, 2N+1$, can be realized precisely by D . In other words, Q is the union of all possible quantized coefficient vectors associated with D .

Comment: It should be noted that the set Q corresponding to any physically realizable digital device for implementing $D(z)$ will be a finite and bounded set in E_{2N+1}^β . Further note that Q is the set of candidate points on which $J(\underline{\beta})$ is to be minimized.

Definition III-3: Let $\underline{\beta}$ and $\underline{\beta}'$ be points of Q . The statement that " $\underline{\beta}$ and $\underline{\beta}'$ are adjacent points in Q " means that there exists one integer $j \in [1, 2N+1]$ such that

- (1) $\beta_i = \beta'_i$ for all $i \leq 2N+1$ except for $i = j$
- (2) $\beta_j \neq \beta'_j$, and
- (3) there is no member $\underline{\beta}'' \in Q$ with the property that $\beta_j < \beta''_j < \beta'_j$ or $\beta_j > \beta''_j > \beta'_j$.

Definition III-4: Let $\underline{\beta}'$ be a member of Q . The statement that "the performance index J has a local minimum at $\underline{\beta}'$ in Q " means that if $\underline{\beta}'' \in Q$ and $\underline{\beta}''$ is adjacent to $\underline{\beta}'$, then $J(\underline{\beta}') \leq J(\underline{\beta}'')$.

Definition III-5: The statement that " $\underline{\beta}'$ is an optimal set of

quantized coefficients of $D(z)$ " means that $\underline{\beta}' \in Q$ and $J(\underline{\beta}')$ is a local minimum in Q .

On the basis of the above definitions, a numerical technique for optimizing $J(\underline{\beta})$ may now be developed.

C. Minimization of $J(\underline{\beta})$

There are a number of algorithms available for minimizing functions of several variables [7-9]. However, the minimization of $J(\underline{\beta})$ does not lend itself readily to any of these methods. There are essentially two reasons for this: (1) most of the existing minimization procedures require that the partial derivatives of $J(\underline{\beta})$ be derived with respect to each of the components of $\underline{\beta}$, which is extremely impractical for anything other than very low order forms of $D(z)$, and (2) the search for minima of $J(\underline{\beta})$ must be confined to the set $Q \subseteq E_{2N+1}^{\beta}$, whereas constraints of this type are not incorporated in existing techniques. It is therefore evident that in order to make use of the available methods, they must be modified to circumvent the above problems.

The two procedures which have proven to be most successful in minimizing the performance index are: (1) a modification of the steepest descent method, and (2) a modification of a method described by Resenbrock [8] in which one quantized coefficient is varied at a time. The first technique converges more rapidly than the second; however, due to the nature of the performance index, the second method is sometimes required for higher precision. This aspect will be discussed in more detail as the minimization techniques are evolved.

1. Modification of the steepest descent method

The steepest descent method in its normal form consists of finding the direction of steepest descent of the function $X(\alpha_1, \alpha_2, \dots, \alpha_N)$, which is to be minimized, as $\underline{\alpha}$ varies from some initial starting point $\underline{\alpha}^0$ in E_N . The direction of steepest descent from $\underline{\alpha}^0$ is given by

$$\underline{\zeta}^0 = -\text{grad} [X(\underline{\alpha}^0)] , \quad (\text{III-2})$$

where $\underline{\zeta}^0$ is a vector in the required direction. The value of $X(\underline{\alpha})$ is then calculated along a line from the starting point parallel to $\underline{\zeta}^0$ until the least value is attained. There are no restrictions on the incremental step-length to be used. Starting from the point of least $X(\underline{\alpha})$ on this line, the process is repeated and a new direction of steepest descent is determined. The procedure continues until $X(\underline{\alpha})$ can be decreased no further.

In order to apply the steepest descent principle to $J(\underline{\beta})$, the following modifications must be effected: (1) the partial derivatives must be numerically approximated, and (2) the choice of step-lengths in the directions of steepest descent must be made such that $J(\underline{\beta})$ is minimized only on Q , the set of permissible quantized coefficient vectors of $D(z)$.

The first modification may be easily implemented by means of the relationship

$$\frac{\partial J(\underline{\beta})}{\partial \beta_1} \cong \frac{J(\beta_1 + \Delta, \beta_2, \dots, \beta_{2N+1}) - J(\beta_1, \beta_2, \dots, \beta_{2N+1})}{\Delta} , \quad (\text{III-3})$$

where Δ is chosen sufficiently small.

A convenient method for accomplishing the second modification is to prohibit step-lengths along any coordinate axis of E_{2N+1}^β of anything

other than integral multiples of one quantization level. This of course eliminates the possibility of $J(\underline{\beta})$ being examined at any point other than points in Q , with one possible exception; the case wherein the step-length when added to a point in Q results in a point whose components exceed the bounds of Q along one or more coordinate axes. This is a highly unlikely occurrence in most practical problems; however, it may be circumvented if the need arises by the addition of specialized functions to $J(\underline{\beta})$ which become extremely large as the bounds on Q are approached [8].

One obvious effect of the proposed technique for selecting step-lengths is that it becomes unlikely that the directions of the steps at the initial starting points in Q will be parallel to those of the corresponding gradient vectors at these points. This affects, in varying degrees, the rate of convergence of the method. At first glance, the apparent solution to this problem is to make the number of quantization levels per step along each axis of E_{2N+1}^{β} proportionally equal (as nearly as possible) to the corresponding components of the gradient vector. However, this is not always the best policy, since the minimum total step-length by this approach could become prohibitively large if one of the components of $-\text{grad}[J(\underline{\beta})]$ were much smaller than another component.

The most successful step-length algorithm which has been investigated thus far is as follows: let h_i denote the quantization granularity associated with the coefficient β_i ; then the vector $\underline{s}$ representing the directed step-length to be used in moving from some initial point $\underline{\beta}^0 \in Q$ toward a local minimum of $J(\underline{\beta})$ in Q is

$$\underline{s} = \begin{bmatrix} -h_1 \operatorname{sgn} \left[\frac{\partial J(\underline{\beta}^0)}{\partial \beta_1} \right] \\ -h_2 \operatorname{sgn} \left[\frac{\partial J(\underline{\beta}^0)}{\partial \beta_2} \right] \\ \vdots \\ -h_{2N+1} \operatorname{sgn} \left[\frac{\partial J(\underline{\beta}^0)}{\partial \beta_{2N+1}} \right] \end{bmatrix}, \quad (\text{III-4})$$

where sgn is defined for this case as

$$\begin{aligned} \operatorname{sgn} [f] &= +1 \text{ if } f > 0, \\ \operatorname{sgn} [f] &= 0 \text{ if } f = 0, \text{ and} \\ \operatorname{sgn} [f] &= -1 \text{ if } f < 0. \end{aligned}$$

It can be seen in (III-4) that the components of $\underline{s}$ will never be greater than one quantization level in magnitude, which eliminates the possibility of the prohibitively large minimum step-lengths mentioned previously. Furthermore, since $\underline{s}$ and $-\operatorname{grad} [J(\underline{\beta}^0)]$ must lie in the same "quadrant" of E_{2N+1}^β , the difference in their directions will not in general be great enough to drastically affect the rate of convergence.

The only undesirable aspect of the modified steepest descent method outlined above is that the iterative process may terminate before a local minimum of $J(\underline{\beta})$, as given by Definition (III-4), is reached. This usually occurs when there are more than one nonzero element of the final directed step-length $\underline{s}$ of the process, and in order to reach a local minimum, it is necessary to perturb only one of the coefficients.

For this reason, a second, more refined, and somewhat slower stage was incorporated in the minimization algorithm.

2. Second Stage

After the modified steepest descent method has terminated, the following technique may be utilized to guarantee that a local minimum of $J(\underline{\beta})$ is reached.

The last point of Q which is attained in the steepest descent procedure is employed as a starting point in the second stage. Each of the coefficients $\beta_1, \beta_2, \dots, \beta_{2N+1}$ is then varied in turn, one quantization level at a time; $J(\underline{\beta})$ is reduced as far as possible with the first coefficient varying and then control passes on to the next. This process continues until no further reduction of $J(\underline{\beta})$ is possible and a local minimum of $J(\underline{\beta})$ is found.

This quantized coefficient vector is therefore the desired optimal set of quantized coefficients of $D(z)$ based on the performance index $J(\underline{\beta})$.

A simple numerical example will now be considered for the purpose of demonstrating the details of implementing the proposed quantized coefficient selection technique.

3. Example

Consider the hybrid system configuration represented in Figure 14, which consists of an unstable continuous-time plant and a digital controller in the feedback loop for the purpose of stabilizing the system. It will be assumed that T , the system sampling period, is 0.04 seconds and that the D/A converter functions as an ideal zero-order

