LIMIT CYCLE OSCILLATIONS IN DIGITAL FILTERS

Parviz Rashidi B.Sc. (Eng), A.C.G.I.

A thesis submitted for the
degree of doctor of philosophy

Department of Electrical Engineering
Imperial College of Science and Technology
University of London

OCTOBER 1977

ABSTRACT

In the realization of digital filters on a computer or by special purpose hardware, errors are generated because of finite word-length arithmetic. These errors, which are called quantization errors, are of special concern in recursive digital filters where the error can build up to intolerable levels. Under the condition of deterministic inputs, these errors appear as sustained oscillations known as limit cycles.

This thesis presents several ideas on the studies of limit cycle oscillations. Zero input response of second-order recursive filters with finite wordlength arithmetic are considered in the phase plane. The results show that the trajectory of the transient response of these filters can go through a process which is referred to as merging. From this behaviour, limit cycles are categorised according to their accessibility from certain initial conditions.

Properties of the quantization processes are formulated and used to transform the structure of a second-order digital filter into a form which facilitates the study of limit cycle oscillations. The conditions for the occurrence of oscillations in the transformed structure are investigated and used to derive the conditions for the existence of some of the limit cycle oscillations.

From the studies based on the transformation technique, a suitable structure has been synthesised in which the method of signal stabilization is used to suppress limit cycle oscillations. Various results concerning the effects of the signal stabilization technique are presented.

Simulations were also conducted to find more about the behaviour of limit cycle oscillation and to verify the theoretical results.

ACKNOWLEDGEMENTS

I wish to express my warmest thanks to Professor R.E. Bogner for his supervision of this research project and for many helpful discussions and original suggestions relevant to the material presented in this thesis. I am also indebted for his assistance in overcoming many of my difficulties throughout the period of this research.

My thanks also go to Professors M. Nashed, C. Cherry, Doctors V. Lawrence, L.T.M. Berry, B. Davis, A.G. Constantinides and many of my colleagues both at Imperial College and Adelaide University for helpful discussions and their contributions towards the development of the ideas presented in this thesis.

I am particularly indebted to Messrs. M. Pearce and N.C. Quynh for their help in the preparation of this thesis. My thanks also go to Ms. L. MacGillivray, Messrs. B. Ackland, B.A. Hamidon and N. Smith.

For financial support during part of the period of study and for travel assistance to Australia, I am indebted to London University.

Finally I wish to express my thanks to Mrs. J. Darwent for her fast and efficient typing of this thesis.

CONTENTS

		Page
ABSTRACT		2
ACKNOWLEDGEMENTS		4
CHAPTER 1	INTRODUCTION	9
	1.1 The structure of digital filters	12
	1.1.1 Direct form realization	13
	1.1.2 Canonic realization	14
	1.1.3 Parallel and cascade realizations	15
	1.1.4 Other realization techniques	16
	1.1.5 The biquadratic section	18
	1.2 Quantization errors in digital filters	18
	1.2.1 Quantization of the input	18
	1.2.2 Quantization of the coefficients	19
	1.2.3 Quantization of the result of multiplication	20
	1.2.4 Adder overflow	22
	1.3 Quantization techniques	23
	1.4 The properties of the quantization processes	26
	1.5 Implementation of the quantization in a digital filter	28
	1.6 Limit cycle oscillations in recursive digital filters	30
	1.7 A survey of the literature on the effects of quantization in fixed-point recursive digital filters	31
	1.8 Outline and the contributions of this thesis	42

CHAPTER 2 RECURSIVE DIGITAL FILTERS WITH FINITE WORDLENGTH ARITHMETIC 46

2.1 Introduction 46

2.2 Realization of the state variables of a second-order recursive filter 46

2.3 Some remarks on the multiplication quantization error 51

2.4 Some remarks on limit cycle oscillations in second-order recursive filters 57

2.4.1 Limit cycles with period of one sample 57

2.4.2 Limit cycles with period of two samples 58

2.4.3 Limit cycles with period greater than two samples 62

2.5 Conclusions 63

CHAPTER 3 A STUDY OF LIMIT CYCLE OSCILLATIONS IN THE PHASE PLANE 65

3.1 Introduction 65

3.2 The transient response of an infinite precision second-order filter 66

3.2.1 The case of real poles 67

3.2.2 The case of complex poles 69

3.3 Response of a finite precision filter in the phase plane 71

3.4 Conditions for the existence of merging trajectories 74

3.4.1 Rounding R[ ] 75

3.4.2 Value truncation T2[ ] 78

3.4.3 Magnitude truncation T1[ ] 79

3.4.4 Generalised results 80

3.5 Boundaries for the points of convergence when $k \geq 2$ 81

3.5.1 Rounding 82

3.5.2 Value truncation 86

3.5.3 Magnitude truncation 88

3.6 Characterisation of limit cycle oscillations 90

3.6.1 Rounding 92

3.6.2 Value or magnitude truncation 93

3.7 Conclusions 93

CHAPTER 4 A TRANSFORMATION TECHNIQUE FOR THE STUDY OF ZERO INPUT LIMIT CYCLE OSCILLATIONS 95

4.1 A new representation of a second-order filter 95

4.1.1 Type-A realization 97

4.1.2 Type-B realization 98

4.2 Studies of the model 101

4.3 Limit cycle oscillations in the type-A realization 103

4.3.1 Limit cycles with period greater than two samples 104

4.3.2 Limit cycles with period of one or two samples 107

4.4 Limit cycle oscillations in the type-B realization 111

4.4.1 Limit cycles with period greater than two samples 114

4.4.2 Limit cycles with period of one or two samples 114

4.5 Conclusions 119

CHAPTER 5 SUPPRESSION OF LIMIT CYCLE OSCILLATIONS 120

5.1 Introduction 120

5.2 The method of signal stabilization 121

5.3 The amplitude of the modified limit cycles for the type-A realization 124

5.3.1 Limit cycles with period greater than two samples 125

5.3.2 Limit cycles with period of one or two samples 126

5.4 Increase in the noise level 133

5.5 Suppression of limit cycle oscillations in the type-B realization 136

5.6 Remarks on the method of signal stabilization 140

5.7 Conclusions 144

CHAPTER 6 SIMULATION STUDIES 145

6.1 Introduction 145

6.2 The programming techniques 146

6.2.1 Choice of quantization steps for the signal and the coefficients 150

6.2.2 Detection and identification of limit cycle oscillations 150

6.2.3 Random number generation 151

6.3 Scan of initial conditions 154

6.4 Suppression of limit cycle oscillations 176

6.4.1 Amplitude of the suppressed limit cycle oscillations 177

6.4.2 Increase in the noise level due to signal stabilization 178

6.4.3 Simulation of the simple implementation of signal stabilization 180

6.5 Conclusions 193

CHAPTER 7 CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDIES 194

APPENDIX A 199

APPENDIX B 200

REFERENCES 202

CHAPTER 1

INTRODUCTION

In general many types of numerical iteration techniques (for example, Simpson's Rule, Interpolation, etc.) can be characterised by the process of digital filtering. The term "digital filtering" usually refers to the numerical operations on a discrete digital signal to shape the spectrum of that signal [33]. Examples of these methods can range from numerical techniques, such as the Discrete Fourier Transform, to a combination, in a certain orderly manner, of simple arithmetical operations such as, addition, multiplication and delaying [6-12]. The term "digital filter" is, therefore, the name given to the piece of hardware or the algorithm which can perform these operations on an input signal.

The digital filter can be realized by programming a general purpose computer, or designing special purpose hardware, to perform the above operations. In general, hardware implementation is preferable because the cost of the equipment is much less, and the speed of operation can be made much higher.

Digital filtering of an analog signal may follow the sequence of operations shown in Fig. 1.1. This process normally is equivalent to the operation of an analog filter which uses components such as capacitors, inductors and resistors.

Because of the fixed arithmetical operations which digital filters perform, they have a number of important advantages over their counterpart analog filters. These advantages may be summarised as

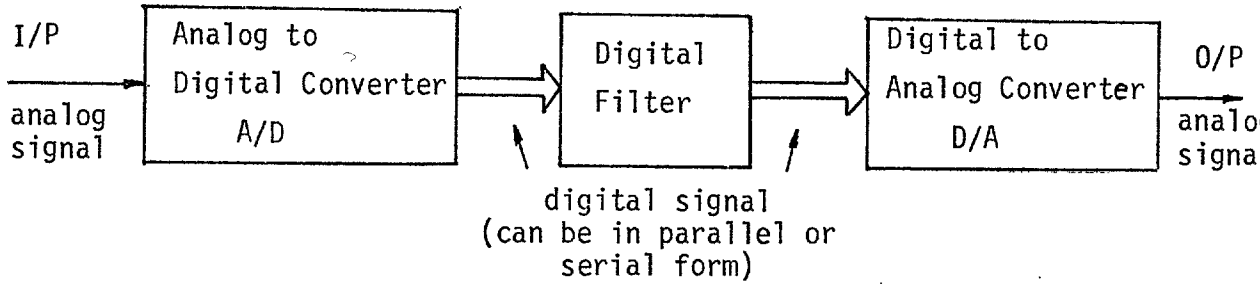


FIG. 1.1

follows:

i) Digital filters are predictable in their performance. For example, temperature variation, old age and similar features do not alter their characteristics.

ii) Because of recent developments in the field of semiconductor devices and Large Scale Integration (LSI), digital filters have a great potential in miniaturisation and reduction in the cost of production.

iii) The precision (i.e. the performance) of digital filters can be chosen to be as high as required, according to the cost versus performance trade-off.

iv) Digital filters can be designed to be programmable, which renders the choice of their characteristics flexible. This feature enables the filter to be used on a time-shared basis, and enables the design of complex filters using a minimal amount of hardware.

v) Digital filters have the potential of being produced in a universal building block form, which enables mass production (see Section 1.1).

vi) Digital filters do not possess the restriction of impedance

matching which is inherent in analog circuits.

vii) Digital filters can be designed to operate at extremely low frequencies without the extra difficulties inherent in analog circuits (i.e. the size of components, drift of signal, etc.).

As in any practical system, digital filters also have a few disadvantages which may be listed as follows:

i) Numerical representations in digital filters have to be restricted to a finite number of bits or digits for hardware implementation. Because of this restriction it is necessary to quantize the input signal and to limit the accuracy for the representation of the filter coefficients. In many cases the accuracy of the arithmetical operations for the filtering process also has to be limited. A full discussion of this problem is given in Sections 1.2 to 1.7, and specific areas of this problem are investigated in the following chapters.

ii) Because they operate on sampled signals, digital filters have a limitation on the maximum frequency of the signal which they can handle. This limitation arises from the restriction on the speed of operation of the filter components. For example, if addition takes 1 micro-second (μs) and multiplication 2 μs, then for a filter with two additions and two multiplications, using multiplexed adders and multipliers, each iteration takes at least 6 μs (we have ignored the time required for data handling). Therefore, the sampling period for this example must be greater than 6 μs, and according to the sampling principle of Nyquist, the frequency of the signal must be less than 83 kHz (preferable by a factor several times greater than one).

In this regard, digital filters have great potential for improvement, because of the recent development in semiconductor technology, and the development of new devices with even higher speeds

of operation.

iii) Normally a digital filter has to be used in conjunction with simple analog filters (analog filters are used in the A/D and D/A converters). This requirement, therefore, makes the use of digital filters uneconomical for simple filtering operations.

These disadvantages of digital filters have the potential of being improved. In this thesis the problem of quantization in digital filters is considered. In particular, the aim of this study has been to investigate the oscillation conditions resulting from the nonlinearities of the quantization processes, and to study a technique for their suppression.

1.1 The structure of digital filters

Digital filters in general can be divided into two basic categories [6-11], [33]: i) the recursive or Infinite Impulse Response (IIR) digital filter, and ii) the non-recursive or Finite Impulse Response (FIR) digital filter .

The term "recursive" describes a process which has its output dependent on the past values of its output. Therefore, a recursive digital filter has its output dependent on the past values of its output and the present and/or the past values of its input. This gives a recursive digital filter an impulse response of infinite duration.

This description of a recursive digital filter may be expressed by a difference equation given as

$$y(n) + \sum_{i=1}^{N} b_i\, y(n-i) = \sum_{i=0}^{M} a_i\, u(n-i) \qquad (1.1)$$

where, $\{u(n)\}$ is the input sequence, and $\{y(n)\}$ is the output sequence of the filter. The coefficients a_i and b_i are constants which determine

the response of the filter.

A non-recursive digital filter on the other hand has its output dependent on only the present and the past values of the input sequence. Consequently, a digital filter of this type has only a finite duration impulse response, and its difference equation is of the form:

$$y(n) = \sum_{i=0}^{M} a_i \, u(n-i) \quad . \tag{1.2}$$

Because of its finite impulse response, a non-recursive digital filter cannot accommodate any steady state oscillation, regardless of the coefficients or the quantization technique used. Therefore the non-recursive digital filter will not be considered in this thesis.

As digital filters operate in the discrete-time domain, their transfer function can be derived using the z-transformation technique [1-3], [6-12], [34-39]. For example, a general transfer function for a recursive digital filter is

$$H(z) = \frac{\sum_{i=0}^{M} a_i \, z^{-i}}{1 + \sum_{i=1}^{N} b_i \, z^{-i}} \quad , \tag{1.3}$$

for $M \leq N$.

Equation (1.3) can be realized in terms of the digital filter components (i.e. delays, multipliers, and adders) in a number of different ways [7-14], [40-47]. The following sub-sections describe some of the more important realization techniques.

1.1.1 Direct form realization

In the direct form realization the structure is constructed by directly simulating the operation of the difference equation (1.1). This

structure is shown in Fig. 1.2. It is well known that the direct form realization uses more delay elements than necessary [7-11].

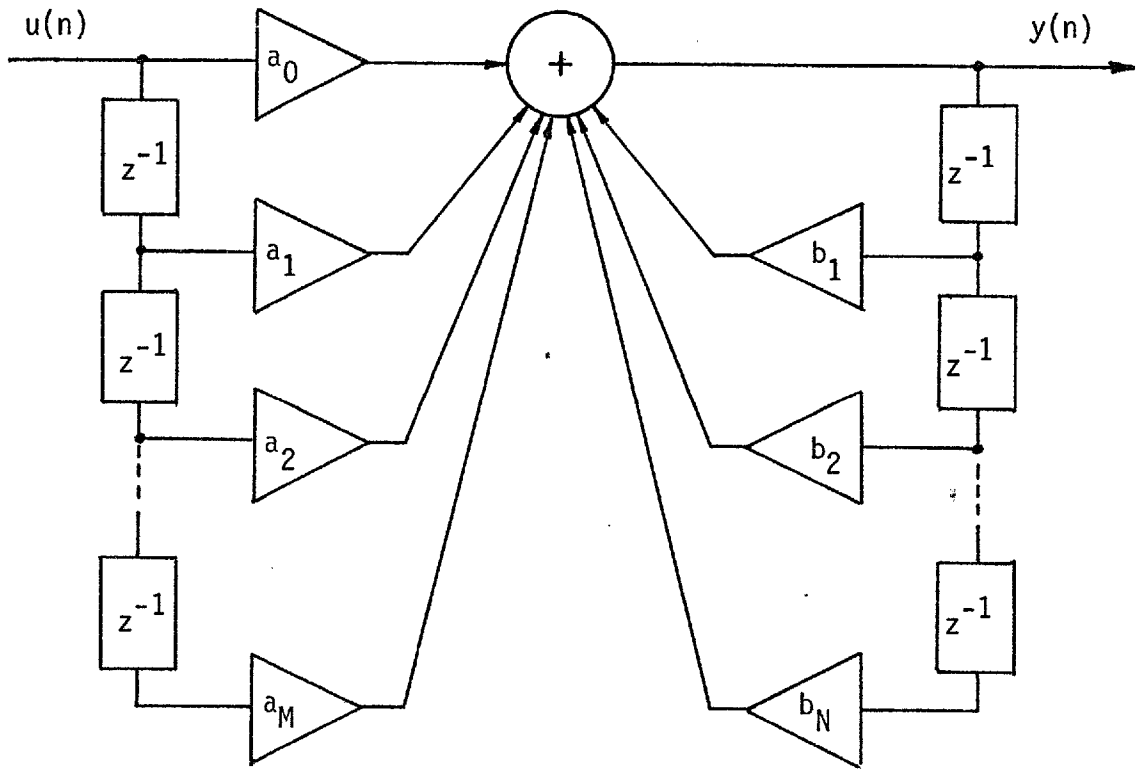


FIG. 1.2 Direct form realization

1.1.2 Canonic realization

By the term "Canonic realization", one normally is referring to a structure which can realize the transfer function of the filter using a minimum number of delay elements. There are a large number of canonic realizations. The most direct form can be derived by expressing the transfer function (1.3) as

$$H(z) = H_1(z) \sum_{i=0}^{M} a_i z^{-i} , \qquad (1.4)$$

where $$H_1(z) = \frac{1}{1 + \sum_{i=1}^{N} b_i z^{-i}} .$$

The structure representing this realization is shown in Fig. 1.3 .

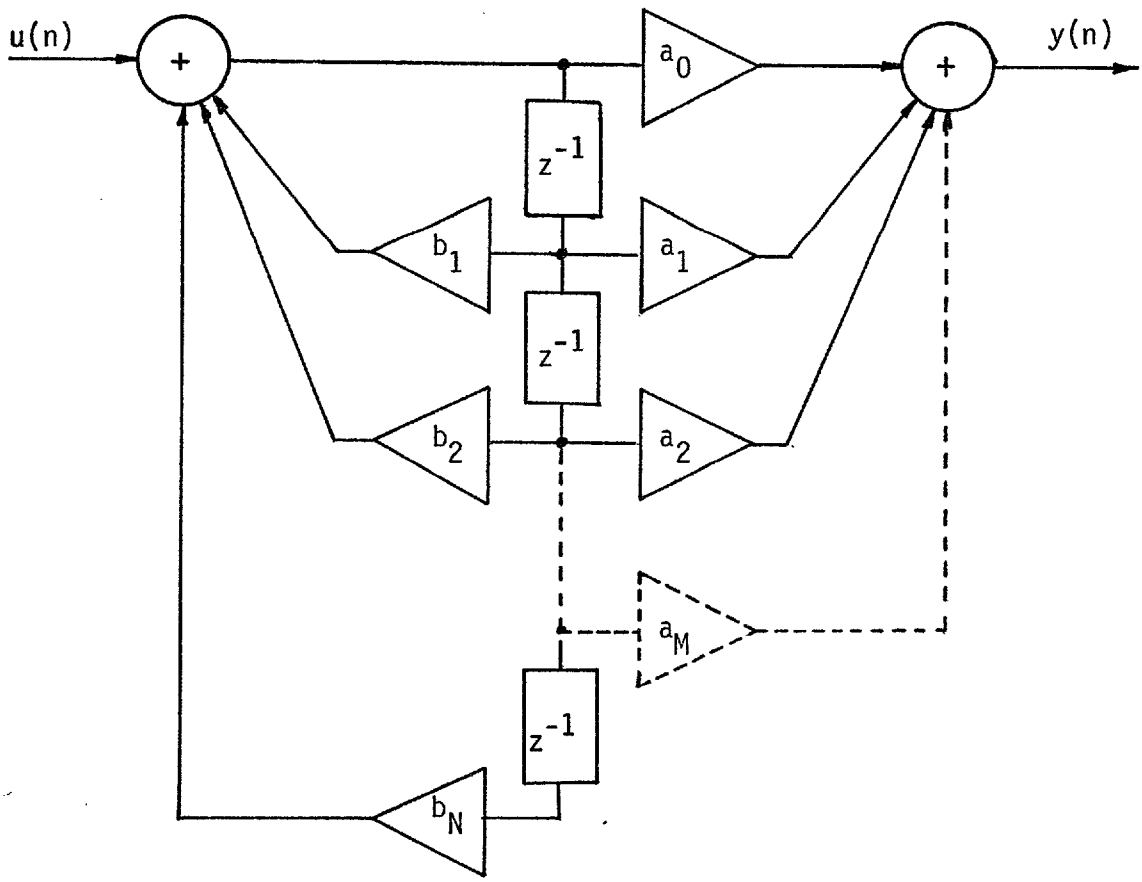


FIG. 1.3 The canonic realization

1.1.3 Parallel and cascade realizations

Two of the more common forms of realizations of digital filters in the canonic form are the parallel and the cascade realizations. In these realizations, the transfer function of the digital filter is realized by the parallel or the cascade connection of first- and second-

order sub-filters respectively.

For example, in the case of parallel form realization, the structure is determined by expanding the transfer function (1.3) in the partial fraction form, such as

$$H(z) = a_0 + \sum_{i=1}^{L} \frac{a_{0i} + a_{1i}\, z^{-1}}{1 + b_{1i}\, z^{-1} + b_{2i}\, z^{-2}} \quad . \tag{1.5}$$

In the case of cascade form realization, the transfer function is expanded in a factorised form as given in equation (1.6).

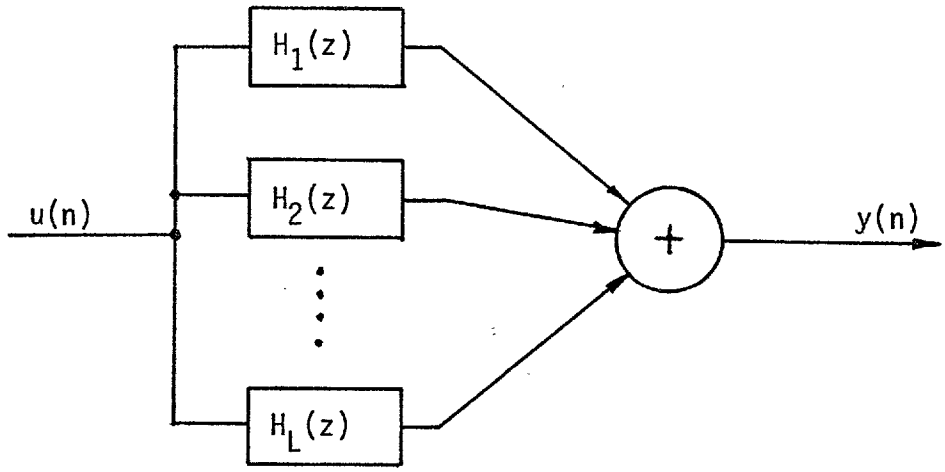
$$H(z) = \prod_{i=1}^{L} \frac{a_{0i} + a_{1i}\, z^{-1} + a_{2i}\, z^{-2}}{1 + b_{1i}\, z^{-1} + b_{2i}\, z^{-2}} \quad . \tag{1.6}$$

A diagrammatic representation of these realizations is shown in Fig. 1.4, where $H_i(z)$ is used to represent the sub-filters. Here we note that for cascade realization we have the choice of the order in which the sub-filters can be cascaded, and the choice of pairing zeros and poles of the filter arbitrarily. This enables the designer to choose a representation with a minimum amount of quantization noise [11], [25], [81-82].

