QUANTIZATION ERRORS AND RECONSTRUCTION ERRORS IN DIGITAL FILTERS

by

Fuyuan Simon Chang

A Thesis Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 7 2

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Fengyuan S Chang

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Theodore L Williams

THEODORE L. WILLIAMS
Associate Professor of Electrical Engineering

December 8, 1972

Date

ACKNOWLEDGMENTS

Sincere gratitude must be extended to Dr. Theodore L. Williams for his help and guidance throughout the preparation of this thesis. His numerous suggestions were invaluable.

Thanks must also be given to Dr. Donald G. Schultz for his continued support to help make it possible.

TABLE OF CONTENTS

TABLE OF CONTENTS--Continued

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

This thesis is an analytical investigation of errors in the implementation of digital filters due to reconstruction errors and quantization errors. Reconstruction errors in both input and output are due to the inter-sample difference between the continuous-time signals and the signals reconstructed from the discrete-time output. Quantization errors are due to the limited number of bits used in storing the variables and the coefficients of the discrete state equations. For fixed-point computation, scale factors are used to decrease the round-off error of the product of the control transition matrix and the input by scaling the elements of the control transition matrix up and by scaling the input down. The effects of the round-off errors on the eigenvalues of the digital filter are expressed in terms of the root sensitivity function. The effects of the sampling interval on the tradeoffs in the reconstruction and quantization errors are discussed.

CHAPTER 1

INTRODUCTION

With the advent of smaller, faster digital processing equipment, digital filtering is becoming a more attractive alternative to analog filtering.

In the engineering literature, the term digital filter is usually applied to a filter operating on a sequence of samples of a continuous-time signal of limited bandwidth. Quantization errors are introduced by the limited number of bits used in storing the variables and the coefficients of the discrete state equations. Truncation errors are introduced by truncating the series expansion in the calculation of the state transition matrix and the control transition matrix. By taking enough terms in the series approximation, the truncation errors are usually limited to the same order as the quantization errors. Reconstruction errors are introduced in the output and the input by the intersample difference between the continuous time signals and the signals reconstructed from the corresponding discrete-time output in the input approximation and the output reconstruction. Chapter 2 includes the effects of the truncation errors and the round-off errors in the calculation of the state transition matrix and the

control transition matrix from the state variable description of the corresponding analog filter. The errors can be limited to the same order as the round-off errors by selecting enough terms in the series representation of the state transition matrix and the control transition matrix. The round-off errors are kept reasonable if the elements in the series are nearly one or smaller. In Chapter 3, the effects of the round-off errors in the elements of the state transition matrix on the eigenvalues of the digital filter are studied by root locus method. The root locus is used to determine the motion of the poles of the digital filter as the round-off errors in the elements of the state transition matrix are varied. The effects of the round-off errors on the eigenvalues also can be expressed in terms of a root sensitivity function. In Chapter 3, the root sensitivity function is used to show that the round-off errors in the diagonal elements of the state transition matrix have a greater effect on the roots than the round-off errors in the off-diagonal elements. Since the input to the continuous-time system u(t) is needed on the whole sample interval in order to calculate the convolution integral in the state equation, u(t) is approximated by a polynomial expansion which is a fit to the samples of the input. In particular, the number of terms in the polynomial expansion required to meet the prescribed accuracy is investigated in Chapter 4 for a sinusoidal input. A factor

of ten reduction in the sampling interval is required to achieve the same accuracy as taking one more term in the polynomial approximation of the input. In Chapter 5, it is shown that the values of the control transition matrix are relatively small with respect to the values of the inputs for the small sampling interval. Since the elements in the control transition matrix are small, for a fixed-point computation the round-off errors in the product of the input and the control transition matrix are improved if the control transition matrix is scaled up and the input is scaled down by an equal factor. Two sets of scale factors are investigated. The first one is a separate scale factor for every row in the control transition matrix which corresponds to one scale factor for each state. The second one is just one scale factor to scale up the entire control transition matrix and to scale down all the inputs. The separate scale factor minimizes the percentage error of the product for each state; however, the single scale factor is easier to implement. In Chapter 6, the time delay inherent in the digital-analog conversion is discussed and two methods to overcome this delay are compared. For a high speed computer the next output can be calculated in less than one sample interval so that the output may be reconstructed early. The other method adds an extra zero to the transfer function being simulated which compensates for the delay. For high speed

computation it is often possible to compute the output samples by one-half sample early to compensate for one-half sample delay in the conversion. The second method is realizable only if the transfer function of the continuous-time filter has a pole-zero excess. In Chapter 7, the effect of the sampling interval on the tradeoffs between the system reconstruction errors and the round-off errors are discussed. As the sampling interval is decreased the amount of computation per second is increased and the effect of the round-off errors on the stability is more pronounced. As the sampling interval is increased the reconstruction errors are increased.

CHAPTER 2

ERRORS IN $\Phi(T)$

2.1 General Approach

This chapter discusses numerical problems associated with the off-line calculation of the state transition matrix (STM) prior to implementing a digital filter. Since the calculation of the STM is very tedious when done by hand a general purpose computer is used; however, some precautions must be taken because even small errors in the components of the STM matrices can result in substantial errors in the output. Heuristically the errors may be built up due to the recursive nature of the filter equations which corresponds to continuous-time counterpart to the filter. The most popular computational method simply sums the first N terms in the matrix exponential series. Two errors cause all the trouble in this method. The least troublesome is the errors caused by truncation of the exponential series. The second is the round-off error (ROE) associated with the word length of the computer. This chapter shows that both errors are tolerable if the time interval is not longer compared to any of the time constants of the filter. The final time interval can be increased by successive squaring of the STM with less

error than attempting a single calculation for a relatively long sample time.

2.2 Problem Statement

A general state variable representation for a system governed by an n^{th} order, linear, time-invariant differential equation consists of a set of n simultaneous first-order differential equations. If such a system has m inputs and n outputs, usually $m \leq n$, this representation takes on the more general form

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.1}$$

$$y(t) = Cx(t) + Du(t) \tag{2.2}$$

Here x is an n-dimensional state vector, u is an m-dimensional control vector, y is an r-dimensional output vector, A is an nXn system matrix, B is an nXm control matrix, and C is an rXm transmission matrix. Here the Du(t) term indicates a direct coupling of the input to the output. Since the direct coupling of the input to the output is rare in control systems, we shall use

$$y(t) = Cx(t) \tag{2.3}$$

as an output expression.

The fundamental principle of a state variable representation lies in the fact that if one is given the initial conditions, $x(t_o)$, and knowledge of the input u(t) for time $t \geq t_o$, the values of x(t) and y(t) for time

$t \geq t_o$ are completely determined. The state of the system at time t is given by

$$x(t) = \Phi(t-t_o)x(t_o + \int_{t_o}^{t} \Phi(t-\xi)Bu(\xi)d\xi \qquad (2.4)$$

Here $\Phi(t)$ is called the STM of A and is defined as the sum of an infinite series,

$$\Phi(t) = \sum_{k=0}^{\infty} (A^k t^k)k! \quad \text{with } (A^o t^o)/0! = I \qquad (2.5)$$

The state at sample time KT+T is given by

$$x(KT+T) = \Phi(T)x(KT) + \int_{0}^{T} \Phi(T-\tau)Bu(\tau+KT)d\tau,$$

$$0 \leq \tau < T \qquad (2.6)$$

In addition, if the system is sampled every T seconds and if the input u(t) is not varied drastically during the sampling interval, then the input can be approximated by u(KT) for $KT \leq t \leq KT+T$; otherwise, the input u(t) will be treated later by a Taylor series approximation. For this case Equation (2.6) can be reduced to

$$x(KT+T) = \Phi(T)x(KT) + \left(\int_{0}^{T} \Phi(\zeta)d\zeta \right) Bu(KT), \qquad (2.7)$$

where $\zeta = T - \tau$

Usually this solution is simplified and written as follows:

$$x(KT+T) = \Phi(T)x(KT) + H_o(T)u(KT) \tag{2.8}$$

Here $\Phi(T)$ is as previously defined, and the control transition matrix is defined by

$$H_o(T) = h_o(T)B, \tag{2.9}$$

where

$$h_o(T) = \int_0^T \Phi(\zeta)d\zeta \tag{2.10}$$

In digital filtering, Equation (2.10) may be expressed in a series form by integrating each term of the series for $\Phi(T)$. The results are

$$h_o(T) = IT + \frac{AT^2}{2!} + \frac{A^2T^3}{3!} + \ldots + \frac{A^kT^{k+1}}{(k+1)!} + \ldots \tag{2.11}$$

Now, if each side is premultiplied by A and then I is added to each side, $\Phi(T)$ and $h_o(T)$ are related by

$$\Phi(T) = I + Ah_o(T) \tag{2.12}$$

Equation (2.12) is a useful result for digital filtering, since only the series for $h_o(T)$ need to summed. The values for $H_o(T)$ and $\Phi(T)$ can be determined from $h_o(T)$ by Equations (2.9) and (2.12) respectively.

Since the digital filter output is the result of numerous matrix multiplications, it is certainly possible that the filter could contain so many errors that it would not be an accurate model. To ensure the needed accuracy of the filter, generally both $\Phi(T)$ and $H_o(T)$ are calculated with the greatest possible accuracy within the particular computer involved for the off-line simulation.

Two kinds of error are considered here: the truncation error, which is sue to a finite approximation of the infinite expansion, and the ROE, which is due to the finite bit capacity of the digital computer. Each is considered separately.

2.3 Truncation Error

The first of the techniques is due to Liou (1966). From Equation (2.5), the series is uniformly convergent in a finite interval (Varga, 1962). It is, therefore, possible to limit the truncation error within any prescribed limit. If the series is truncated at k = N, then Equation (2.5) can be rearranged into two terms M and R,

$$\Phi(T) = \Sigma_{k=0}^{N} \frac{(AT)^k}{k!} + \Sigma_{k=N+1}^{\infty} \frac{(AT)^k}{k!} = M + R \qquad (2.13)$$

The first term in Equation (2.13) represents its series approximation, while the second term corresponds its remainder.

If each element in $\Phi(T)$ is required to be within an accuracy of at least d significant figures, then

$$|r_{ij}| \leq 10^{-d} \ |m_{ij}| \tag{2.14}$$

where r_{ij} and m_{ij} correspond to the elements of R and M respectively. Let the norm of A be

$$||A|| = \left(\Sigma_{i=1}^{n} \ \Sigma_{j=1}^{n} \ (a_{ij})^2 \right)^{1/2} \tag{2.15}$$

Then, it can be easily seen that

$$||A^k|| \leq ||A|| \ , \ k=1, \ 2, \ \ldots \tag{2.16}$$

Hence, each element of A^k is less than or equal to $||A||^k$. It follows that

$$|r_{ij}| \leq \Sigma_{k=N+1}^{\infty} \frac{||A||^k T^k}{k!} \tag{2.17}$$

Let the ratio of the second term to the first term of the series, Equation (2.17), be ϵ, that is

$$\epsilon = \frac{||A||T}{N+2} \tag{2.18}$$

from which we conclude that

$$\frac{||A||T}{2} \leq \epsilon \tag{2.19}$$

Substituting the inequality (2.19) into the inequality (2.17), we have

$$|r_{ij}| \leq \frac{||A||^{N+1} T^{N+1}}{(N+1)!} \frac{1}{1-\epsilon} \tag{2.20}$$

With the help of Equations (2.13), (2.14), (2.18), and (2.20), $\Phi(T)$ can be evaluated to a prescribed accuracy so that by including enough terms we may enforce the truncation error to be smaller than the round-off error.

2.4 Round-Off Error

Equation (2.5) can be calculated to a prescribed accuracy without the truncation error, that is

$$e^{AT} = \sum_{k=0}^{N_1} \frac{(AT)^k}{k!} \tag{2.21}$$

$$= \left(\sum_{k=0}^{N_2} \frac{(A\sigma)^k}{k!} \right)^{2^\alpha} \tag{2.22}$$

$$= (e^{A\sigma})^{2^\alpha} \tag{2.23}$$

where

$$\sigma = 2^{-\alpha} T \tag{2.24}$$

From Equation (2.17), the smaller the value of AT, the lower the upper bound of the remainder would be. The matrix series would converge faster. For σ small, N_2 may be small, and only α additional multiplications are required to get e^{AT}. That is

$$e^{AT} = \underbrace{\left(\left(\left| (e^{A\sigma})^2 \right|^2 \right) \right)^2}_{\alpha \text{ times}} \tag{2.25}$$

Consider calculating $e^{A\sigma}$ by the finite series

$$e^{A\sigma}\Big|_{\text{calculated}} = \Sigma_{k=0}^{N} \frac{(A\sigma)^k}{k!} \tag{2.26}$$

One can find an N for any σ such that only round-off error, and no truncation error, occurs and the least significant term can be reduced to and within the prescribed bound.

In particular,

$$\frac{n\sigma^N(\max|a_{ij}|)^N}{N!} \leq n\eta \max_{\text{all } 1 < N} \left(\frac{(|a_{ij}|\sigma)^1}{1!} \right)$$

$$\eta = \gamma^{-\beta} \tag{2.27}$$

where n = order of A, γ = numerical base, and β = number of digits used.

Define the following terms:

K_k = ROE in finding A^k

L_k = ROE in finding $(A\sigma)^k/k!$

M_k = ROE in adding $(A\sigma)^k/k!$ to $\Sigma_{1=0}^{k-1} (A\sigma)^1/1!$

Then, the N$^{\text{th}}$ order series approximation for $e^{A\sigma}$ is given by

$$e^{A\sigma}\Big|_{\text{calculated}} = \Sigma_{k=0}^{N} \frac{(A^k+K_k)\sigma^k}{k!} + L_k + M_k \tag{2.28}$$

If the higher order terms of η are neglected, Mankin and Hung (1969) show that

$$||K_k|| < \left[(1+n\eta)^k - 1\right] ||A||^k \tag{2.29}$$

and Simes and Mastascusa (1970) show that

$$||L_k|| \leq \frac{||(A\sigma)^k||}{k!} 2n\eta k \tag{2.30}$$

$$||M_k|| \leq \left(\Sigma_{l=0}^{k} \frac{||(A\sigma)^l||}{l!} \right) n\eta \tag{2.31}$$

where $||K_k||$, $||L_k||$, and $||M_k||$ are norms of K_k, L_k, and M_k respectively. Let $||R_\sigma||$ be norms of round-off error R_σ in calculating $e^{A\sigma}$. Then,

$$||R_\sigma|| \leq n\eta \; e^{||A||\sigma} \; (3||A||\sigma + N) \tag{2.32}$$

Hence, $e^{A\sigma}$ can be expressed in the form

$$e^{A\sigma}| \text{ calculated} = \Phi_\sigma + R_\sigma \tag{2.33}$$

where $\Phi_\sigma = e^{A\sigma}$ to the prescribed accuracy. e^{AT} can be calculated in the similar method shown in Equation (2.33) by

$$e^{AT}| \text{ calculated} = (\Phi_\sigma + R_\sigma)^{2^\alpha} \tag{2.34}$$

where $T = 2^\alpha \sigma$.