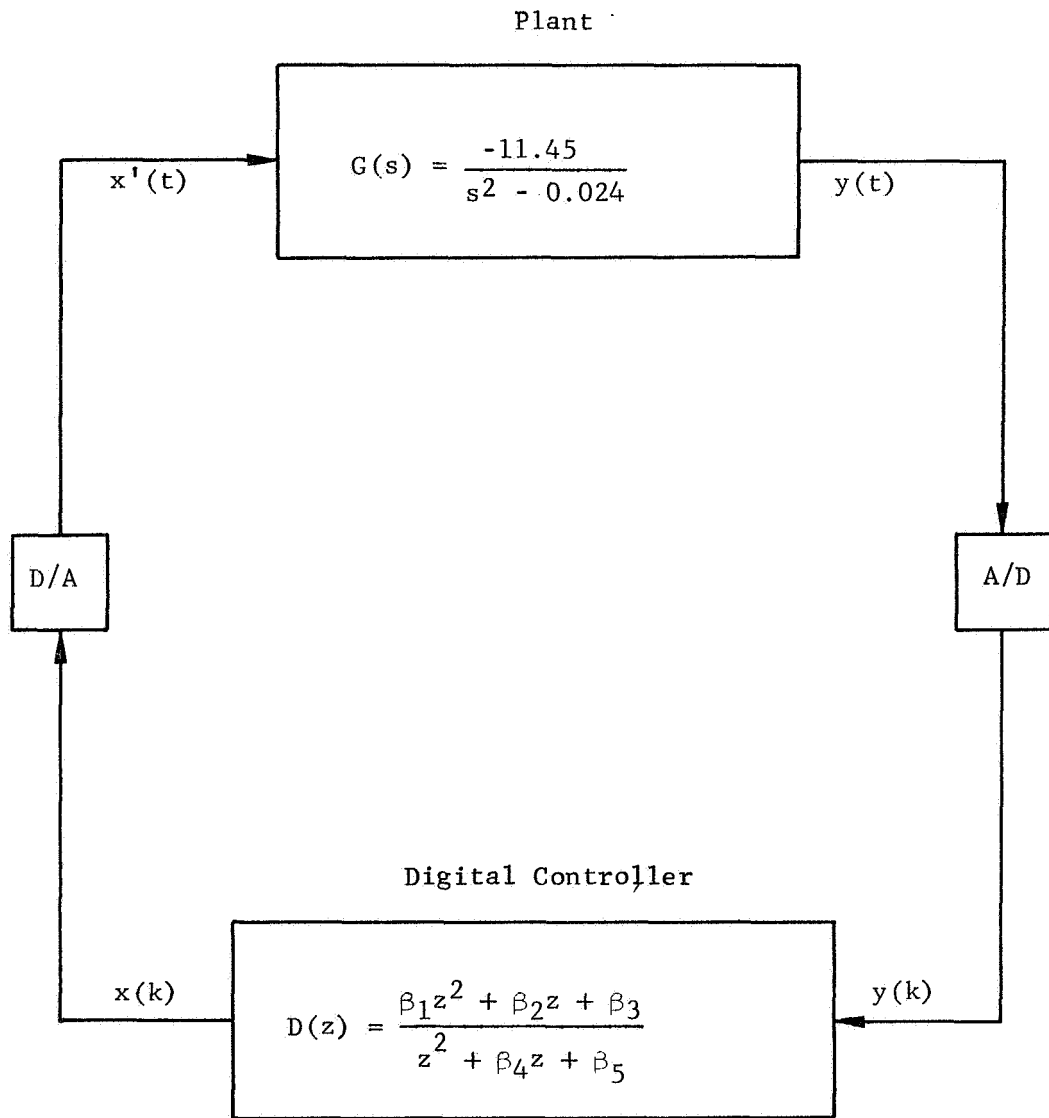


Fig. 14--Example.

data-hold. Classical frequency domain techniques may be used to show that the following digital controller transfer function adequately stabilizes the system:

$$D(z) = \frac{\beta_1 z^2 + \beta_2 z + \beta_3}{z^2 + \beta_4 z + \beta_5}, \quad (\text{III-5})$$

where

$$\beta_1 = 1.0000,$$

$$\beta_2 = -1.9400,$$

$$\beta_3 = 0.9405,$$

$$\beta_4 = -1.8150,$$

and

$$\beta_5 = 0.8159.$$

For the purpose of the following discussion, it will be assumed that the digital device used in implementing $D(z)$ is capable of realizing both positive and negative coefficients having magnitudes from 0 through $2047/1024$, in increments of $1/1024$. This defines the set Q of permissible quantized coefficient vectors in E_5^β and the corresponding quantization granularities; i.e., $h_1 = h_2 = h_3 = h_4 = h_5 = 2^{-10}$.

The design specifications which will be emphasized in this example are the system phase margin and the static accuracy of the system; moreover, the objective will be to select a quantized coefficient vector from Q which minimizes the deviation of these performance characteristics from those of the nominal system.

To proceed further, a knowledge of the nominal magnitude-phase plot of the system is required. The magnitude-phase plot associated with the open-loop z -transfer function of the system in Figure 14 may be generated by

conventional sampled-data techniques [1]. This information has been computed and plotted in the form of a Nyquist diagram in Figure 15. It is apparent from this plot that the performance index must include terms which reflect deviations in $D(z)$ in the vicinity of $z = 1.0000 + j0.0000$ and $z = 0.9950 + j0.0998$. These requirements may be incorporated in (III-1) to obtain a suitable performance index for the problem under consideration; i.e., let

$$J(\underline{\beta}) = \sum_{i=1}^2 \left\{ \lambda_i \frac{\operatorname{Re} [D(z_i) - \bar{D}(z_i)]}{\operatorname{Re}[D(z_i)]} \right\}^2 + \sum_{i=1}^2 \left\{ \lambda_i \frac{\operatorname{Im} [D(z_i) - \bar{D}(z_i)]}{\operatorname{Im} [D(z_i)]} \right\}^2, \quad (\text{III-6})$$

where $z_1 = 1.0000 + j0.0000$, $z_2 = 0.9950 + j0.0998$, and $\lambda_1 = \lambda_2 = 1.0$.

All that remains to be done before minimization of $J(\underline{\beta})$ may be carried out is the selection of an initial starting point, $\underline{\beta}^0$, in Q . Experience has shown that a convenient choice for the initial starting point is the member of Q whose components deviate from those of the nominal coefficient vector by the least amount. However, in order to effectively demonstrate the convergence properties of the two-stage minimization technique in this example, a starting point having components that deviated several quantization levels from those of the nominal coefficient point was chosen. This point is given by (III-7).

$$\begin{aligned} \beta_1^0 &= \frac{1024}{1024} \\ \beta_2^0 &= -\frac{1987}{1024} \\ \beta_3^0 &= \frac{968}{1024} \end{aligned}$$

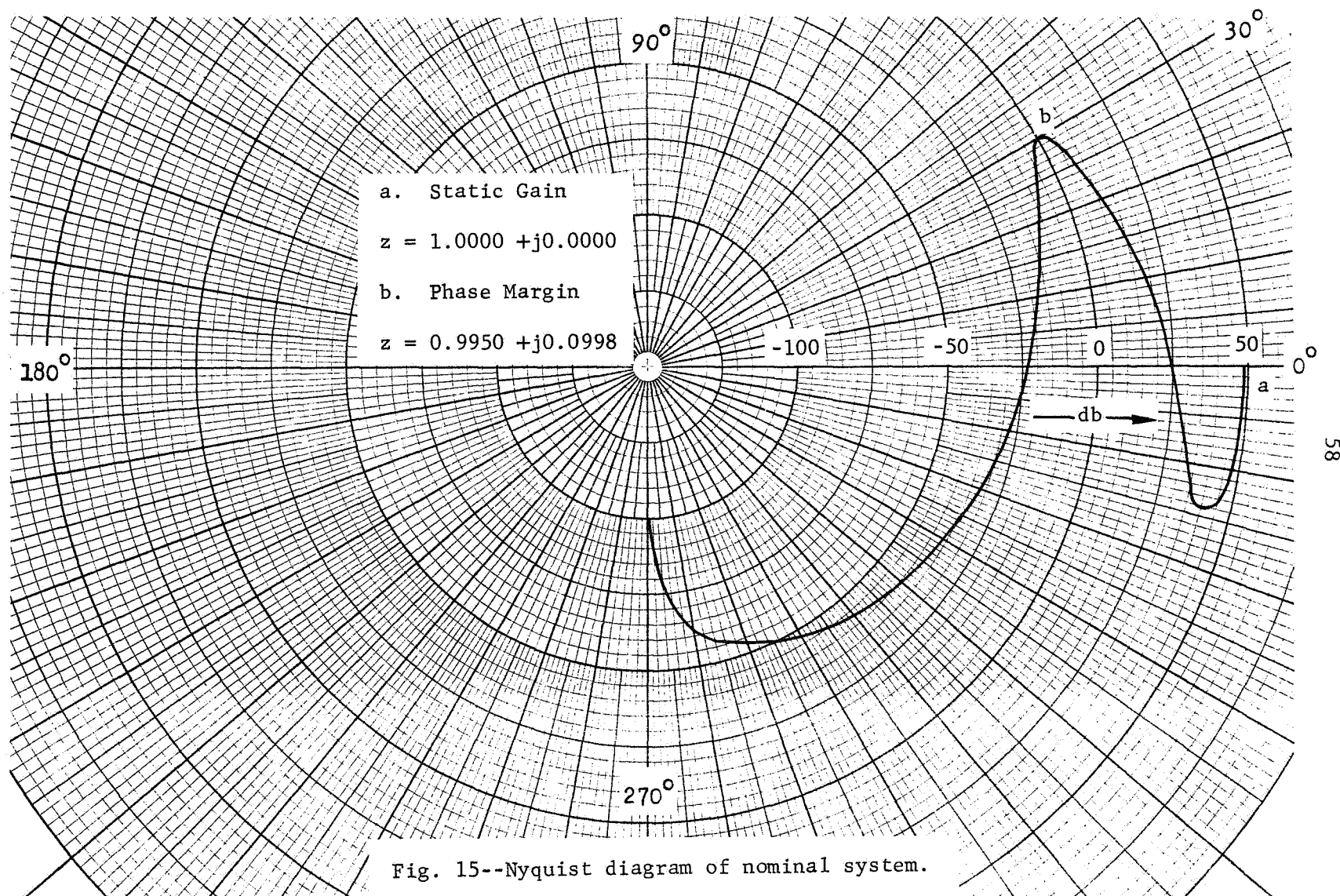


Fig. 15--Nyquist diagram of nominal system.

$$\begin{aligned}\beta_4^0 &= -\frac{1859}{1024} \\ \beta_5^0 &= \frac{833}{1024},\end{aligned}\tag{III-7}$$

which corresponds to $J(\underline{\beta}^0) = 30.94133$.

A digital computer program (see Appendix) was written for implementing the previously described two-stage minimization technique, with $\underline{\beta}^0$ as the starting point, and the results of each successive iteration of this program are tabulated in Table 1. As a note of practical interest, the results presented in Table 1 required 14 seconds of execution on an IBM model 7040 digital computer.

The system magnitude-phase plot associated with the optimal quantized coefficients from Table 1 is presented in Figure 16 as an indication of the effectiveness of the quantized coefficient selection technique. A comparison of this plot with that of Figure 15 reveals that the deviation from nominal of the system phase margin due to the selection of quantized coefficients is approximately five degrees. Moreover, the static gain deviation is nearly undetectable on the plots. Whether or not these deviations are tolerable depends of course upon the designer's judgement. If, however, they are unacceptable, the designer has the following alternatives: (1) he may adjust the weighting factors λ_1 and λ_2 in order to effect trade-offs between static accuracy and phase margin deviations, (2) he may select different initial starting points in Q in an attempt to locate relative minima of $J(\underline{\beta})$ in Q which are less than that obtained in Table 1, or (3) he may simply require that coefficient word-lengths be increased within the digital hardware.