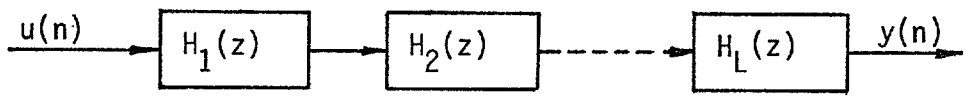
1.1.4 Other realization techniques

There are several other forms of structures for realizing a digital filter which represents a given transfer function [40-47]. The realization based on continued fraction expansion of the transfer function [41], the ladder structure realization [42], and the wave digital filter [46], are some examples.

There are also various methods for forming new structures. Jackson [25], [81-82] has presented a method which can be used to generate a new structure from a given one, by transposing the state equations of that structure. It has been shown [25] that this process



(a) Parallel realization



(b) Cascade realization

FIG. 1.4

is equivalent to reversing the direction of all the delay and multiplier elements, interchanging the summing nodes with the signal nodes, and interchanging the input with the output of the filter.

Mitra and Hirano [43-44] and others [45] have developed another method for generating new realizations. This method is referred to as the "multiplier extraction method". In this technique, firstly all the multipliers necessary for the representation of the filter are assumed, then the structures with only delay and adder elements that relate the inputs and the outputs of the multipliers are considered.

1.1.5 The biquadratic section

In the cascade and the parallel realizations we noted that any complex transfer function can be realized by a combination of simpler sub-structures. Each of these sub-structures can be defined in the general form by the transfer function

$$H_b(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \quad . \tag{1.7}$$

This basic form of a digital filter is referred to as the "biquadratic section". In this thesis we concentrate on the behaviour of this type of filter under the restriction of finite wordlength arithmetic.

1.2 Quantization errors in digital filters

In the implementation of digital filters by hardware, or on a digital computer, one is restricted to the use of a finite number of bits or digits. Consequently, errors will be generated in the overall process of digital filtering. These errors may be categorised into four different groups, as described below.

1.2.1 Quantization of the input

For filtering analog signals, the samples of the input to the digital filter have to be converted to digital samples which, because of the wordlength limitation of the hardware, are quantized. Therefore, a process of quantization is necessary at the input stage of the digital filter. This quantization process can be considered equivalent to adding a noise (or error) source to the input of the digital filter. This error is referred to as the "input quantization error".

Normally the input quantization error can be considered to be of a random nature with uniform probability distribution and power spectrum.

This assumption has been shown valid, by Bennet [76] for a wide range of input signals. Therefore, the properties of the output error due to the input quantization can readily be calculated [7-14].

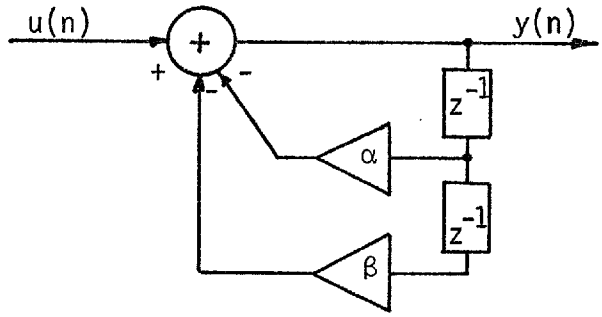
To reduce the effects of the input quantization, one can increase the wordlength used for signal representation in the digital filter. Therefore this limitation need not be considered any further.

1.2.2 Quantization of the coefficients

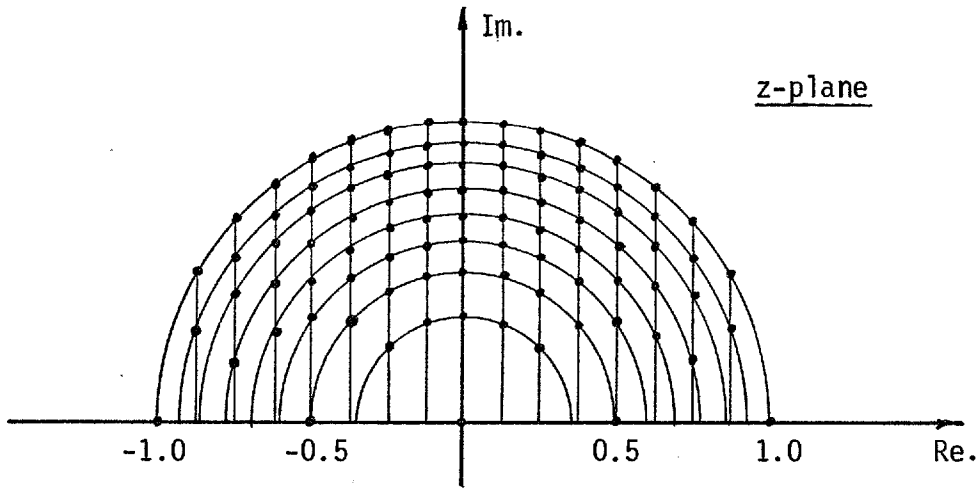
The representation of the coefficients in a digital filter is also subject to errors because of wordlength limitations. These errors cause a difference in the performance of the digital filter from that which was intended by the designer. In some special cases this effect can cause instability.

The severity of the effects of the coefficient quantization depend on the pole and zero sensitivity of the digital filter structure [10-11] [56-58]. For example, let us consider the pole sensitivity of two quadratic sections of different structures, as shown in Figures 1.5(a) and 1.6(a). Their corresponding pole positions, assuming three bit wordlength, are shown in Figures 1.5(b) and 1.6(b) respectively. We note that the pole grid in the region of the real axis is very course for the structure of Fig. 1.5(a) as compared with that of Fig. 1.6(a). However, the better pole sensitivity for this example is achieved at the expense of using two extra multipliers.

In more complex filters, for a given choice of a structure, the effect of the coefficient quantization is minimized by the optimum choice of the different ways in which the coefficients can be quantized [59-67].



(a) Direct form realization of the quadratic section

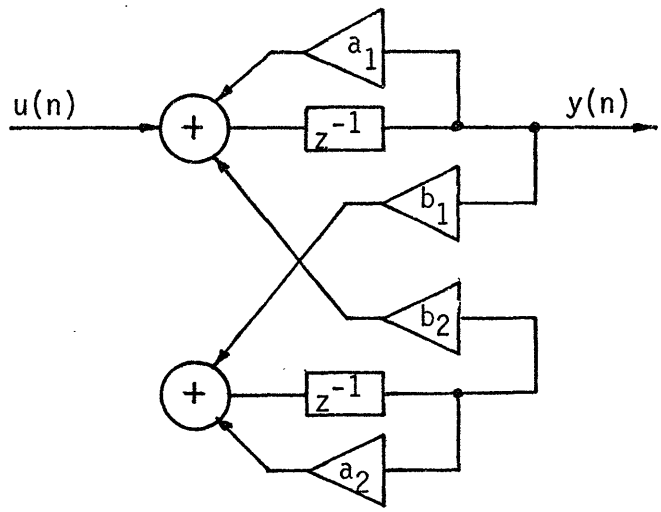


(b) Allowable pole positions for coefficients of 3 bits

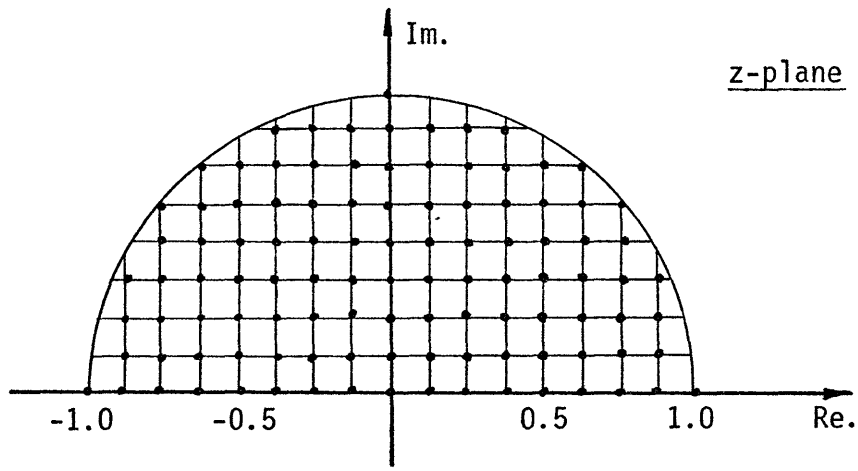
FIG. 1.5.

1.2.3 Quantization of the result of multiplication

In a recursive digital filter it is found necessary to quantize all multiplication products for each cycle of iteration either before or after the addition. For example, consider the product of two numbers with a wordlength of i bits resulting in a number with a wordlength of 2i bits. If this word is not quantized back to i bits in the next iteration, the result of the product will have 3i bits, and



(a) Coupled form realization of the quadratic section



(b) Allowable pole positions for coefficients of 3 bits

FIG. 1.6

so on for the subsequent iterations. Hence the wordlength requirement for a recursive digital filter will grow indefinitely if no quantization is allowed.

The error introduced by the quantization of the results of multiplication is normally referred to as the "round-off" or "truncation" error, or simply the "quantization error" (short form for "multiplication quantization error").

The quantization error partly depends on the number representation used. There are basically three different ways of representing numbers in a digital filter [7-11], [23-24], [49]. These are known as fixed-point, floating-point, and block floating-point representations. The simplest of these is the fixed-point number system. This is therefore more desirable for use in hardware implementation of digital filters. In this thesis we consider only the fixed-point number representation.

1.2.4 Adder overflow

A further limitation that can cause severe oscillation conditions in digital filters is the effects of adder overflow. This occurs when the result of addition exceeds the wordlength limit of the registers. The effect of this is the loss of the most significant bit of the number. Thus, the result can have a significant error which can cause large amplitude oscillations [8-11], [68-75]. This severe performance can generally be reduced by use of "saturation adders" [33], which produce a limiting effect when overflow occurs.

One of the first studies of overflow oscillations was presented by Ebert et al. [68], and also by Jackson [25], [81]. Ebert et al. derived the conditions for zero input oscillations, ignoring the quantization errors. They also showed that, limit cycle oscillations due to adder overflow can be stopped by the use of saturation adders (also called

"saturation arithmetic"). Sandberg [69] considered the overflow problem in digital filters with saturation adders, allowing for quantization errors. He showed that any self-sustained limit cycle oscillation can be reduced to an arbitrarily small amplitude by increasing the number of bits used for the arithmetical operations.

Works of Willson [70-72] and Classen and Kristianson [73] give more detailed studies of the limit cycle oscillations resulting from overflow effects in adders of various characteristics. The advantage of a normal adder over the saturation adder is pointed out, because overflows in the intermediate results of a multi-input addition can cause no errors in normal addition as long as the final result is within the capacity of the register.

Kristianson [74] reported a jump phenomenon in the performance of a second-order recursive filter with saturation arithmetic, having a sinusoidal input. He reported a jump in the amplitude of the output of the filter caused by small changes in the frequency or the amplitude of the input. Landauro and Lienard [75] study the hardware for overflow detection.

1.3 Quantization techniques

In a digital machine there are two basic ways of discarding the unwanted bits of an arithmetical result. These two ways are truncation and rounding. In the case of truncation two different characteristics can be obtained, depending on the number representation employed. Therefore, there are total of three characteristics for the quantization processes. These are referred to as "Magnitude Truncation" T1[ ], "Value Truncation" T2[ ] and "Rounding" R[ ], [33].

Magnitude Truncation is achieved by discarding the lower significant bits of a number represented in sign-magnitude or one's

complement forms. This is also sometimes called "sign-magnitude truncation". The function representing the characteristics of this process is shown in Fig. 1.7, where the crosses represent the values of the function at the transition regions. We shall use the symbol $T1_q[\ ]$ to represent this function when the quantization step is q.

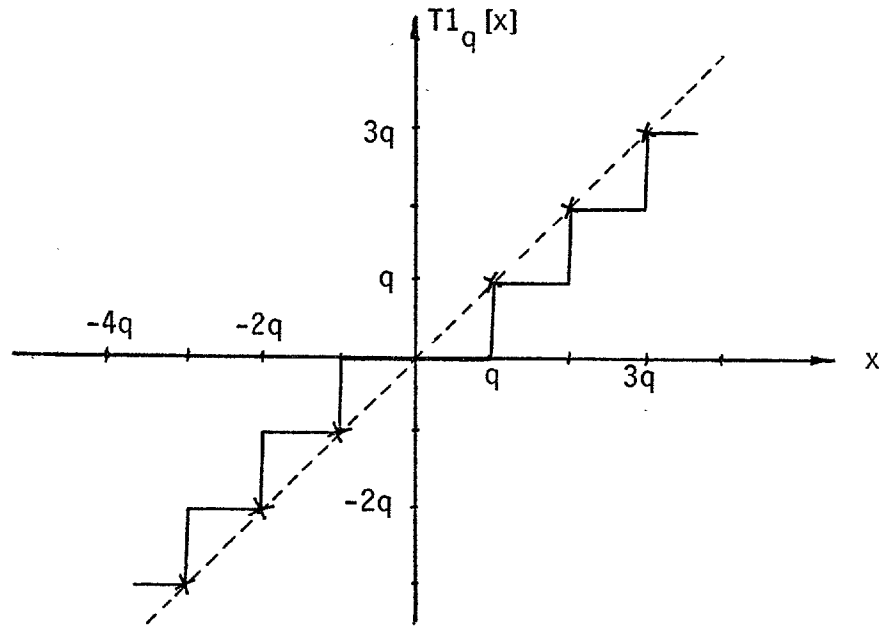


FIG. 1.7 Magnitude truncation

Value Truncation is the result of discarding the lower significant bits of a two's-complement number. This process results in a number which is always smaller in value than the original number (hence the term "Value Truncation"). The function representing the characteristics of this process is shown in Fig. 1.8, where, as before, the crosses denote the values of the function at the transition regions. We shall use the symbol $T2_q[\ ]$ to represent this function when the quantization step is q.

The rounding process is achieved also by discarding the unwanted

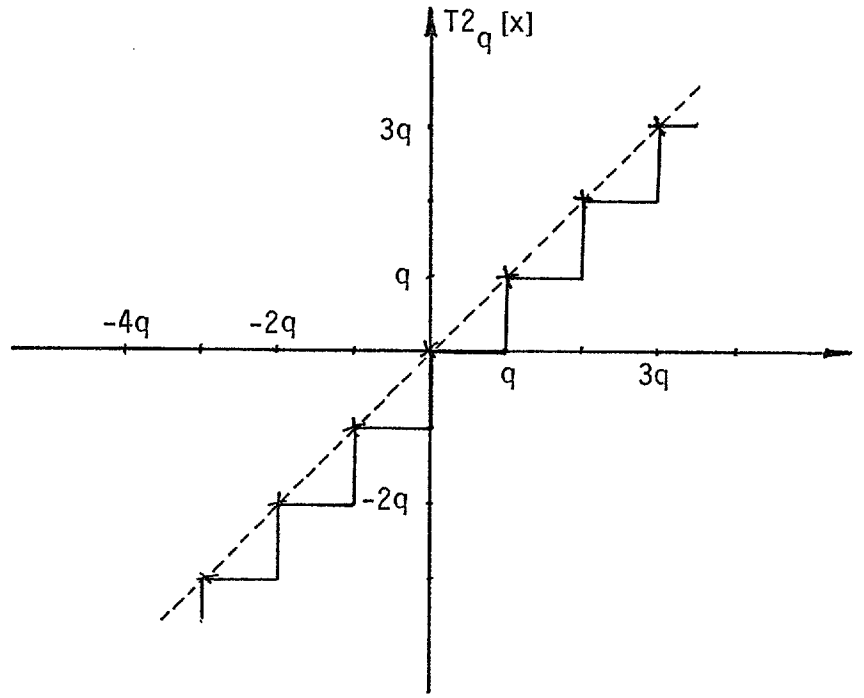


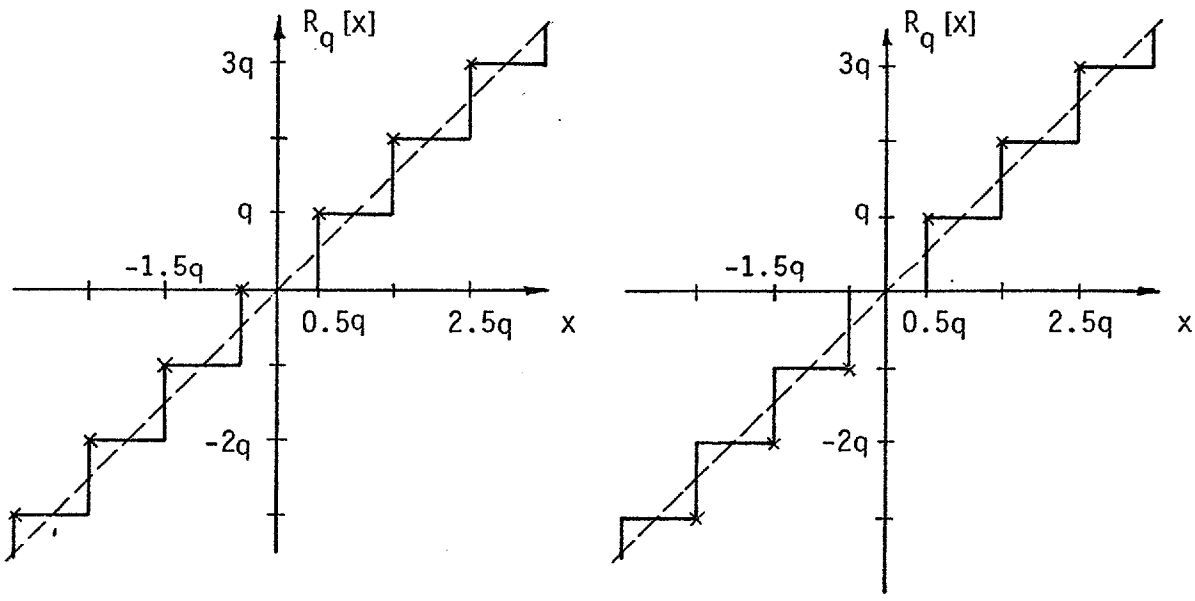
FIG. 1.8 Value truncation

bits of the magnitude of the number. However, if the most significant bit of the discarded part is a one, then a one is added to the least significant bit of the magnitude of the remaining part. This operation will result in a lower error compared to the error in truncation.

The rounding can also be achieved by adding a one to the largest significant bit of the unwanted part of a two's-complement number, and discarding the unwanted bits.

The two different techniques for rounding a number are similar except when the number is of value $(-i + 0.5)q$, where i is any positive integer. Therefore, the rounding process should be described correctly by referring to it as either "magnitude", or "value" rounding. The functions for these two cases are shown in Figures 1.9(a) and 1.9(b) respectively, where the points marked by the crosses represent the values

of the function at the transition regions. We note that the points where the two functions do not correspond occur in very rare cases. Thus, magnitude and value rounding are almost equivalent, and we can use the symbol $R_q[\ ]$ to represent the rounding function in general.



(a) Value rounding (b) Magnitude rounding

FIG. 1.9

1.4 The properties of the quantization processes

From the definitions of the quantization techniques described in Section 1.3, we can derive the following properties.

a) Let $Q_q[\ ]$ represent a general quantization function with the quantization step of q. Define the error of quantization e as

$$Q_q[x] = x + e \quad , \tag{1.8}$$

and use an appropriate subscript for e to distinguish between the quantization processes. Then

i) for value truncation we have

$$-q < e_{T2} \leq 0 \quad , \tag{1.9}$$

ii) for magnitude truncation equation (1.9) applies only when $x \geq 0$, however, for $x < 0$, we have

$$0 \leq e_{T1} < q \quad . \tag{1.10}$$

iii) for rounding, the limits for the error are given by

$$-0.5q < e_R < 0.5q \quad , \tag{1.11a}$$

with the possible equality

$$e_R = 0.5q \quad , \tag{1.11b}$$

for value rounding, and

$$e_R = 0.5q \cdot \mathrm{sgn}(x) \quad . \tag{1.11c}$$

for magnitude rounding.

b) Comparing the various quantization functions it can be shown that the following relations exist.

$$T2_q[x] = T1_q[x] + \frac{\mathrm{sgn}(x) - 1}{2} \quad , \tag{1.12}$$

and

$$R_q[x] = T2_q[x + 0.5q] \quad , \tag{1.13}$$

where $R_q[x]$ represents the value rounding.

c) If p and q are two different steps of quantization, then for any quantization process we have

$$Q_q[x] = Q_p\left[\frac{xp}{q}\right]\left[\frac{q}{p}\right] \quad . \tag{1.14}$$

d) For value truncation and (value) rounding, the quantization of the sum of a quantized variable, iq, and a non-quantized variable, x,

can be expressed as,

$$T2_q[iq + x] = iq + T2_q[x] \quad , \tag{1.15}$$

and

$$R_q[iq + x] = iq + R_q[x] \quad . \tag{1.16}$$

However, with magnitude truncation, we can write

$$T1_q[iq + x] = iq + T1_q[x] \quad , \tag{1.17}$$

only if sgn(x) = sgn(iq + x).

e) In terms of the modulo function, x mod q, the quantization error resulting from the various quantization techniques can be represented as

$$e_R = -[(x + 0.5q) \bmod q] + 0.5q \quad , \tag{1.18}$$

$$e_{T2} = -x \bmod q \quad , \tag{1.19}$$

and

$$e_{T1} = -(x \bmod q)\ \mathrm{sgn}(x) \quad . \tag{1.20}$$

1.5 Implementation of the quantization in a digital filter

The quantization processes discussed in the previous sections can be used in two different configurations. One method is to quantize the output of each multiplier before they are added together. A simple example is shown in Fig. 1.10.

The second method is to apply quantization after the outputs of the multipliers are added together. An example is shown in Fig. 1.11.