As shown in Equation (2.25), this takes a multiplication and when round-off errors are considered, Simes and Mastascusa (1970) show that

$$||e^{AT}| \text{ calculated} - (\Phi_\sigma + R_\sigma)^{2^\alpha}|| \leq (2^\alpha - 1)n\eta \; ||\Phi_\sigma||^{2^\alpha} \tag{2.35}$$

If $||R_\sigma|| << ||\Phi_\sigma||$, then Equation (2.34) can be reduced as follows.

$$(\Phi_\sigma+R_\sigma)^{2^\alpha} \simeq \Phi_\sigma^{2^\alpha} + 2^\alpha \Phi_\sigma^{2^\alpha-1} R_\sigma \tag{2.36}$$

When Equations (2.32), (2.35), and (2.36) are combined together, the round-off error in e^{AT}, ROE_T is given by

$$||ROE|| \leq 2^\alpha \Phi^{2^\alpha-1} n\eta \; e^{||A||\sigma} (3||A||\sigma+N) + (2^\alpha-1)n\eta \; ||\Phi_\sigma||^{2^\alpha} \tag{2.37}$$

In general, $||\Phi_\sigma|| < e^{||A||\sigma}$ and $2^\alpha - 1 < 2^\alpha$, Equation (2.37) can be reduced further

$$||ROE_T|| \leq n\eta \; e^{||A||T} \left(3||A||(N+1)/\sigma\right)T \tag{2.38}$$

Equation (2.27) can be approximated as

$$\frac{n\sigma^N(\max|a_{ij}|)^N}{N!} \leq n\eta,$$

or

$$\sigma \leq \frac{(\eta N!)^{1/N}}{\max|a_{ij}|} \tag{2.39}$$

Substituting Equation (2.39) into Equation (2.38), we get

$$||ROE_T|| \leq n\eta \; Te^{||A||T}\left(3||A|| + \frac{(N+1)\max|a_{ij}|}{(\eta N!)^{1/N}}\right) \tag{2.40}$$

Obviously, the smaller the term $(N+1)\max|a_{ij}|/(\eta N!)^{1/N}$, the smaller the $||ROE_T||$. Table 2.1 shows the value of $(N+1)/(\eta N!)^{1/N}$ for various values of N when $\eta = 10^{-7}$. It is easily seen that when N is greater than 11, the accuracy is hardly improved. At the same time when $\eta N!$ is approximately equal to the unity, that is, when

$$\eta^{-1} = N! \tag{2.41}$$

then Equations (2.39) and (2.41) are good figures for a high accuracy calculation.

Table 2.1[a] Values of $(N+1)/(\eta N!)^{1/N}$ with $\eta = 10^{-7}$

N	$(N+1)/(\eta N!)^{1/N}$	$\eta N!$
1	2×10^{14}	10^{-7}
5	5.66×10	1.2×10^{-5}
11	1.06×10	3.99
20	5.6	2.4×10^{11}
30	4.1	3.0×10^{57}

[a]Simes and Mastascusa (1970, p. 127).

2.5 Summary of the Recommended Procedure

Select σ such that the reciprocal of σ is less than or equal to the maximal element of the A-matrix. Calculate $e^{A\sigma}$ by taking N terms to a prescribed accuracy, and then square $e^{A\sigma}$ α times so that $T = \sigma 2^{\alpha}$ and $\eta = 1/N!$ always hold. It is better to calculate e^{AT} by squaring $e^{A\sigma}$ α times which costs an additional α bits, than to calculate e^{AT} directly by series which loses even more bits. By taking 11 terms to calculate $e^{A\sigma}$, we can obtain the results up to 7 digit accuracy. A further increase in the number of terms does not help much because the ROE dominates.

The ROE discussed in this chapter is applicable to the calculation of the state transition matrix and the control transition matrix. Usually the state transition matrix is calculated from Equation (2.12) by calculating $h_o(T)$ first according to Equation (2.11).

In Section 2.3 the implementation of Liou's method forms a recursive procedure to calculate $\Phi(T)$ to a prescribed accuracy if AT is not too large. However, if the sampling interval T is large compared to the frame time σ in Equation (2.24), then the number of terms N_1 required in Equation (2.21) becomes larger than the number of terms N_2 required in Equation (2.22) for Simes and Mastascusa's method. If the sampling interval is equal to or less than the frame time, both methods reach to the same accuracy. If the sampling time interval is quite large compared to

the frame time, then only Simes and Mastascusa's method can converge the series before the ROE dominates. As a result, Simes and Mastascusa's method is more efficient.

CHAPTER 3

ROOT SENSITIVITY

3.1 General Approach

Once the STM matrix is calculated, it must be stored with limited precision in the digital processor. The ROE connected with the truncation of the elements of the STM shifts the roots of the discrete time transfer function. The roots of the z-domain transfer function are

$$\det(zI - \Phi) = 0 \tag{3.1}$$

The subsequent discussion relates the motion of these roots to the magnitude of the ROE in the elements of the STM. One particularly annoying situation occurs when the roots are shifted to the outside of the unit circle, resulting in an unstable response. The root locus is a classical approach which is amenable to both hand and computational methods. The ROE is envisioned as a gain change and the resulting motion of the roots may be charted by applying root locus methods.

3.2 Particular Case in the Second-Order Filter

To analyze the results of error in the Φ-matrix, let us consider only a 2 x 2 matrix. Without error the eigenvalues are found as follows:

$$\Phi = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \tag{3.2}$$

Solve for the characteristic equation

$$(z-\Phi_{11})(z-\Phi_{22}) - \Phi_{12}\Phi_{21} = 0, \tag{3.3}$$

or

$$(z-z_1)(z-z_2) = 0 \tag{3.4}$$

Where the two eigenvalues z_1 and z_2 are determined by solving the quadratic equation. But if one of the diagonal elements contains some error ϵ,[1] then

$$\Phi = \begin{pmatrix} \Phi_{11}-\epsilon & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \tag{3.5}$$

The resulting eigenvalues are

$$(z-\Phi_{11})(z-\Phi_{22}) - \Phi_{12}\Phi_{21} + \epsilon(z-\Phi_{22}) = 0 \tag{3.6}$$

The first two terms in Equation (3.6) are the same as Equation (3.3); that is,

$$(z-z_1)(z-z_2) + \epsilon(z-\Phi_{22}) = 0 \tag{3.7}$$

---

1. Here ϵ is arbitrary defined to be a number such that if the number is positive, it will cause a reduction of the Φ-element. In this case the root locus is known as a π-degree root locus. On the other hand, if ϵ is negative, then it is a zero-degree root locus.

To see the effect of ϵ upon the roots of Equation (3.7) it is put into the root locus form

$$\frac{\epsilon(z-\Phi_{22})}{(z-z_1)(z-z_2)} = -1 \qquad (3.8)$$

Where z_1 and z_2 are the eigenvalues of the characteristic equation without error.

In the drawing of the root locus shown in Fig. 3.1, z_1 and z_2 are the poles, Φ_{22} is a zero and ϵ is the gain. As can be seen, the error ϵ effectively shifts each of the roots to some new locations. Although the poles may lie anywhere within the unit circle centered at the origin, the closer they are to the circumference of the unit circle the smaller the tolerance of error will have to shift one of them outside of the unit circle. If ϵ is positive, there is little danger of shifting the poles far enough to cause the system to become unstable. On the other hand, if ϵ is negative, then a zero-degree root locus will result. From Fig. 3.1(b), one of the loci will start on the pole z_2 and will terminate at positive infinity on the real axis. In this case, any (negative) infinitesimal error will cause z_2 to be shifted toward the edge of the unit circle. Hence a sufficient error will cause the stable continuous system to be unstable when it is modeled as a digital filter. In this case, the amount of error which will just shift the

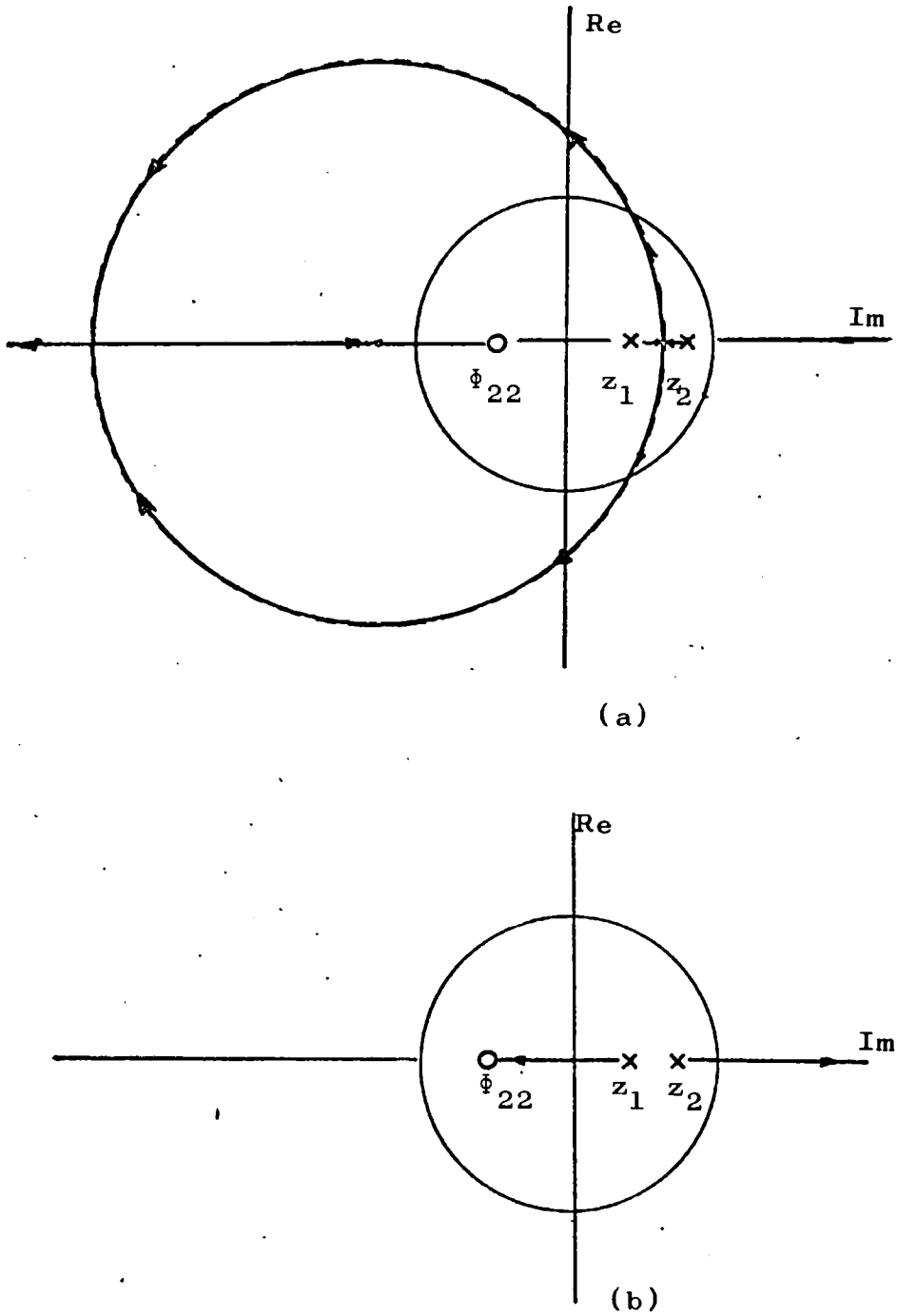


Fig. 3.1. Root locus of a second-order digital filter with ROE in the diagonal element -- (a) $\epsilon > 0$; (b) $\epsilon < 0$.

pole to the edge of the unit circle is

$$\epsilon = - \frac{(1-z_1)(1-z_2)}{(1-\Phi_{22})} \tag{3.9}$$

Equation (3.9) says that the error needed to shift one pole to the edge is the product of the distances of each pole to the edge divided by the distance of the zero to the edge. If the error is considered in Φ_{22}, the effect is exactly the same except the zero is at $z = \Phi_{11}$ instead of $z = \Phi_{22}$.

Now consider the case where the error ϵ is in one of the off-diagonal elements. Then

$$\Phi = \begin{pmatrix} \Phi_{11} & \Phi_{12}-\epsilon \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \tag{3.10}$$

Solve for the characteristic equation

$$(z-z_1)(z-z_2) + \epsilon\Phi_{21} = 0, \tag{3.11}$$

where z_1 and z_2 are the eigenvalues of the equation without error. Set Equation (3.11) into the root locus form

$$\frac{\epsilon\Phi_{21}}{(z-z_1)(z-z_2)} = 1 \tag{3.12}$$

For which the root locus is shown in Fig. 3.2.