Several observations are appropriate at this point concerning the

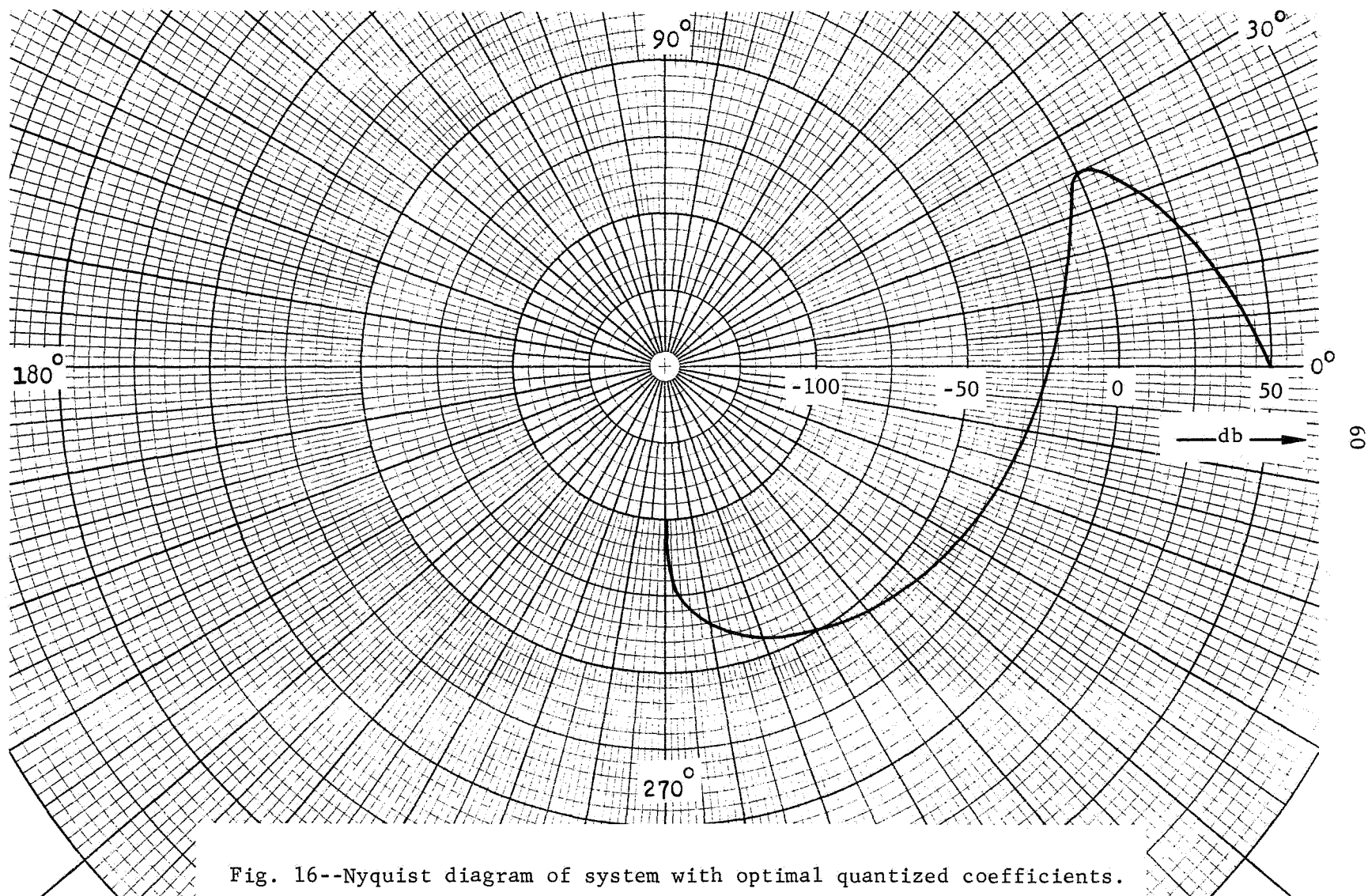


Fig. 16--Nyquist diagram of system with optimal quantized coefficients.

TABLE 1 - MINIMIZATION OF PERFORMANCE INDEX

Modified Steepest Descent

$J(\underline{\theta})$	β_1	β_2	β_3	β_4	β_5
30.94133	1024/1024	-1987/1024	968/1024	-1859/1024	833/1024
3.89818	1023/1024	-1988/1024	967/1024	-1860/1024	832/1024
0.58223	1022/1024	-1989/1024	966/1024	-1861/1024	831/1024
0.04723	1021/1024	-1990/1024	965/1024	-1863/1024	830/1024

Stage Two

0.02668	1021/1024	-1990/1024	965/1024	-1861/1024	830/1024
---------	-----------	------------	----------	------------	----------

second alternative stated above. It has been observed in several practical examples that there may exist more than one relative minimum of $J(\underline{\beta})$ in Q in the immediate vicinity of the nominal coefficient point. The number of these relative minima of course depends almost entirely on the character of $D(z)$ and the coefficient word-lengths being considered. For instance, under the conditions imposed in the previous example, it was found that by choosing different starting points in Q , relative minima of $J(\underline{\beta})$ in Q were attained ranging from $J(\underline{\beta}) = 1.00083$ (with $\beta_1 = 1024/1024$, $\beta_2 = -1987/1024$, $\beta_3 = 963/1024$, $\beta_4 = -1859/1024$, and $\beta_5 = 835/1024$) to $J(\underline{\beta}) = 0.01123$ (with $\beta_1 = 1024/1024$, $\beta_2 = -1986/1024$, $\beta_3 = 963/1024$, $\beta_4 = -1858/1024$, and $\beta_5 = 836/1024$).

It is especially interesting to note that the relative minimum $J(\underline{\beta}) = 1.0083$ corresponds to the point in Q whose components deviated by the least amount from the respective components of the nominal coefficient point. It is difficult to attach any significance to this occurrence, aside from the fact that the absolute minimum of $J(\underline{\beta})$ in Q is not limited to points of close proximity to the nominal coefficient point.

IV. AN ALGORITHM FOR COMPUTING FREQUENCY- RESPONSE BOUNDS FOR SYSTEMS SUBJECT TO PARAMETER ANOMALIES

A problem of considerable interest in effecting a control system evaluation is that of determining in some measure the effect of parameter anomalies upon the system performance. This problem arises both in continuous-time systems, where component tolerances may introduce parameter uncertainties, and in digital systems, where equipment malfunctions might conceivably result in erroneous realizations of $D(z)$ coefficient magnitudes. Several techniques have been advanced for treating the parameter anomaly problem, as evidenced by [10-12].

In this chapter, a new solution to the parameter anomaly problem is developed based on frequency-domain design techniques. An algorithm is introduced which enables the designer to generate a set of absolute bounds on the magnitude and the phase plots of a system transfer function subject to a given set of parameter error ranges. These envelopes may then be used to determine the effects of parameter variations upon system stability.

The technique is developed initially for the case of continuous-time systems wherein the parameters are allowed to assume a continuum of values within their respective tolerance ranges. Then, with these results as a basis, the method is extended to the magnitude and the phase plots of $D(z)$, where bounds must be obtained relative to a finite

set of coefficient magnitude errors.

Central in the development of the algorithm are certain definitions and theorems related to the calculus of extrema for functions of several variables. The following is a compilation of these theorems and definitions.

A. Definitions and Theorems

Definition IV-1: Let $\underline{x}$ and $\underline{y}$ be points in the N -dimensional Euclidean space E_N . The distance between $\underline{x}$ and $\underline{y}$ is defined as

$$|\underline{x} - \underline{y}| = \left[\sum_{i=1}^N (x_i - y_i)^2 \right]^{1/2} .$$

Definition IV-2: Let r be a number greater than zero. The statement that " $N_r(\underline{x}_0)$ is a neighborhood of $\underline{x}_0$ in E_N " means that $N_r(\underline{x}_0)$ is the set of all points $\underline{x}$ in E_N such that

$$|\underline{x} - \underline{x}_0| < r .$$

Definition IV-3: The statement that "the set of points $S \subset E_N$ is an open set" means that if $\underline{x}$ is a point in S , then there exists a neighborhood $N_r(\underline{x}) \subset S$.

Definition IV-4: Let $f(\underline{x})$ be a continuous function defined on an open set $S \subset E_N$ with function values in E_1 . Then $f(\underline{x})$ is said to have an absolute maximum at the point $\underline{a} \in S$ if

$$f(\underline{x}) \leq f(\underline{a}) , \text{ for all } \underline{x} \in S .$$

If $\underline{a} \in S$ and if there exists a neighborhood $N_r(\underline{a}) \subset S$ such that

$$f(\underline{x}) \leq f(\underline{a}), \text{ for all } \underline{x} \in N_r(\underline{a}),$$

then $f(\underline{x})$ is said to have a relative maximum at the point $\underline{a}$. Absolute minima and relative minima are similarly defined.

Theorem IV-1: Let $f(\underline{x})$ be a function on an open set $S \subset E_N$ with finite first-order partial derivatives at a point $\underline{x}_0 \in S$. If $f(\underline{x})$ has a relative minimum or relative maximum at $\underline{x}_0$, then the gradient of $f(\underline{x})$ at $\underline{x}_0$ is $\underline{0}$ [13].

Theorem IV-2: Let $f(\underline{x})$ have continuous second-order partial derivatives on an open set $S \subset E_N$, and let $\underline{x}_0 \in S$ be a point having the property that the gradient of $f(\underline{x})$ at $\underline{x}_0$ is $\underline{0}$. Let $H(\underline{x}_0)$ be the $N \times N$ symmetric matrix whose ij th entry is $h_{ij} = \partial^2 f(\underline{x}_0) / \partial x_i \partial x_j$ and let $\Delta = \det[H(\underline{x}_0)]$. Let $\Delta_0 = 1$ and let Δ_{N-k} be the determinant obtained from Δ by deleting the last k rows and columns.

- i. A sufficient condition for $f(\underline{x})$ to have a relative minimum at $\underline{x}_0$ is that the $N+1$ numbers $\Delta_0, \Delta_1, \dots, \Delta_N$ be positive; a sufficient condition for $f(\underline{x})$ to have a relative maximum at $\underline{x}_0$ is that $\Delta_0, \Delta_1, \dots, \Delta_N$ be alternately positive and negative. That is, if $H(\underline{x}_0)$ is positive (negative) definite, then $f(\underline{x}_0)$ is a relative minimum (maximum).
- ii. A necessary condition for $f(\underline{x})$ to have a relative minimum (maximum) at $\underline{x}_0$ is that $H(\underline{x}_0)$ be positive (negative) semidefinite [13,14].

Comment: Theorem IV-2 provides a useful analytical method for locating relative extrema of functions of several variables. It furnishes a "combined" set of necessary and sufficient conditions for the existence of relative extrema of $f(\underline{x})$ at any point in S , with one

notable exception. This exception is the case wherein the H matrix associated with a given point in S is semidefinite. In this case, Theorem IV-2 does not guarantee that $f(\underline{x})$ achieves a relative extrema. Therefore, when this situation arises, one must resort to Definition IV-4 to determine whether or not $f(\underline{x})$ attains a relative extremum; and if so, the type of extremum.