We note that in the structure of Fig. 1.11, the addition has to be carried out with a greater precision than that of the structure of Fig. 1.10. If the wordlength of the coefficients and the signal to the multipliers are the same, the precision of the addition for the second method needs to be twice that of the first method. These methods are,

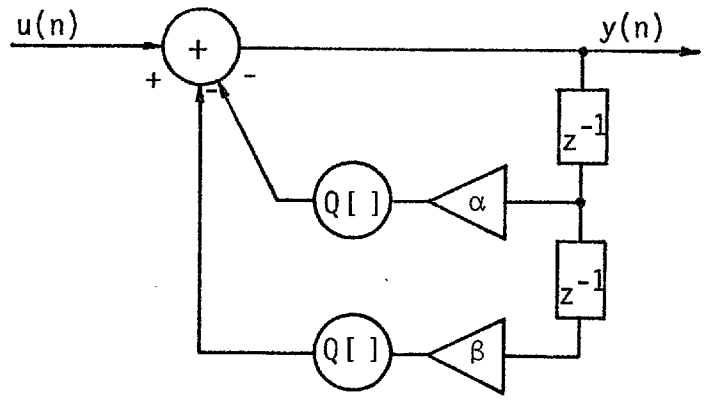


FIG. 1.10 Single precision digital filter

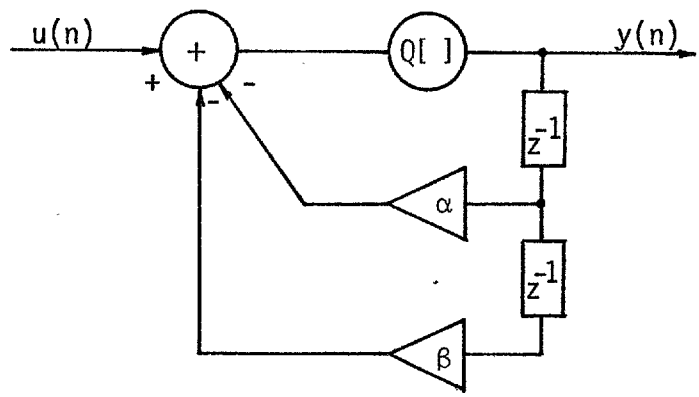


FIG. 1.11 Double precision digital filter

therefore, referred to as "single-precision" and "double-precision" respectively.

The hardware for single-precision filters would be more economical, since addition would be simpler, and also multipliers can be

designed to evaluate only the required (higher significant) bits of the product (i.e. giving us faster and simpler multipliers).

1.6 Limit cycle oscillations in recursive digital filters

For the special case when the input to a recursive digital filter is deterministic, for example, an impulse or a step function, the quantization effects can cause a stable digital filter to go into a steady state oscillation. These are generally known as "limit cycle oscillations" (or, as used by earlier authors, the "dead-band effect" [5]).

A simple demonstration of limit cycle oscillation can be shown by considering, for example, a first order recursive filter using rounding with quantization step of 1 and having coefficient of say 0.9. The difference equation defining this filter is given by

$$y(n) = R_1[0.9\ y(n - 1)] + u(n) \quad ,$$

where u(n) is the input and y(n) is the output of the filter. If an impulse of say amplitude 10 is applied at time zero, then the response of the filter will be given by the sequence {10, 9, 8, 7, 6, 5, 5, 5, . . . etc.}. The DC steady state of value 5 is the limit cycle oscillation caused by the rounding process. If, for example, magnitude truncation had been used, the limit cycle could not exist.

In fact, the properties of limit cycle oscillations depend on the following factors:

i) the quantization technique,

ii) the precision of the arithmetic (i.e. single or double precision arithmetic),

iii) the order of the filter (higher order filters have more

complex limit cycle oscillations),

iv) the coefficients of the filter,

v) the type of the input,

vi) the initial conditions,

vii) the configuration of the structure.

The limit cycle oscillations in general can be categorised into three different forms:

i) limit cycles with period of one sample (this oscillation is a constant DC signal as in the example above),

ii) limit cycles with period of two samples,

iii) limit cycles with period of three or more samples.

The techniques for the analysis of these limit cycles generally differ because of the different complexities of the limit cycles. Some of these techniques will be investigated in this thesis.

There have been a large number of publications on the studies of the quantization noise and the limit cycle problem in digital filters. In the following section an overview of the literature on the problems of multiplication quantization in fixed-point recursive digital filters will be given.

1.7 A survey of the literature on the effects of quantization in fixed-point recursive digital filters

Since the advent of digital computers and their application in signal processing and control systems, many authors have become interested in the study of the various aspects of the errors caused by the inherent effects of the quantization in digital machines. In recent years, as digital filters have become more competitive with other forms of

filtering, the interest in this field has increased.

In this section, first the basic ideas used for the analysis of the quantization errors will be introduced, then the individual works of the various authors will be discussed.

The effects of quantization errors, in general, are studied by considering two different basic points of view:

i) The errors due to the quantization effects are assumed to be of random nature. A stochastic approach may then be used to analyse the statistical properties (i.e. variance σ^2, correlation coefficient, etc.) of the error on the output of the filter.

Bennet [76] was the first author to look at the quantization errors in this way. He concluded that for most input conditions the assumption that the quantization noise has random properties with a uniform probability density is valid. In general, the results based on this assumption give a very good estimate when the signal that is being processed is of random or near random nature.

ii) The other point of view is to consider the quantization error to be of a deterministic nature. For example, the error due to quantization could be considered to be the worst case possible. This type of analysis normally results in an estimate of the absolute bound for the error on the output of the filter. In general the results which are based on the worst case error analysis are pessimistic.

The model which is often used for the analysis of the quantization error is similar to the general model shown in Fig. 1.12 (first used by Jackson [25], [81] in the form shown). In this model the error due to the quantization of the products is given by a source of noise at the subsequent summation point, and the multiplier is assumed to be ideal. The signal $e_i(n)$ represents the total error

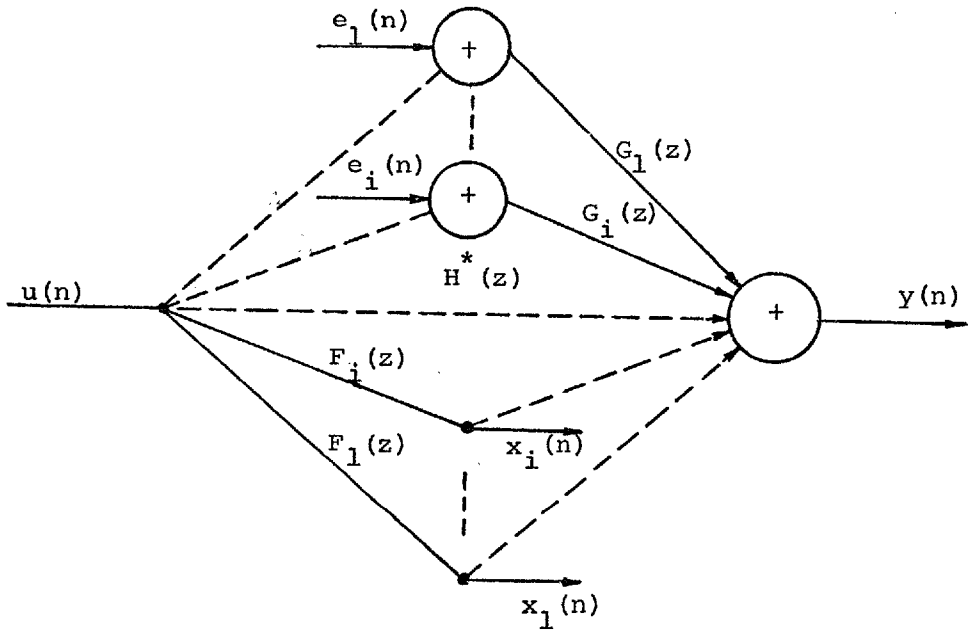


FIG. 1.12 A general model of a finite wordlength digital filter

which is needed at the summation node i to allow for the quantization effects of all multipliers feeding that node. The transfer function of the filter from the i^{th} summation node to the output of the filter is $G_i(z)$, and the transfer function from the input to the i^{th} branch node is $F_i(z)$. $H^*(z)$ represents the ideal transfer function of the overall system.

An estimate of the variance of the error on the output, $\sigma^2_{\varepsilon_i}$, due to the noise source $e_i(n)$, is then given by:

$$\sigma^2_{\varepsilon_i} = \sigma^2_{e_i} \sum_{n=0}^{\infty} g_i^2(n) \quad , \tag{1.21}$$

where, $g_i(n)$ is the inverse z-transform of the transfer function $G_i(z)$. We can also express equation (1.21) as

$$\sigma^2_{\varepsilon_i} = \frac{\sigma^2_{e_i}}{2\pi j} \oint G_i(z)\, G_i^*(z)\, \frac{dz}{z} \quad , \tag{1.22}$$

where $G_i^*(z)$ is the complex conjugate of $G_i(z)$. The integration in equation (1.22) is evaluated around a path which contains all the poles of $G_i(z)$ (normally this path is the unit circle).

The variance σ^2_ε of the total error on the output due to all the sources of noise $e_i(n)$, assuming uncorrelated sources, is given by

$$\sigma^2_\varepsilon = \sum_i \sigma^2_{\varepsilon_i} \quad . \tag{1.23}$$

The limits of the quantization error on the output of the filter can also be calculated using the same model. This limit, which is normally based on the worst case analysis, is as follows:

$$\varepsilon_i(n) = \sum_{m=0}^{n} g_i(n - m)\, e_i(m) \quad , \tag{1.24}$$

Therefore,

$$\left|\varepsilon_i(n)\right| \leq \sum_{m=0}^{n} \left|g_i(n - m)\, e_i(m)\right| \quad . \tag{1.25}$$

We can also write

$$\left|\varepsilon(n)\right| \leq \sum_i \left|\varepsilon_i(n)\right| \quad . \tag{1.26}$$

The evaluation of equations (1.21) - (1.26) for a complex filter is generally difficult. A number of authors [77-111] have presented solutions for the analysis of noise in recursive digital filters, based on similar equations to those presented above. The following is a summary of the individual works cited in the literature. The works considered in this survey are mainly on the analysis of various aspects of the multiplication quantization errors and the limit cycle oscillations in a recursive digital filter with fixed point arithmetic.

One of the first analyses of the quantization noise in digital control systems was carried out by Bertram [77], and later by Slaughter [78]. In these analyses a model similar to that of Fig. 1.12 was considered. The worst case error on the output was estimated by using the state equations of the filter. Knowles and Edwards [79] derived the bounds for the expected mean-square output noise for the cases of direct, cascade, and parallel realizations. Their analyses were based on equations (1.24) to (1.26).

Jackson [25], [81], [82] used the stochastic approach to give a comprehensive analysis of the quantization noise in various cascade and parallel realizations. He analysed the relation between the quantization noise and the dynamic range of the digital filter based on several scaling methods. He indicated that a good sequential ordering of the stages of the cascade realization can result in low round-off noise. The importance of the correct choice of the filter structure and its scaling was also indicated by Edwards et al. [83]. Their results, which are based on spectral estimation of the quantization noise, show that the ratio of the noise levels for canonic and parallel realizations for a particular eleventh order filter could be of the order of 10^{13} : 1. Kan and Aggarwal [84] considered the dead-band of the basic realizations (direct, parallel and cascade), and suggested some design rules for choosing a structure to give a minimum dead-band effect.

Sandburg and Kaiser [85] considered the existence of a general limit cycle in a digital filter, and derived a bound for the rms. of the amplitude of oscillation. Their analysis was based on consideration of the worst case spectral transformation of the quantization noise. Long and Trick [86] also considered the existence of limit cycle oscillations, and by considering the worst case errors, similar to the arguments for equation (1.25), they solved an absolute bound for the amplitude of limit

cycles in a second-order recursive filter. In a later work [87] they considered the absolute bound for a general recursive digital filter, and showed that their result was similar to that of Yakowitz and Parker [88].

Yakowitz and Parker used the state space representation of the digital filter, and by using the spectral theory of operators, they found a bound for the quantization error in a general digital filter. The exact solution to their equation proved difficult to solve in general. Therefore, they also presented an approximate solution which was easier to use. Ünver and Abdullah [89], Fink [90], and Belal [91] pursued their results and presented other approximations which were somewhat better.

Parker and Yakowitz [92-93] and Hwang [94] presented other studies of the round-off noise based on using the state space representation. In [92] and [94] these authors considered the statistical properties of the quantization noise, and derived bounds for the quantization noise in a general digital filter. In [93] Parker and Yakowitz used a state variable formulation of the quantization error in a multivariable digital filter and, based on the worst case error, they derived a bound for the quantization error.

Mitra et al. [95] presented a method for simplifying the evaluation of equation (1.22). This method was based on the partial fraction expansion of the product $G(z)\ G^*(z)$.

Other approximate solutions of the method of worst case error, i.e. equation (1.25), were also presented by Appel [96] and Chang [97]. They claimed better accuracy for their approximations.

Jenkins and Leon [98] used the theories of interval algebra to derive a solution for the bound of the output quantization error. Their method was based on analysing the interval convolution.

A new approach for the analysis of the statistical properties of the quantization noise was presented by Yarlagadda and Massad [99]. They used the bilinear transformation to transform equation (1.22) into an equivalent one in the s-domain and then find its solution.

The direct method of Liapunov was used by Johnson [100], and later by Hess [26], to derive an upper bound for the quantization error. However, the results derived using the Liapunov method in general are very pessimistic, and up to now no solutions based on this method have been found which give near true estimates.

A modification to the noise model was suggested by Fettweis [101]. He showed that a more appropriate way of defining the noise samples e(n) due to the quantization of the variable x is

$$e = k\, Q[x] - x \quad , \tag{1.27}$$

as opposed to equation (1.8). k here is a constant which depends on the energy of the quantization noise.

Halyo and McAlpine [102] studied the discrete properties of the quantization error. Their result showed that, when an M bit number is quantized to M' bits, if M-M' $\geq$ 6, the difference between the rms errors, assuming continuous and discrete models, is no greater than 0.1 dB. These studies are extended by Thompson [103], and Eckhardt and Schussler [104].

Parker and Girard [105] considered the correlation which exists between the round-off noise from different multipliers in a digital filter. This correlation was shown to be significant even when the filter is driven by a random signal. Their result also showed that the cross-correlation between the error sources is minimum when the multiplying constants are relatively prime.

The output round-off noise due to the multipliers of a digital

filter was shown by Fettweis [106-107] to be closely related to the sensitivity of the digital filter structure toward coefficient variations under the special input condition of a sine wave. Bonzanigo [108] followed up Fettweis's result and showed that low coefficient sensitivity does not necessarily mean low round-off noise under all input conditions. Jackson [109] derived a lower bound for the round-off noise in digital filters considering the relation between the output quantization noise and the coefficient sensitivity.

The analysis of the output noise in digital filters adopting magnitude truncation was considered by Liu and van-Valkenburg [110], and Claasen et al. [111]. The analysis of the noise in this case is more complex when compared with the rounding and value truncation. This is because there exists a high order of correlation between the signal and the quantization noise. However, assuming simple statistical properties of the signal, a solution was made possible.

Other studies of the quantization errors when they have developed to a deterministic stage (i.e. self-sustained or limit cycle oscillations) have also been extensive in the literature. Blackman [5] was one of the first to consider the limit cycle problem. He has referred to it as the "dead-band effect".

Many other authors have since studied the properties of limit cycle oscillations such as the amplitude, frequency, and other characteristics. These works have been based on considering the limit cycle oscillations as the worst case error conditions, or by developing some novel technique for their study. The studies which consider the limit cycle oscillations to be the worst case error have already been covered in this section. In the literature mentioned below, the authors have either considered the limit cycle from a deterministic point of view other than the worst case, or presented other techniques for

their consideration.

Jackson [25], [112] presented an effective value linear model for the study of limit cycle oscillations. He suggested that a digital filter under limit cycle oscillation is equivalent to a filter having effective coefficients of a digital oscillator. Using this idea he presented a simple formula for the amplitude of the limit cycles. His results normally work out to be correct. However, for some coefficients limit cycles with larger amplitude are observed.

Parker and Hess [26], [113] considered Jackson's model and other techniques to analyse the amplitude of limit cycle oscillations. They also presented the concept of a successive value phase plane, equivalent to the phase plane in continuous systems for the display of limit cycle oscillations. Hess [26] also noted several types of symmetries for limit cycle oscillations.

White [114] considered the describing function analysis of the limit cycles in a double precision digital filter. Ojongbede [29], [115-116] considered the state equation analysis of the properties of limit cycle oscillations in various realizations.

By assuming sinusoidal oscillation conditions for limit cycles, Hess and Parker [117] derived a heuristic bound for the frequency of digital oscillators with roundoff error. Claasen et al. [118] introduced the concept of accessible and inaccessible limit cycle oscillations. This idea is similar to that presented in Chapter 3 of this thesis.

Jatnieks and Shenoi [120] considered the limit cycle oscillation in a second-order digital filter of the coupled form. Their analysis was based on the effective value linear model of Jackson. Claasen et al. [119] derived the conditions for the absence of limit cycle oscillations in digital filters having one or more non-linear elements. Their analysis was based on considering the sufficient conditions in the frequency domain for the absence of any self-sustained oscillations.

The conditions for limit cycle oscillations can also occur under deterministic inputs such as a step function or a sine wave. Hess [26] studied this phenomenon briefly by considering simulation results. He found that for low input conditions the zero input limit cycle dominated the output. However, under large input conditions the forced output was the dominating factor. The period of the forced response limit cycles was studied by Storzbach [121]. He showed that the period of the limit cycle had to be an integral multiple of the period of the input signal. Kan and Aggarwal [122], Brubaker and Gowdy [123], and Maria and Fahmy [124] also considered limit cycle oscillations in digital filters under forced input conditions. These considerations are particularly useful for estimating properties of limit cycle oscillations in a cascade of first-and second-order structures. Results for these cases have been presented as a recursion formulas relating the bounds for limit cycle oscillations in a cascade structure.

Work on limit cycles in digital filters with sign-magnitude truncation have been presented by Bonzanigo [125], Kao [126], and Claasen et al. [127-129]. Kao showed that the only limit cycle oscillations which can exist in single precision digital filters with magnitude truncation are of the period of one or two samples. He therefore suggested the use of magnitude truncation as a simple way to eliminate complex limit cycles. This study was later taken up by Claasen et al., who indicated that the use of double precision arithmetic with magnitude truncation resulted in a digital filter with a very small possibility of limit cycle oscillations.

Thong and Liu [130] considered limit cycle oscillations in the combinatorial implementation of digital filters [53]. These are digital filters in which the multiplier is replaced by a look-up table. They used the effective value linear model of Jackson, and results from

computer simulations.

Results on the behaviour of limit cycle oscillations based on experimental verification have been presented by Büttner [131] and Kieburtz [132]. Kieburtz considered a tenth order cascade realization of a particular transfer function. He indicated the importance of the ordering of the sections of the cascade realization for minimum limit cycle amplitude. He also indicated that the use of dithering of the least significant bit of the quantized word could completely suppress all limit cycle oscillations.

Published literature on the study of quantization noise and limit cycle oscillations in digital filters also includes the study of the techniques for reducing the effects of these phenomena. Some of these techniques have already been mentioned in this section.

One of the basic methods used for reducing the quantization noise was to choose the cascade structure with the optimum ordering of its sub-sections. Howang [133] used dynamic programming techniques for the optimization, and Lee [27] [134] considered a minimax optimization method.

Mitra et al. [135] presented a technique for designing digital filters with low round-off noise. They suggested the use of all-pass sections which would ensure low quantization noise. By feeding the input signal at the appropriate internal nodes of the filter, the required transfer could then be achieved.

Mullis and Roberts [136-137] also considered the optimization of the cascade realization. They presented an optimum second-order structure for low round-off noise. However, their second-order structure used nine multipliers as compared with the minimum of only 6.

Suppression of limit cycle oscillation was first studied for

control systems. Oldenburger et al. [138-140] considered the use of an external signal of random or other nature to stabilize a non-linear control system. Blackman [5] applied this method to digital filtering. He suggested that the addition of a dither signal could suppress the dead-band effects (limit cycle oscillations). Wong and King [141] suggested the use of multi-rate digital filters with periodically varying coefficients to suppress limit cycle oscillations. Their conclusions were based on simulation results.

The idea of controlled quantization to suppress limit cycle oscillations was presented by Butterweck [142]. He defined an energy term, or pseudo-power, for the digital filter. By considering the conditions for the non-existence of limit cycles based on his pseudo-power term, he suggested that the systematic control of the rounding process could eliminate the existence of limit cycle oscillations. The idea of pseudo-power for wave digital filters was also studied by Fettweis and Meerkötter [143-144]. They derived a criterion for the absence of limit cycles in wave digital filters. By applying this method Meerkötter and Wegener [145] and Meerkötter [146] developed second-order filters with no limit cycle oscillations.

1.8 Outline and the contributions of this thesis

In this thesis several original studies of the finite wordlength digital filters, which are believed to advance the study of limit cycle oscillations, are presented. In this section the outline and the contribution of this thesis are discussed.

Chapter 2 deals with the background theoretical development for the studies in this thesis. It is argued (Section 2.2) that, for the study of the existence of limit cycle oscillations in digital filters, it is sufficient to only consider the behaviour of the state variables

of that filter. Using the method suggested by Jackson [25] and Hess [26] it is shown that there are only two basic realizations for the state variables of a second-order filter. These are called "type-A" and "type-B" realizations.

Further, in Chapter 2, the noise properties of the individual multiplier branches and some of the properties of the limit cycle oscillations are considered. The noise analysis, which is similar to the works of Halyo and McAlpine [102], considers the quantization noise to be a discrete variable. This consideration is used in Chapter 5 to analyse and compare the increase in the noise level caused by the signal stabilization method.

The development of the properties of limit cycle oscillations deals with the proof of some of the assumptions which are relevant for the analysis of limit cycle oscillations. These assumptions for the one- and two-sample limit cycles are widely used by other authors but, to the knowledge of this author, have not been proved.

In Chapter 3 the phase plane behaviour of second-order recursive filters of the type-A realization is considered. The full analysis and the results presented in this chapter are original and were developed prior to the publication of Claassen et al. [118].

Firstly in Chapter 3 the behaviour of infinite wordlength filters is considered. From these results, a suggestion for the definition of the energy (or Lyapunov) function of the filter is presented. This is used in Chapter 6 to select a suitable search area for the test cases.

For the finite wordlength filter, the trajectory of the zero input response is found to go through a process called merging. Various

results relating to the merging properties of the transient response of the second-order filter with the type-A realization have been developed.

Further in Chapter 3 it is shown that, because of the process of merging, limit cycle oscillations for these filters can be categorised into two forms. These are called accessible and inaccessible limit cycle oscillations.

The aim of the study presented in Chapter 4 has been to present some conditions by which the occurrence of limit cycle oscillations can be predicted. This idea is used later in Chapter 5 for developing techniques to suppress limit cycle oscillations. The arguments used are original and believed to be new.

The study carried out in Chapter 4 is based on a technique for transforming a digital filter into two parts; one which has integer coefficients and therefore no quantization effect, and the other which has coefficients of magnitude less than 0.5 and therefore has quantization effects. This consideration permit analysis of the properties of some of the limit cycle oscillations. The results for limit cycle oscillations with period greater than two samples are similar to those of Jackson [112] which have been shown to be wrong in some cases. However, this study has proved useful for the synthesis of a suitable structure for applying the method of signal stabilization (discussed in Chapter 5).

In Chapter 5 the method of signal stabilization is used to modify the conditions for the limit cycle oscillations derived in Chapter 4. The analysis of the amplitude of the suppressed limit cycle oscillations, increase in noise level, and the amplitude of a suitable signal for the stabilization method are presented. In Section 5.6 a discussion of a suitable signal for the stabilization is

given. The arguments and the results presented in Chapter 5 are original and are new results.