As in the case of error in a diagonal element, if ϵ is positive, it is not likely to be large enough to shift the poles into an unstable region. But if ϵ is negative,

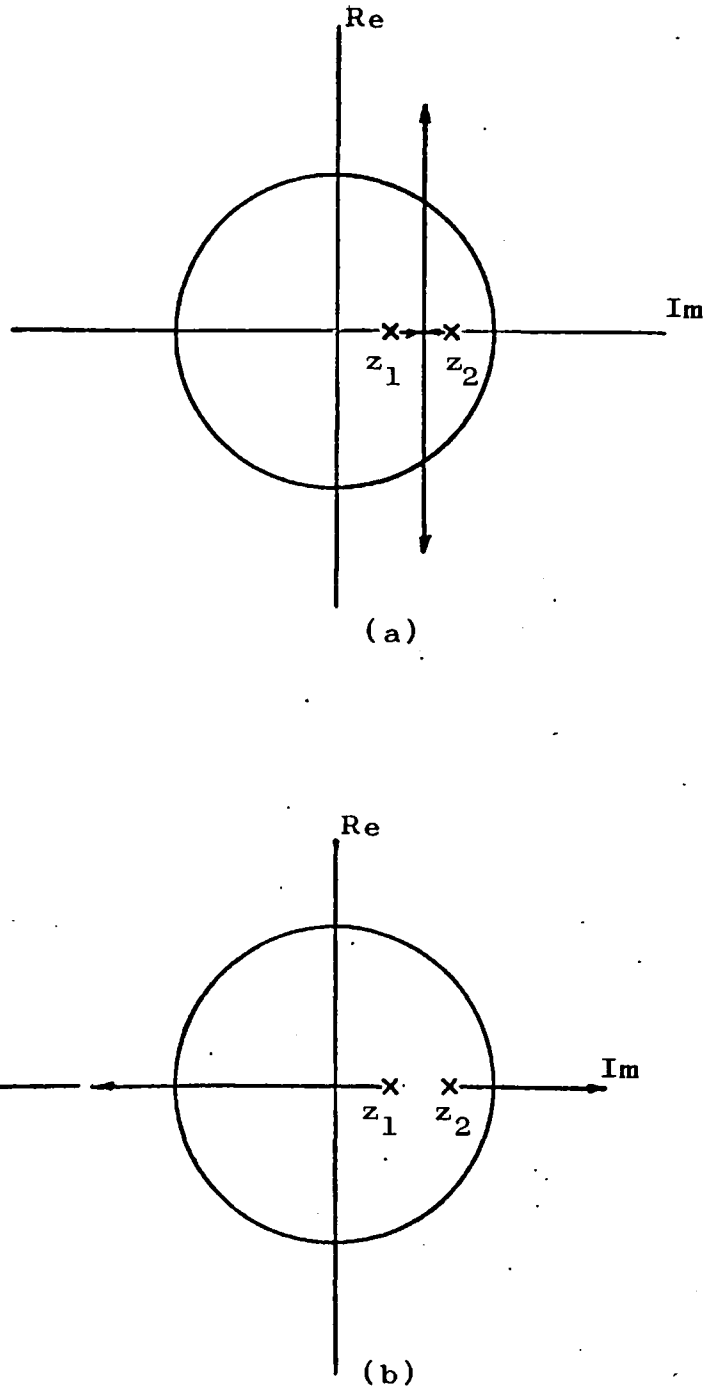


Fig. 3.2. Root locus of a second-order digital filter with ROE in the off-diagonal element -- (a) $\epsilon > 0$; (b) $\epsilon < 0$.

one of the poles is shifted toward the edge of the unit circle along the real axis. Again the amount of error needed to shift one of the poles to the edge is

$$\epsilon = - \frac{(1-z_1)(1-z_2)}{\Phi_{21}} \tag{3.13}$$

The analysis of error in the Φ_{21} element is exactly the same except the denominator is Φ_{12} instead of Φ_{21}.

3.3 General Case in the Second-Order Filter

Now consider the case where errors occur in each of the Φ-elements. Then

$$\Phi = \begin{pmatrix} \Phi_{11}-\epsilon_{11} & \Phi_{12}-\epsilon_{12} \\ \Phi_{21}-\epsilon_{21} & \Phi_{22}-\epsilon_{22} \end{pmatrix} \tag{3.14}$$

Solve for the characteristic equation

$$(z-\Phi_{11})(z-\Phi_{22}) + \epsilon_{11}(z-\Phi_{22}) + \epsilon_{22}(z-\Phi_{11}) + \epsilon_{11}\epsilon_{22}$$

$$- \Phi_{12}\Phi_{21} + \epsilon_{12}\Phi_{21} + \epsilon_{21}\Phi_{12} - \epsilon_{12}\epsilon_{21} = 0 \tag{3.15}$$

Since the ϵ's are small with respect to the Φ-element, the products of the ϵ's are negligible. That is

$$\epsilon_{11}\epsilon_{22} = \epsilon_{12}\epsilon_{21} = 0 \tag{3.16}$$

Then Equation (3.15) can be reduced to

$$(z-z_1)(z-z_2) + \epsilon_{11}(z-\Phi_{22}) + \epsilon_{22}(z-\Phi_{11}) + \epsilon_{12}\Phi_{21} + \epsilon_{21}\Phi_{12} = 0 \qquad (3.17)$$

There are two ways to analyze Equation (3.17); first let us suppose that we consider the worst case where $\epsilon = \max(\epsilon_{11}, \epsilon_{12}, \epsilon_{21}, \epsilon_{22})$. Then

$$(z-z_1)(z-z_2) + \epsilon(2z-\Phi_{11}+\Phi_{12}+\Phi_{21}-\Phi_{22}) = 0. \qquad (3.18)$$

Equation (3.18) in the root locus form is

$$\frac{2\epsilon\left|z-(\Phi_{11}-\Phi_{12}-\Phi_{21}+\Phi_{22})/2\right|}{(z-z_1)(z-z_2)} = -1 \qquad (3.19)$$

This case is essentially the same as the case where the error existed only in one of the diagonal elements. The only difference is that each of the four Φ-elements affects the location of the zero.

The second way of analyzing Equation (3.17) is to put it into the root locus form for leaving only one of the ϵ's in the numerator. That is

$$\frac{\epsilon_{11}(z-\Phi_{22})}{(z-z_1)(z-z_2) + \epsilon_{12}\Phi_{21} + \epsilon_{21}\Phi_{12} + \epsilon_{22}(z-\Phi_{11})} = -1 \qquad (3.20)$$

It has already been shown how ϵ_{11} will shift the roots of the denominator. However, at this point, the roots of the

denominator can also be put into the root locus form

$$\frac{\epsilon_{12}\Phi_{21}}{(z-z_1)(z-z_2) + \epsilon_{21}\Phi_{12} + \epsilon_{22}(z-\Phi_{11})} = -1 \qquad (3.21)$$

But the roots of this denominator are also unknown, so it is also put into the root locus form

$$\frac{\epsilon_{21}(z-\Phi_{12})}{(z-z_1)(z-z_2) + \epsilon_{22}(z-\Phi_{11})} = -1 \qquad (3.22)$$

Do the same sort of forming

$$\frac{\epsilon_{22}(z-\Phi_{11})}{(z-z_1)(z-z_2)} = -1 \qquad (3.23)$$

This last equation can now be plotted and new root locations found as a result of ϵ_{22}. By knowing these, it gives us the root locations of the denominator of Equation (3.22). As a result, the new root locations can be found which are the result of both ϵ_{22} and ϵ_{21}. Continuing in this manner, Equation (3.21) can be solved and then Equation (3.20).

From the above result, it should be fairly obvious that the errors in each of the Φ-elements shift the pole locations by some amount in addition to shift resulting from the errors in the remaining Φ-elements. Therefore, these errors are additive, or shown in Fig. 3.3. Furthermore, from the first method, assuming that $\epsilon = \max(\epsilon_{11}, \epsilon_{12}, \epsilon_{21}, \epsilon_{22})$, we can say that the errors occurred in the

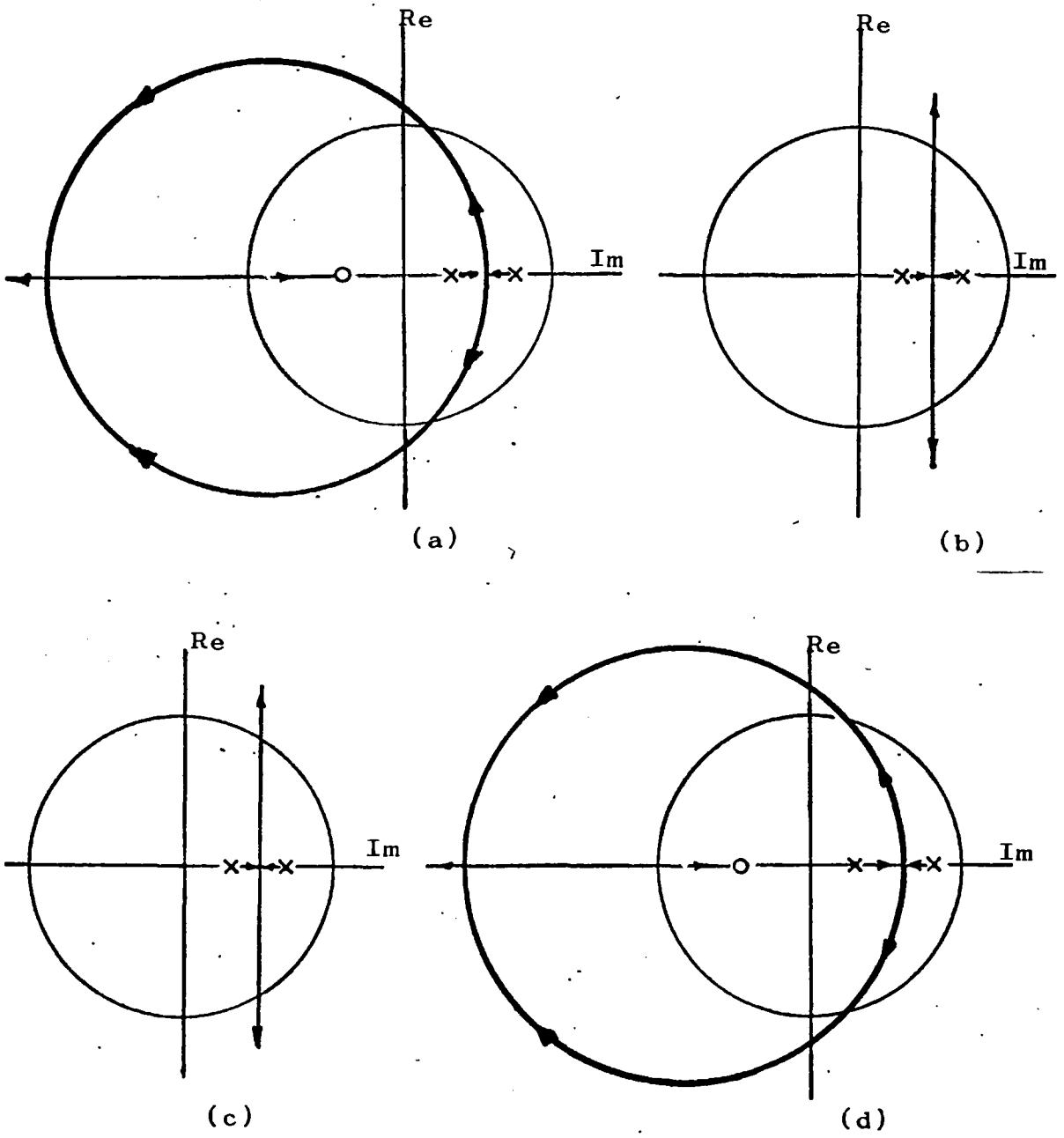


Fig. 3.3. Root locus of a second-order digital filter with ROE in every Φ-element, where ROE reduces the Φ-element -- (a) ε_{22} only, (b) ε_{21} only, (c) ε_{12} only, (d) ε_{11} only.

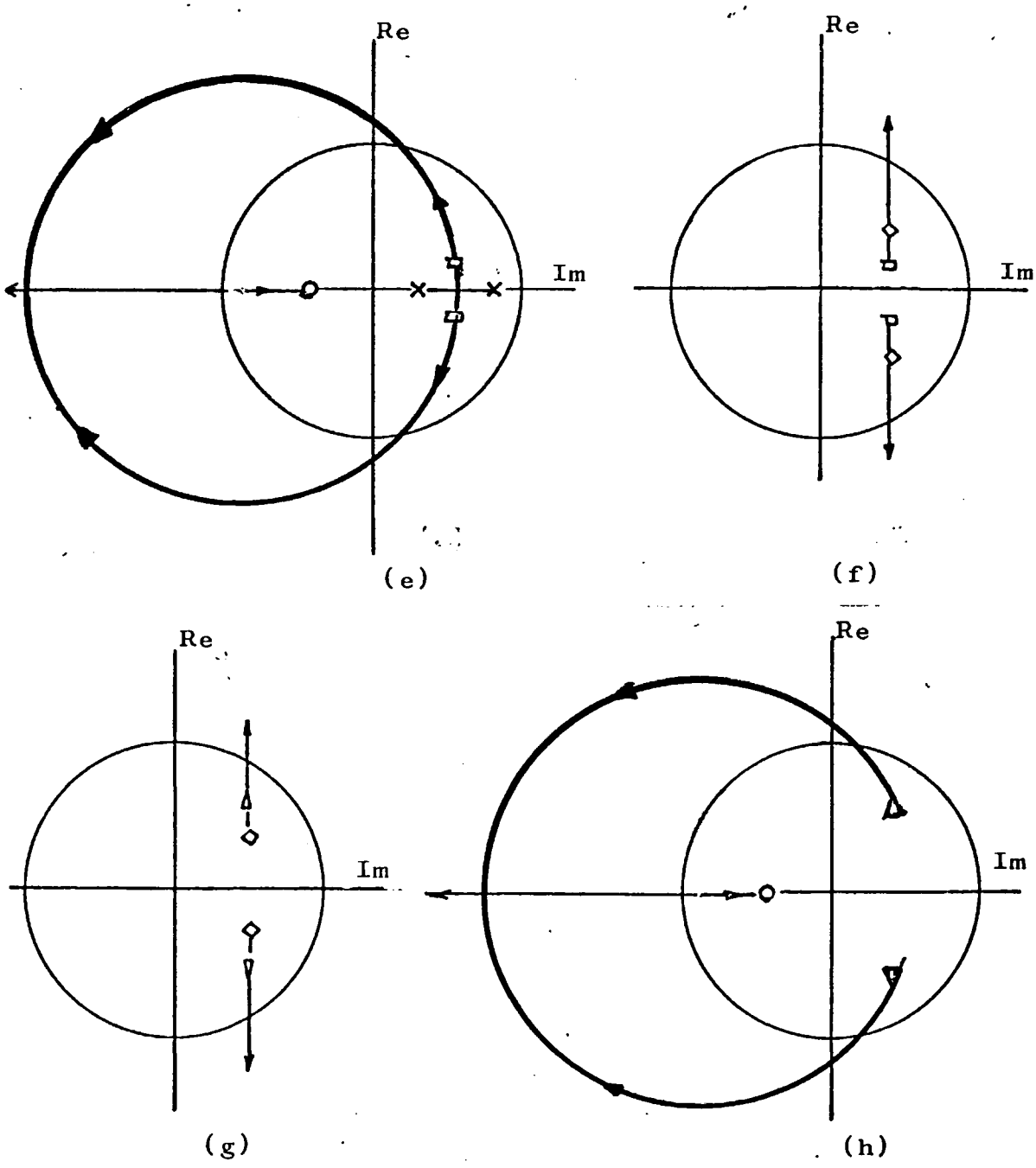


Fig. 3.3.--Continued -- (e) conjugate poles □ shifted by ϵ_{22}, (f) ◇ shifted by ϵ_{22} and ϵ_{21}, (g) △ shifted by ϵ_{22}, ϵ_{21}, and ϵ_{12}, (h) locus formed by ϵ_{22}, ϵ_{21}, ϵ_{12}, and ϵ_{11}.

diagonal elements dominate those in the off-diagonal elements.