Definition IV-5: Let $G(s, \bar{a}_1, \bar{a}_2, \dots, \bar{a}_M)$ denote the generalized transfer function of the continuous-time system to be considered, where $\bar{a}_1, \bar{a}_2, \dots, \bar{a}_M$ represent the tolerated parameters whose values are not known precisely but are known to be within a prescribed set of tolerance ranges.

Definition IV-6: Let a_i denote the nominal value of $\bar{a}_i$, for $i = 1, 2, \dots, M$.

Definition IV-7: Let $a_i + \Delta_{ui}$ and $a_i - \Delta_{li}$ denote the maximum and minimum permissible values of $\bar{a}_i$, respectively, where $\Delta_{ui} \geq 0$ and $\Delta_{li} \geq 0$; i.e. $a_i + \Delta_{ui} \geq \bar{a}_i \geq a_i - \Delta_{li}$, for $i = 1, 2, \dots, M$.

Definition IV-8: Let $G(j\omega, \bar{a}) = U(\omega, \bar{a}) + jV(\omega, \bar{a})$.

Definition IV-9: Let $D(z, \bar{b}_1, \dots, \bar{b}_N)$ denote the generalized transfer function of the digital system under investigation, where $\bar{b}_1, \bar{b}_2, \dots, \bar{b}_N$ denote the coefficients whose incorrect representations within the digital system are to be considered.

Definition IV-10: Let b_i represent the nominal value of $\bar{b}_i$, for $i = 1, 2, \dots, N$.

Definition IV-11: Let Q_i denote the set of erroneous representations of $\bar{b}_i$ which are to be considered in generating the envelopes on

the magnitude and the phase plots of $D(e^{j\omega T}, \underline{\bar{b}})$. Further, let $b_i + \Delta_{ui}$ be the greatest member of Q_i and let $b_i - \Delta_{li}$ be the least member of Q_i , where $\Delta_{ui} \geq 0$, $\Delta_{li} \geq 0$ and $i = 1, 2, \dots, N$.

Comment: Note that since the coefficients of $D(z, \underline{\bar{b}})$ must be realized by words of finite length in a physically realizable digital system, there are a finite number of misrepresentations of b_i which can occur. Consequently, Q_i must be a finite set.

Definition IV-12: Let $D(e^{j\omega T}, \underline{\bar{b}}) = U(\omega, \underline{\bar{b}}) + jV(\omega, \underline{\bar{b}})$.

With the above definitions and theorems as a basis, the development of the algorithm may now proceed.

B. Development of the Algorithm for Toleranced Parameters in Continuous-Time Systems.

Consider the system represented by $G(s, \underline{\bar{a}})$. Due to the assertion that each of the toleranced parameters may assume a continuum of values within its respective tolerance range, the magnitude and phase associated with the complex number $G(j\omega, \underline{\bar{a}})$ become M -parameter families of curves when plotted versus ω in the conventional frequency response format. Each permissible combination of the M parameter values will in general result in a different set of frequency-response characteristics. Furthermore, each of U and V is an M -parameter family of curves in the U - ω and V - ω planes, respectively. (In some cases, the arguments of U and of V are omitted for notational convenience). It will be shown that by determining the envelopes on the U and the V families of curves, one can establish a set of absolute bounds on the phase and the magnitude families of characteristics of $G(j\omega, \underline{\bar{a}})$.

At this point it is convenient to establish an M-dimensional Euclidean space E_M^a in which the coordinates are defined by the tolerated system parameters $\bar{a}_1, \bar{a}_2, \dots, \bar{a}_M$. Consequently, the parameter tolerances describe a set of permissible "operating points" in E_M^a . The objective now is to absolutely bound the magnitude and the phase of $G(j\omega, \bar{a})$ at each frequency of interest for any permissible set of parameter values. Since

$$\angle G(j\omega, \bar{a}) = \tan^{-1} \left[\frac{V(\omega, \bar{a})}{U(\omega, \bar{a})} \right], \quad (\text{IV-1})$$

one possible approach to the problem is to extremize each of U and V independently at the frequencies of interest and then select the greatest and least values of $\angle G(j\omega, \bar{a})$ from the set of values corresponding to the four combinations of the extrema of U and V. Furthermore, since

$$|G(j\omega, \bar{a})| = \left[U(\omega, \bar{a})^2 + V(\omega, \bar{a})^2 \right]^{1/2}, \quad (\text{IV-2})$$

a similar argument applies to the magnitude of $G(j\omega, \bar{a})$. It is evident from (IV-1) and (IV-2) that this procedure would in fact yield a set of absolute bounds on the gain and the phase of $G(j\omega, \bar{a})$ in every case except one; the case wherein the absolute extrema of U and (or) V are of opposite sign. Therefore, this eventuality must be taken into account if the foregoing procedure is employed to generate frequency-response bounds. This problem will be considered in more detail as the development progresses.

1. Change of variables.

It would be convenient to employ Theorem IV-1 and Theorem IV-2 in extremizing U and V. However, since the set of permissible operating points in E_M^a is not an open set, these theorems are not directly applicable. For this reason, it is advantageous to perform the following changes of variables. Let

$$\bar{a}_i = a_i + \frac{\Delta_{ui} - \Delta_{li}}{2} + \frac{\Delta_{ui} + \Delta_{li}}{2} \sin \bar{\alpha}_i, \quad (\text{IV-3})$$

where $i = 1, 2, \dots, M$.

Note that the transformation described by (IV-3) in effect maps the set of operating points, a closed and bounded subset of E_M^a , onto E_M^α , an M-dimensional Euclidean space whose coordinates are the $\bar{\alpha}_i$ variables. Therefore, since the transformed set of operating points is an open set, the above theorems are now applicable, assuming of course that the requirements on the partial derivatives are satisfied.

It is important to note that $\sin \bar{\alpha}_i$ is not the only function which might be employed in (IV-3) to attain the desired transformation; for example, $\cos \bar{\alpha}_i$ would perform equally well.

2. Absolute bounds on U and V.

In the subsequent discussion, it is assumed that both U and V are continuous and have continuous first- and second-order partial derivatives at all points in E_M^α and for all frequencies of interest. This is not a severe limitation since the transfer functions encountered in the modeling of continuous-time systems usually possess this property.

Consider now the real part U of $G(j\omega, \bar{\underline{\alpha}})$. A set of absolute bounds on U at any desired frequency ω may be obtained by the following procedure. The solution of the equations $\partial U(\omega, \bar{\underline{\alpha}}) / \partial \bar{\alpha}_i = 0$ for the variables $\bar{\alpha}_i$, $i = 1, 2, \dots, M$, yields a set of points $T \subset E_M^\alpha$ having the property that each point $\bar{\underline{\alpha}} \in T$ is, by Theorem IV-1, a candidate for extremizing U .

To determine the points of T at which U does in fact achieve relative extrema, it is necessary to apply Theorem IV-2 and to examine the signature of the sequence $\Delta_0, \Delta_1, \dots, \Delta_M$ which results for each point in T . If none of the members of T yield a semidefinite form of H , then Theorem IV-2 may be used exclusively to isolate extrema of U . If, however, certain members of T lead to semidefinite H matrices, then it is necessary to numerically evaluate U in the neighborhoods of these points and to employ Definition IV-4 to investigate extremal behavior of U .

The next step in the algorithm, after the points in T where U has relative maxima and where U has relative minima have been isolated, is to select from these sets the points which produce the greatest relative maximum and least relative minimum of U . The values of U corresponding to these points are of course the desired absolute bounds on U at the frequency ω .

In the preceding discussion, only the real part of $G(j\omega, \bar{\underline{\alpha}})$ was considered. However, it is evident that the same procedure is equally effective in the problem of generating absolute bounds on $V(\omega, \bar{\underline{\alpha}})$.

It is important to note that the aforementioned procedure for bounding

U and V must be reapplied for each frequency of interest. However, this step of the algorithm is easily implemented with a digital computer program once the forms of the necessary partial derivatives have been established for evaluating the sequence $\Delta_0, \Delta_1, \dots, \Delta_M$ in Theorem IV-2.

3. Bounds on magnitude and phase.

After the absolute extrema of U and of V have been established for a given frequency of interest, it is possible to obtain a set of absolute bounds on the magnitude and the phase of $G(j\omega, \bar{\alpha})$ as follows. It can be seen in Figure 17 that the bounds on U and on V generate a rectangular region in the U-V plane within which $G(j\omega, \bar{\alpha})$ must lie for any $\bar{\alpha}$ in E_M^α . As mentioned previously in connection with (IV-1) and (IV-2), an upper and a lower bound on the magnitude and on the phase of $G(j\omega, \bar{\alpha})$ can be easily determined by consideration of only the four vertices of this rectangular region. However, the case wherein the absolute bounds on U and (or) on V are of opposite sign must be treated as a special case. If the rectangular boundary intersects one of the coordinate axes, as is depicted for example in Figure 18-a, the lower bound on the magnitude of $G(j\omega, \bar{\alpha})$ is computed at this intersection rather than at a vertex of the rectangle. If both coordinate axes of the U-V plane are intersected, as in Figure 18-b, only the upper bound on the magnitude of $G(j\omega, \bar{\alpha})$ is computed at the vertices. The three remaining bounds must be chosen as