Chapter 6 covers the results for simulation studies. Collective presentation of these results is found to be convenient because some of these could be related to the works of more than one chapter.

Some original new results, based on simulation studies are presented. In addition, the theoretical results of Chapters 3, 4, and 5 are compared with the results from the simulations.

Chapter 7 concludes the results of this thesis and makes some suggestions for further studies.

CHAPTER 2

RECURSIVE DIGITAL FILTERS WITH FINITE WORDLENGTH ARITHMETIC

2.1 Introduction

In the previous chapter it was pointed out that a second-order recursive digital filter (or the biquadratic form) can be the basic building block for much more complex filters. In fact, it is more economical to design general purpose second-order recursive filters and use them in the parallel or the cascade forms, than to realize the complex transfer functions directly. Therefore, in this thesis we only consider the (commonly used) second-order recursive filters for the study of limit cycle oscillations.

In this chapter different aspects of the finite wordlength digital filters will be developed as background to the following chapters. First we will consider the simplification of the structure of the recursive second-order filter to a form suitable for the study of limit cycle oscillations. Then, in Section 2.3, the discrete behaviour of the quantization noise in the multiplier branches will be considered (similar results to Halyo and McAlpine [102]). Finally, some of the properties of the limit cycles in second-order filters will be developed to simplify the analyses for these oscillations.

2.2 Realization of the state variables of a second-order recursive filter

For studies of limit cycle behaviour of second-order filters it is usually sufficient to know the limit cycle behaviour of the filter's

state variables. This is because the output of the filter is related to the state variables by a non-recursive transfer function which does not contribute towards the build-up of the limit cycle oscillations.

In this section we use the method suggested by Jackson [25] and Hess [26] to determine all of the different structures which realize the state variables of the second-order recursive filter .

Consider the state equations of an ideal second-order digital filter with the state variables $x_1(n)$ and $x_2(n)$, as given in equation (2.1).

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} -a_{11} & -a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} u(n) \quad , \qquad (2.1)$$

where a_{ij}, for i and j = 1 or 2, are the coefficients of the filter, and u(n) is the input sequence. The signs for a_{ij} are chosen in the above form for convenience.

The structure represented by equation (2.1) is shown in Fig. 2.1.

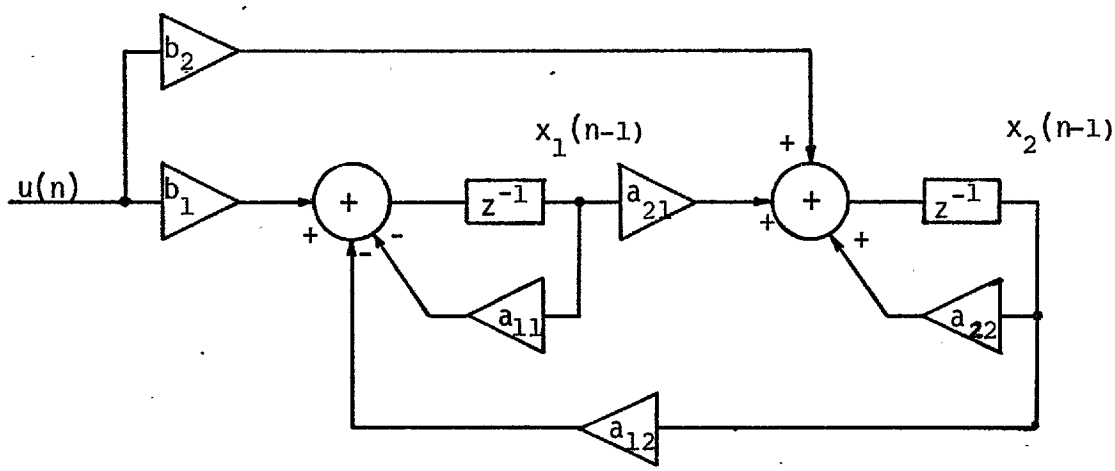


FIG. 2.1 A general structure for the realization of the state variables $x_1(n)$, $x_2(n)$.

In this structure there are a number of redundant coefficients which can therefore be chosen arbitrarily. In the following discussion we consider the choices involved, and thus derive all of the basic structures for the realization of the state variables of these filters.

To study the zero input limit cycle oscillations the choice of the b coefficients in equation (2.1) is immaterial. Therefore, to simplify the realization, these coefficients can be chosen as: $b_1 = 1$, and $b_2 = 0$.

The transfer functions relating the state variables and the input to the filter can now be written in the following forms:

$$H_1(z) = \frac{x_1(z)}{u(z)} = \frac{1 - a_{22}\, z^{-1}}{1 + (a_{11} - a_{22})\, z^{-1} + (a_{12}\, a_{21} - a_{11}\, a_{22})\, z^{-2}} \quad , \qquad (2.2)$$

and

$$H_2(z) = \frac{x_2(z)}{u(z)} = \frac{a_{21}\, z^{-1}}{1 + (a_{11} - a_{22})\, z^{-1} + (a_{12}\, a_{21} - a_{11}\, a_{22})\, z^{-2}} \quad . \qquad (2.3)$$

To realize a general characteristic equation defined by

$$1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} = 0 \quad , \qquad (2.4)$$

we require

$$a_{11} - a_{22} = \alpha_1 \quad , \qquad (2.5)$$

and

$$a_{12}\, a_{21} - a_{11}\, a_{22} = \alpha_2 \quad . \qquad (2.6)$$

There are two redundancies in the choice of the coefficients a_{ij}. The arbitrary values of these coefficients can be chosen such that the hardware requirements for the implementation of the resulting structure are minimized.

For example, coefficients of values zero or one do not require any multipliers, and the coefficients of values 2, 4, 8, . . . etc. or $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ etc., for a binary machine, can be implemented by simple shift operations. However, in the case of coefficients with values of 2, 4, 8, . . . etc., there is a larger chance for overflow which can limit the performance of the filter. On the other hand, coefficients with values of $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ etc. cause extra noise to be generated because of the loss of the Least Significant Bit (LSB). Thus, the values of the arbitrary coefficients are chosen as either zero or one. In particular the choice of zero is preferred since then the resultant structure will involve one less addition than the case with one as the coefficient.

Considering equations (2.5) and (2.6), for the unique representation of the values of α_1 and α_2, the arbitrary choices of one or zero can be made for either a_{11} or a_{22}, and the arbitrary choice of one for either a_{12} or a_{21}(see references [25] and [26] for more details). We also note that if, for example, a_{11} is chosen to be arbitrary, and $a_{22} = a_{11} - \alpha_1$, the resultant structure would be equivalent to choosing a_{22} to be arbitrary and letting $a_{11} = \alpha_1 + a_{22}$. Therefore, there are only two basic structures for the realization of the state variables of a second-order recursive filter. The one which uses a minimum number of components is obtained by choosing $a_{22} = 0$ and $a_{21} = 1$. The other structure is obtained by choosing $a_{22} = 1$. The former structure will be called "basic second-order section" or "type-A realization". The latter structure will be called "type-B realization".

Figures 2.2 and 2.3 show these structures when infinite precision arithmetic is used. In Fig. 2.2 we have used $\alpha = a_{11} = \alpha_1$, $\beta = a_{12} = \alpha_2$, and in Fig. 2.3 we have used $a = a_{11} = \alpha_1 + 1$, $b = a_{12} = 1 + \alpha_1 + \alpha_2$.

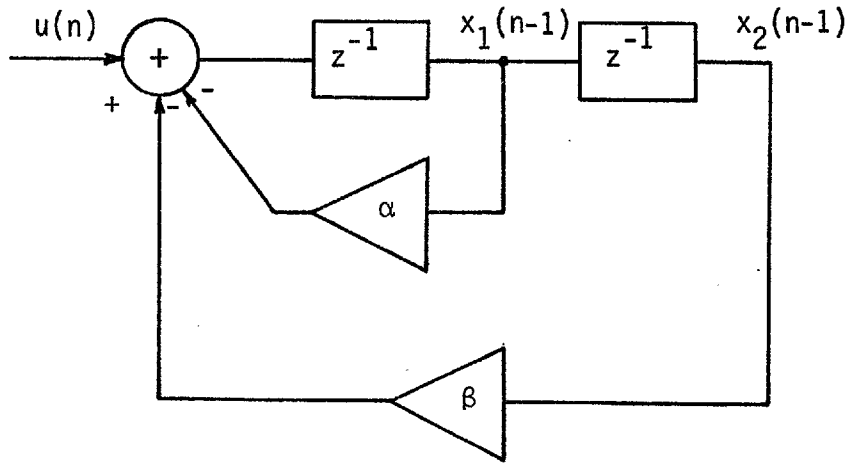


FIG. 2.2 Type-A realization

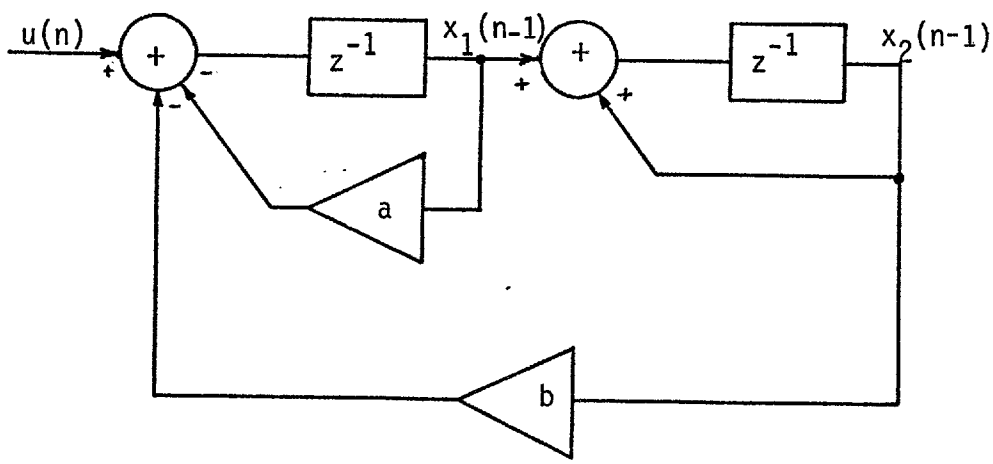


FIG. 2.3 Type-B realization

The difference equations for these basic structures are as follows:

i) for type-A realization

$$x_1(n) = -\alpha x_1(n-1) - \beta x_2(n-1) + u(n) \quad , \qquad (2.7a)$$

$$x_2(n) = x_1(n-1) \quad ; \tag{2.7b}$$

and ii) for the type-B realization

$$x_1(n) = -ax_1(n-1) - bx_2(n-1) + u(n) \quad , \tag{2.8a}$$

$$x_2(n) = x_1(n-1) + x_2(n-1) \quad . \tag{2.8b}$$

For finite wordlength arithmetic with single precision addition the above difference equations can be expressed as:

i) for type-A realization

$$x_1(n) = -Q_q[\alpha x_1(n-1)] - Q_q[\beta x_2(n-1)] + u(n) \quad , \tag{2.9a}$$

$$x_2(n) = x_1(n-1) \quad ; \tag{2.9b}$$

and ii) for type-B realization

$$x_1(n) = -Q_q[ax_1(n-1)] - Q_q[bx_2(n-1)] + u(n) \quad , \tag{2.10a}$$

$$x_2(n) = x_1(n-1) + x_2(n-1) \quad . \tag{2.10b}$$

Comparing the structures shown in Figs. 2.2 and 2.3 with the comprehensive range of second-order structures given by Jackson [25] and Hess [26] we note that if we ignore the input and the output of these structures, and assume infinite wordlength arithmetic, they can all be simplified to the forms shown. However the simplification for some of the structures require the combination of two parallel multipliers, which for finite wordlength single precision arithmetic is not possible. Thus, the results of this thesis only apply to those structures with finite wordlength arithmetic which can be simplified to the forms shown in Figs. 2.2 and 2.3. Further studies are necessary to find the behaviour of the other class of filters. However, they can be assumed to have a worse noise (and therefore limit cycle) performance, because, these structures use a larger number of multipliers in their recursive section (note that a larger number of multipliers cannot give a better noise performance).

2.3 Some remarks on the multiplication quantization error

The quantization error from a multiplier branch with finite wordlength coefficient is a quantized variable [102-104]. Normally this

error can be approximated to be a continuous variable. However, for comparison of the noise level of the "signal stabilization" technique discussed in Chapter 5, this approximation will not be sufficient.

In this section we consider the average error due to quantization when the coefficient has a finite wordlength. The model used for representing the quantization error is the noise source model [7-11] of Fig. 2.4. This model can also be described as

$$Q_q[\gamma x(n)] = \gamma x(n) + e(n) \quad . \tag{2.11}$$

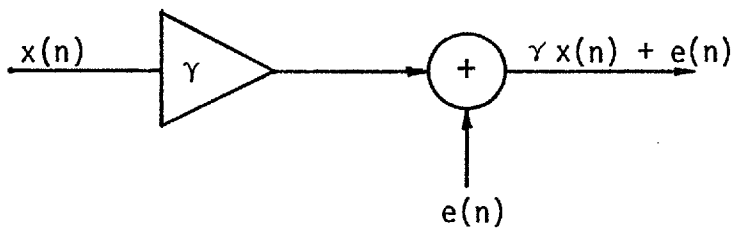


FIG. 2.4

Let the quantization step of the state variable x(n) be q, and let the quantization step of the coefficient γ be q_c. Then the quantization step of the error e(n) will be given by $q_e = qq_c$, provided that q_c is defined appropriately. For example, let $\gamma = \frac{6}{32}$, then q_c can be $\frac{1}{32}$, $\frac{1}{16}$ or $\frac{3}{16}$. If $q_c = \frac{1}{32}$, then for q = 1, the possible values of e(n) for the case of rounding can be any of the following: $\pm\frac{2}{32}$, $\pm\frac{4}{32}$, $\pm\frac{6}{32}$, . . . $\pm\frac{14}{32}$, $-\frac{16}{32}$. This indicates that the quantization step of e(n) is $\frac{1}{16}$, which is not the same as qq_c. However, if $q_c = \frac{1}{16}$, then the quantization step of e(n) will be correctly equal to qq_c.

To find the function for evaluating q_c, consider for example the case of rounding, and let x(n) = jq and $\gamma = iq_c$. Then from the properties of the rounding function as given in Section 1.4, we have

$$- e(n) = kqq_c = (ijqq_c + 0.5q) \bmod q - 0.5q \quad , \tag{2.12}$$

where i, j, k are integers, and "mod" is the modulo function. Dividing both sides of equation (2.12) by q and simplifying, we have

$$(j\gamma + 0.5) \bmod 1 - 0.5 = kq_c \quad , \tag{2.13}$$

where q_c has a value such that k is an integer for all the integer values of j. The condition expressed by equation (2.13) is equivalent to

$$\frac{j\gamma \bmod 1}{q_c} = k' \quad , \tag{2.14}$$

where k' is an integer for all the integer values of j.

A further condition for e(n) to have the quantization step of qq_c is

$$\frac{1}{q_c} = M \quad , \tag{2.15}$$

where M is an integer. This property of q_c can be seen from the modulo function, $(m\gamma) \bmod 1$, as shown in Fig. 2.5. That is, for $(iq_c) \bmod 1 \big|_{i = \pm M,\ \pm 2M,\ \ldots \text{ etc.}} = 0$, q_c has to satisfy equation (2.15).

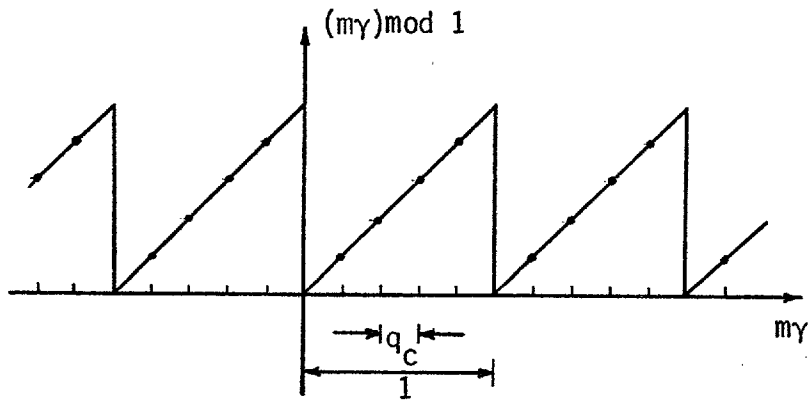


FIG. 2.5 The modulo function

If the signal x(n) is random and has standard deviation much larger than the quantization step q, the probability density of e(n) will be nearly uniform and independent of the probability density and spectral distribution of the signal x(n). Thus for many different types of input, the probability density of the error e(n) will be of the forms given by Fig. 2.6 for the case of rounding, Fig. 2.7 for the case of value truncation and Fig. 2.8 for the case of magnitude truncation. In each of these cases N_e is defined as $N_e = \frac{1}{2q_c}$.

Consider the case of rounding, i.e. Fig. 2.6, the variance $\sigma^2_{e_R}$ and the mean square error $\rho^2_{e_R}$ are given by

$$\rho^2_{e_R} = \sigma^2_{e_R} = \frac{1}{2N_e} \sum_{i=-N_e}^{N_e-1} (iqq_c)^2 \quad . \tag{2.16}$$

i.e.

$$\rho^2_{e_R} = \left[\frac{1 + 2q_c^2}{12}\right] q^2 \quad . \tag{2.17}$$

If $q_c \ll 1$ then the above equation can be approximated to

$$\rho_{e_R} = \sigma_{e_R} \approx \frac{1}{\sqrt{12}}\, q \quad . \tag{2.18}$$

In the case of value truncation the mean square error can be written as

$$\rho^2_{e_{T2}} = \frac{1}{2N_e} \sum_{i=-(2N_e-1)}^{0} (iqq_c)^2 \quad , \tag{2.19}$$

where N_e is as defined above. Therefore,

$$\rho^2_{e_{T2}} = \frac{(1 - q_c)(2 - q_c)\, q^2}{6} \quad . \tag{2.20}$$

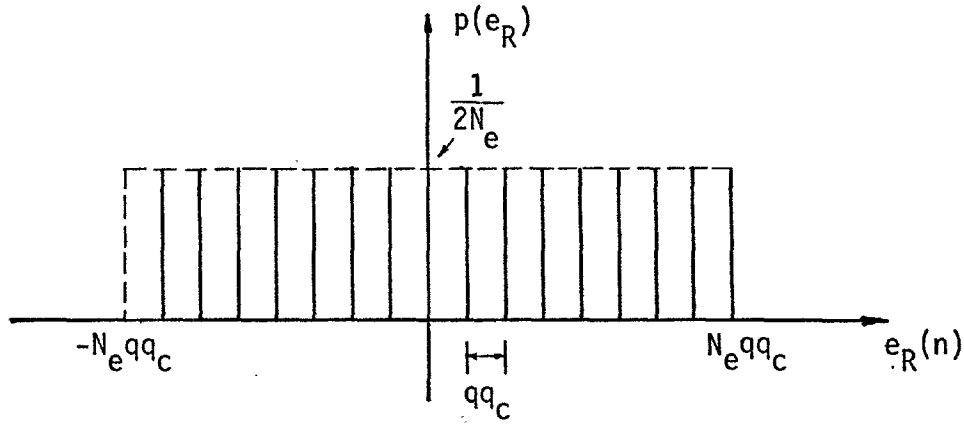


FIG. 2.6 Probability density of error due to rounding

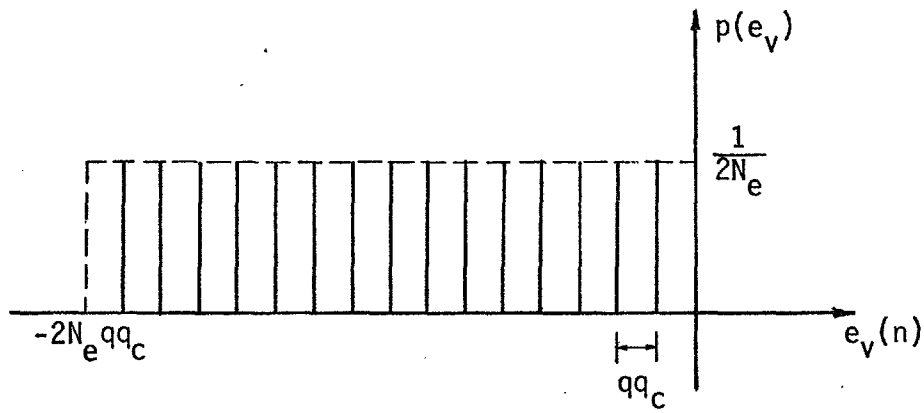


FIG. 2.7 Probability density of error due to value truncation

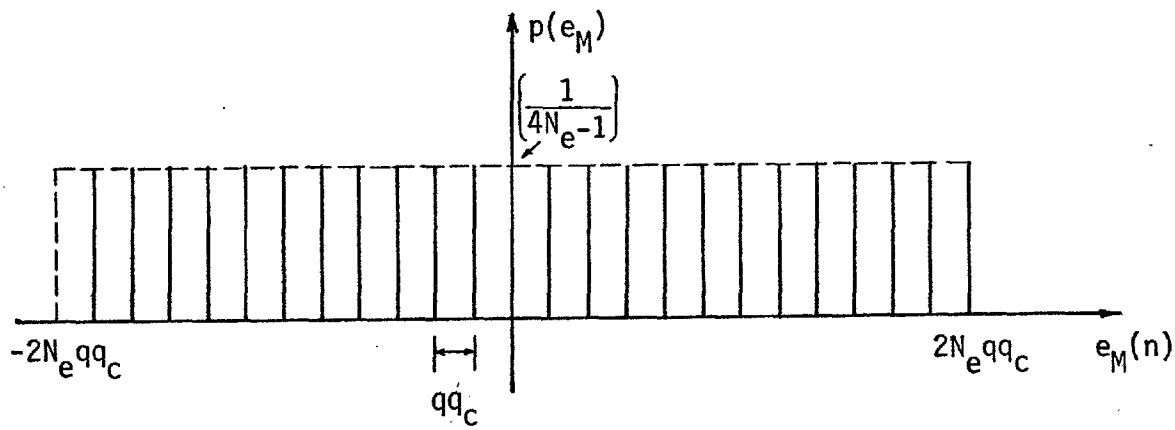


FIG. 2.8 Probability density of error due to magnitude truncation

If $q_c << 1$, then the above equation can be approximated to

$$\rho_{e_{T2}} \approx \frac{1}{\sqrt{3}} q \quad . \tag{2.21}$$

In the case of magnitude truncation the mean square error is given by

$$\rho^2_{e_{T1}} = \frac{1}{(4N_e - 1)} \sum_{i=-(2N_e-1)}^{(2N_e-1)} (iqq_c)^2 \quad , \tag{2.22}$$

where N_e is as defined above. Therefore,

$$\rho^2_{e_{T1}} = \frac{1}{3}(1 - q_c)\, q^2 \quad , \tag{2.23}$$

and if $q_c << 1$, then we have

$$\rho_{e_{T1}} \approx \frac{1}{\sqrt{3}} q \quad . \tag{2.24}$$

The above results show that the approximation

$$\rho_{e_{T1}} \approx \rho_{e_{T2}} \approx 2\rho_{e_R} \approx \frac{1}{\sqrt{3}} q$$

is only true if $q_c << 1$.