3.4 Particular Case in the Higher Order Filter

The next case considered would logically be a 3 x 3 Φ-matrix. However, in using the same procedure as before, most useful results become lost in algebraic complexity. For this reason the 3 x 3 case is not considered here. Next we cascade two 2 x 2 filters expressed in terms of the phase variable form (Schultz and Melsa, 1967), the resulting continuous state equations will be

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -a & -b & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -c & -d \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} u(t) \qquad (3.24)$$

And the resulting Φ-matrix will be[1]

$$\Phi(T) = \begin{pmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} & \Phi_{14} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} & \Phi_{24} \\ 0 & 0 & \Phi_{33} & \Phi_{34} \\ 0 & 0 & \Phi_{43} & \Phi_{44} \end{pmatrix} \qquad (3.25)$$

---

1. Finding the inverse of (sI-A) and the inverse of LaPlace Transformation of that is a tedious process. However, to verify this $\Phi(T)$, it is only necessary to determine which of the elements of the adjoint of (sI-A) are zero.

Solve for the characteristic equation

$$\left[(z-\Phi_{11})z-\Phi_{22}) - \Phi_{12}\Phi_{21}\right]\left[(z-\Phi_{33})(z-\Phi_{44}) - \Phi_{34}\Phi_{43}\right] = 0,$$

or

$$\left[(z-z_1)(z-z_2)\right]\left[(z-z_3)(z-z_4)\right] = 0 \qquad (3.26)$$

The above solution produces an interesting result. Namely, that the eigenvalues for this 4 x 4 case are divided into two sets. From Φ_{11}, Φ_{12}, Φ_{21}, and Φ_{22}, z_1 and z_2 are determined; while z_3 and z_4 are determined from Φ_{33}, Φ_{34}, Φ_{43}, and Φ_{44}. As a result, errors in Φ_{11}, Φ_{12}, Φ_{21}, or Φ_{22} will only affect z_1 and z_2; while any error in Φ_{33}, Φ_{34}, Φ_{43}, or Φ_{44} will only affect z_3 and z_4. However, any error in Φ_{13}, Φ_{14}, Φ_{23}, or Φ_{24} will not affect any of the eigenvalues of this composite filter. The above results are another supporting matter to ensure that the errors in the diagonal elements dominate those in the off-diagonal elements.

3.5 Root Sensitivity

In general, for a n x n Φ-matrix its characteristic equation is an n^{th} order rational function in z and its n eigenvalues have nothing to do with the appearance of the zeros caused by the presence of the errors to some certain Φ-elements. Thereby, ϵ can be well defined as the ratio of the product of the distances from poles to point z' on

the root locus to the product of the distances from zeros to point z' on the root locus. This can be expressed symbolically[1]

$$\epsilon = \frac{\pi_{k=1}^{n} |z'-p_k|}{(\pi_{k=1}^{m} |z'-z_k|)C_{\Phi}} ,$$

$$n - 2 \leqq m \leqq n - 1, \text{ and } m = 0, 1, 2, \ldots, n - 1 \qquad (3.27)$$

Where z' is the new pole location shifted by the presence of errors, p_k and z_k are poles and zeros in the z-plane respectively, n is the order of the filter and m is the number of zeros caused by the presence of errors in the Φ-element, and C_{Φ} is a constant.

If error occurs only in any of the diagonal elements, the order of m is n - 1; while error occurs only in any of the off-diagonal elements, the order of m is n - 2. Of course, if error occurs in both of them, the order of m is n - 1.

Next, we consider the sensitivity of the pole with respect to the error of the elements of the Φ-matrix, $S_{\epsilon}^{z'}$,

$$S_{e}^{z'} = \frac{\epsilon dz'}{z' d\epsilon} \qquad (3.28)$$

In the above equation, the reciprocal of $dz'/d\epsilon$ can be calculated from Equation (3.27)

---

1. The sign convention on ϵ defined as before.

$$\frac{d\epsilon}{dz'} = \frac{d}{dz'}\left[\frac{\pi_{k=1}^{n} |z'-p_k|}{(\pi_{k=1}^{m} |z'-z_k|)C_\Phi}\right] = \frac{1}{(\pi_{k=1}^{m} |z'-z_k|)^2 C_\Phi^2}$$

$$\cdot \left[(\Sigma_{k=1}^{n} \pi'^{n}_{\substack{l=1 \\ l\neq k}} |z'-p_l|)\left((\pi_{k=1}^{m} |z'-z_k|)C_\Phi\right)\right.$$

$$\left. - \left(\Sigma_{k=1}^{m} (\pi'^{n}_{\substack{l=1 \\ l\neq k}} |z'-z_l|)C_\Phi\right)(\pi_{k=1}^{n} |z'-p_k|)\right]$$

$$= \epsilon(\Sigma_{k=1}^{n} \frac{1}{|z'-p_k|} - \Sigma_{k=1}^{m} \frac{1}{|z'-z_k|}) \qquad (3.29)$$

where $\pi'^{n}_{\substack{l=1 \\ l\neq k}}$ is defined as the deleted product of $|z'-p_l|$.

Substituting Equation (3.29) into Equation (3.28), we get the root sensitivity function

$$S_\epsilon^{z'} = \frac{1}{z'}\left[\frac{1}{\Sigma_{k=1}^{n} \frac{1}{|z'-p_k|} - \Sigma_{k=1}^{m} \frac{1}{|z'-z_k|}}\right] \qquad (3.30)$$

Equation (3.27) shows that the effect of zeros on the root sensitivity due to the errors in the elements of the state transition matrix. From Equation (3.30), the root sensitivity function has one fewer zero for the case that the round-off errors occur in the off-diagonal elements, with the same location of the poles as those for the case that the round-off errors occurred in the diagonal elements. Thus, the denominator of Equation (3.30) is larger for the case that the errors in the off-diagonal

elements; consequently, the value of its root sensitivity function is smaller.

3.6 Summary of Results

The root locus is used to determine the motion of the poles of the digital filter as the round-off errors in elements of the state transition matrix are varied. The effects of the round-off errors on the eigenvalues also can be expressed in terms of a root sensitivity function, which shows that the eigenvalues of the digital filter are more sensitive to the round-off errors in the diagonal elements than those in the off-diagonal elements.

CHAPTER 4

INPUT APPROXIMATION ERROR

4.1 General Approach

The output of the continuous-time system could be reproduced exactly at the sampling instants by the digital filter, provided that the convolution integral in Equation (2.6) could be evaluated exactly; however, in the digital simulation the input is not known over the whole interval, but just at the discrete sampling instants. Although a spline[1] may be used to form a minimum mean square error approximation of the input from the power spectral density of this input, we will just approximate the input by a Taylor series. Thereby, extra storage of input samples and undesirable phase shifts will be averted. As more terms are taken in the expansion of the input the accuracy improves, but estimates of higher order differences of the input are required. Often accurate estimates of these differences are not available; nevertheless, these terms may be used as estimates of the error when the input is approximated by a step approximation.

---

1. A kind of input approximation that minimizes, with respect to a set of discrete linear structures, the maximum possible error between the samples of the output of the analog filter and the corresponding output of the digital simulator.

4.2 Input Approximation

Equation (2.4) is referred to as the state transition equation of the system described by Equation (2.1). In the sampled-data control system the sampling interval is relatively small between two sampling instants kT + T and kT, the input u(t) can be obtained by Taylor series approximation (Wait, 1969)

$$u(t) = u(\tau+kT)$$

$$= u(kT) + \frac{\tau}{T}\Delta u(kT) + \frac{\tau^2}{2!T^2}\Delta^2 u(kT) + \ldots \qquad (4.1)$$

where $kT \leq \tau < kT + T$, $\Delta u(kT) = u(kT) - u(kT - T)$, and $\Delta^2 u(kT) = u(kT) - 2u(kT - T) + u(kT - 2T)$. Substitute Equation (4.1) into Equation (2.6), we get

$$x(kT + T) = \Phi(T)x(kT + \int_0^T \Phi(T - \tau)Bu(kT)d\tau$$

$$+ \int_0^T \Phi(T - \tau)B\frac{\tau}{T}\Delta u(kT)d\tau$$

$$+ \int_0^T \Phi(T - \tau)B\frac{\tau^2}{2!T^2}\Delta^2 u(kT)d\tau + \ldots \qquad (4.2)$$

Define

$$H_m = H_m(T) = \left(\int_0^T \left(\frac{\sigma}{T}\right)^m \Phi(\sigma)d\sigma\right)B, \quad m = 0, 1, 2, \ldots \qquad (4.3)$$

Then Equation (4.2) can be reduced to the form

$$x(kT + T) = \Phi(T)x(kT) + H_o u(kT)$$

$$+ (H_o - H_1)\Delta u(kT) + \frac{1}{2}(H_o - 2H_1 + H_2)\Delta^2 u(kT)$$

$$\simeq \Phi(T)x(kT) + I \qquad (4.4)$$

where the integral $I = H_o u(kT) + (H_o - H_1)\Delta u(kT)$

$$+ \frac{1}{2}(H_o - 2H_1 + H_2)\Delta^2 u(kT) + \ldots$$

In order to investigate these effects upon the input approximation, three examples of digital filters are presented. For the comparison, two second-order filters with different time constants are given. Then the two second-order filters are cascaded into one fourth-order filter. The block diagrams for each are illustrated in Figs. 4.1(a), (b), and (c). In the form of Equation (2.1), the A and B matrices for Example 1 are

$$A_1 = \begin{pmatrix} 0 & 1 \\ -13.7 & -3 \end{pmatrix} \qquad \text{and } B_1 = \begin{pmatrix} 0 \\ 13.7 \end{pmatrix}$$

For Example 2,

$$A_2 = \begin{pmatrix} 0 & 1 \\ -1 & -.5 \end{pmatrix} \qquad \text{and } B_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

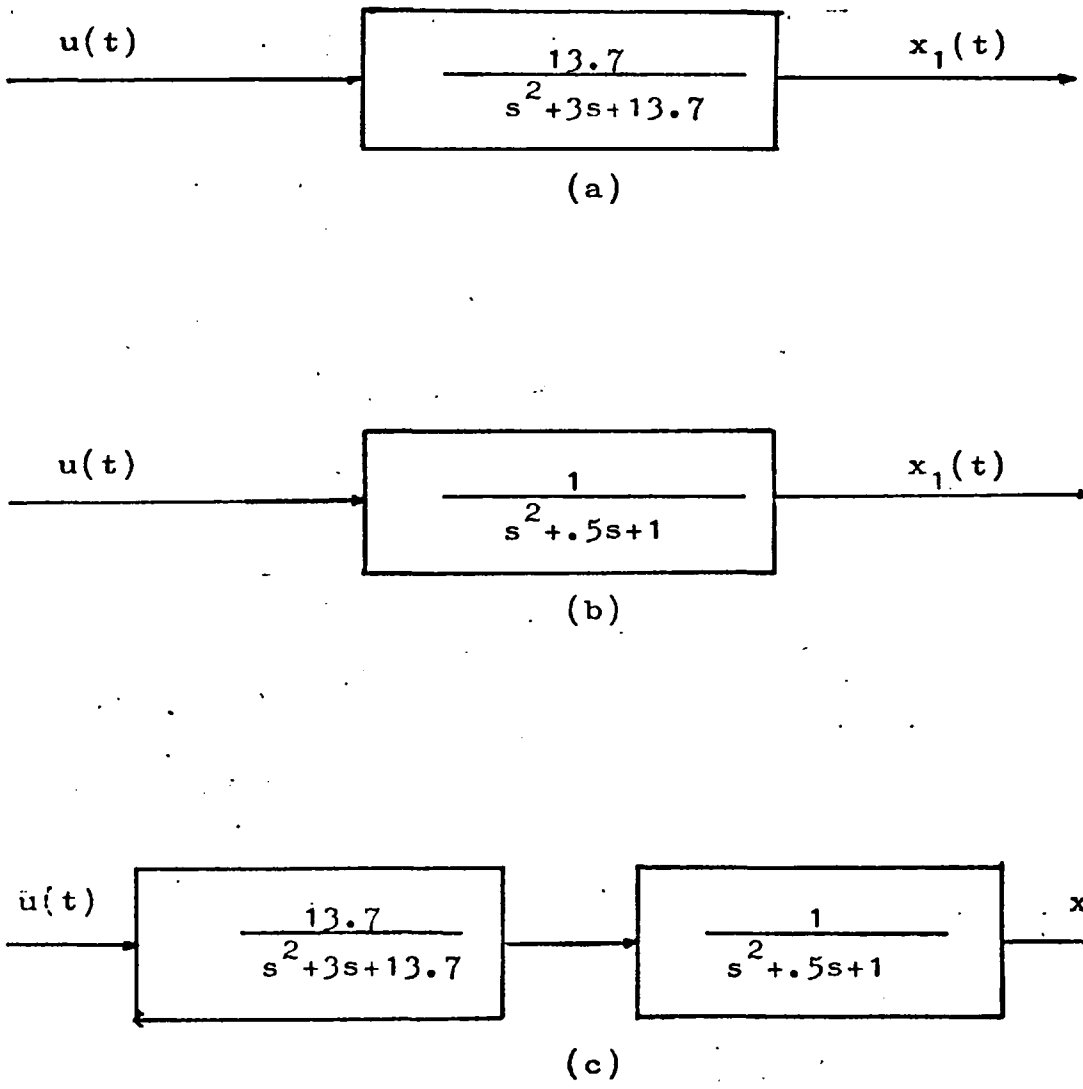


Fig. 4.1. The linear, continuous filters (a), (b), and (c) used for Examples 1, 2, and 3 respectively.