$$-\pi \leq \angle G(j\omega, \bar{\alpha}) < \pi \quad , \quad (\text{IV-4})$$

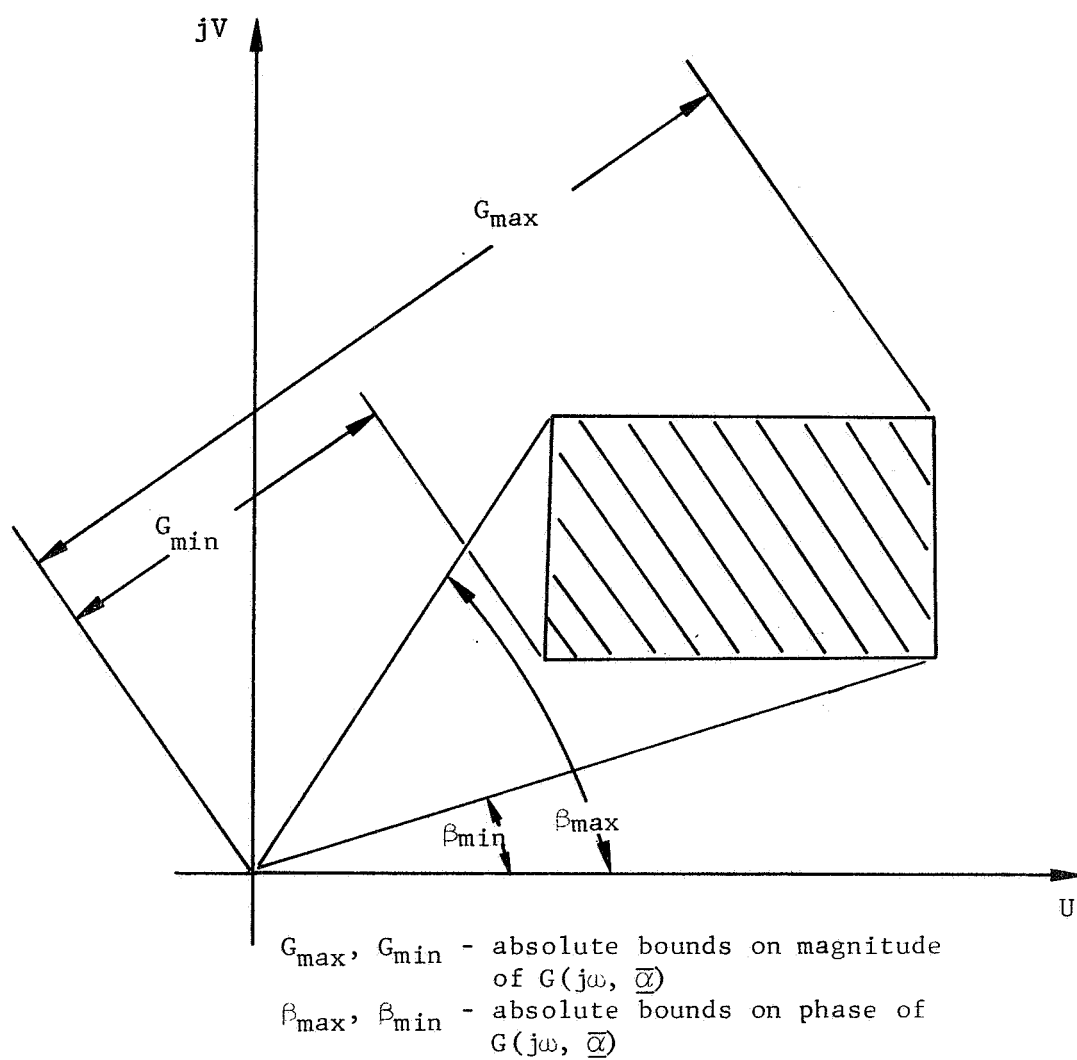
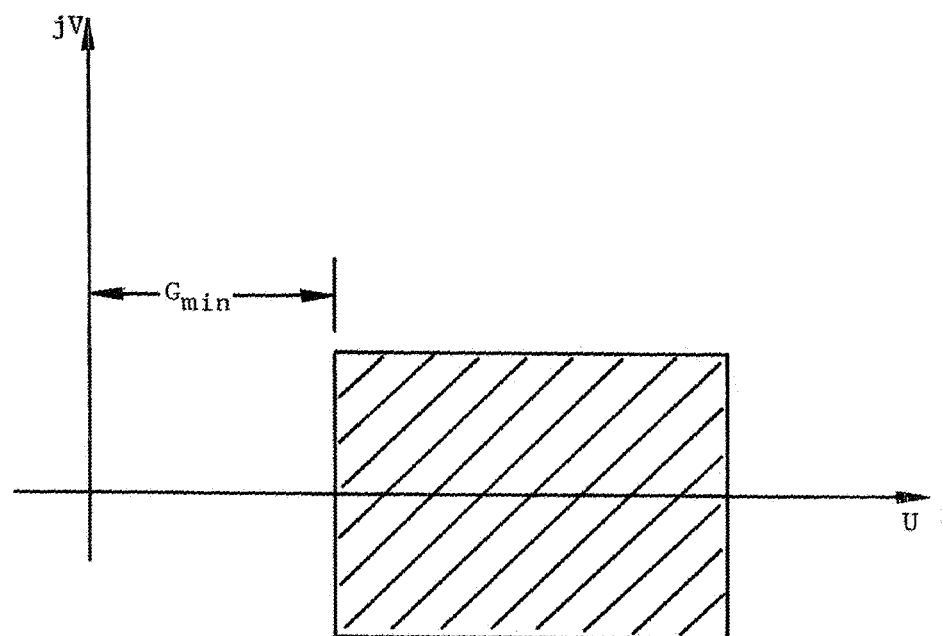
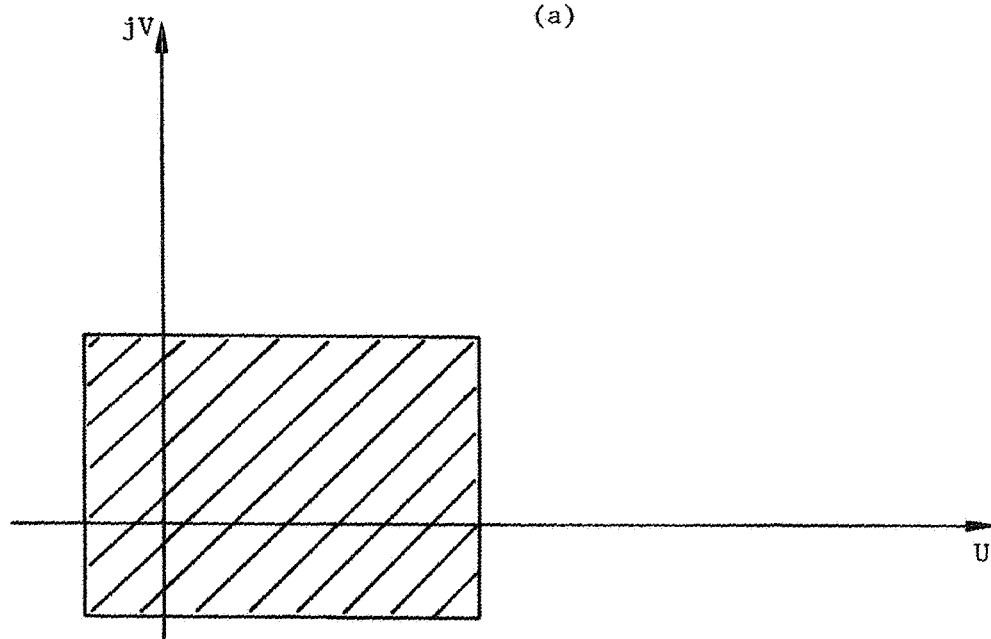


Fig. 17.--Typical rectangular region in the U-V plane.



(a)



(b)

Fig. 18.--Special cases where magnitude or phase bounds are not obtained at a vertex of the rectangular region.

and

$$0 \leq |G(j\omega, \bar{\alpha})|. \quad (\text{IV-5})$$

The repeated application of the above procedure for all frequencies of interest results in a set of absolute bounds, or an envelope, on the system frequency-response characteristics. This step, like the preceding steps, is well suited to digital computer implementation.

4. Simplifications.

The procedure for extremizing U and V is not as involved in many cases as it might first appear. It is simplified considerably by certain properties which are frequently exhibited by the U and the V functions. Several of these properties will now be considered and their effect on the application of the algorithm will be noted.

Property 1: The trigonometric terms which enter into the changes of variables in (IV-3) usually result in $\cos \bar{\alpha}_1$ being a factorable multiplier in the expressions for $\partial V(\omega, \bar{\alpha}) / \partial \bar{\alpha}_i$, $i = 1, 2, \dots, M$. Consequently, due to the unbounded number of zeroes of $\cos \bar{\alpha}_1$, the set T of candidate points will be infinite. However, due to the cyclic nature of the function $\cos \bar{\alpha}_1$, only a finite subset T' of T corresponding to $\bar{\alpha}_1 \in [-\pi, \pi)$, $i = 1, 2, \dots, M$, need be considered in extremizing U or V . The remaining members of T will yield no extrema of U or V not already produced by members of T' .

Property 2: In most cases the matrix H of second-order partial derivatives associated with U and with V is a diagonal matrix when

evaluated at points in T . In terms of the application of Theorem IV-2, this means that H is positive definite if and only if each of the elements h_{ii} , $i = 1, 2, \dots, M$, are positive. Moreover, H is negative definite if and only if h_{ii} , $i = 1, 2, \dots, M$, are negative. This property substantially simplifies the use of Theorem IV-2 in the search for extrema of U and of V .

Property 3: Another useful property is that in many cases every principal minor of H is nonzero when evaluated at points in T . In the case of a diagonal H matrix, this means that $\frac{\partial^2 U(\omega, \bar{\alpha})}{\partial \bar{\alpha}_i^2} \neq 0$ or $\frac{\partial^2 V(\omega, \bar{\alpha})}{\partial \bar{\alpha}_i^2} \neq 0$, for $i = 1, 2, \dots, M$ and $\bar{\alpha} \in T$. Consequently, Property 3 eliminates the possibility of H being semidefinite in those cases, which means that Theorem IV-2 provides a combined set of necessary and sufficient conditions for the extremization of U or V . Thus, one needs only to select from the permissible candidates for $\bar{\alpha}_i$, $i = 1, 2, \dots, M$, all of the combinations which produce positive definite H matrices, and Theorem IV-2 guarantees that there exist no other points of relative minima of U (or V). The same argument applies to relative maxima of functions having Property 3.

A numerical example which is treated in the following section will provide further clarification of the above statements and will indicate in a somewhat heuristic manner the reasons for the existence of these properties in the U and the V functions.

5. Example

The following numerical example is given to demonstrate the application of the foregoing algorithm. Consider the system represented by

the transfer function

$$G(s, \bar{a}_1, \bar{a}_2, \bar{a}_3) = \frac{\bar{a}_1(\bar{a}_2 s + 1)}{s(\bar{a}_3 s + 1)} , \quad (\text{IV-6})$$

where $\bar{a}_1$, $\bar{a}_2$, and $\bar{a}_3$ are tolerated parameters having nominal values given by

$$a_1 = 1.0, \quad (\text{IV-7})$$

$$a_2 = 2.0, \quad (\text{IV-8})$$

and

$$a_3 = 1.0 . \quad (\text{IV-9})$$

Further, suppose it is desired to establish a set of absolute bounds on the system frequency-response characteristics which result from a ± 10 percent variation of these parameters about their nominal values. Then, following the algorithm, the variable parameters are defined as

$$\bar{a}_1 = 1.0 + 0.1 \sin \bar{\alpha}_1 , \quad (\text{IV-10})$$

$$\bar{a}_2 = 2.0 + 0.2 \sin \bar{\alpha}_2 , \quad (\text{IV-11})$$

and

$$\bar{a}_3 = 1.0 + 0.1 \sin \bar{\alpha}_3 . \quad (\text{IV-12})$$

The next step is that of determining the bounds on the U and on the V families of curves. In the discussion which follows, only the development of the envelope on the U family will be considered in detail. Since the procedure for generating the envelope on V is based on the same arguments as for U, the discussion of V will be limited to

a brief summary.