The results expressed by equations (2.18), (2.21) and (2.24) are well known [7-11]. However, the above analysis indicates that if q_c is not much smaller than unity, the correct estimates for the mean square error are given by the results presented above. For example, the difference between the mean square value of the rounding error due to two coefficients with the values of $\frac{4}{8}$ and $\frac{3}{8}$ for the same quantization level q, is 1.63 dB .

2.4 Some remarks on limit cycle oscillations in second-order recursive filters

In Chapter 1 it was pointed out that, depending on their period of oscillation, limit cycles in digital filters can be categorised into three different types. These are limit cycles with the period of one sample, two samples, or greater than two samples. In this section we consider each of these types separately, and derive new properties which will be useful in the analysis.

2.4.1 Limit cycles with period of one sample

In general, limit cycle oscillations of this type can be described by the sequence {Kq, Kq, Kq, . . . etc.}, where K is an integer and q is the quantization step. The condition for this type of limit cycle oscillation in the type-A realization is therefore given by

$$Kq + Q_q[\alpha Kq] + Q_q[\beta Kq] = 0 \quad . \tag{2.25}$$

The solutions to equation (2.25) give all the possible amplitudes for this type of limit cycle oscillation. In Chapter 4 a method for solving this equation will be discussed.

For the type-B realization we note that

$$x_2(n) = x_1(n-1) + x_2(n-1) \quad ,$$

i.e.

$$x_2(n) = \sum_{i=0}^{n-1} x_1(i) \quad . \tag{2.26}$$

Hence, for the state variable $x_2(n)$ to be bounded in the case of a filter with stable response

$$\text{ave}[x_1(n)] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} x_1(i) = 0 \quad , \tag{2.27}$$

where ave[ ] represents the time average of the signal. It follows that

limit cycles with period of one sample cannot exist on the state variable $x_1(n)$. However, the state variable $x_2(n)$ can accommodate limit cycles of this type. Therefore the limit cycles of period one and amplitude Kq, on the state variable $x_2(n)$, have to satisfy

$$Q_q[bKq] = 0 \quad , \tag{2.28}$$

i.e. for the case of rounding

$$|K| \leq \frac{0.5}{|b|} \quad , \tag{2.29a}$$

for the case of value truncation

$$0 \leq K < \frac{1}{|b|} \quad , \tag{2.29b}$$

and for the case of magnitude truncation

$$|K| < \frac{1}{|b|} \quad . \tag{2.29c}$$

2.4.2 Limit cycles with period of two samples

In general limit cycle oscillations with period of two samples can be described by the sequence {Kq, -Lq, Kq, -Lq, . . . etc.}, where K and L are integers. The conditions for this type of oscillation in the type-A realization can be written as

$$Kq + Q_q[-\alpha Lq] + Q_q[\beta Kq] = 0 \quad , \tag{2.30a}$$

and

$$-Lq + Q_q[\alpha Kq] + Q_q[-\beta Lq] = 0 \quad . \tag{2.30b}$$

From equations (2.30a) or (2.30b), we note that:

i) if $\alpha > 0$, $\text{sgn}(K) = \text{sgn}(L)$,

and ii) if $\alpha < 0$, $\text{sgn}(K) \neq \text{sgn}(L)$.

Considering rounding or value truncation, equations (2.30a and b) can be rewritten as

$$Q_q[(1+\beta)Kq] + Q_q[-\alpha Lq] = 0 \quad , \tag{2.31a}$$

and

$$Q_q[-(1+\beta)Lq] + Q_q[\alpha Kq] = 0 \quad . \tag{2.31b}$$

However, the condition for the stability of the digital filter is given by

$$(1+\beta) > |\alpha| \quad . \tag{2.32}$$

Therefore, if $K > 0$,

$$Q_q[(1+\beta)Kq] \geq Q_q[\alpha Kq] \quad . \tag{2.33}$$

From equations (2.31a and b) it follows that

$$Q_q[-(1+\beta)Lq] \geq Q_q[-\alpha Lq] \quad . \tag{2.34}$$

But if $L > 0$, for non-zero solutions $\alpha > 0$, and since $(1+\beta) > 0$, then

$$Q_q[-(1+\beta)Lq] \not\geq Q_q[-\alpha Lq] \quad . \tag{2.35}$$

From inequalities (2.34) and (2.35) it follows that

$$Q_q[-(1+\beta)Lq] = Q_q[-\alpha Lq] \quad , \tag{2.36a}$$

and

$$Q_q[(1+\beta)Kq] = Q_q[\alpha Kq] \quad . \tag{2.36b}$$

Similarly for all possible values of L and K we can derive similar equations, all of which may be expressed in the form

$$Kq + d\, Q_q[\alpha Kq] + Q_q[\beta Kq] = 0 \tag{2.37}$$

where $d = +1$ or -1 for $\alpha < 0$ or $\alpha > 0$ respectively. Equation (2.37) describes the general condition for limit cycle oscillations of period one-or two-samples in the cases of rounding or value truncation.

The above results lead to the following theorem concerning limit cycles of period two in the case of rounding.

Theorem 2.1

In a basic second-order recursive digital filter (type-A realization) with single precision arithmetic and rounding for the

quantization, if limit cycle of the sequence {Kq, -Lq, Kq, -Lq, . . . etc.}, where sig(K) = sig(L), is possible, then the limit cycles with the sequences {Kq, -Kq, Kq, -Kq, . . . etc.} and {Lq, -Lq, Lq, -Lq, . . . etc.} will also be possible. Similarly if limit cycle with the sequence {Kq, Lq, Kq, Lq, . . . etc.} is possible, then the limit cycles with the sequences {Kq, Kq, Kq, . . . etc.} and {Lq, Lq, Lq, . . . etc.} will also be possible.

For the type-B realization the limit cycles with the period of two samples can be described as {Kq, -Kq, Kq, -Kq, . . . etc.} for the state variable $x_1(n)$, and as {-Lq, Mq, -Lq, Mq, . . . etc.} for the state variable $x_2(n)$. In these descriptions K, L and M are integers, where

$$K = L + M \quad . \tag{2.38}$$

The conditions for the existence of this type of limit cycle oscillation are given by

$$Kq + Q_q[-aKq] + Q_q[bMq] = 0 \quad , \tag{2.39a}$$

and

$$-Kq + Q_q[aKq] + Q_q[-bLq] = 0 \quad . \tag{2.39b}$$

For the case of rounding, equations (2.39a and b) can be rewritten as

$$R_q[(a-1)Kq] = R_q[bMq] \quad , \tag{2.40a}$$

and

$$R_q[(a-1)Kq] = R_q[bLq] \quad . \tag{2.40b}$$

To simplify equations (2.30a and b), we note that

$$(bM - 0.5)q \leq R_q[bMq] < (bM + 0.5)q \quad , \tag{2,41a}$$

and

$$(bL - 0.5)q \leq R_q[bLq] < (bL + 0.5)q \quad . \tag{2.41b}$$

Adding inequalities (2.41a) and (2.41b) and from equations (2.38) and (2.40a and b) it follows that

$$\left(\frac{bK}{2} - 0.5\right)q \leq R_q[(a-1)Kq] < \left(\frac{bK}{2} + 0.5\right)q \quad , \tag{2.42}$$

i.e.

$$R_q[(a-1)Kq] = R_q\left[\frac{bKq}{2}\right] \quad , \tag{2.43a}$$

and

$$R_q[bLq] = R_q[bMq] = R_q\left[\frac{bKq}{2}\right] \quad . \tag{2.43b}$$

The amplitude of limit cycles with the period of two samples on the state variables $x_1(n)$ and $x_2(n)$ of the type-B realization for the case of rounding can be found by solving equations (2.43a) and (2.43b) respectively. The solutions to equation (2.43a) will be investigated in Chapter 4, and the solutions to equation (2.43b) are then a simple extension when $R_q\left[\frac{bKq}{2}\right]$ is known.

These results show that, for the study of the amplitude of limit cycles with the period of two samples on the state variable $x_1(n)$ for the type-B realization in the case of rounding, the sequences for the state variables $x_1(n)$ and $x_2(n)$ can be chosen as {Kq, -Kq, Kq, . . . etc.} and $\{-\frac{K}{2}q, \frac{K}{2}q, -\frac{K}{2}q, \ldots \text{etc.}\}$ respectively.

For the case of value truncation a similar method may be used to simplify equation (2.39a) and (2.39b). In this case we have

$$T2_q[-(a-1)Kq] + T2_q[bMq] = 0 \quad , \tag{2.44a}$$

and

$$T2_q[(a-1)Kq] + T2_q[-bLq] = 0 \quad . \tag{2.44b}$$

Equations (2.44a and b) can be rewritten as

$$T2_q[(a-1)Kq] - T2_q[bMq] = -q \quad , \tag{2.45a}$$

and

$$T2_q[(a-1)Kq] - T2_q[bLq] = q \quad . \tag{2.45b}$$

However,

$$bMq \leqslant T2_q[bMq] < (bM+1)q \quad , \tag{2.46a}$$

and

$$bLq \leqslant T2_q[bLq] < (bL+1)q \quad . \tag{2.46b}$$

Adding inequalities (2.46a) and (2.46b), and from equations (2.38) and (2.45a and b), it follows that

$$\frac{bKq}{2} \leqslant T2_q[(a-1)Kq] < \left(\frac{bK}{2}+1\right)q \quad , \tag{2.47}$$

i.e.

$$T2_q[(a-1)Kq] = T2_q\left[\frac{bKq}{2}\right] \quad . \tag{2.48}$$

Equation (2.48) describes the condition for limit cycles with period of two samples and amplitude of Kq, on the state variable $x_1(n)$ of the type-B realization in the case of value truncation. The solutions to this equation will be discussed in Chapter 4.

Comparing equations (2.43a) and (2.48) we note that a general condition for limit cycles of amplitude Kq and period of two samples on the state variable $x_1(n)$ of the type-B realization, is given by

$$Q_q[(a-1)Kq] = Q_q\left[\left(\frac{b}{2}\right)Kq\right] \quad , \tag{2.49}$$

where $Q_q[\ ]$ represents either rounding or value truncation.

2.4.3 Limit cycles with period greater than two samples

Limit cycle oscillations with period greater than two samples in the type-A realization have to satisfy

$$x_1(n) + Q_q[\alpha x_1(n-1)] + Q_q[\beta x_2(n-1)] = 0 \quad , \tag{2.50}$$

i.e.

$$x_1(n) + \alpha x_1(n-1) + \beta x_1(n-2) = e_1(n) + e_2(n) \quad ,$$

where $e_1(n)$ and $e_2(n)$ are the quantization errors of the products $\alpha x_1(n)$ and $\beta x_2(n)$ respectively. Therefore, we can write

$$\text{ave}[x_1(n)] = \text{ave}[x_2(n)] = \frac{\text{ave}[e_1(n)] + \text{ave}[e_2(n)]}{1 + \alpha + \beta} \quad . \tag{2.51}$$

Thus, for rounding

$$\text{ave}[x_1(n)] \approx 0 \quad , \tag{2.52}$$

and for value truncation

$$\text{ave}[x_1(n)] \approx \frac{q}{1+\alpha+\beta} \quad . \tag{2.53}$$

For type-B realization, in Section 2.4.1 it was shown that

$$\text{ave}[x_1(n)] = 0 \quad . \tag{2.54}$$

By considering the difference equations for this case it can be shown that

$$\text{ave}[x_2(n)] = \frac{\text{ave}[e_1(n)] + \text{ave}[e_2(n)]}{b} \quad , \tag{2.55}$$

i.e. for rounding

$$\text{ave}[x_2(n)] \approx 0 \quad , \tag{2.56}$$

and for value truncation

$$\text{ave}[x_2(n)] \approx \frac{q}{b} \quad . \tag{2.57}$$

2.5 Conclusions

In this chapter several aspects of the finite wordlength second-order recursive digital filters were investigated. The results found are as follows.

i) It was suggested that for the study of limit cycle oscillations, in most cases, it is sufficient to consider the behaviour of the state variables of the filter and therefore ignore the zeros of the transfer function.

The method suggested by Jackson [25] and Hess [26] was used to show that there are only two basic structures for the realization of the state variables of a second-order recursive filter. It was also mentioned that the basic second-order section, or the type-A realization, is the most economical in terms of hardware.

ii) In Section 2.3 some of the properties of noise in digital filters were considered. It was shown that the effective noise introduced by a quantizer depended on the quantization step q_c of the coefficients (similar result to Halyo and McAlpine [102]). This value, as defined by equations (2.14) and (2.15), should be considered when comparing noise from the coefficients with different quantization steps.

iii) Finally some of the properties of limit cycle oscillations in the two structures of type-A and type-B realizations were investigated. It was shown that for the study of the amplitude of two-sample limit cycle oscillations it is sufficient to consider the solutions to equation (2.37) for the type-A realization and equation (2.49) for the type-B realization.

Limit cycle oscillations with period greater than two samples were considered by an approximate study of their time average. These results are given by equations (2.52) and (2.56) for rounding and equations (2.53) and (2.57) for value truncation.

CHAPTER 3

A STUDY OF LIMIT CYCLE OSCILLATIONS IN THE PHASE PLANE

3.1 Introduction

For a second-order analog system the phase plane is generally defined as the plane with the cartesian coordinates representing the state variables of that system. Normally these variables are $x(t)$ and $\dot{x}(t)$, where $\dot{x}(t)$ represents the derivative of the state variable $x(t)$.

In this study we shall represent the phase plane for the second-order recursive digital filter also in terms of its state variables $x_1(n)$ and $x_2(n)$. This representation is equivalent to the successive value phase plane suggested by Hess [26] [113]. He suggested that for a basic second-order section of Fig. 2.2, a suitable definition for the phase plane is the one having the cartesian coordinates representing the present and the immediately past values of the output.

In this chapter we present new results derived from the zero input phase plane response of the basic second-order section of Fig. 2.2. Initially the transient response of a filter with infinite precision arithmetic is considered. From the trajectory of the infinite precision filter a term for the energy (or Lyapunov) function in the digital filter has been suggested. The result for the infinite precision filter is then modified to represent the general behaviour of the finite precision filter.

In Sections 3.3 to 3.5 the properties of the trajectories for the

finite precision filter are considered, and used to derive new results on the behaviour of the limit cycle oscillations.

3.2 The transient response of an infinite precision second-order filter

The difference equation for the infinite precision digital filter of Fig. 2.2 is given by

$$x_1'(n) = -\alpha x_1'(n-1) - \beta x_2'(n-1) + u(n) \quad , \tag{3.1a}$$

$$x_2'(n) = x_1'(n-1) \quad , \tag{3.1b}$$

where $x_1'(n)$ and $x_2'(n)$ represent the state variables of the ideal filter.

If the input to the filter is zero, and any transient response is assumed to be due to the initial conditions $x_1(0)$, $x_2(0)$ for the state variables $x_1'(n)$ and $x_2'(n)$ respectively, we can write the z-transform for $x_1'(n)$ as

$$X_1'(z) = \frac{x_1(0) - \beta x_2(0)\, z^{-1}}{1 + \alpha z^{-1} + \beta z^{-2}} \quad ,$$

or

$$X_1'(z) = [x_1(0) - \beta x_2(0)\, z^{-1}]\, H(z) \quad , \tag{3.2}$$

where

$$H(z) = \frac{1}{1 + \alpha z^{-1} + \beta z^{-2}}$$

is the transfer function of the filter with $x_1'(n)$ as the output.

If $h(n)$ is the inverse z-transform of $H(z)$ (i.e. the impulse response), then we can write

$$x_1'(n) = x_1(0)\, h(n) - \beta x_2(0)\, h(n-1) \quad , \tag{3.3a}$$

and

$$x_2'(n) = x_1(0)\, h(n-1) - \beta x_2(0)\, h(n-2) \quad , \tag{3.3b}$$

for $n \geq 1$.

In general there are two types of responses described by equations (3.3a) and (3.3b). These depend on whether the poles of the filter are complex or real. The trajectory for each case is described below.

3.2.1 The case of real poles

Let the poles of H(z) be represented by the real quantities λ_1 and λ_2. Then the impulse response h(n) is given by

$$h(n) = \frac{\lambda_1^{n+1} - \lambda_2^{n+1}}{\lambda_1 - \lambda_2} \quad . \tag{3.4}$$

Substituting for h(n), h(n - 1) and h(n - 2) in equation (3.3) we obtain

$$x_1'(n) = \left[\frac{\lambda_1 x_1(0) - \beta x_2(0)}{\lambda_1 - \lambda_2}\right] \lambda_1^n + \left[\frac{\beta x_2(0) - \lambda_2 x_1(0)}{\lambda_1 - \lambda_2}\right] \lambda_2^n \quad , \tag{3.5a}$$

and

$$x_2'(n) = \left[\frac{\lambda_1 x_1(0) - \beta x_2(0)}{\lambda_1 - \lambda_2}\right] \lambda_1^{n-1} + \left[\frac{\beta x_2(0) - \lambda_2 x_1(0)}{\lambda_1 - \lambda_2}\right] \lambda_2^{n-1} \quad . \tag{3.5b}$$

Writing equations (3.5a and b) in the vector form, we have

$$\underline{X}'(n) = \begin{bmatrix} x_1'(n) \\ x_2'(n) \end{bmatrix} = \left[\frac{\lambda_1 x_1(0) - \beta x_2(0)}{\lambda_1 - \lambda_2}\right] \lambda_1^n \begin{bmatrix} 1 \\ \frac{1}{\lambda_1} \end{bmatrix} +$$

$$\left[\frac{\beta x_2(0) - \lambda_2 x_1(0)}{\lambda_1 - \lambda_2}\right] \lambda_2^n \begin{bmatrix} 1 \\ \frac{1}{\lambda_2} \end{bmatrix} \quad . \tag{3.6}$$

It follows that the sample representing the transient response of the filter in the phase plane, at the time index n, is the sum of two vectors,

$$A\lambda_1^n \begin{bmatrix} 1 \\ \frac{1}{\lambda_1} \end{bmatrix} \quad \text{and} \quad B\lambda_2^n \begin{bmatrix} 1 \\ \frac{1}{\lambda_2} \end{bmatrix} \quad ,$$

where

$$A = \frac{\lambda_1 x_1(0) - \beta x_2(0)}{\lambda_1 - \lambda_2} \quad ,$$

and

$$B = \frac{\beta x_2(0) - \lambda_2 x_1(0)}{\lambda_1 - \lambda_2} \quad .$$

These vectors are in the direction of $\left[1, \frac{1}{\lambda_1}\right]$ and $\left[1, \frac{1}{\lambda_2}\right]$ respectively, having magnitudes of λ_1^n and λ_2^n respectively, which are functions of the time index n. An example of this trajectory for the case when both λ_1 and λ_2 are positive is shown in Fig. 3.1. The different trajectories

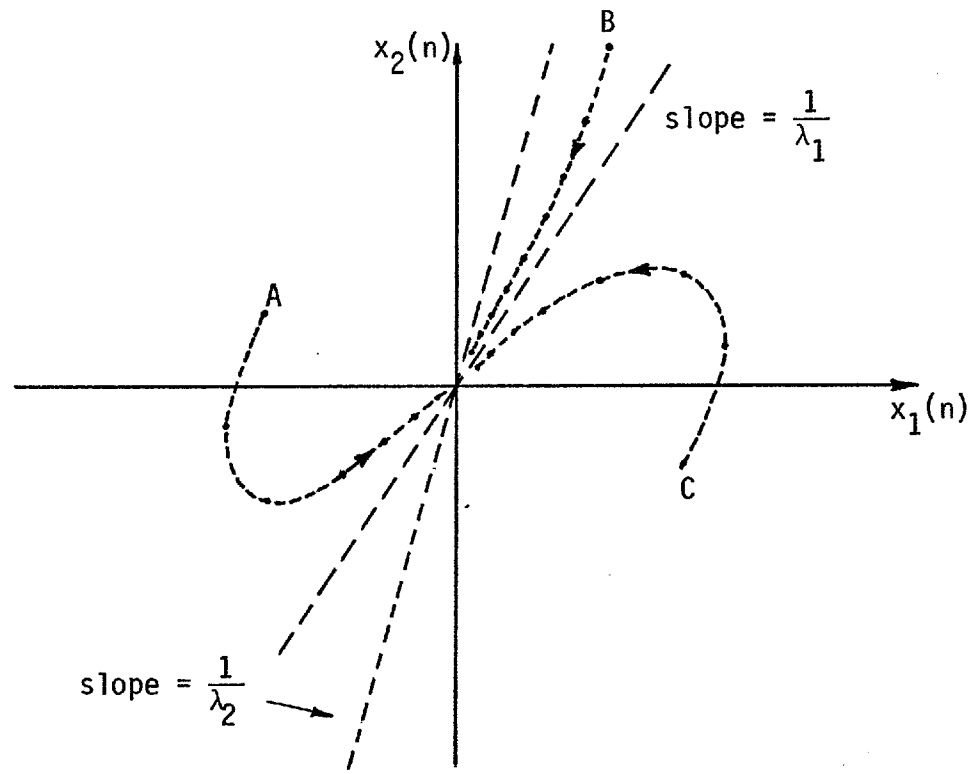


FIG. 3.1 Typical phase plane transient trajectories of an infinite precision filter with real poles

given in this diagram show the behaviour of the transient response with different initial conditions A, B and C.

3.2.2 The case of complex poles

If the poles of the transfer function $H(z)$ are complex, then they can be represented by the conjugate pair $re^{j\theta}$ and $re^{-j\theta}$, where r and θ are the magnitude and the angle of the complex poles (i.e. $r = \sqrt{\beta}$ and $\theta = \cos^{-1}\frac{\alpha}{2\sqrt{\beta}}$).

By taking the inverse z-transform of $H(z)$ we can show that

$$h(n) = r^n \frac{\sin(n+1)\theta}{\sin\theta} \quad . \tag{3.7}$$

From equation (3.3) we, therefore, have

$$x_1'(n) = x_1(0)r^n \frac{\sin(n+1)\theta}{\sin\theta} - \beta x_2(0)r^{n-1} \frac{\sin n\theta}{\sin\theta} \quad , \tag{3.8a}$$

for $n \geq 0$, and

$$x_2'(n) = x_1(0)r^{n-1} \frac{\sin n\theta}{\sin\theta} - \beta x_2(0)r^{n-2} \frac{\sin(n-1)\theta}{\sin\theta} \quad , \tag{3.8b}$$

for $n > 0$.