For Example 3,

$$A_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & -.5 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -13.7 & -3 \end{pmatrix} \quad \text{and} \quad B_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 13.7 \end{pmatrix}$$

The sampling interval T is chosen less than any time constant of the system to be simulated by the digital filter and less than the period of the highest frequency component in the input to the filter. Suppose that the term containing the highest frequency in the input u(t) is

$$u(t) = \cos(w_o t)$$

If the first-order and second-order differences of the input are substituted into Equation (4.4), we get

$$I = H_o\cos(kw_oT) - (H_o - H_1)(w_oT)\sin(kw_oT)$$

$$- \frac{1}{2}(w_oT)^2(H_o - 2H_1 + H_2)\cos(kw_oT) \qquad (4.5)$$

For small enough T, the magnitudes of the sinusoids of Equation (4.5) are arranged into

$$H_o > (H_o - H_1)(w_oT) > \frac{1}{2}(w_oT)^2(H_o - wH_1 + H_2) \qquad (4.6)$$

Particularly, the integral of Equation (4.5) for $x_1(kT)$ in Example 1 is obtained from Table 4.1,

$$(2.50504\text{x}10^{-2}) > (1.65564\text{x}10^{-2})(w_oT) > (6.17620\text{x}10^{-3})(2_oT)^2$$

Table 4.1. Coefficients of u(kT), $\Delta u(kT)$, $\Delta^2 u(kT)$.

	Example 1 T=0.0625 sec	Example 2 T=0.1 sec	Example 3 T=0.0625 sec
H_o	2.50504×10^{-2}	4.91351×10^{-3}	8.32528×10^{-6}
	7.73826×10^{-1}	9.73787×10^{-2}	5.26606×10^{-4}
			2.50504×10^{-2}
			7.73826×10^{-1}
$H_o - H_1$	1.65564×10^{-2}	3.26839×10^{-3}	6.64731×10^{-6}
	3.73019×10^{-1}	4.82425×10^{-2}	3.93402×10^{-4}
			1.65564×10^{-2}
			3.73019×10^{-1}
$\frac{H_o - 2H_1 + H_2}{2}$	6.17620×10^{-3}	1.22398×10^{-3}	2.76587×10^{-6}
	1.22010×10^{-1}	1.60054×10^{-2}	1.56946×10^{-4}
			6.17620×10^{-3}
			1.22010×10^{-1}

For Example 2,

$$(4.91361\text{x}10^{-3}) > (3.26839\text{x}10^{-3})(\ _{o}T) > (1.22398\text{x}10^{-3})(w_{o}T)^{2}$$

For Example 3,

$$(8.32528\text{x}10^{-6}) > (6.64731\text{x}10^{-6})(w_{o}T) > (2.76587\text{x}10^{-6})(w_{o}T)^{2}$$

It is obvious that unless the term $w_{o}T$ is sufficiently less than unity, then the first three terms in Equation (4.1) must be considered. In the case above the highest frequency in the input must be less than 16 radians per second when the sampling interval is .0625 second in Examples 1 and 3, and less than 10 radians per second when the sampling interval is .1 second in Example 2.

If the term $w_{o}T$ is much smaller than unity, we can neglect the squared term of $w_{o}T$. In other words, we only consider the first two terms in Equation (4.1). If the term $w_{o}T$ is much greater than unity, then we know that the sampling interval is so large that the first few terms of Equation (4.5) are not a very accurate representation of the integral I. It is better to reduce the sampling interval so that the term $w_{o}T$ is much less than unity or approximately equal to unity.

4.3 Summary of Results

To reduce the round-off errors, the sampling interval T must be less than any time constant of the

system to be simulated. From Table 4.1, $H_o > (H_o - H_1) > \frac{1}{2}(H_o - 2H_1 + H_2)$ holds for most cases, which indicates that the higher order differences of Taylor series approximation in the input quantization are of much less importance. When the highest frequency component of the input is known, then we can determine the number of terms required in the expansion of the input to the digital filter. From Table 4.2, for 10^{-3} accuracy in the integral I of Equation (4.5), then the first-order difference should be considered in the input approximation for w_oT equal to 10^{-2}; the second-order difference should be considered when w_oT equal to 10^{-1}. The step approximation of the input is normally used where the product of the highest frequency and the sampling interval should be no larger than 10^{-3}. As a result, a factor of ten reduction in the sampling interval is required to achieve the same accuracy as taking one more term in the series approximation of the input.

Table 4.2. Ratios of the magnitudes of the sinusoids with respect to its step approximation.

	$(w_oT)\frac{H_o-H_1}{H_o}$	$\frac{1}{2}(w_oT)^2\frac{H_o-2H_1+H_2}{H_o}$
Example 1 $w_oT = 10^{-1}$	6.60923×10^{-2}	2.46550×10^{-3}
Example 2 $w_oT = 10^{-1}$	6.65170×10^{-2}	2.49099×10^{-3}
Example 3 $w_oT = 10^{-1}$	7.98448×10^{-2}	3.32225×10^{-3}
Example 1 $w_oT = 10^{-2}$	6.60923×10^{-3}	2.46550×10^{-4}
Example 2 $w_oT = 10^{-2}$	6.65170×10^{-3}	2.49099×10^{-4}
Example 3 $w_oT = 10^{-2}$	7.98448×10^{-3}	3.32225×10^{-4}

CHAPTER 5

ROUND-OFF ERROR IN $H_o(T)$

5.1 General Approach

For small sample time $H_o(T)$ is nearly proportional to the sample time T; thus the values of $H_o(T)$-elements are relatively small with respect to those of inputs. For a fixed-point representation of a small number, many significant bits are lost. Often it is better to scale $H_o(T)$ up and to scale u(kT) down in order to minimize the round-off error while still keeping the product of $H_o(T)u(kT)$ unchanged.

5.2 Scale Factors

A block diagram of the digital filter is shown in Fig. 5.1, where z^{-1} represents one sampling interval delay. The state equation of the digital filter is

$$x(kT+T) = \Phi(T)x(kT) + H_o(T)u(kT)$$

A single scale factor could be used to scale all the elements in the H_o-matrix or a separate scale factor could be used for every element in a particular row of $H_o(T)$. If the latter course is taken every input must be scaled into every new one corresponding to that particular scale factor with respect to a specific state. For

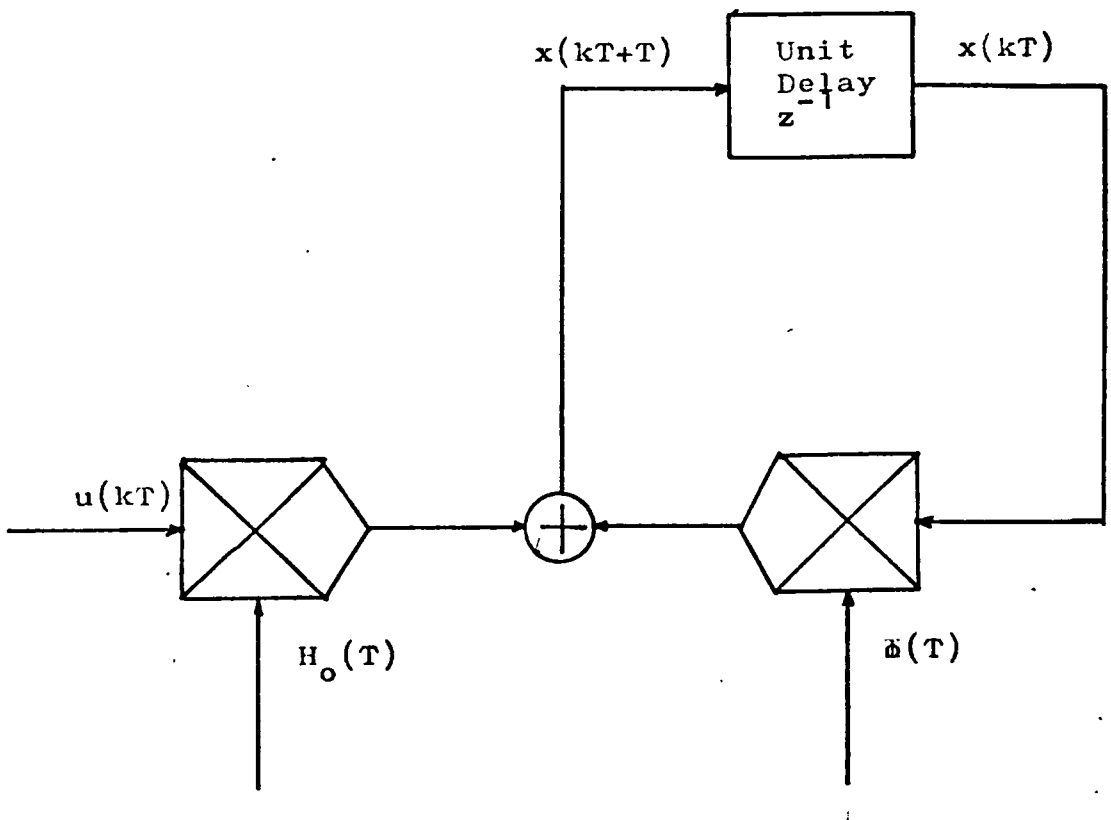


Fig. 5.1. Block diagram of digital filter.

example, we have a second-order multiple-input-multiple-output filter illustrated in Fig. 5.2. We wish to select a scale factor for each state which minimizes the composite round-off errors in $H_o(T)u(kT)$, where n scale factors correspond to n states. To do this, define a percentage error of the i^{th} output

$$\%Er = \frac{\Sigma_{j=1}^{m}(h_{ij}P_i+\epsilon)\left|u_j(kT)/P_i+\epsilon\right| - \Sigma_{j=1}^{m}\left|h_{ij}u_u(kT)\right|}{\Sigma_{j=1}^{m} h_{ij}u_j(kT)}$$

$$= \frac{\Sigma_{j=1}^{m}\left|h_{ij}P_i+u_j(kT)/P_i\right| + m\epsilon^2}{\Sigma_{j=1}^{m} h_{ij}u_j(kT)} \qquad (5.1)$$

where ϵ is the round-off error. We minimize the error in Equation (5.1) with respect to the i^{th} scale factor P_i

$$\frac{d}{dP_i}(\%Er) = \frac{\Sigma_{j=1}^{m}\left|h_{ij}-u_j(kT)/P_i^2\right|}{\Sigma_{j=1}^{m} h_{ij}u_j(kT)} = 0 \qquad (5.2)$$

Solve for P_i

$$P_i = \left|\Sigma_{j=1}^{m} u_j(kT)/\Sigma_{j=1}^{m} h_{ij}\right|^{1/2} \qquad (5.3)$$

On obtaining the scale factor P_i, we can modify the state equation of the filter

$$x(kT+T) = \Phi(T)x(kT) + H_o'(T)u'(kT) \qquad (5.4)$$

where $H_o'(T)$, nX(nXm) matrix, and $u'(kT)$, (nXm)X1 vector, are the modified $H_o(T)$, nXm matrix, and the modified $u(kT)$,

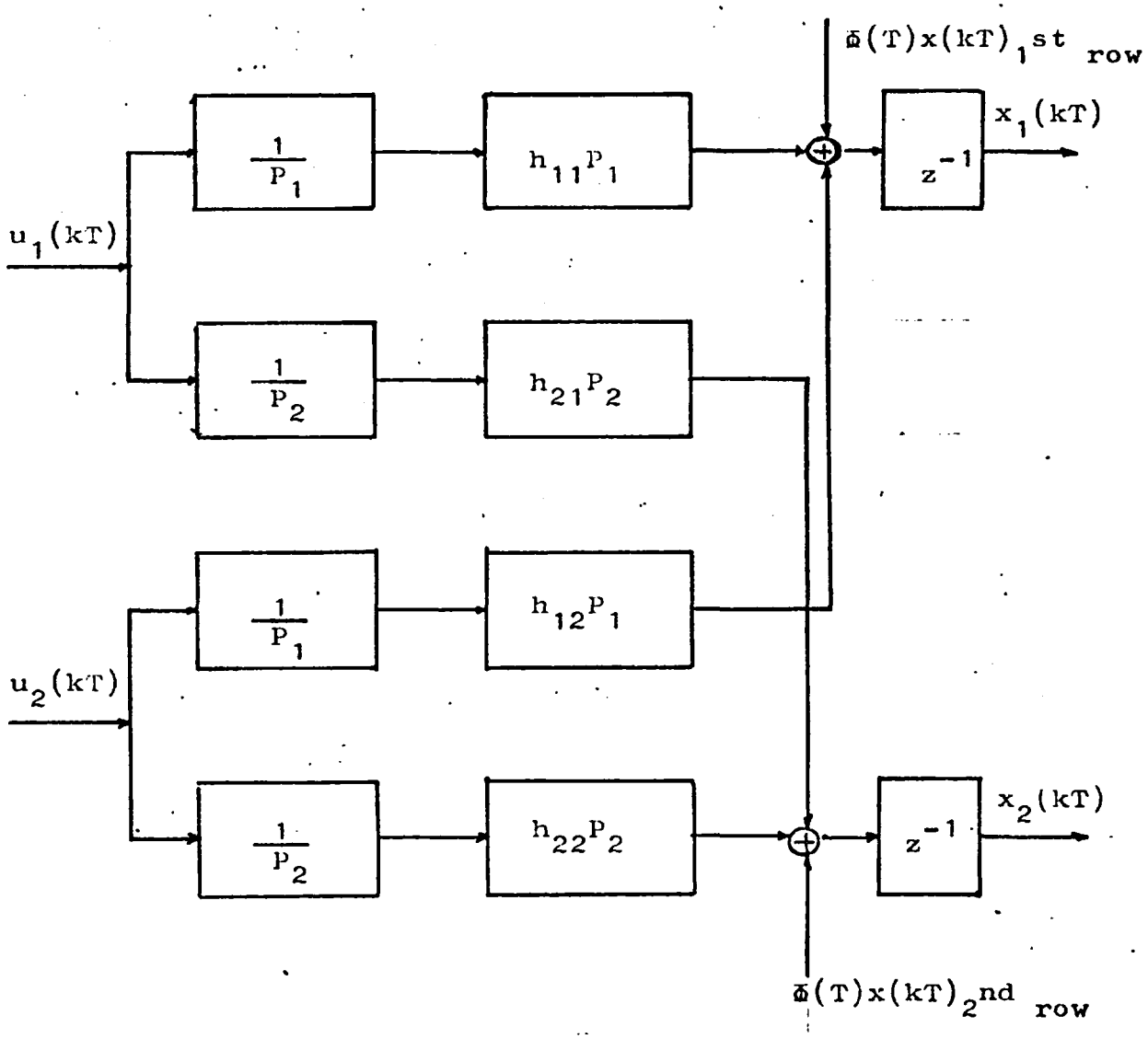


Fig. 5.2. Block diagram of a second-order multiple-input-multiple-output digital filter with separate scale factors.

mX1 vector, respectively. This notation can be illustrated in Fig. 5.3, where $h'_{ij} = h_{ij}P_i$, and $u_{ji}(kT) = u_j(kT)/P_i$ for the i^{th} state $x_i(kT+T)$.