The real part of $G(j\omega, \underline{\bar{\alpha}})$ is

$$U(\omega, \underline{\bar{\alpha}}) = \frac{(1.0 + 0.1 \sin \bar{\alpha}_1) (1.0 - 0.1 \sin \bar{\alpha}_3 + 0.2 \sin \bar{\alpha}_2)}{(1.0 + 0.1 \sin \bar{\alpha}_3)^2 \omega^2 + 1.0} \quad (\text{IV-13})$$

and the subset T_U of points in E_3^α which are candidates for extremizing $U(\omega, \underline{\bar{\alpha}})$ may be determined by Theorem IV-1. To be more explicit T_U is comprised of points which satisfy the equation

$$\text{grad } [U(\omega, \underline{\bar{\alpha}})] = \underline{0}. \quad (\text{IV-14})$$

The components of this gradient vector are listed below for convenience:

$$\frac{\partial U(\omega, \underline{\bar{\alpha}})}{\partial \bar{\alpha}_1} = \frac{0.1(1.0 - 0.1 \sin \bar{\alpha}_3 + 0.2 \sin \bar{\alpha}_2) \cos \bar{\alpha}_1}{(1.0 + 0.1 \sin \bar{\alpha}_3)^2 \omega^2 + 1.0}, \quad (\text{IV-15})$$

$$\frac{\partial U(\omega, \underline{\bar{\alpha}})}{\partial \bar{\alpha}_2} = \frac{0.2(1.0 + 0.1 \sin \bar{\alpha}_1) \cos \bar{\alpha}_2}{(1.0 + 0.1 \sin \bar{\alpha}_3)^2 \omega^2 + 1.0}, \text{ and} \quad (\text{IV-16})$$

$$\begin{aligned} \frac{\partial U(\omega, \underline{\bar{\alpha}})}{\partial \bar{\alpha}_3} = & - \frac{0.1(1.0 + 0.1 \sin \bar{\alpha}_1) \cos \bar{\alpha}_3}{(1.0 + 0.1 \sin \bar{\alpha}_3)^2 \omega^2 + 1.0} \\ & - \frac{0.2(1.0 + 0.1 \sin \bar{\alpha}_1)(1.0 - 0.1 \sin \bar{\alpha}_3 + 0.2 \sin \bar{\alpha}_2)(1.0 + 0.1 \sin \bar{\alpha}_3) \cos \bar{\alpha}_3}{[(1.0 + 0.1 \sin \bar{\alpha}_3)^2 \omega^2 + 1.0]^2}. \end{aligned} \quad (\text{IV-17})$$

Therefore, from (IV-15) through (IV-17), it is apparent that T_U consists of points having the property that

$$\bar{\alpha}_i = \frac{(2n + 1)\pi}{2} \quad ; \quad n = 0, \pm 1, \pm 2, \dots, \quad (\text{IV-18})$$

for $i = 1, 2, 3$. Note that in this case, T_U is invariant with ω . This of course means that T_U must be determined only once and is applicable at all frequencies. Note also that T_U has Property 1.

Theorem IV-2 may now be employed to isolate points of T_U at which $U(\omega, \bar{\alpha})$ is extremized. In this example it is evident that each one of the off-diagonal elements in the matrix H of second-order partial derivatives of $U(\omega, \bar{\alpha})$ contains at least one of the terms $\cos \bar{\alpha}_i$; $i = 1, 2, 3$. Hence, H is a diagonal matrix for any $\bar{\alpha} \in T_U$ and therefore possesses Property 2. Furthermore, it is easily shown that none of the diagonal elements of H are zero for $\bar{\alpha} \in T_U$; i.e., H has Property 3. Therefore, only Theorem IV-2 is required to isolate the extrema of $U(\omega, \bar{\alpha})$ in T_U ; and, the application of Definition IV-4 is not necessary.

Since H is a diagonal matrix, the sequence of determinants in Theorem IV-2 reduces to the form

$$\begin{aligned}
 \Delta_0 &= 1 \\
 \Delta_1 &= \frac{\partial^2 U(\omega, \bar{\alpha})}{\partial \bar{\alpha}_1^2} \\
 \Delta_2 &= \Delta_1 \frac{\partial^2 U(\omega, \bar{\alpha})}{\partial \bar{\alpha}_2^2} \\
 \Delta_3 &= \Delta_2 \frac{\partial^2 U(\omega, \bar{\alpha})}{\partial \bar{\alpha}_3^2} .
 \end{aligned} \tag{IV-19}$$

From the sequence of expressions given by (IV-19), it is a simple matter to select the components $\bar{\alpha}_1$, $\bar{\alpha}_2$, and $\bar{\alpha}_3$ of the points in T_U such

that the criteria set forth in Theorem IV-2 are satisfied and the points of relative minima and of relative maxima of $U(\omega, \underline{\alpha})$ in T_U are obtained. Consider Δ_1 for example:

$$\frac{\partial^2 U(\omega, \underline{\alpha})}{\partial \bar{\alpha}_1^2} = - \frac{0.1(1.0 - 0.1 \sin \bar{\alpha}_3 + 0.2 \sin \bar{\alpha}_2) \sin \bar{\alpha}_1}{(1.0 + 0.1 \sin \bar{\alpha}_3)^2 \omega^2 + 1.0} \quad (\text{IV-20})$$

It is evident from (IV-20) and Theorem IV-2 that $\bar{\alpha}_1 = \frac{(4n - 1)\pi}{2}$, $n = 0, \pm 1, \pm 2, \dots$, are the only values of $\bar{\alpha}_1$ which need be considered in the minimization of $U(\omega, \underline{\alpha})$. Further, because of Property 1, this set of candidates may be reduced to the single value $\bar{\alpha}_1 = -\pi/2$. Moreover, $\bar{\alpha}_1 = \pi/2$ is the only component value of $\bar{\alpha}_1$, that need be considered in the maximization of $U(\omega, \underline{\alpha})$ in T_U .

Similar arguments may be employed to show that $\bar{\alpha}_2 = -\pi/2$ and $\bar{\alpha}_3 = \pi/2$ are the only values of the remaining components of $\underline{\alpha} \in T_U$ which must be considered in minimizing $U(\omega, \underline{\alpha})$; and $\bar{\alpha}_2 = \pi/2$ and $\bar{\alpha}_3 = -\pi/2$ are the only values that must be employed in maximizing $U(\omega, \underline{\alpha})$.

Substitution of the above results into (IV-13) results in a set of boundary equations which absolutely bound $U(\omega, \underline{\alpha})$ for at any frequency of interest and at any operating point $\underline{\alpha} \in E_3^\alpha$. These equations are

$$U(\omega) = \frac{1.43}{0.81 \omega^2 + 1.0} \quad (\text{upper bound}) \quad (\text{IV-21})$$

and

$$U(\omega) = \frac{0.63}{1.21\omega^2 + 1.0} \quad (\text{lower bound}). \quad (\text{IV-22})$$

The V boundary equations, which are determined by the same procedure as for the U family, are

$$V(\omega) = - \frac{1.78\omega^2 + 0.9}{\omega(1.21\omega^2 + 1.0)} \quad (\text{upper bound}) \quad (\text{IV-23})$$

and

$$V(\omega) = - \frac{2.178\omega^2 + 1.1}{\omega(0.81\omega^2 + 1.0)} \quad (\text{lower bound}). \quad (\text{IV-24})$$

The final step of the algorithm can now be applied; that is, the U and V boundary equations are used to define, for each frequency of interest, a rectangular region in the U-V plane which bounds $G(j\omega, \overline{\mathcal{Q}})$, for the prescribed set of parameter tolerance ranges. Then, from the boundaries of these rectangular regions, the maximum and minimum magnitude and phase are determined for each frequency of interest by the method illustrated in Figure 17 and Figure 18. This step was implemented with a digital computer program which generated the boundary regions for each desired frequency and systematically checked the boundaries for the extremes of phase and gain.

The resulting phase and gain envelopes for this particular example are illustrated in Figure 19 and Figure 20.

One additional observation should be made before the above example is completed. It is quite possible that, at a given frequency, a vertex of the rectangular region in the U-V plane which is used to