Combining equations (3.8a) and (3.8b) we get

$$x_1'^2(n) + \beta x_2'^2(n) + \alpha x_1'(n)\, x_2'(n) = \beta^n [x_1^2(0) + \beta x_2^2(0) + \alpha x_1(0)\, x_2(0)] \quad , \tag{3.9}$$

for $n > 0$. We note that equation (3.9) is also true for time index $n = 0$. This case describes the initial conditions of the filter.

To simplify equation (3.9), we define a constant $\Gamma(0)$ which depends only on the initial conditions, and the coefficients of the filter. This is given by

$$\Gamma(0) = x_1^2(0) + \beta x_2^2(0) + \alpha x_1(0)\, x_2(0) \quad . \tag{3.10}$$

Therefore, equation (3.9) can be written as

$$x_1'^2(n) + \beta x_2'^2(n) + \alpha x_1'(n)\, x_2'(n) = \beta^n \Gamma(0) \quad . \tag{3.11}$$

The relation between $x_1(n)$ and $x_2(n)$ described by equation (3.11) represents the trajectory of the transient response of the filter in the phase plane. We note that, at the time index n, the sample of the trajectory in the phase plane lies on an ellipse with its centre at the origin, and its size dependent on the quantity $\beta^n\Gamma(0)$. According to the definition of $\Gamma(0)$ we can also define

$$\Gamma(n) = x_1^2(n) + \beta x_2^2(n) + \alpha x_1(n)\, x_2(n) \quad , \tag{3.12}$$

which can be written as

$$\Gamma(n) = \beta^n\Gamma(0) \quad .$$

Therefore, we can conclude that the phase plane trajectory of a stable filter with complex poles is an elliptical spiral as shown in Fig.3.2. Each

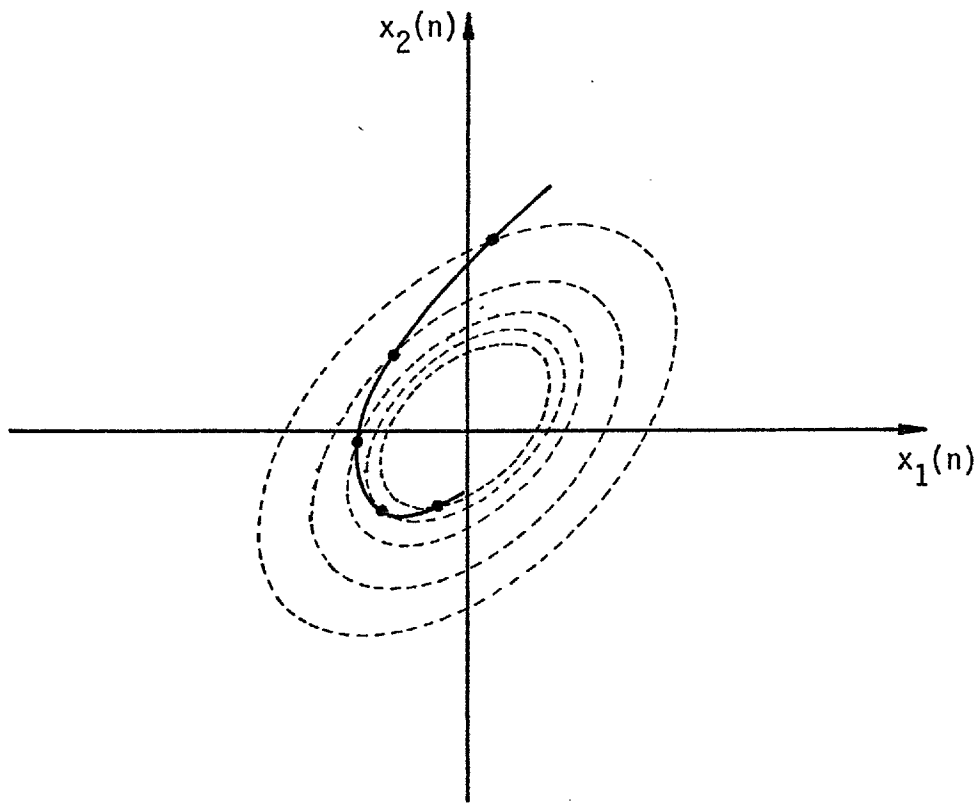


FIG. 3.2 Typical phase plane transient trajectories of an infinite precision filter with complex poles

sample of this trajectory lies on an ellipse of constant shape but of diminishing size at the rate of β per sample.

The variable $\Gamma(n)$ as defined above can also be defined in the vector form as

$$\Gamma(n) = \underline{X}'^{T}(n)\ P\ X'(n) \quad , \tag{3.13}$$

where

$$P = \begin{bmatrix} 1 & \frac{\alpha}{2} \\ \frac{\alpha}{2} & \beta \end{bmatrix} \quad .$$

P is a positive definite matrix when $4\beta > \alpha^2$ (i.e. condition for complex poles). Therefore, for a second-order filter with complex poles we can consider $\Gamma(n)$ as a Lyapunov or energy function. According to this definition for energy the rate of loss of energy per sample in a stable filter, having complex poles and infinite precision arithmetic, is the factor $(1 - \beta)$ per sample.

3.3 Response of a finite precision filter in the phase plane

The analysis of finite wordlength digital filters can normally be carried out using two basic modelling techniques: the equivalent coefficient model of Jackson [112], and the equivalent noise source model [7-11]. A survey of these techniques was presented in Chapter 1. In this section, to estimate the general trend for the trajectory of a finite precision filter, we consider the equivalent noise source model and apply the analysis technique discussed in the previous section.

If the total quantization error in the evaluation of $x_1(n)$ is $e(n)$, then the total deviation of $x_1(n)$ from the ideal response $x_1'(n)$ will be given by

$$\varepsilon(n) = \sum_{i=0}^{n} e(n)h(n-i) \quad , \tag{3.14}$$

where h(n) is defined by equation (3.7). The absolute maximum bound ($\hat{\varepsilon}(n)$) for the error $\varepsilon(n)$, considering the worst possible error sequence for e(n), for the case of rounding, is given by

$$\hat{\varepsilon}(n) = q \sum_{i=0}^{n-1} |h(i)| \quad . \tag{3.15}$$

Therefore, we can represent the phase plane trajectory of the finite wordlength digital filter as shown in Fig. 3.3. In this diagram the

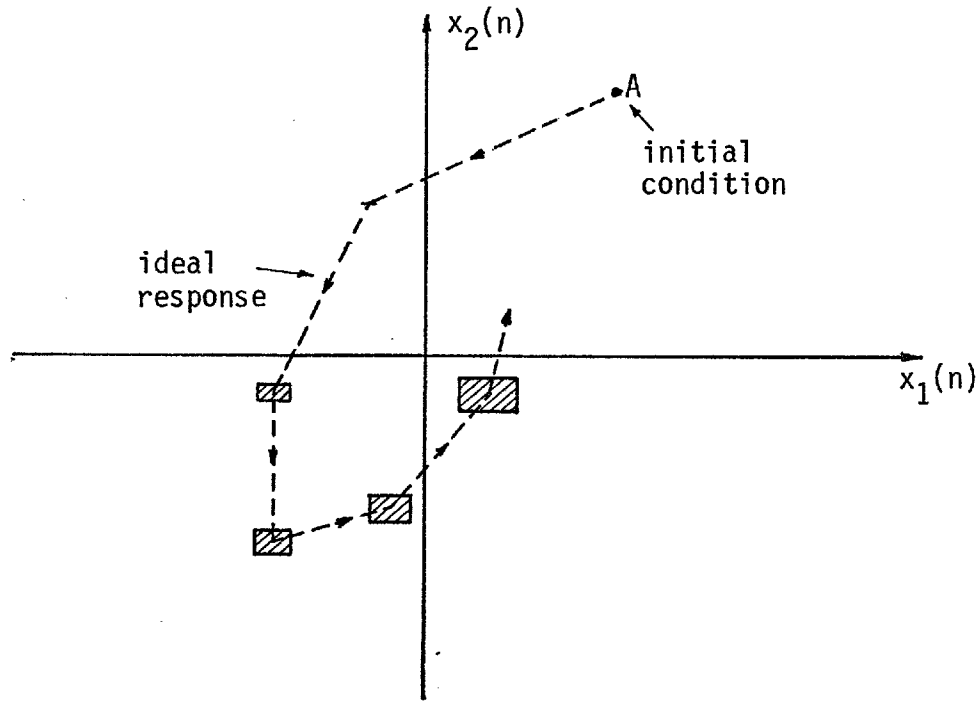


FIG. 3.3 Possible areas in the phase plane where the samples of a finite precision filter may exist

shaded areas represent the (worst case) bounds around the ideal trajectory where the response of the finite precision filter (with the initial Condition "A") can exist.

It follows that a trajectory of a finite wordlength digital filter would be similar to that of an infinite wordlength filter (equation (3.11)),

except that each sample has a fluctuating error, which grows in magnitude towards a limit (for a stable filter) as n increases. For n approaching infinity any error will be in the form of a cyclic oscillation, which is the limit cycle oscillation.

In the time domain a limit cycle is a sequence of numbers which are repeated periodically. In the phase plane a limit cycle, therefore, corresponds to a trajectory which is a closed loop. Thus, the phase plane trajectory of a stable basic second-order recursive digital filter with complex poles and finite wordlength is expected to be of an approximately spiralling elliptical shape. The trajectory is expected to terminate on a closed loop or at a single point. The closed loop corresponds to a limit cycle oscillation of period greater than one. The single point corresponds to a limit cycle with period of one sample, or to the zero steady state response. The zero steady state would be the termination of the trajectory at the origin of the phase plane.

This behaviour implies that there exists a point or points in the phase plane where two or more trajectories can merge to form a single trajectory. These points will be called "points of convergence" or simply "merge-points", M. Fig. 3.4 shows a typical phase plane behaviour of the transient response of a basic second-order recursive filter with finite wordlength arithmetic. The points immediately prior to convergence, i.e. the points represented by P in Fig. 3.4, will be called "P-points".

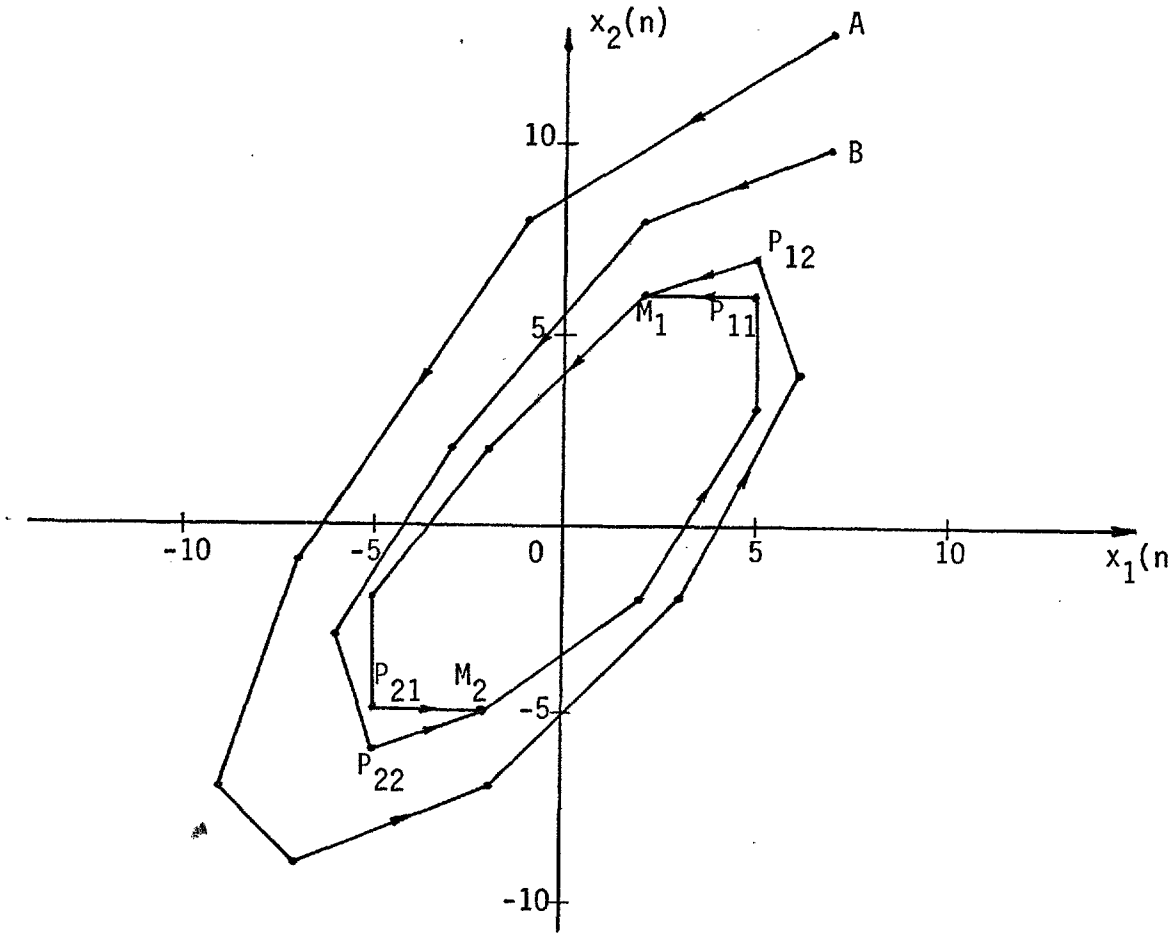


FIG. 3.4 A typical response of a finite wordlength filter in the phase plane

3.4 Conditions for the existence of merging trajectories

Let the point of convergence M for K trajectories be defined by the coordinates $[x_1(n), x_2(n)]$ in the phase plane, where K is an integer greater than zero. Define the point P immediately prior to point M (i.e. the P-point) by the coordinates $[x_1(n-1), x_2(n-1)]$. The coordinates of M and P must then satisfy the state equations of the finite wordlength digital filter as defined by equation (2.9). To have

K points such as P which can have a common point of convergence M, the state equations must be satisfied by K different values of $x_2(n-1)$, for a given $[x_1(n), x_2(n)]$. On the other hand, because we have a quantized system, the variables $x_1(n)$, $x_2(n)$, $x_1(n-1)$ and all possible $x_2(n-1)$ must be integer multiples of the quantization step q. Therefore, the conditions for these restrictions can be formulated as follows:

$$Q_q[\beta x] = Q_q[\beta(x + kq)] \quad , \tag{3.16}$$

where x represents a general $x_2(n-1)$, and k is an integer that can have K different values including zero.

The limitation on the possible values of k helps us find the conditions for the possibility of K trajectories merging at a single point, and the limitation on the values of x helps us find the boundaries in the phase plane where the points of convergence can exist.

To find the conditions for K trajectories merging at a single merge-point the boundaries for the values of k for different types of quantization processes will be analysed.

3.4.1 Rounding R[ ]

Consider, for example, the magnitude rounding as defined in Section 1.3 (similar arguments to the following could also be used for value rounding). The conditions for the values of k and x to satisfy equation (3.16) are given by

$$R_q[\beta x] - 0.5q < \beta(x + kq) < R_q[\beta x] + 0.5q \quad . \tag{3.17}$$

Equation (3.16) is also satisfied if the following equality conditions exist:

$$R_q[\beta x] - 0.5q = \beta(x + kq) \quad , \tag{3.18a}$$

for $R_q[\beta x] > 0$, and

$$R_q[\beta x] + 0.5q = \beta(x + kq) \quad , \tag{3.18b}$$

for $R_q[\beta x] < 0$.

Rearranging inequality (3.17) and equations (3.18a and b) we have

$$\frac{R_q[|\beta|x] - 0.5q}{|\beta|q} - \frac{x}{q} < k < \frac{R_q[|\beta|x] + 0.5q}{|\beta|q} - \frac{x}{q} \quad , \tag{3.19}$$

and

$$k = \frac{R_q[|\beta|x] - 0.5q\ \mathrm{sgn}(x)}{|\beta|q} - \frac{x}{q} \quad , \tag{3.20}$$

where equation (3.20) applies when $|R_q[\beta x]| > 0$.

The conditions given by inequality (3.19) and equation (3.20) can be expressed graphically as shown in Fig. 3.5. The shaded areas show the

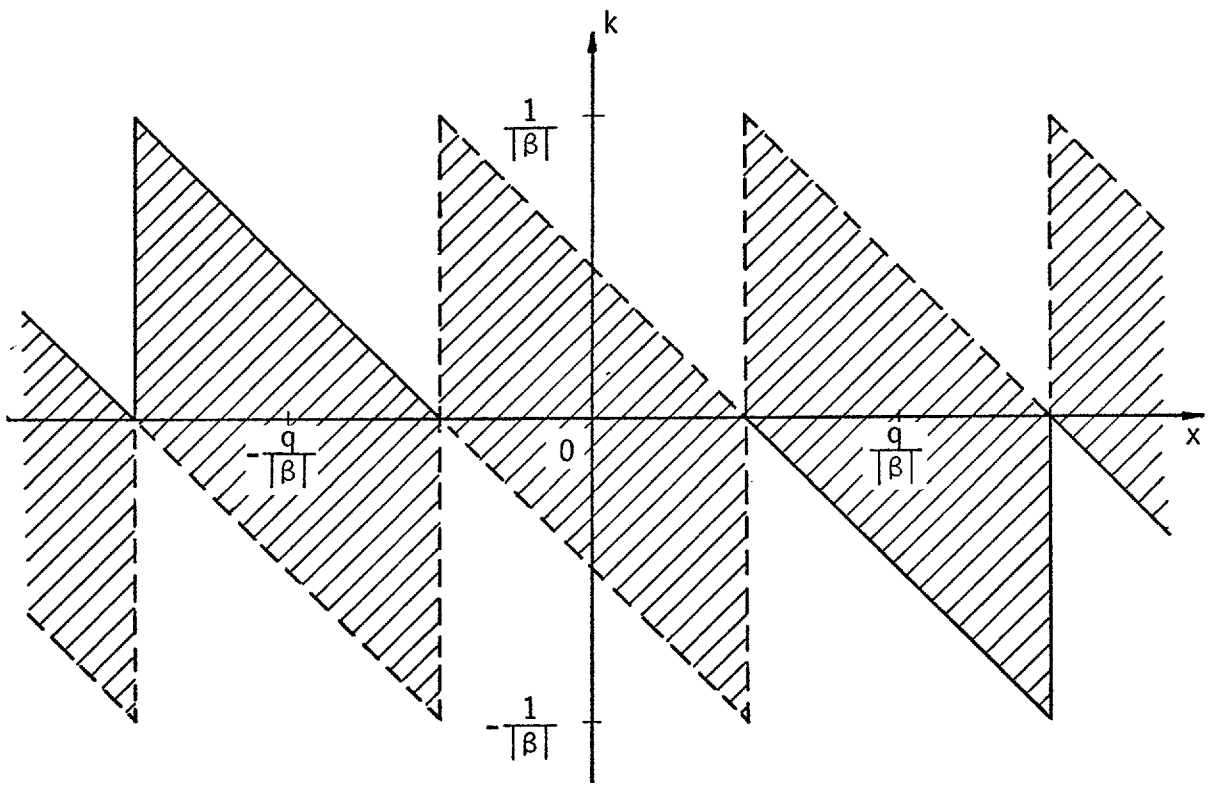


FIG. 3.5 Possible solutions to equation (3.16) for rounding

regions where k and x satisfy equation (3.16), and the solid line graphs express the condition of equation (3.20). It should be noted that k and

x are discrete variables, although for the convenience of drawing they are shown as continuous variables.

To have K possible solutions to equation (3.16) for a given x, there must exist K different integer values for k which lie within the shaded areas of Fig. 3.5. Therefore, the condition for the possibility of only K trajectories converging at a single merge-point is given by

$$\frac{1}{K} \leq |\beta| < \frac{1}{K-1} \quad . \tag{3.21}$$

The value of K is thus independent of the coefficient α. Hence, we can specify different regions in the coefficient plane where the values of K can be equal to 1, 2, 3, etc.

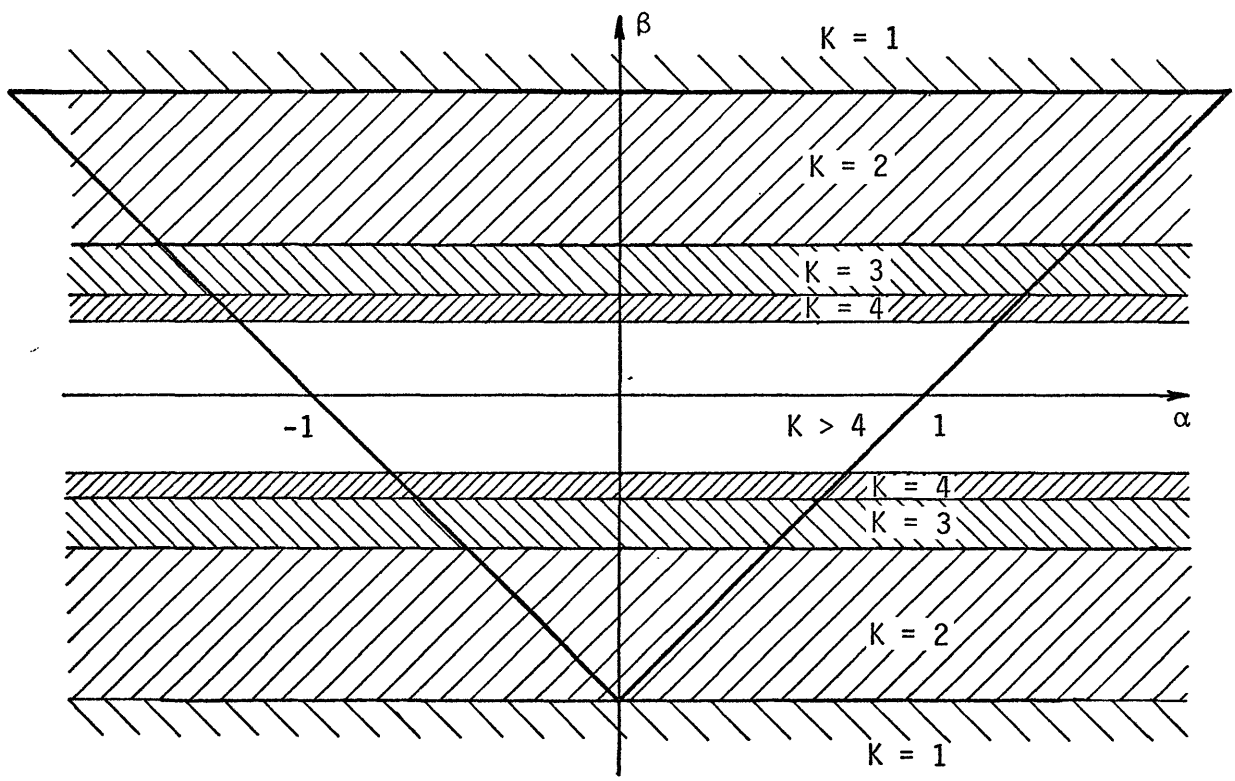


FIG. 3.6 Boundaries in the coefficient plane where maximum of K trajectories can merge

3.4.2 Value truncation T2[ ]

Using the properties of value truncation the conditions for the solutions to equation (3.16) can be expressed as

$$T2_q[\beta x] \leq \beta(x + kq) < T2_q[\beta x] + q \quad . \tag{3.22}$$

These conditions can be rearranged to

$$\frac{T2_1[|\beta|x]}{|\beta|q} - \frac{x}{q} \leq k < \frac{T2_1[|\beta|x] + q}{|\beta|q} - \frac{x}{q} \quad . \tag{3.23}$$

The inequality (3.23) represents the boundaries for the values of k which satisfy equation (3.16). These results are shown by the shaded areas in Fig. 3.7.