These enlarged matrix manipulations can be easily calculated by the computer. For the second-order multiple-input-multiple-output filter, the product $H'_o(T)u'(kT)$ can be expressed in the form

$$H'_o(T)u(kT) = \begin{pmatrix} h'_{11} & h'_{12} & 0 & 0 \\ 0 & 0 & h'_{21} & h'_{22} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{pmatrix}$$

$$= \begin{pmatrix} h_{11}P_1 & h_{12}P_1 & 0 & 0 \\ 0 & 0 & h_{21}P_2 & h_{22}P_2 \end{pmatrix} \begin{pmatrix} u_1/P_1 \\ u_2/P_1 \\ u_1/P_2 \\ u_2/P_2 \end{pmatrix} \qquad (5.5)$$

The block diagram of Equation (5.5) is illustrated in Fig. 5.2.

The single scale factor could be simply determined by considering their geometric mean values of $H_o(T)$ and $u(kT)$. We can sort the minimum value of H_o-elements and the maximum value of $u(kT)$, and then evaluate their geometric mean value, $(h_{ij}u_k)^{1/2}$. Hence, the scale

$$H_o'(T)u'(kT) = \begin{pmatrix} h_{11}' \cdots h_{1m}' & 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & h_{21}' \cdots h_{2m}' & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & h_{n1}' \cdots h_{nm}' \end{pmatrix} \begin{pmatrix} u_{11} \\ \vdots \\ u_{m1} \\ u_{12} \\ \vdots \\ n_{m2} \\ \vdots \\ u_{1n} \\ \vdots \\ u_{mn} \end{pmatrix}$$

Fig. 5.3. $H_o'(T)u'(kT)$, the modified $H_o(T)u(kT)$.

factor can be determined by

$$P = \frac{(h_{ij}u_k)^{1/2}}{h_{ij}} \tag{5.6}$$

or

$$P = \frac{u_k}{(h_{ij}u_k)^{1/2}} \tag{5.7}$$

Equations (5.6) and (5.7) can be combined into the form

$$P = (u_k/h_{ij})^{1/2} \tag{5.8}$$

where u_k is the maximum value of the u(kT), and h_{ij} is the minimum value of the H_o-elements. For this case the block diagram is illustrated in Fig. 5.4.

5.3 Summary of Results

There are two kinds of scale factors. One is a single scale factor to scale all the H_o-elements, whose value is determined by $(\max_j u_j/\min_{i,j} h_{ij})^{1/2}$, the other is a separate scale factor P_i corresponding to every state $x_i(kT+T)$, whose value is determined by $(\Sigma_{j=1}^{m} u_j(kT)/\Sigma_{j=1}^{m} h_{ij})^{1/2}$. The separate scale factor minimizes the percentage error of the product of the input and the control transition matrix for each state; however, the single scale factor is easier to implement.

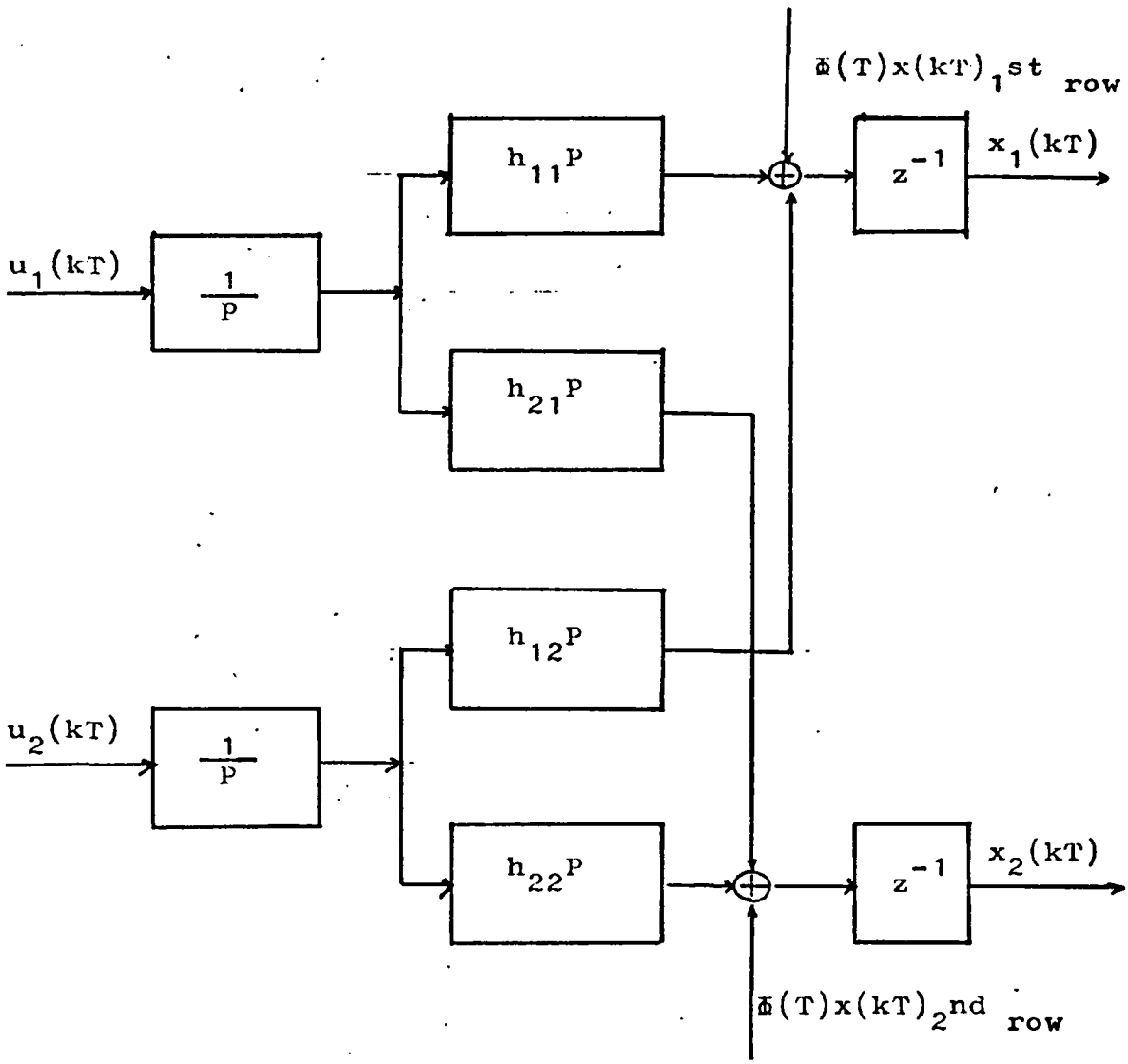


Fig. 5.4. Block diagram of a second-order multiple-input-multiple-output digital filter with one scale factor.

CHAPTER 6

DATA RECONSTRUCTION

6.1 General Approach

Hybrid computation entails conversion from analog to digital signals and back again which inserts an error into the filtering process. The major component of this error is a time delay of about one-half sample interval. The high frequency error components are usually attenuated by the subsequent equipment. After a brief discussion of time delays in reconstruction, this chapter presents two methods of compensating for time delays. The first method is early reconstruction of the output; the second method is adding a zero into the transfer function.

6.2 Time Delay and Reconstruction

A generalized configuration of the hybrid loop is illustrated in Fig. 6.1, where τ_1 is the time delay inherent in analog-digital conversion and usually related to the word length of the digital output, τ_2 is the time required for the digital computations and usually related to the complexity of computations, and τ_3 is the time delay inherent in digital-analog conversion and approximately equal to one-half the sampling interval (Karplus, 1966).

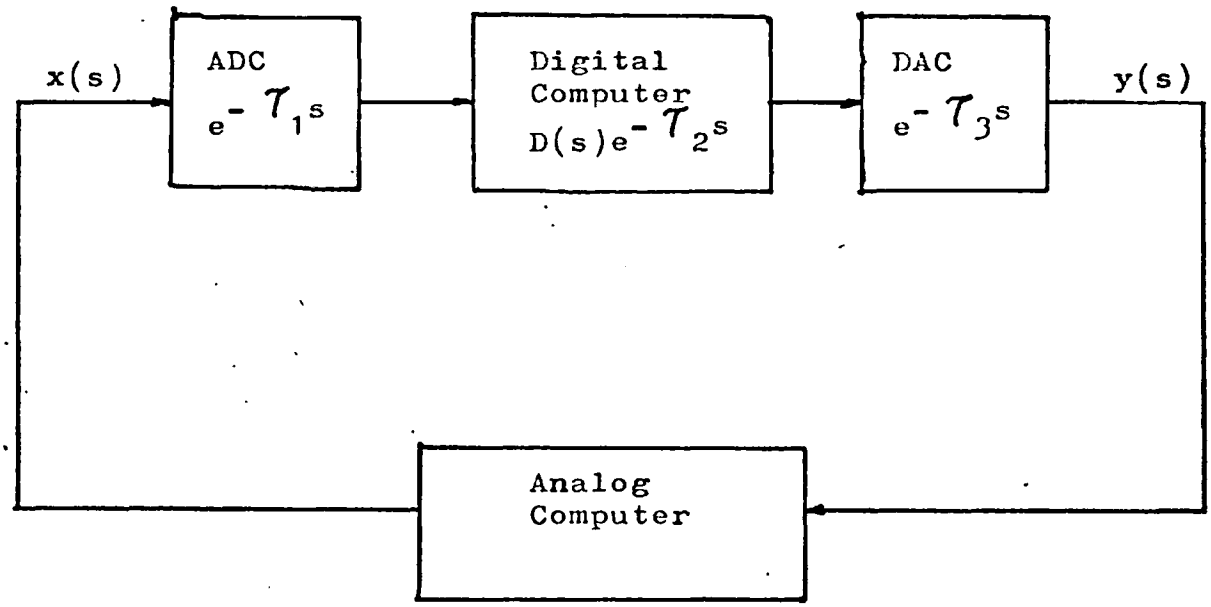


Fig. 6.1. Time delay in hybrid loop.

For instance, consider a zero-order hold realization of the digital-analog converter in Fig. 6.2.

It can be seen that the total time delay through the system is $\tau = \tau_1 + \tau_2 + T/2$. Depending on the analog-digital conversion, digital calculations, and the sampling interval, any one of them may be dominant. Usually the relationship $T/2 > \tau_2 > \tau_1$ holds.

From Figs. 6.1 and 6.2, we have

$$Y(s) = X(s)D(s)e^{-\tau s} \tag{6.1}$$

It is assumed that some sort of prediction circuit is employed to compensate the time delay in the hybrid computation. Time delay compensation can be implemented by introducing a series compensator with a transfer function $e^{-\tau s}$, which corresponds to a time advance or predictor. This term can be approximated by the zero-pole approximation, $e^{-\tau s} \simeq 1 + \tau s$. Then the compensated system yields

$$Y_c(s) = P(s)X(s)D(s)e^{-\tau s} = (1+\tau s)X(s)D(s)e^{-\tau s} \tag{6.2}$$

Expanding $e^{-\tau s} = 1 - \tau s + (\tau s)^2/2! + \ldots$ and substituting into Equation (6.2), we have

$$Y_c(s) = X(s)D(s)\left[1-(\tau s)^2/2 + \ldots\right] \tag{6.3}$$

Neglecting these terms containing the square or higher power of τ, Equation (6.3) becomes the desired output.