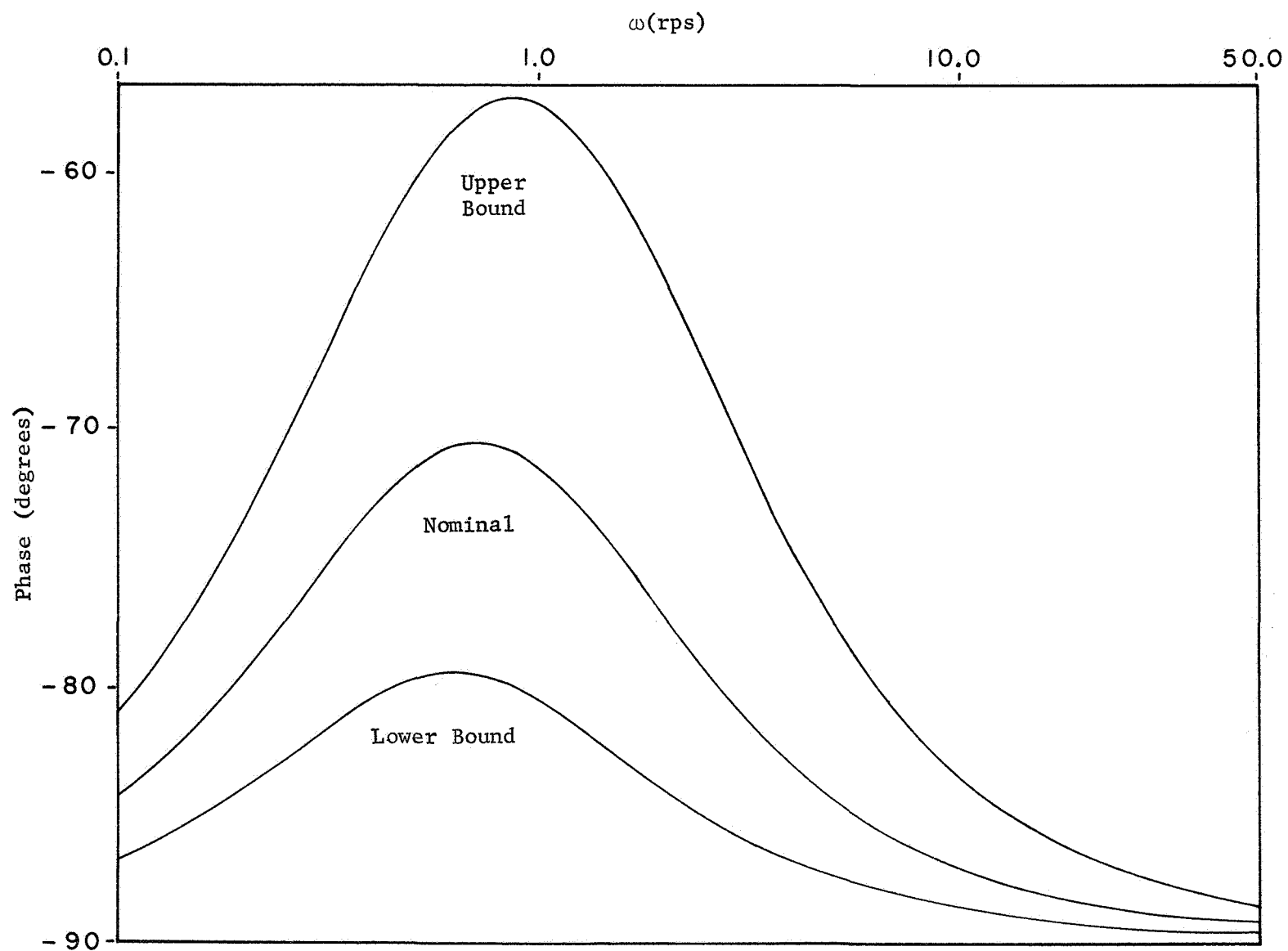


Fig. 19--Envelope on the phase characteristic of $G(j\omega, \bar{a})$ for $0.9 \leq \bar{a}_1 \leq 1.1$, $1.8 \leq \bar{a}_2 \leq 2.2$, and $0.9 \leq \bar{a}_3 \leq 1.1$.

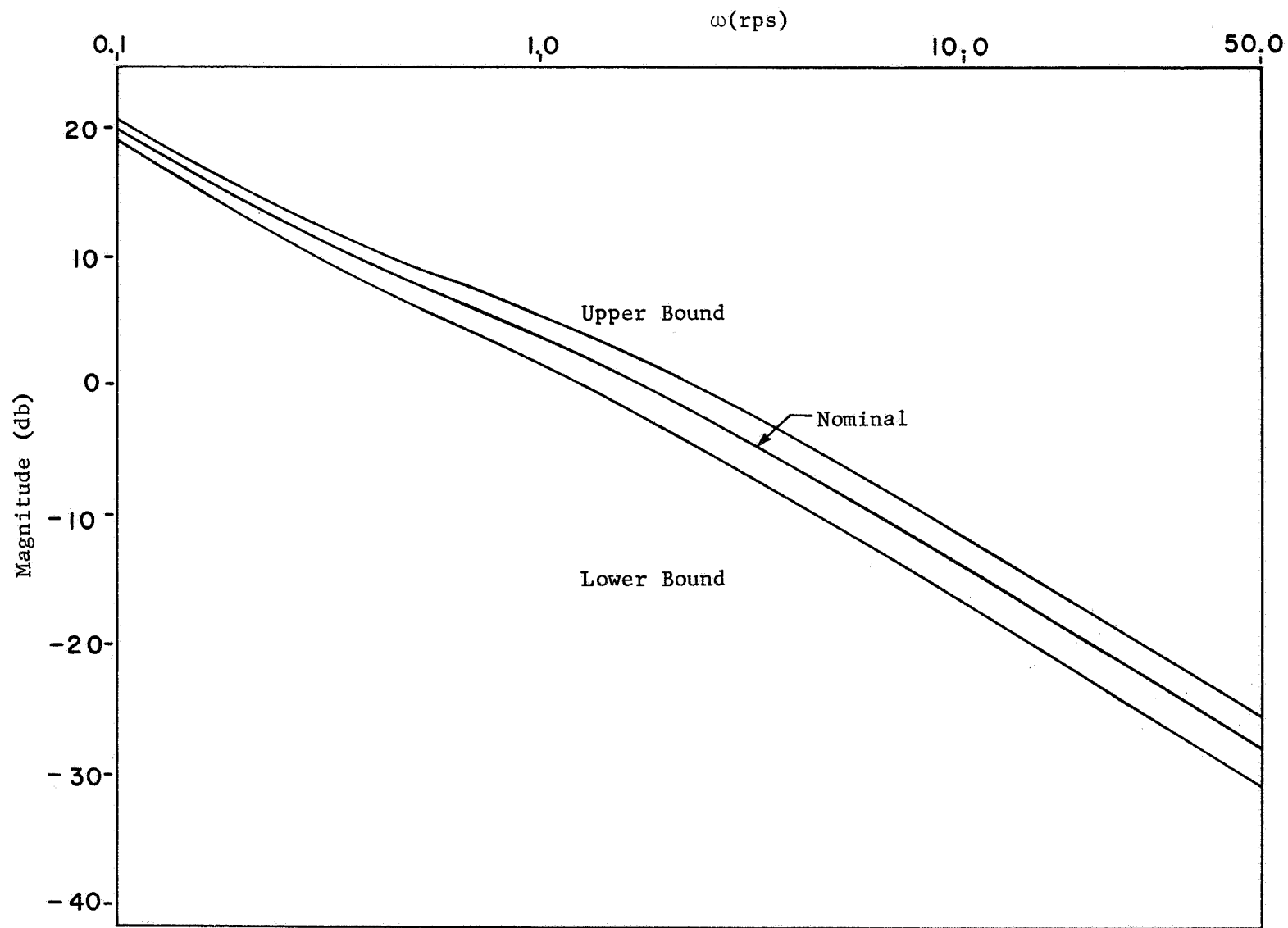


Fig. 20--Envelope on the gain characteristic of $G(j\omega, \bar{a})$ for $0.9 \leq \bar{a}_1 \leq 1.1$,
 $1.8 \leq \bar{a}_2 \leq 2.2$, and $0.9 \leq \bar{a}_3 \leq 1.1$.

obtain one of the magnitude or the phase bounds is itself unattainable for any $\underline{\alpha}$ in E_3^α . The bound obtained in this case is, however, an absolute bound, but it may be rather conservative. Consequently, the envelopes depicted in Figure 19 and Figure 20 may be conservative over certain frequency ranges.

C. Extention to Digital Systems with Coefficient Anomalies

With only minor modifications, the algorithm presented in the foregoing discussion may be used to generate envelopes on the magnitude and the phase plots of the complex number $D(e^{j\omega T}, \underline{b})$ subject to a given set of erroneous representations of the digital filter coefficients.

The changes of variables which are necessary for the application of Theorem IV-1 and Theorem IV-2 in the extremization of $U(\omega, \underline{b})$ and $V(\omega, \underline{b})$ are

$$\bar{b}_i = b_i + \frac{\Delta_{ui} - \Delta_{li}}{2} + \frac{\Delta_{ui} + \Delta_{li}}{2} \sin \bar{\beta}_i, \quad (\text{IV-25})$$

$$i = 1, 2, \dots, N.$$

Consequently, the N-dimensional space E_N^β replaces E_M^α of the previous development.

The magnitude and the phase characteristics associated with $D(e^{j\omega T}, \underline{\beta})$ are N-parameter families of curves when plotted versus ω . Each possible set of coefficient misrepresentations will in general result in a different magnitude-phase plot. However, unlike the parameter tolerance problem in continuous-time systems, there are

only a finite set of coefficient anomalies which might occur; i.e., Q_i , $i = 1, 2, \dots, N$, are finite sets. Thus, when the sets T_U and T_V of candidate points in E_N^{β} for extremizing U and V are generated, any point with components not associated with Q_i , $i = 1, 2, \dots, N$, must be neglected.

Aside from the above stated changes, the technique previously developed for generating frequency-response bounds in continuous-time systems is directly applicable to digital systems with coefficient anomalies.

V. CONCLUSIONS

Several of the problems associated with coefficient quantization in hybrid control systems were considered and solutions to these problems were advanced.

A technique for precisely correcting the time-response of a hybrid system containing quantization approximations of nominal digital filter coefficients was developed in Chapter II. Central in the technique is the use of a Taylor series expansion of the nominal system response about the approximated system response. In order to implement the correction technique, it is necessary to realize a separate auxiliary correction equation for each coefficient to be corrected. However, the coefficients of these equations are precisely realizable by the word-lengths of the digital filter. Further, the orders of the correction equations are the same as the order of the difference equation being realized by the digital filter. Consequently, the technique has the desirable property that it may be implemented by time-sharing the digital filter or by using duplicate digital filters. The coefficient correction scheme increases considerably the system hardware requirements, and therefore, it is recommended only in cases where infinite word-length precision of coefficient realizations is essential.

In cases where coefficient quantization errors are tolerable, the

procedure which was outlined in Chapter III may be employed to optimize the selection of quantized coefficients. The procedure minimizes a performance index which may be fashioned by the designer to reflect deviations from nominal of a wide variety of design specifications. The minimization technique is implemented by a two-stage digital computer program which locates relative minima of the performance index in the set of permissible quantized coefficient combinations. However, since the results are not global, it is sometimes necessary to locate more than one relative minimum of the performance index and select as the optimal set of coefficients the set with least relative minimum of the performance index.

A useful algorithm was presented in Chapter IV for establishing, subject to a given set of parameter tolerance ranges, an envelope that bounds the frequency-response characteristics of a linear, stationary, continuous-time system. The bounds obtained from this technique are in some cases conservative; however, they provide information which is indicative of the worst possible performance of the system for a given set of parameter tolerances. The method also exhibits the desirable feature that the effects of many parameter tolerances may be investigated without a significant increase in the difficulty of obtaining the envelope. Thus, the algorithm is a useful analytical method for effecting a system parameter tolerance study. The algorithm also provides, with minor modifications, an attractive method for investigating the effects of coefficient anomalies in digital systems. Given a set of bounds within which the coefficient errors must lie, one can generate via the algorithm

an envelope on the gain and the phase plots of the digital system z -transfer function.