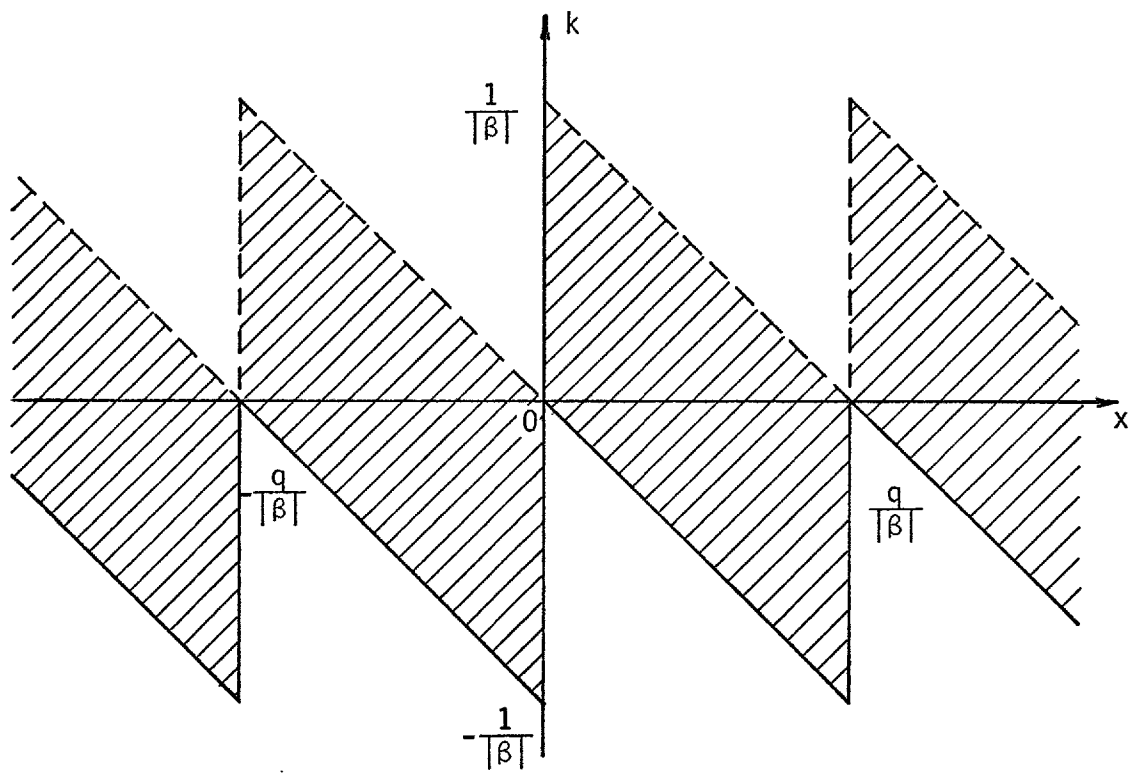


FIG. 3.7 Possible solutions to equation (3.16) for value truncation

Fig. 3.7 indicates that the condition for having a maximum of only K trajectories merging at a single point of convergence is similar to the case of rounding and, therefore, is given by the expression (3.21) or Fig. 3.6.

3.4.3 Magnitude truncation T1[ ]

Using the properties of magnitude truncation the conditions for the solutions to equation (3.16) can be expressed as:

$$T1_q[\beta x] \leqslant \beta(x+kq) < T1_q[\beta x] + q \quad , \tag{3.24a}$$

for $T1_q[\beta x] > 0$, and

$$T1_q[\beta x] - q < \beta(x+kq) \leqslant T1_q[\beta x] \quad , \tag{3.24b}$$

for $T1_q[\beta x] < 0$.

If $T1_q[\beta x] = 0$, then we have

$$-q < \beta(x+kq) < +q \quad . \tag{3.24c}$$

We can also express the above conditions in the following forms:

i) if $T1_q[|\beta|x] > 0$

$$\frac{T1_q[|\beta|x]}{|\beta|q} - \frac{x}{q} \leqslant k < \frac{T1_q[|\beta|x] + q}{|\beta|q} - \frac{x}{q} \quad , \tag{3.25a}$$

ii) if $T1_q[|\beta|x] < 0$

$$\frac{T1_q[|\beta|x] - q}{|\beta|q} - \frac{x}{q} < k \leqslant \frac{T1_q[|\beta|x]}{|\beta|q} - \frac{x}{q} \quad , \tag{3.25b}$$

and iii) if $T1_q[|\beta|x] = 0$

$$-\frac{1}{|\beta|} - \frac{x}{q} < k < \frac{1}{|\beta|} - \frac{x}{q} \quad . \tag{3.25c}$$

These inequalities specify the range of values of k which satisfy equation (3.16) for magnitude truncation. A graphical representation of these results is shown in Fig. 3.8.

A study of Fig. 3.8 or inequalities (3.25a - c) indicate that for magnitude truncation, if $|\beta| < 1$ and satisfies the relation (3.21), there can exist a maximum of $(2K+1)$ merging trajectories at a single merge-point. However, if $|\beta| \geqslant 1$, then no merging can take place.

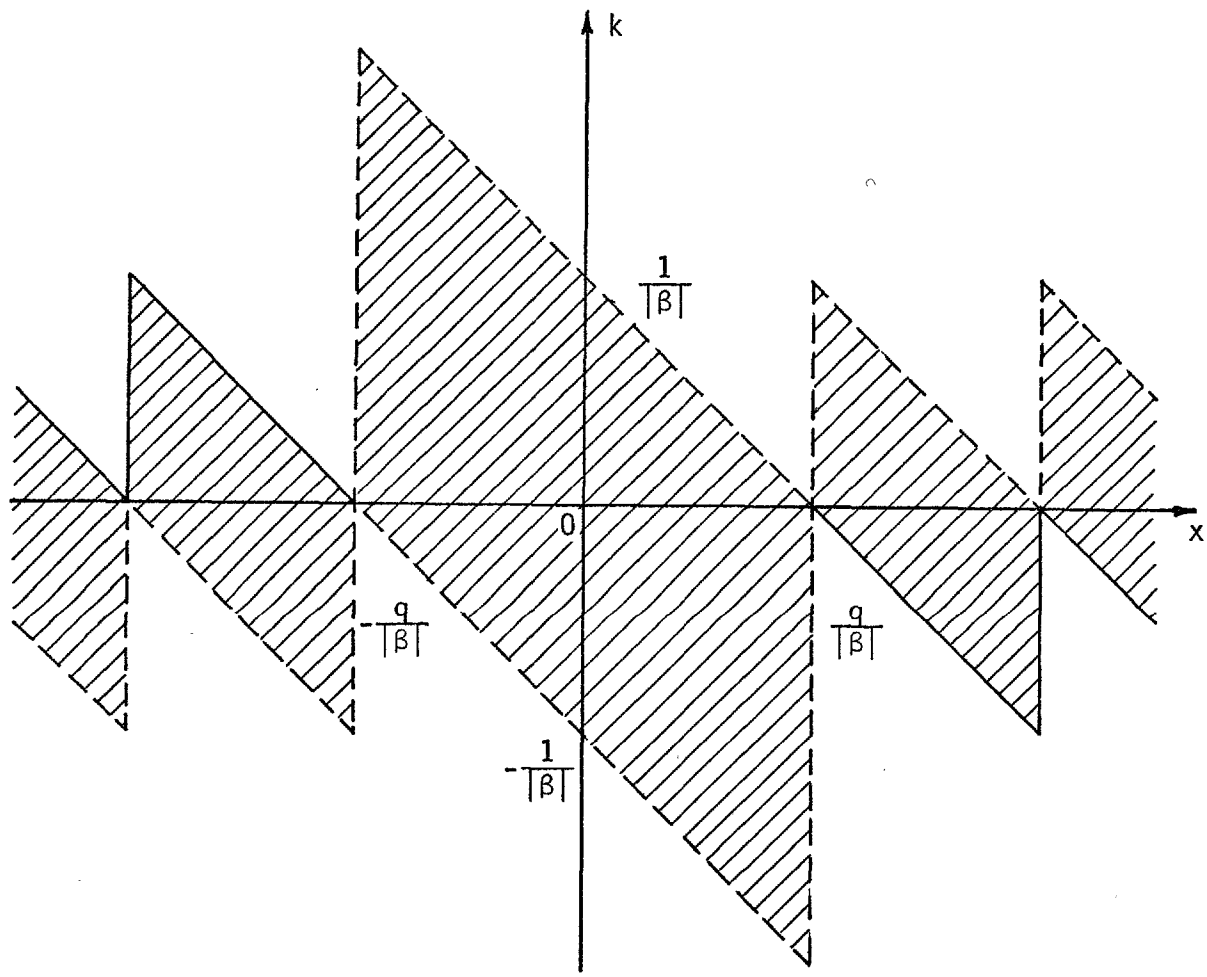


FIG. 3.8 Possible solutions to equation (3.16) for magnitude truncation

3.4.4 Generalised results

The above results for all the cases of quantization can be summarised in the following theorem.

Theorem 3.1:

For a basic second-order recursive digital filter with single precision finite wordlength arithmetic and coefficients of α and β using rounding or value truncation, if the coefficient β satisfies

$$\frac{1}{K} \leq |\beta| < \frac{1}{K-1} \quad ,$$

where K is a positive integer, there cannot exist any point in the phase

plane which is a merge-point for more than K trajectories, but it is always possible to find a merge-point for K or fewer trajectories. If the quantization process is magnitude truncation then the maximum number of merging trajectories at some merge-points can be up to (2K+1) except when $|\beta| \geq 1$, in which case it will be only one. These results are independent of the coefficient α.

The above theorem applies to all coefficient pairs, whether corresponding to a filter with a stable response or an unstable response. For example, when K = 1, i.e. the condition for no merging, $|\beta| \geq 1$. This corresponds to a digital oscillator or an unstable filter.

A particular characteristic of finite wordlength digital oscillators with complex poles, using any type of quantization process, may be observed at this point. For these "filters", $\beta = 1$ and $-2 < \alpha < 2$. Since no merging can take place under these conditions, then all initial conditions must be part of their subsequent oscillation trajectories, or the response will be unstable. But, the case of unstable response is not true for these coefficients. Therefore, for any finite wordlength digital filter with the coefficients of a digital oscillator having complex poles, all the initial conditions will be part of a cycle of oscillation.

3.5 Boundaries for the points of convergence when $K \geq 2$

In the previous section we studied the conditions for having the maximum of only K possible merging trajectories. However, we still have not considered where in the phase plane merging can occur. In this section we will consider this problem and derive the boundaries in the phase plane where the points of convergence can exist.

Firstly we will consider the case when the maximum number K of the merging trajectories at a single merge-point is 2. For rounding and value truncation the condition for this was shown (inequality (3.21)) to be

$$0.5 \leq |\beta| < 1 \quad . \tag{3.26}$$

Then we will show that, if $K > 2$ is possible, every point in the phase plane will be a merge-point for at least two trajectories.

To derive these results each type of quantization process will be considered separately.

3.5.1 Rounding

Consider the solutions of equation (3.16) as expressed by Fig. 3.5, and introduce the restriction that k can only have integer values (these values can be 0, 1 or 0, -1, for K = 2). These solutions are restricted to the values shown by the heavy lines in Fig. 3.9.

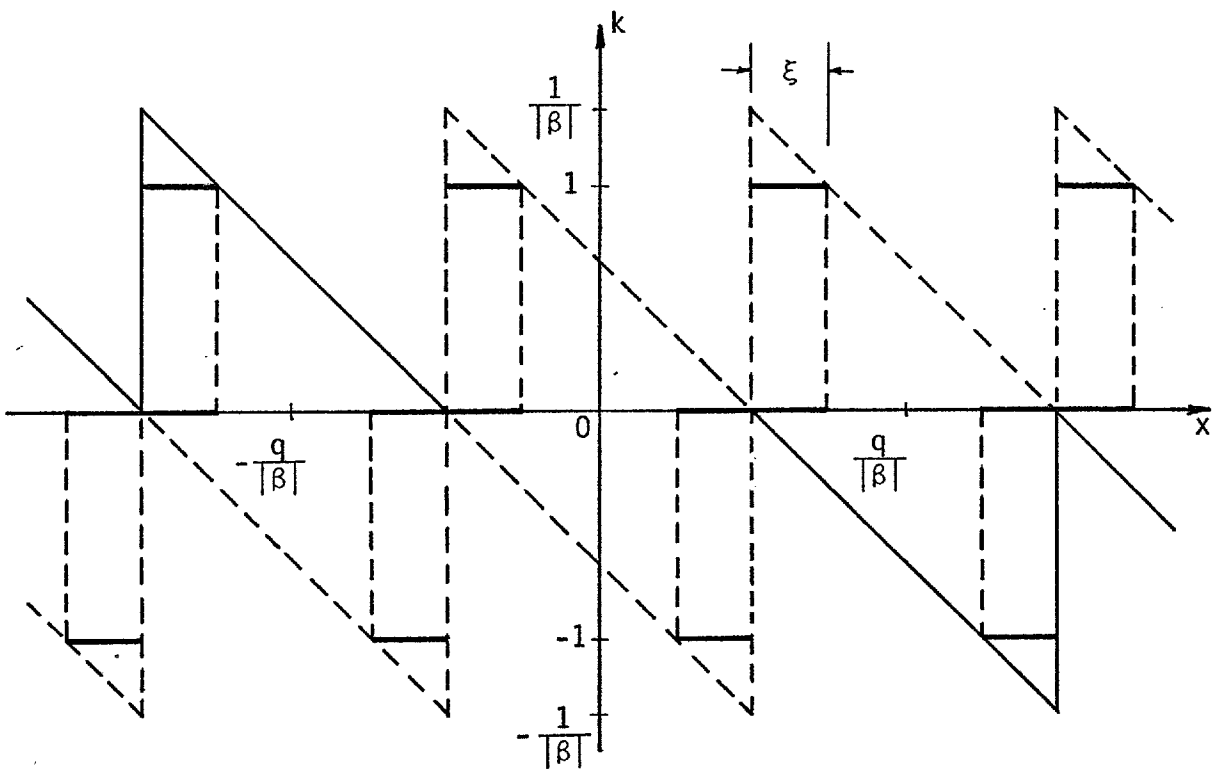


FIG. 3.9

By inspection of Fig. 3.9 we can write the following conditions for the values of x which satisfy equation (3.16).

$$\left(\frac{i+0.5}{|\beta|}\right)q - \xi < x < \left(\frac{i+0.5}{|\beta|}\right)q + \xi \quad , \tag{3.27a}$$

and for $i > 0$

$$x = \left(\frac{i+0.5}{|\beta|}\right)q - \xi \quad , \tag{3.27b}$$

or $i < 0$

$$x = \left(\frac{i+0.5}{|\beta|}\right)q + \xi \quad , \tag{3.27c}$$

where i is any integer, and ξ is the quantity shown in Fig. 3.9, i.e.

$$\xi = \left(\frac{1}{|\beta|} - 1\right)q \quad . \tag{3.28}$$

x is a quantized variable which can be expressed by $x = mq$, where m is an integer. The inequality (3.27a) can be rewritten as

$$\frac{i - 0.5 + |\beta|}{|\beta|} < m < \frac{i + 1.5 - |\beta|}{|\beta|} \quad . \tag{3.29}$$

The solution given by inequality (3.29) is not satisfied for all the values of i if m is to be an integer. For example, if $\beta = 0.9$ and $i = 1$, the inequality (3.29) reduces to

$$1.67 < m < 1.78 \quad ,$$

which is meaningless if m has to be an integer. Therefore, we must transform inequality (3.29) to an expression which is correct for all the possible values of its index variable. For this purpose we define $j = m - i$. Substituting for $i = m - j$ in inequality (3.29), we obtain

$$\frac{j - 1.5 + |\beta|}{1 - |\beta|} < m < \frac{j + 0.5 - |\beta|}{1 - |\beta|} \quad . \tag{3.30}$$

We now have a solution for m for all the values of the integer variable j, since the difference between the upper and the lower limits of inequality (3.30) is 2.

The values of x which can satisfy equation (3.16), therefore, satisfy

$$\left(\frac{j - 1.5 + |\beta|}{1 - |\beta|}\right)q < x < \left(\frac{j + 0.5 - |\beta|}{1 - |\beta|}\right)q \quad , \tag{3.31a}$$

where j is any integer.

Equations (3.27b and c) can also be expressed in terms of the index variable j. The result for this substitution is as follows:

i) for $j > 0$

$$x = \left(\frac{j + 0.5 - |\beta|}{1 - |\beta|}\right)q \quad , \tag{3.31b}$$

and ii) for $j \leq 0$

$$x = \left(\frac{j - 1.5 + |\beta|}{1 - |\beta|}\right)q \quad . \tag{3.31c}$$

We note:

i) for $i > 0$, $m > i$ and therefore $j > 0$, and

ii) for $i < 0$, $m \leq i$ and therefore $j \leq 0$.

The conditions (3.31a, b and c) represent the boundaries for $x_2(n)$ in the phase plane which include all the P-points.

In the above discussion there has been no restriction on the $x_1(n)$ coordinates of a P-point. Therefore, we can represent the regions in the phase plane, where the P-points can exist, as shown in Fig. 3.10. As before, the equality conditions are represented by the solid lines.

By considering conditions (3.31a, b and c) and the difference equation (2.9), we can also derive the phase plane boundaries for the points of convergence M. This result for the case of (magnitude) rounding is given by

$$-0.5q < x_1(n) + \alpha x_2(n) + \eta_j < 0.5q \quad , \tag{3.32a}$$

and

$$x_1(n) + \alpha x_2(n) + \eta_j = 0.5q \, \mathrm{sgn}(\alpha x_2(n)) \quad , \tag{3.32b}$$

where

$$\eta_j = R_1\left[|\beta|\left(\frac{j - 0.5}{1 - |\beta|}\right)\right]q \tag{3.33}$$

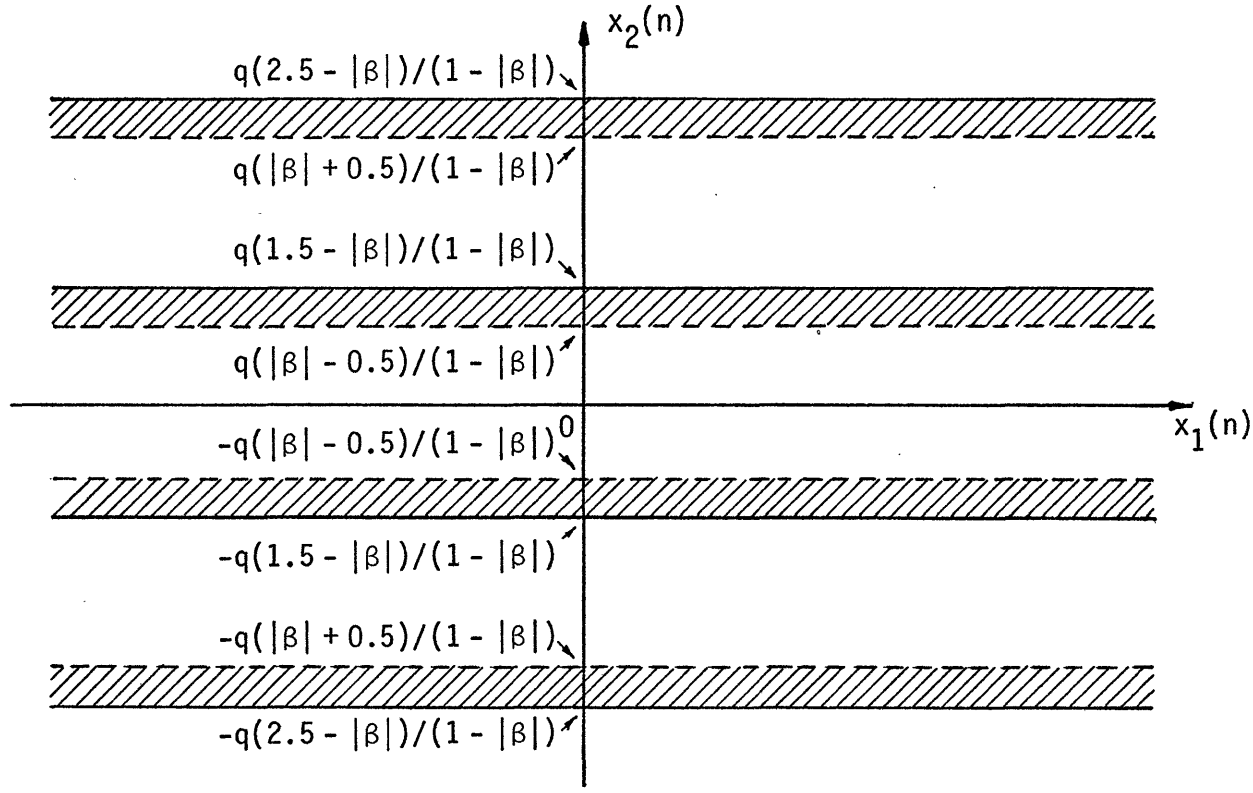


FIG. 3.10 Regions in the phase plane for p-points when K = 2 (case of rounding)

This boundary, for $\alpha < 0$, is shown in Fig. 3.11.

The boundaries for the P-points have a simpler definition from the boundaries for the merge-points. Therefore, for convenience, the boundaries for the P-points will be considered in the rest of this chapter.