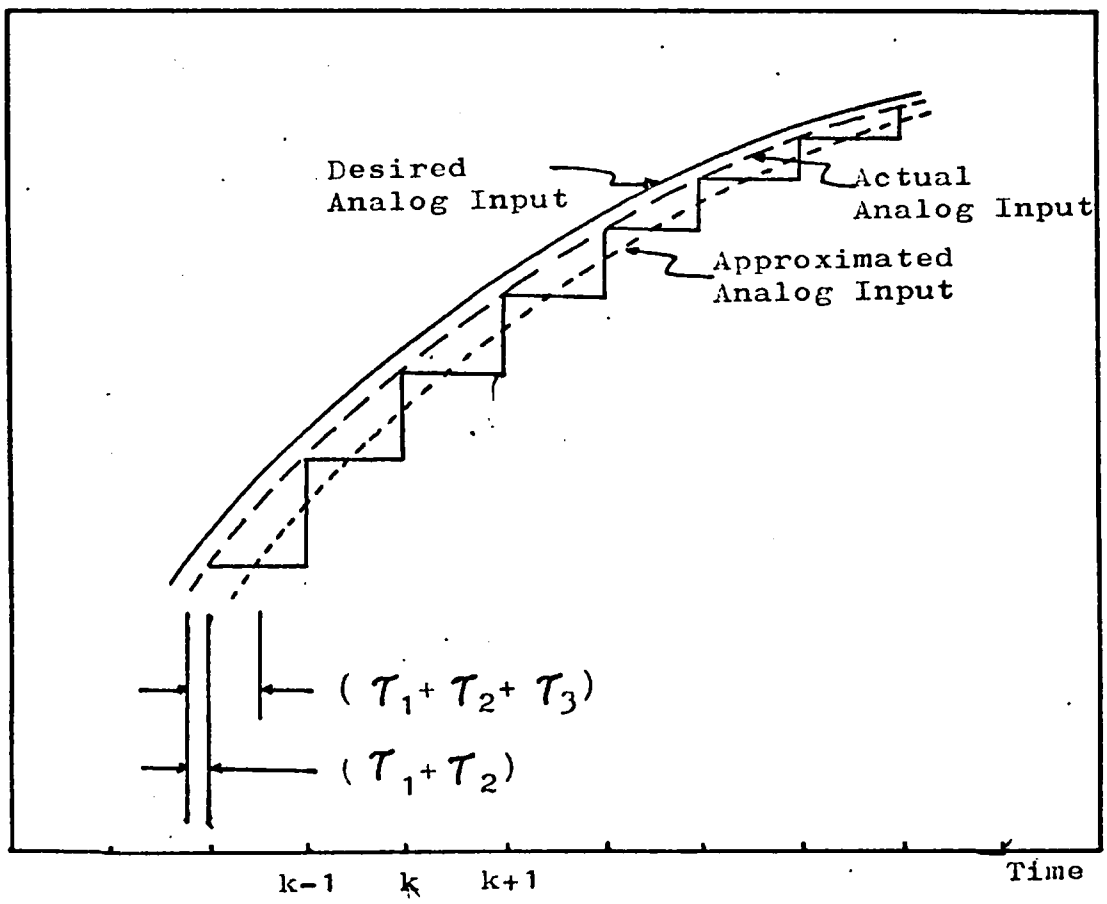


Fig. 6.2. Time delay in quantization.

6.3 Two Methods of Implementation

Method 1 is the early reconstruction of the output by a zero-order hold unit. At the sampling time t = kT all the information is calculated according to Equations (2.1) and (2.2). Generally these calculations take much less than one-half the sampling interval so that we can employ a zero-order hold unit to hold the output signal y(kT+T) at time kT+T/2. The resultant output waveform is depicted in Fig. 6.3.

Compare Fig. 6.2 with Fig. 6.3. The time delay inherent in the hybrid computation is compensated by the early reconstruction of the output.

Method 2 is adding a zero to the transfer function of the digital filter. For example, the transfer function of an n^{th} order digital filter D(s),

$$D(s) = \frac{K(C_m s^{m-1}+C_{m-1}s^{m-2}+\ldots+C_1)}{s^n+a_n s^{n-1}+\ldots+a_1} \tag{6.4}$$

is replaced by modified digital filter $D_m(s)$ with an extra zero,

$$D_m(s) = \frac{K(1+\tau s)(C_m s^{m-1}+C_{m-1}s^{m-2}+\ldots+C_1)}{s^n+a_n s^{n-1}+\ldots+a_1} \tag{6.5}$$

where $m < n$ for the physically realizable digital filter, and this process is illustrated in Fig. 6.4.

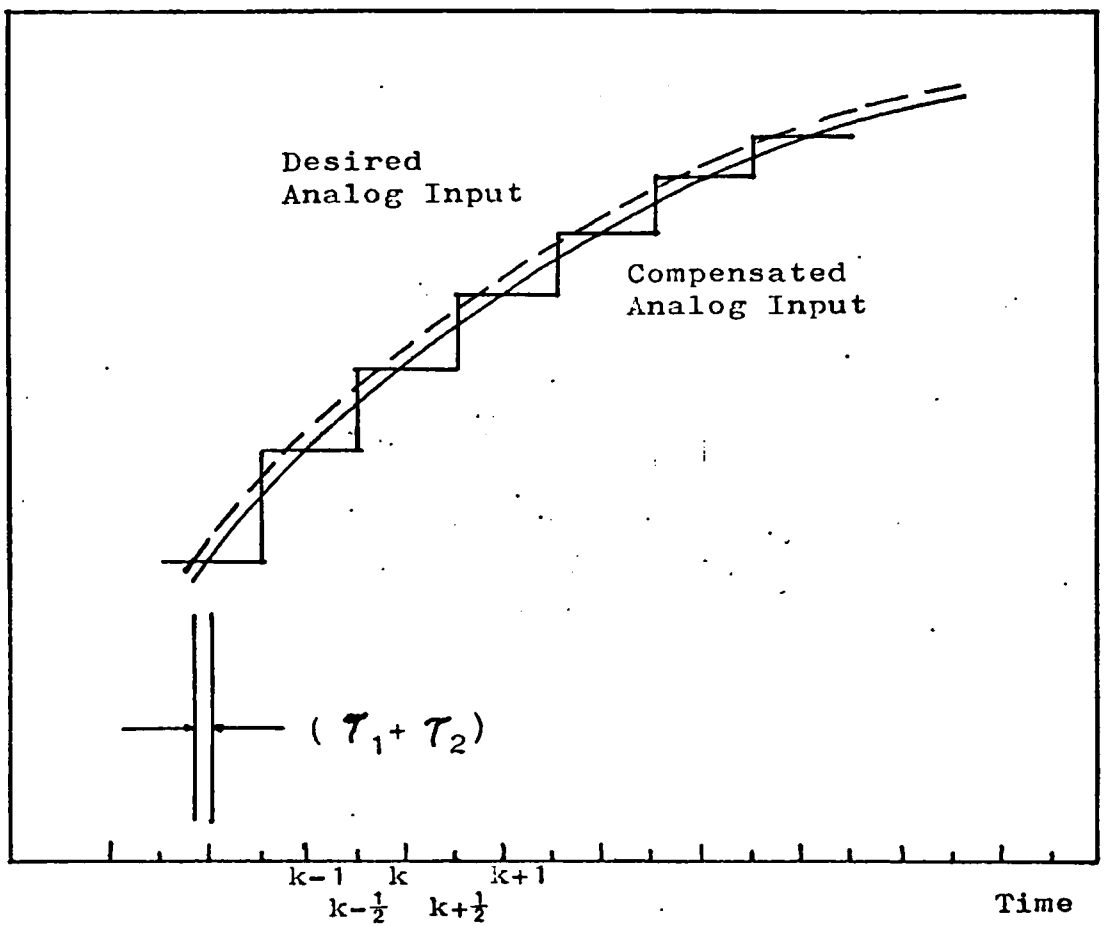


Fig. 6.3. Implementation of Method 1.

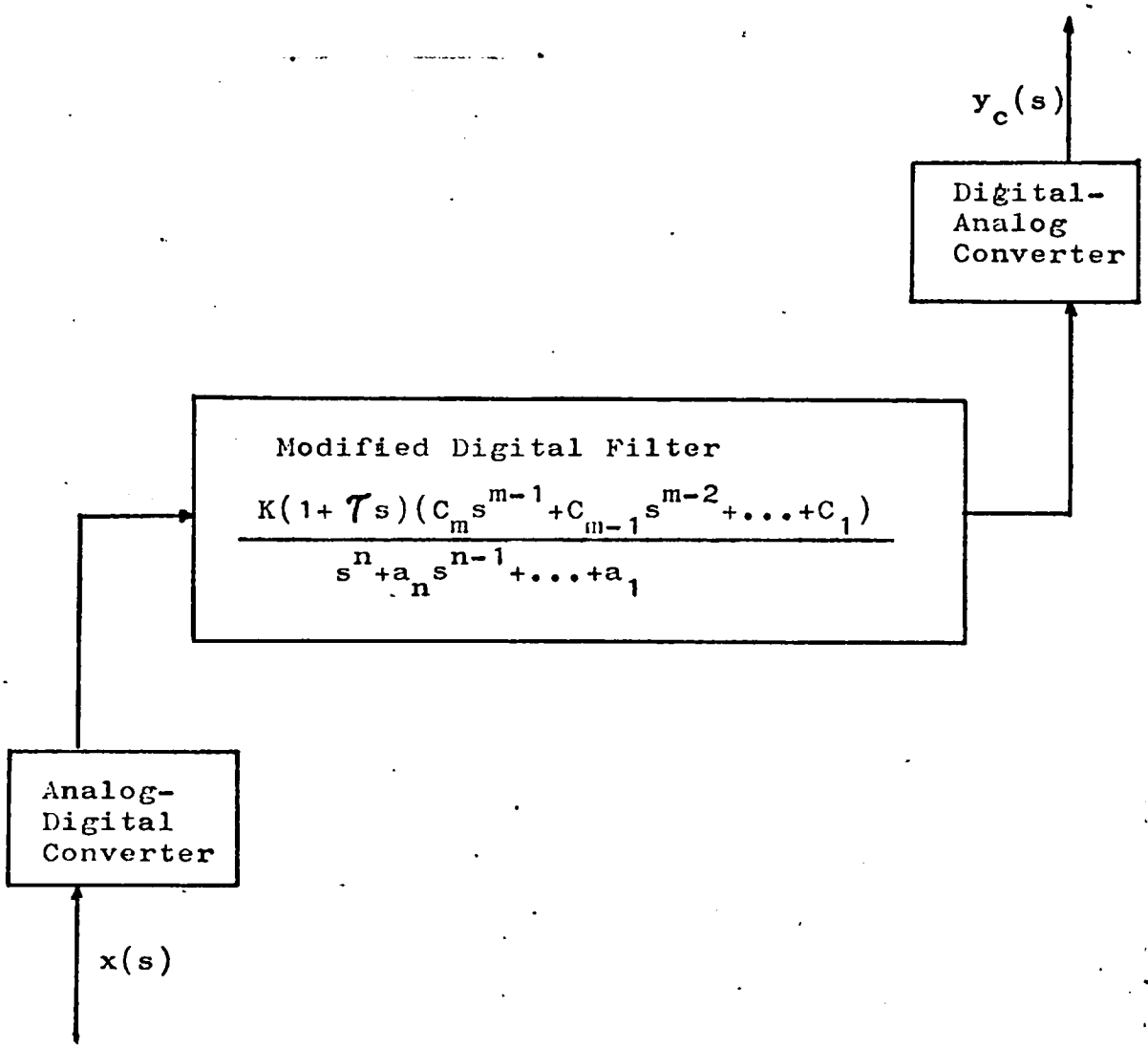


Fig. 6.4. Implementation of Method 2.

The digital filter D(s) can be expressed in terms of phase-variable representation; for example, if we consider the single output case, then $y(t) = z_1(t)$, $\dot{y}(t) = z_2(t)$, and $y_c(t) = y(t) + \tau\dot{y}(t)$; that is,

$$\dot{z}(t) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_n \end{pmatrix} z(t) + \begin{pmatrix} 0 \\ 0 \\ \cdot \\ 0 \\ K \end{pmatrix} u(t)$$

$$y(t) = (1, 0, 0, \ldots, 0)z(t)$$

$$y_c(t) = (1, \tau, 0, \ldots, 0)z(t) \qquad (6.6)$$

The implementation of this method is obvious by taking the compensated filter output $y_c(t)$ through the digital-analog converter to get the desired output in Fig. 6.2.

6.4 Summary of Results

If the speed of the digital computation is relatively faster with respect to the sampling interval, then the early reconstruction of digital output by a zero-order hold unit can be employed in Method 1 to get satisfactorily reconstructed output. Method 2 is good for those higher pole-zero excess filters to ensure the stability of the

digital filters simulating the stable continuous-time systems.

CHAPTER 7

SAMPLING INTERVAL

7.1 General Approach

If the sampling interval T is reduced, the round-off errors are increased and the amount of computation per second is increased. Also, any decrease in the sampling interval would result in very little improvement in the response characteristics. On the other hand, the sampling theorem says that the sampling frequency must be at least twice as large as the frequency of the highest Fourier component of the analog signal being sampled; otherwise, the distortions in the frequency spectrum due to the overlapping of the signal components can not be removed by subsequent filtering or other manipulation. Hence, it is important to recognize that no higher frequencies occur in the analog signal. Accordingly, it is desirable to precede the sampler with a signal conditioner in the form of a low-pass filter.

7.2 Limits on the Sampling Interval

The pole location s in the s-domain (the continuous time system) and its corresponding eigenvalue z in the z-domain (the sampled-data system) are related by

$$z = e^{sT} \tag{7.1}$$

where T is the sampling interval. For example, if we try to simulate a continuous-time system with the pole located at s = -5 by a digital filter and we require that its eigenvalue z should be no larger than z = 0.95. To limit the effect of the ROE on the stability, then the lower bound of the sampling interval is

$$T = \left|\ln(z)\right|/s = 0.01 \text{ sec.} \tag{7.2}$$

If we arbitrarily reduce the lower bound, then in Equation (7.1) the eigenvalue goes to the edge of the unit circle. The digital filter may become unstable due to the ROE. More bits per word would be needed to reduce the ROE and keep the system stable.

In practice system stability of the closed-loop system also dictates the choice of the sampling frequency and makes it necessary to sample at a rate much higher than the theoretical minimum.