REFERENCES

1. B. C. Kuo, Analysis and Synthesis of Sampled-Data Control Systems, Prentice Hall, Englewood Cliffs, New Jersey, 1963.
2. H. T. Nagle and C. C. Carroll, "Organizing a Special-Purpose Computer to Realize Digital Filters for Sampled-Data Systems," IEEE Transactions on Audio and Electroacoustics, Volume AU-16, No. 3, September 1968.
3. L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "Implementation of Digital Filters," IEEE International Convention - 68, New York, N. Y., March 1968.
4. J. F. Kaiser, "Some Practical Considerations in the Realization of Linear Digital Filters," Proceedings of the Third Allerton Conference on Circuit and System Theory, October 1965, p. 621.
5. C. L. Phillips, et. al., "Quantization Error-Bounds For Hybrid Control Systems," Technical Report Number 13, NAS8-11274, Engineering Experiment Station, Auburn, Alabama, September, 1968.
6. C. L. Phillips, et. al., "Simulation of High-Order Hybrid Control Systems," Technical Report Number 12, NAS8-11274, Engineering Experiment Station, Auburn, Alabama, July, 1968.
7. R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," The Computer Journal, Volume 6, 1963, p. 163.
8. H. H. Rosenbrock, "An Automatic Method for Finding the Greatest or Least Value of a Function," The Computer Journal, Volume 5, 1960, p. 175.
9. H. B. Curry, "The Method of Steepest Descent for Non-Linear Minimization Problems," Quarterly of Applied Mathematics, Volume 2, 1944, p. 258.
10. R. M. Stewart, "A Simple Graphical Method for Constructing Families of Nyquist Diagrams," Journal of Aeronautical Sciences, Volume 18, July, 1951, p. 767.
11. F. H. Bletcher, "Transister Multiple Loop Feedback Amplifiers,"

National Electronics Conference, Volume 13, 1957, p. 19.

12. I. M. Horowitz, Synthesis of Feedback Control Systems, Academic Press, New York, N. Y., 1963.
13. T. M. Apostol, Mathematical Analysis, Addison-Wesley, Reading, Massachusetts, 1964.
14. T. Chaundy, The Differential Calculus, Clarendon Press, Oxford, 1935.

APPENDIX

TWO-STAGE PROGRAM FOR MINIMIZATION OF PERFORMANCE INDEX

MODIFIED STEEPEST DESCENT METHOD

```
DIMENSION P1(10), SUMSQ1(20), SUMSQ2(20), X1(20), X2(20), X3(20),  
1 X4(20), X5(20), X6(20), B(20), B1(20), B2(20), A1(20), RE(20), KN(  
220)
```

```
COMPLEX CMPLX, CEXP, G, G1, G2, G3, S(20), Z(20)
```

```
K=0
```

```
N=2
```

```
PRINT 20
```

```
PRINT 19
```

```
M=1
```

```
VAL=.1E20
```

DEFINE COEFFICIENT GRANULARITIES

```
GRAN=1./1024.
```

DEFINE DEL USED IN PARTIAL DERIVATIVE APPROXIMATIONS

```
DEL=1./2048.
```

DEFINE CONSTRAINT POINTS ON UNIT CIRCLE OF Z-PLANE

```
Z(1)=CMPLX(0.10000000E01,0.00000000E00)
```

```
Z(2)=CMPLX(0.99500416E00,0.99833415E-01)
```

NOMINAL COEFFICIENTS

```
B(1)=1.0000
```

```
B(2)=-1.9400
```

```
B(3)=0.9405
```

```
B(4)=-1.8150
```

```
B(5)=0.8159
```

```
DO 1 I=1,N
```

```
G=(B(1)*Z(I)**2+B(2)*Z(I)+B(3))/(Z(I)**2+B(4)*Z(I)+B(5))
```

```
A1(I)=AIMAG(G)
```

```
1 RE(I)=REAL(G)
```

DEFINE INITIAL STARTING POINT

```

B1(1)=1024./1024.
B1(2)=-1987./1024.
B1(3)=968./1024.
B1(4)=-1859./1024.
B1(5)=833./1024.
2 DO 3 I=1,N
  G1=(B1(1)*Z(I)**2+B1(2)*Z(I)+B1(3))/(Z(I)**2+B1(4)*Z(I)+B1(5))
  X1(I)=REAL(G1)
3 X2(I)=AIMAG(G1)
  DO 8 J=1,5
  DO 4 I=1,5
4 B2(I)=B1(I)
  SUMSQ1(J)=0.0
  SUMSQ2(J)=0.0
  B2(J)=B2(J)+DEL
  DO 5 I=1,N
  G2=(B2(1)*Z(I)**2+B2(2)*Z(I)+B2(3))/(Z(I)**2+B2(4)*Z(I)+B2(5))
  X3(I)=REAL(G2)
  X4(I)=AIMAG(G2)
  SUMSQ1(J)=SUMSQ1(J)+((RE(I)-X3(I))/RE(I))**2+((AI(I)-X4(I))/AI(I))
  1**2
5 SUMSQ2(J)=SUMSQ2(J)+((RE(I)-X1(I))/RE(I))**2+((AI(I)-X2(I))/AI(I))
  1**2
  IF (M.GT.1) GO TO 7
  M=M+1
  VAL=SUMSQ2(J)
  DO 6 I=1,5
6 KN(I)=B1(I)/GRAN
  PRINT 21, SUMSQ2(J),KN(1),KN(2),KN(3),KN(4),KN(5)

```

PARTIAL DERIVATIVE APPROXIMATION

```

7 P1(J)=(SUMSQ1(J)-SUMSQ2(J))/DEL
8 CONTINUE
  DO 10 J=1,5
  IF (P1(J).EQ.0.0) GO TO 10
  IF (P1(J).GT.0.0) GO TO 9
  B1(J)=B1(J)+GRAN
  GO TO 10
9 B1(J)=B1(J)-GRAN
10 CONTINUE

```

ITERATE UNTIL SUMSQ EXCEEDS SUMSQ OF PREVIOUS ITERATION

```

11 SUMSQ=0.0

```

```

DO 12 I=1,N
G1=(B1(1)*Z(I)**2+B1(2)*Z(I)+B1(3))/(Z(I)**2+B1(4)*Z(I)+B1(5))
X1(I)=REAL(G1)
X2(I)=AIMAG(G1)
12 SUMSQ=SUMSQ+((RE(I)-X1(I))/RE(I))**2+((AI(I)-X2(I))/AI(I))**2
IF (SUMSQ.GE.VAL) GO TO 16
DO 13 I=1,5
13 KN(I)=B1(I)/GRAN
PRINT 21, SUMSQ,KN(1),KN(2),KN(3),KN(4),KN(5)
VAL=SUMSQ
K=0
DO 15 J=1,5
IF (P1(J).EQ.0.0) GO TO 15
IF (P1(J).GT.0.0) GO TO 14
B1(J)=B1(J)+GRAN
GO TO 15
14 B1(J)=B1(J)-GRAN
15 CONTINUE
GO TO 11
16 DO 18 J=1,5

```

REGRESS ONE ITERATION AND REEVALUATE PARTIAL DERIVATIVES

```

K=K+1
IF (P1(J).EQ.0.0) GO TO 18
IF (P1(J).GT.0.0) GO TO 17
B1(J)=B1(J)-GRAN
GO TO 18
17 B1(J)=B1(J)+GRAN
18 CONTINUE
IF (K.GE.6) CALL STAGE2 (VAL,B1,AI,RE,Z,GRAN,N)
GO TO 2
STOP

19 FORMAT (5X,25HMODIFIED STEEPEST DESCENT,/)
20 FORMAT (1H1)
21 FORMAT (5X,5HP.I.=,F8.5,4X,3HB1=,I5,5H/1024,4X,3HB2=,I5,5H/1024,4X,
1,3HB3=,I5,5H/1024,4X,3HB4=,I5,5H/1024,4X,3HB5=,I5,5H/1024,/)
END

```

SUBROUTINE STAGE2 (VAL,B1,AI,RE,Z,GRAN,N)

BEGIN SECOND STAGE

DIMENSION P1(10), SUMSQ1(20), SUMSQ2(20), X1(20), X2(20), X3(20),
X4(20), X5(20), X6(20), B(20), B1(20), B2(20), AI(20), RE(20), KN(220)

COMPLEX CMPLX,CEXP,G,G1,G2,G3,Z(20),Z1(20)

PRINT 15

K=0

J=1

BEGIN VARIATION OF COEFFICIENTS ONE AT A TIME

```

1 CONTINUE
  DO 2 I=1,N
    G1=(B1(1)*Z(I)**2+B1(2)*Z(I)+B1(3))/(Z(I)**2+B1(4)*Z(I)+B1(5))
    X1(I)=REAL(G1)
  2 X2(I)=AIMAG(G1)
    DO 3 I=1,5
  3 B2(I)=B1(I)
    SUMSQ1(J)=0.0
    SUMSQ2(J)=0.0
    B2(J)=B2(J)+GRAN
    DO 4 I=1,N
      G2=(B2(1)*Z(I)**2+B2(2)*Z(I)+B2(3))/(Z(I)**2+B2(4)*Z(I)+B2(5))
      X3(I)=REAL(G2)
      X4(I)=AIMAG(G2)
      SUMSQ1(J)=SUMSQ1(J)+((RE(I)-X3(I))/RE(I))**2+((AI(I)-X4(I))/AI(I))
1**2
  4 SUMSQ2(J)=SUMSQ2(J)+((RE(I)-X1(I))/RE(I))**2+((AI(I)-X2(I))/AI(I))
1**2
    P1(J)=SUMSQ1(J)-SUMSQ2(J)
    IF (P1(J).EQ.0.0) GO TO 6
    IF (P1(J).GT.0.0) GO TO 5
    B1(J)=B1(J)+GRAN
    GO TO 6
  5 B1(J)=B1(J)-GRAN
  6 CONTINUE

```

ITERATE UNTIL SUMSQ EXCEEDS SUMSQ OF PREVIOUS ITERATION

```

7 SUMSQ=0.0
  DO 8 I=1,N
    G1=(B1(1)*Z(I)**2+B1(2)*Z(I)+B1(3))/(Z(I)**2+B1(4)*Z(I)+B1(5))
    X1(I)=REAL(G1)
    X2(I)=AIMAG(G1)

```

```

8 SUMSQ=SUMSQ+((RE(I)-X1(I))/RE(I))**2+((AI(I)-X2(I))/AI(I))**2
  IF (SUMSQ.GE.VAL) GO TO 12
  DO 9 I=1,5
9 KN(I)=B1(I)/GRAN
  PRINT 16, SUMSQ,KN(1),KN(2),KN(3),KN(4),KN(5)
  VAL=SUMSQ
  IF (P1(J).EQ.0.0) GO TO 11
  IF (P1(J).GT.0.0) GO TO 10
  B1(J)=B1(J)+GRAN
  GO TO 11
10 B1(J)=B1(J)-GRAN
11 CONTINUE
  K=0
  GO TO 7
12 CONTINUE
  REGRESS ONE ITERATION AND BEGIN VARIATIONS OF NEXT COEFFICIENT
  K=K+1
  IF (K.EQ.10) STOP
  IF (P1(J).EQ.0.0) GO TO 14
  IF (P1(J).GT.0.0) GO TO 13
  B1(J)=B1(J)-GRAN
  GO TO 14
13 B1(J)=B1(J)+GRAN
14 CONTINUE
  J=J+1
  IF (J.GT.5) J=1
  GO TO 1
RETURN

```