For example, we consider a second order system with ROE in its diagonal elements,

$$\dot{x}(t) = \begin{pmatrix} -a & 0 \\ 0 & -b \end{pmatrix} x(t) + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u(t)$$

$$y(t) = (1,1)x(t) \tag{7.3}$$

The corresponding sampled-data system is

$$x(kT+T) = \begin{pmatrix} e^{-aT} & 0 \\ 0 & e^{-bT} \end{pmatrix} x(kT) + \begin{pmatrix} (1-e^{-aT})/a \\ (1-e^{-bT})/b \end{pmatrix} u(kT)$$

$$y(kT) = (1,1)x(kT) \qquad (7.4)$$

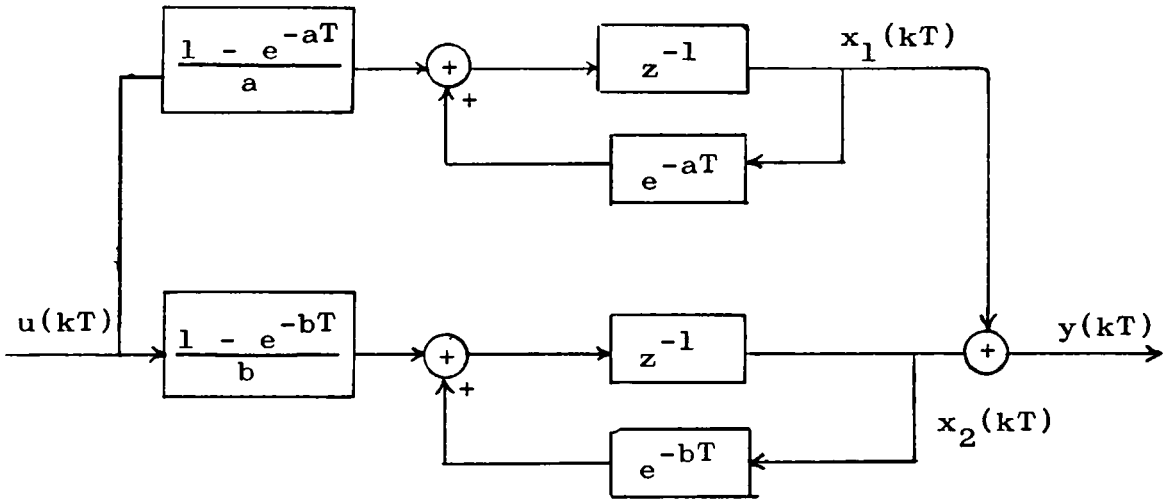
Equation (7.4) can be expressed in terms of the block diagram representation in Fig. 7.1(a) which can be reduced to the form of the transfer function illustrated in Fig. 7.1(b).

The overall z-domain transfer function can be analyzed in the r-domain through the bilinear transformation $z = (1+r)/(1-r)$, and then examined by Routh's stability criterion. The conditions for stable system are

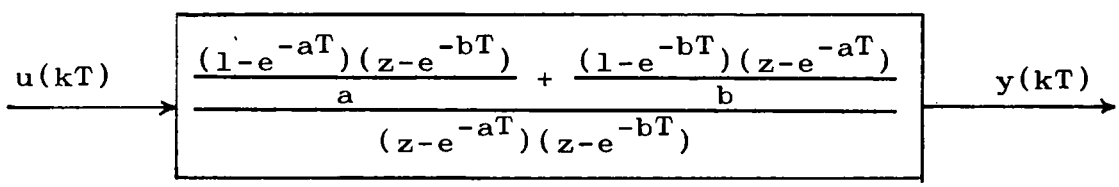
$$1 + e^{-(a+b)T} > e^{-aT} + e^{-bT}, \qquad (7.5)$$

where $a > 0$, and $b > 0$.

If $(a+b)T < 1$, then $T < 2/(a+b)$ satisfies Equation (7.5). More conservatively, pick $T < 1/(a+b)$ to meet the assumption. Hence, the upper bound of the sampling interval for stability is known. Actually, the sampling interval is determined with respect to the calculation of the Φ-matrix and the approximation of the input signal. If we reduce the sampling interval, we can decrease the distortion in the input approximation and we also decrease the



(a)



(b)

Fig. 7.1. (a) Block diagram representation of a second-order sampled-data system; (b) reduction of Fig. 7.1(a).

ROE in calculation of the Φ-matrix according to Equation (2.22).

As long as the limit ensure the system is stable, that limit can be made as the upper bound. The upper and lower bound of the sampling interval determine the permissible range of variation of T. For maximum efficiency the sampling interval can be a signal-dependent variable (Dorf, 1962).

A realistic analysis is made by considering instantaneous error in reconstruction through the zero-order and first-order hold unit. Consider that it is desired to reconstruct a continuous sinusoidal signal

$$y(t) = A \sin(wt) \tag{7.6}$$

which has been sampled with the sampling interval T.

For a zero-order hold unit, the reconstructed signal is held constant between sampling instants so that the absolute error is zero at the sampling instant and varies according to the following equation (Wait, 1970)

$$\varepsilon = A \sin\left[(k+\tau/T)wt\right] - A \sin(kwT) \tag{7.7}$$

where $0 \leq \tau < T$. The maximum instantaneous error can be evaluated at the point where the maximum slope of the continuous signal y(t) in Equation (7.6) occurs. In this case the maximum slope of y(t) occurs in the vicinity of sin(kwT) = 0 which results in the maximum reconstruction error

$$\epsilon_{max} = A \sin(w\tau) \tag{7.8}$$

In general, $0 \leq \tau < T$ and $w << 2\pi$ hold. Therefore,

$$\epsilon_{max} \simeq AwT \tag{7.9}$$

In the case of a first-order hold unit, the output function is in the form of a series of ramps, each ramp commencing at the value of the function at the preceding instant and having a slope proportional to the difference between the values of two consecutive sampling instants according to

$$y(\tau) = y(kT) + (\tau/T)\left[y(kT) - y(kT-T)\right] \tag{7.10}$$

where $0 \leq \tau < T$. If the sinusoidal signal expressed by Equation (7.6) is to be reconstructed, then the instantaneous error is

$$\epsilon = A \sin(kwT) + A(\tau/T\sin(kwT) - A(\tau/T)\sin\left[(k-\tau)wT\right] - A \sin\left[(k+\tau/T)wT\right] \tag{7.11}$$

The maximum instantaneous error can be evaluated at the point where the second derivative of the continuous signal y(t) in Equation (7.6) is maximum. In this case the maximum concavity occurs in the vicinity of $\cos(kwT) = 0$ which results in the maximum reconstruction error

$$\epsilon_{max} = A\left[1 + \frac{\tau}{T} - \frac{\tau}{T}\cos(wT) - \cos(w\tau)\right] \tag{7.12}$$

In general, $0 \leq \tau < T$ and $w\tau < wT << 2\pi$ hold. Therefore,

$$\epsilon_{max} \simeq A(wT)^2 \tag{7.13}$$

For example, if we want to limit the instantaneous reconstruction error to .1% of the full scale, then the sampling interval should be less than the value with 6280 samples per cycle for a zero-order hold unit in Equation (7.9), with 199 samples per cycle for a first-order hold unit in Equation (7.13).

7.3 Summary of Results

There exist the upper and lower bound to the sampling interval. The lower bound is determined from the tradeoffs in the amount of computation per second and the effects of the round-off errors on the stability, while the upper bound is usually determined from the stability of the digital filter as well as the prescribed accuracy of the convolution integral and output reconstruction. The sampling interval of the most efficient digital filter could be a signal-dependent variable between the permissible range of variation of the sampling interval limited by those bounds.

CHAPTER 8

CONCLUSIONS

8.1 Summary

The results discussed in the last six chapters indicate that there are six different kinds of errors in linear digital signal processing.

The state transition matrix can be calculated in such a way as to reduce the truncation errors to any desired level according to Equations (2.24) and (2.41). By taking 11 terms in series approximation 7 digit accuracy can be obtained. The effects of the round-off errors in the elements of the state transition matrix on the eigenvalues of the digital filter can be related to the number of zeros in the root sensitivity function, which shows that the eigenvalues are more sensitive to the round-off errors in the diagonal elements than those in the off-diagonal elements. In the examples shown in Chapter 4, the input was approximated by a step approximation which resulted in 3 digit accuracy in the convolution integral of the state equation for a sinusoidal input with $w_o T = 10^{-3}$ radians (w_o is the highest frequency in the input). For fixed-point computation, extra care must be taken in storing the small numbers in the control transition matrix. By

selecting an appropriate scale factor the round-off errors in the product of the control transition matrix and the input can be minimized. The time delay inherent in the hybrid computation can be compensated for either by early reconstruction of the digital output by computing the digital output one-half the sampling interval early, which is feasible for a high speed computer, or by adding an extra zero to the transfer function of the digital filter, which is feasible for a filter with a pole-zero excess. The effect of the sampling interval on the tradeoffs in the amount of computation per second and the effects of the round-off errors on the stability determine the lower bound to the sampling interval, while the upper bound is usually determined from the stability of the digital filter as well as the prescribed accuracy in the convolution integral and in the output reconstruction. The selection of the sampling interval results in a tradeoff between the reconstruction errors and the round-off errors.

8.2 Area for Future Research

In a digital filter, the pole location can be used to determine thenecessary accuracy for the on-line simulation in the following ways:

1. Estimate the needed accuracy of $\Phi(T)$ and calculate $\Phi(T)$.
2. Determine the eigenvalues of $\Phi(T)$.

3. From the eigenvalues, determine the locations of the continuous-time poles of the system.
4. Compare the pole locations of (3) with the desired locations of the system.
5. Increase or decrease the number of digits in $\Phi(T)$ depending on the results of (4) and recalculate $\Phi(T)$.
6. Go back to (2) until the desired accuracy is obtained.

If the program of this procedure can be developed, then the digital filter can be used to simulate the continuous-time system exactly.

Between the lower and upper bound of the sampling interval, there is a permissible range of variation of the sampling interval such that the system implementation can be performed in the sense of minimum-sample filtering; that is, the signal-dependent sampling interval. The minimum-sample filtering may be considered the most efficient digital filter.

REFERENCES

Dorf, R. C., "Adaptive Sampling Frequency for Sampled-Data Control Systems," IRE Trans. on Automatic Control, Vol. 57, January 1962, p. 42.

Karplus, W. J., "Error Analysis of Hybrid Computer Systems," Simulation, Vol. 6, No. 2, February 1966.

Liou, M. L., "A Novel Method of Evaluating Transition Response," Proceedings of IEEE, Vol. 54, No. 1, January 1966, pp. 20-23.

Mankin, J. B., and Hung, J. C., "On Round-Off Errors in Computation of Transition Matrices," Joint Automatic Control Conference, 1969, p. 61.

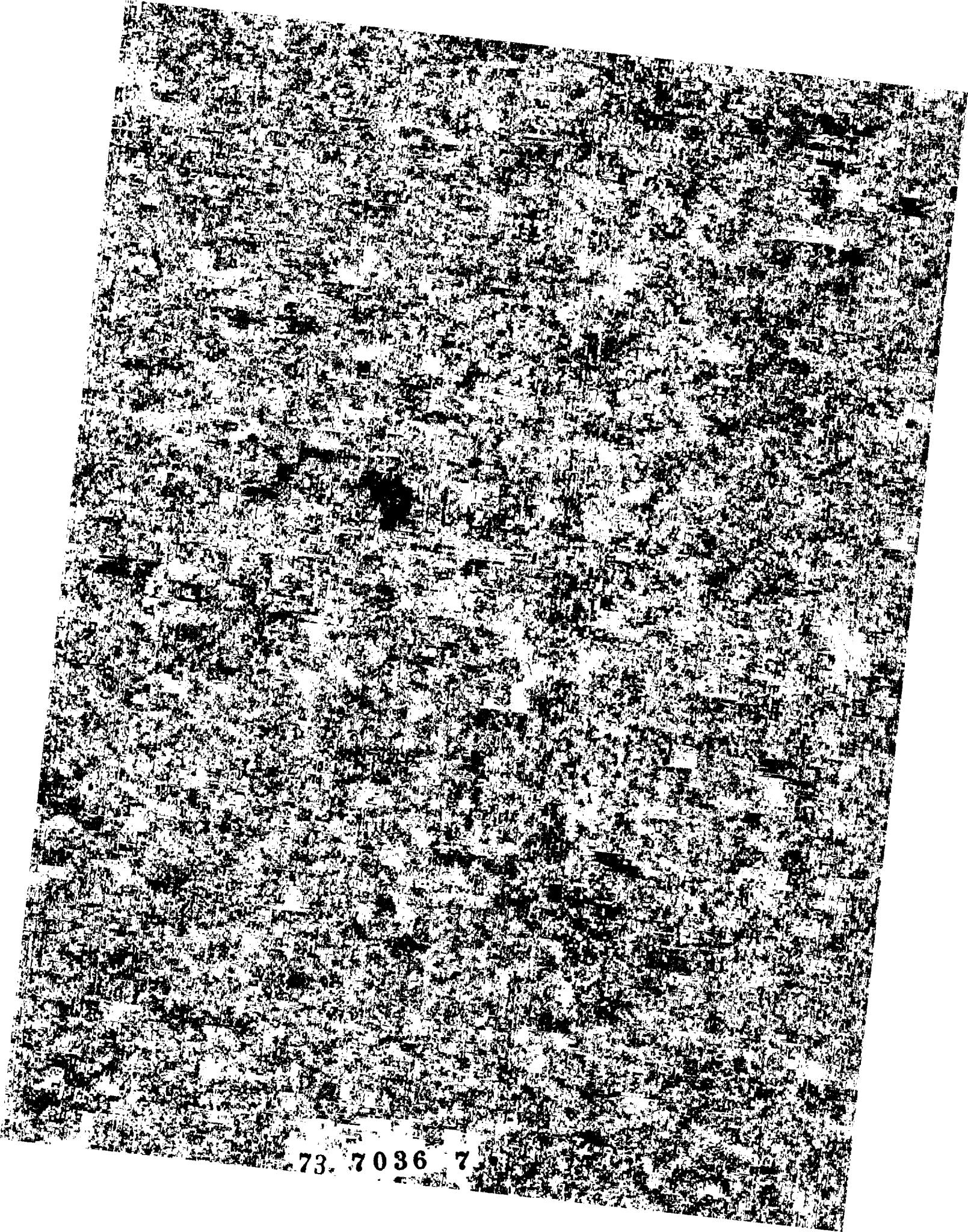
Schultz, D. G., and Melsa, J. L., State Function and Linear Control Systems, New York, McGraw-Hill, 1967.

Simes, J. G., and Mastascusa, E. J., "Round-Off Errors in Calculating exp(AT)," Proceedings on the National Electronics Conference, Vol. 26, 1970, pp. 125-128.

Varga, R. S., Matrix Iterative Analysis, Englewood Cliffs, N. J., Prentice-Hall, 1962.

Wait, J. C., "A Note on Designing Algorithms to Simulate Fixed Linear Systems Using Polynomial Input Approximation," CSRL Memo 196R, Dept. of Electrical Engineering, University of Arizona, October 1969.

Wait, J. C., "Digital Filters," in Active Filters: Lumped, Distributed, Integrated, Digital, and Parametric, edited by L. P. Huelsmann, New York, McGraw-Hill, 1970.



73-7036-7