From the above analysis it follows that, if $|\beta| < 0.5$, then every point in the phase plane is a merge-point for at least two trajectories. For the proof of this, consider Fig. 3.9, where if $\xi > \frac{0.5q}{|\beta|}$, then k can have at least two different integer values for any value of x. Thus, for any x (i.e. $x_2(n-1)$), there exists at least two solutions to

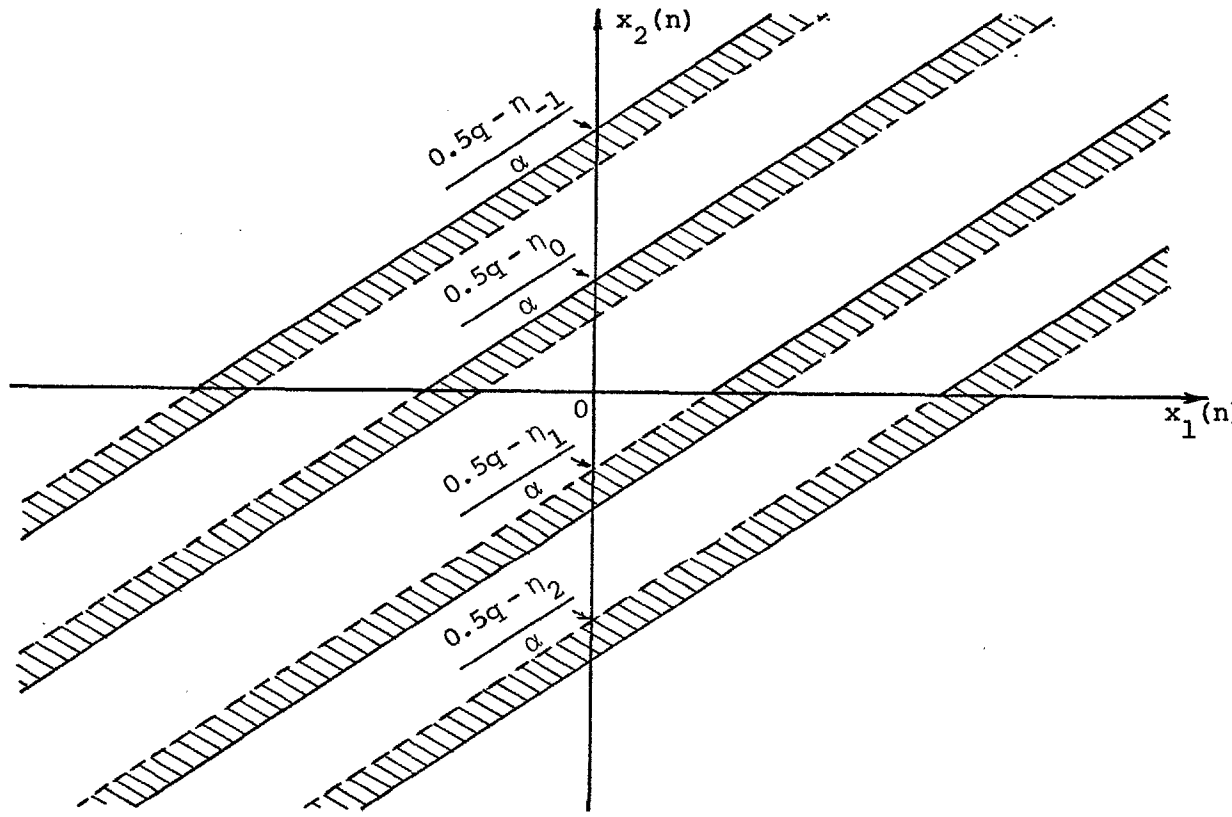


FIG. 3.11 Regions in the phase plane for points of convergence when K = 2 (case of rounding)

equation (3.16). This condition for the value of ξ is equivalent to $|\beta| < 0.5$. Therefore, for this case every point in the phase plane is a P-point and thus a merge-point.

3.5.2 Value truncation

The solutions of equation (3.16) for value truncation can be found in a similar way to above. Considering the conditions when K = 2, i.e. $0.5 \leq |\beta| < 1$, we can express these solutions by the heavy lines shown in Fig. 3.12.

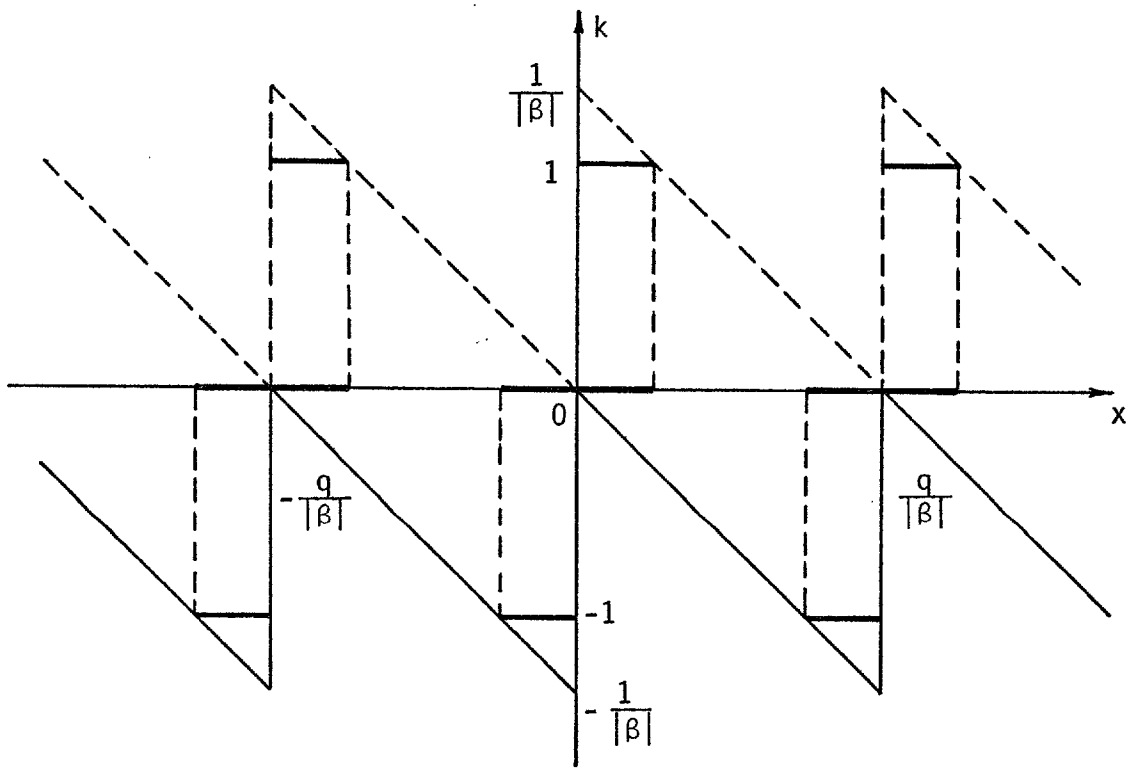


FIG. 3.12

The quantized values of x which satisfy equation (3.16) are therefore given by the limits

$$\left(\frac{j-1+|\beta|}{1-|\beta|}\right)q < x \leq \left(\frac{j+1-|\beta|}{1-|\beta|}\right)q \quad , \tag{3.34}$$

where j is an integer.

The phase plane regions where the P-points can exist for the case of value truncation are as shown in Fig. 3.13. As for rounding, we note that if $|\beta| < 0.5$, every point in the phase plane is a merge-point for at least two trajectories.

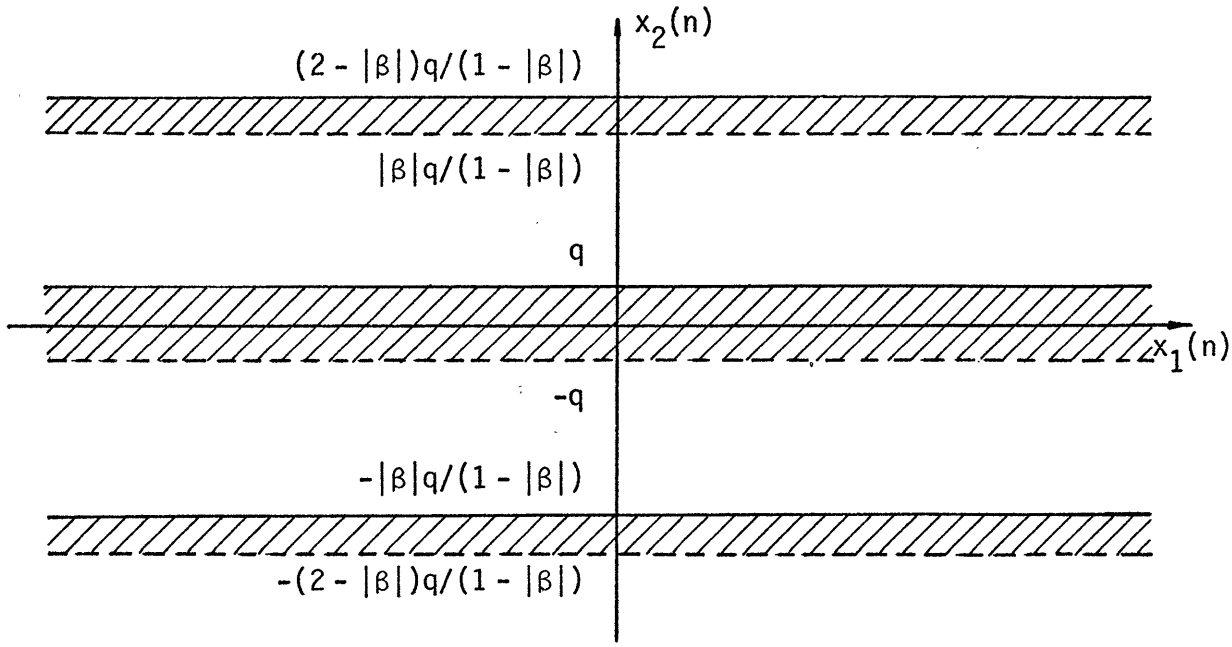


FIG. 3.13 Regions in the phase plane for p-points when K = 2 (case of value truncation)

3.5.3 Magnitude truncation

The preceding method of analysis can also be applied to the case of magnitude truncation. Assuming $0.5 \leq |\beta| < 1$, the solutions to equation (3.16) for having at least two integer values for k are indicated by the heavy lines in Fig. 3.14. From these results we note that the quantized values of x must satisfy

$$\left(\frac{j - 1 + |\beta|}{1 - |\beta|}\right)q < x < \left(\frac{j + 1 - |\beta|}{1 - |\beta|}\right)q \quad , \tag{3.35a}$$

and, for $j \geq 0$

$$x = \left(\frac{j + 1 - |\beta|}{1 - |\beta|}\right)q \quad , \tag{3.35b}$$

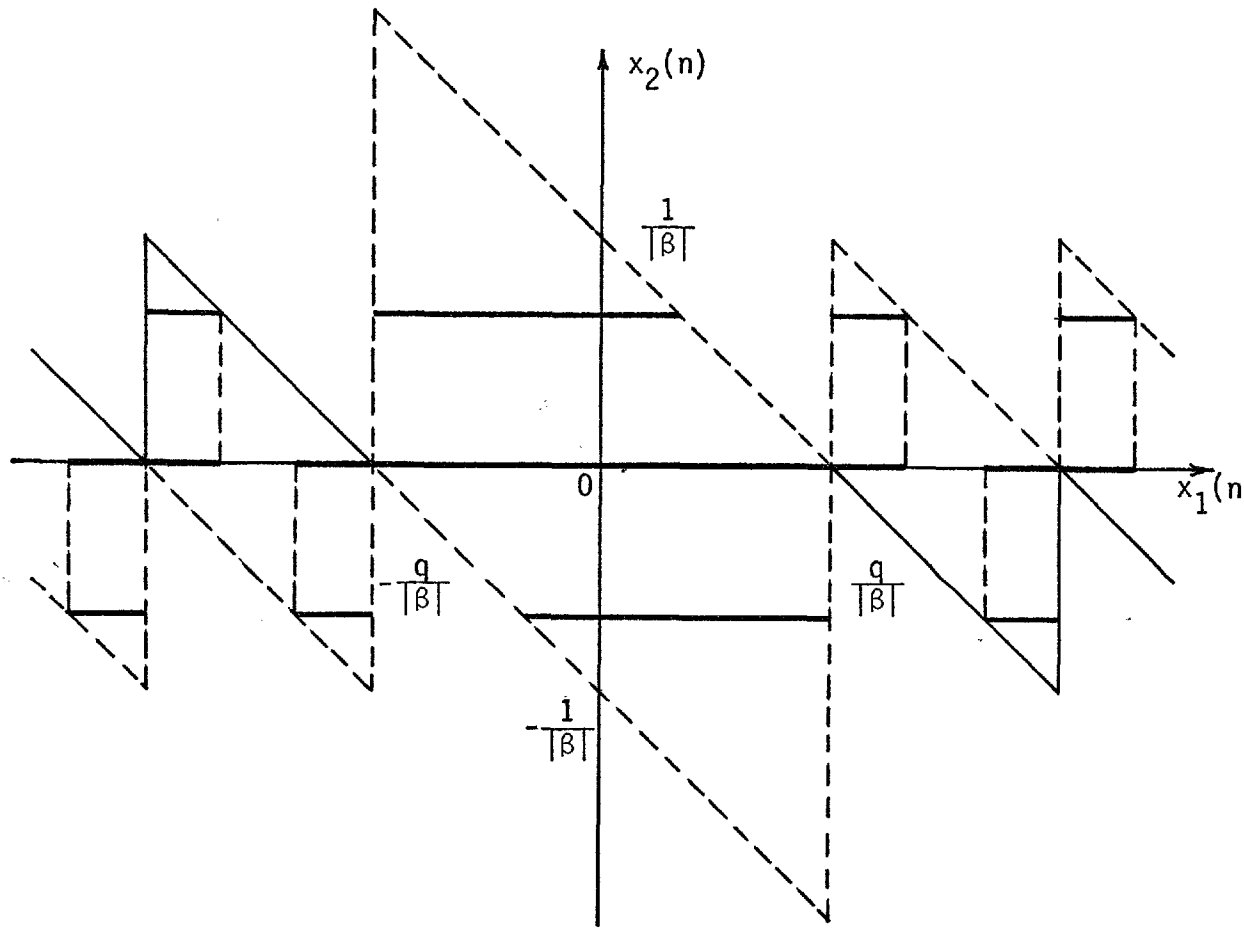


FIG. 3.14

or for $j \leq 0$

$$x = \left(\frac{j - 1 + |\beta|}{1 - |\beta|}\right) q, \tag{3.35c}$$

where, as before, j is an integer.

The regions in the phase plane where the P-points can exist are, therefore, the shaded areas of Fig. 3.15. We note that for $\beta < 0.5$, in the case of magnitude truncation, every point in the phase plane is a merge-point.

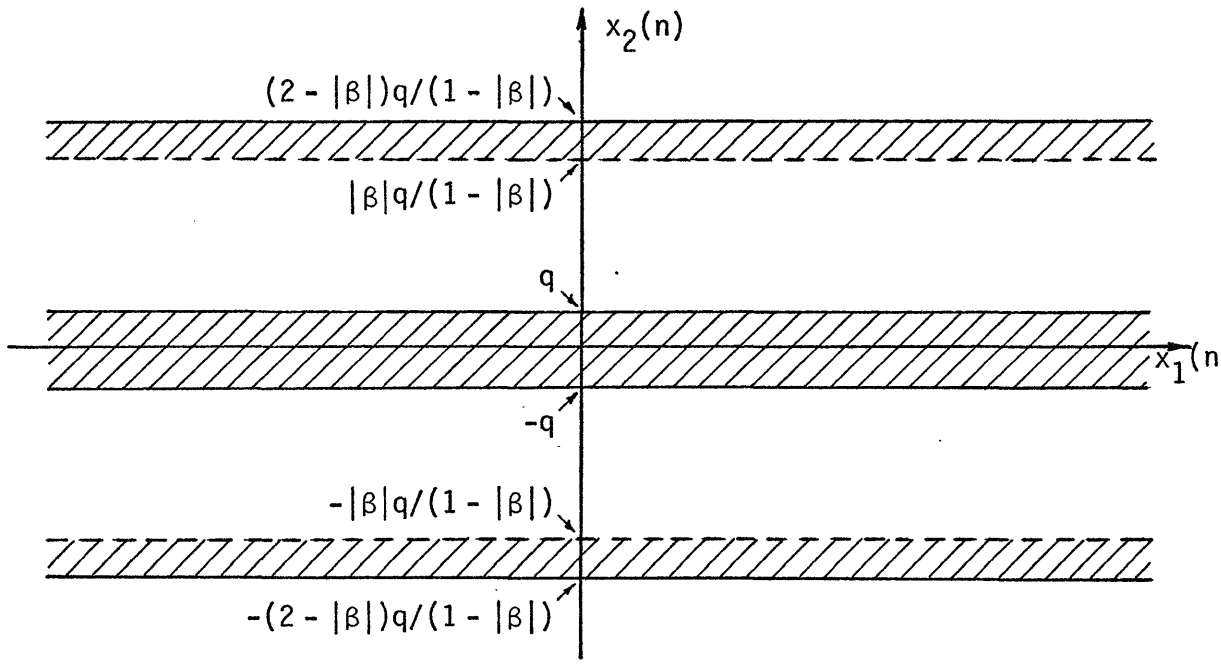


FIG. 3.15 Regions in the phase plane for p-points when K = 2 (case of magnitude truncation)

3.6 Characterisation of limit cycle oscillations

From the simulation studies of zero input limit cycle oscillations it has been observed that a digital filter can accommodate a number of different limit cycle oscillations. The particular limit cycle oscillation which occurs normally depends on the initial conditions. In this section the results on the behaviour of the transient response of the basic second-order section in the phase plane will be used to characterise and develop properties of limit cycle oscillations.

It was pointed out (in Section 3.3) that for a limit cycle to be the steady state response of an initial condition that is not a sample of that limit cycle, the trajectory of the transient response in the phase plane must go through a process called merging. The regions in the phase plane where this process can take place were investigated in the previous

section.

Examining the results obtained (for example Fig. 3.11), we can find trajectories of limit cycle oscillations in the phase plane which do not have any samples in the regions where merge-points can exist. Such limit cycle oscillations have the special property that, they cannot occur unless the initial condition is one of the samples of its trajectory. We can, therefore, categorise limit cycle oscillations into two basic types called "accessible limit cycles" and "inaccessible limit cycles". The definitions of these terms are the following:

i) an "accessible limit cycle oscillation" is defined as the steady state response of a transient from an initial condition which is not a sample of that limit cycle;

ii) an "inaccessible limit cycle oscillation" is defined as the steady state response which can only occur when the initial condition is a sample of the trajectory of that limit cycle.

These terms for the limit cycle oscillations were first used by Claasen et al. [118]. However, all of this chapter is based on works which were fully independent of their work.

From the above discussion it follows that, an inaccessible limit cycle oscillation cannot have a merge-point as one of its samples, whereas an accessible limit cycle oscillation must have at least one merge-point on its trajectory. In the phase plane, however, a point of convergence is the subsequent sample to the P-points. Therefore the trajectory of an accessible limit cycle in the phase plane has to have at least one sample in the regions where the P-points exist. Thus, any limit cycle which is not of large enough amplitude to reach the boundaries for the P-points will be inaccessible.

Because inaccessible limit cycles are initiated by a small and limited number of initial conditions, these modes of oscillation,

therefore, rarely take place in practice. On the other hand, the accessible limit cycles are much more common. This characterisation of limit cycle oscillations has, therefore, enabled us to comment on the probability of occurrence of two categories of limit cycle oscillations.

From the results in the previous section we can specify some properties of the limit cycles in the basic second-order filter .

Theorem 3.2

For a basic second-order recursive filter with $|\beta| < 0.5$, all limit cycle oscillations, as well as the zero steady state, are always accessible.

The proof of this theorem follows from the result that if $|\beta| < 0.5$, then all points in the phase plane are merge-points.

3.6.1 Rounding

From the results presented by Fig. 3.10 or conditions (3.31a, b and c) we note that the amplitude of the smallest accessible limit cycle oscillation can be expressed by

$$A_{min} = R_1\left[\frac{|\beta| - 0.5}{1 - |\beta|} + 0.5\right]q \quad ,$$

i.e.

$$A_{min} = R_1\left[|\beta|\frac{0.5}{1 - |\beta|}\right]q \quad . \tag{3.36}$$

The following theorem, therefore, follows from this discussion.

Theorem 3.3

For the basic second-order digital filter, with coefficients α and β which correspond to a stable response, with $|\beta| \geq 0.5$, and using rounding, the transient response of the filter for all initial conditions which are not part of a subsequent steady state response will result in a limit cycle oscillation (an accessible limit cycle) with

amplitude A which is given by

$$A \geq A_{min} = R_1 \left| |\beta| \frac{0.5}{1 - |\beta|} \right| q \quad .$$

Any limit cycle with amplitude less than A_{min} (including the zero steady state response) will be inaccessible and therefore rarely occurs.

3.6.2 Value or magnitude truncation

The results of Figures 3.13 and 3.15 indicate that the theorem 3.3 does not apply when either value or magnitude truncation is used. In fact, there would probably be a large number of initial conditions in the phase plane which can result in a zero steady state response for any stable transfer function.

3.7 Conclusions

In this chapter the properties of the trajectories of the phase plane response of finite wordlength basic second-order recursive digital filters were considered. It was shown that for these filters, with coefficients corresponding to a stable response, K trajectories (where $K \geq 2$) can merge at a single point which was referred to as the merge-point. Results giving the regions in the phase plane where merging can take place have been analysed. Figures 3.10, 3.13 and 3.15 show the boundaries for the points immediately prior to merge-points (i.e. P-points).

The concept of merging trajectories was applied to the trajectories of limit cycle oscillations to show that these oscillations can be categorised as either accessible or inaccessible.

Some properties of these modes of oscillation were considered. In particular, results show that for the case of rounding, if the

coefficient β satisfies,

$$0.5 \leq |\beta| < 1 \quad ,$$

then there is a minimum amplitude for all accessible limit cycle oscillations, and therefore, the zero steady state will be inaccessible (theorem 3.3). For value or magnitude truncation the zero steady state was found to be always accessible for stable filters.

CHAPTER 4

A TRANSFORMATION TECHNIQUE FOR THE STUDY OF ZERO INPUT LIMIT CYCLE OSCILLATIONS

In this chapter a new technique is introduced for modifying the structure of a digital filter to facilitate the prediction of the mechanism by which some of the limit cycles occur. This information is then used in Chapter 5 to synthesize a suitable structure to enable suppression of limit cycle oscillations by the method of "signal stabilization" [138-140].

This technique is based on a property of the rounding or the value truncation functions, and it does not necessarily apply to the magnitude truncation function. Therefore, the studies presented in this chapter and Chapter 5 are only useful for these two quantization methods.

For reasons described in Chapter 2, the digital filter considered is the all-pole second-order recursive filter with finite single precision fixed-point arithmetic. Both of the realizations for these types of filter are studied.

4.1 A new representation of a second-order filter

Consider a general multiplier branch of a finite precision digital filter, consisting of a multiplier with coefficient γ, followed by a quantizer $Q_q[\ ]$ of quantization step q, as shown in Fig. 4.1a. Let the input to the multiplier be x, where x is a quantized variable of quantization step q. Then the output of that branch will be $Q_q[\gamma x]$.

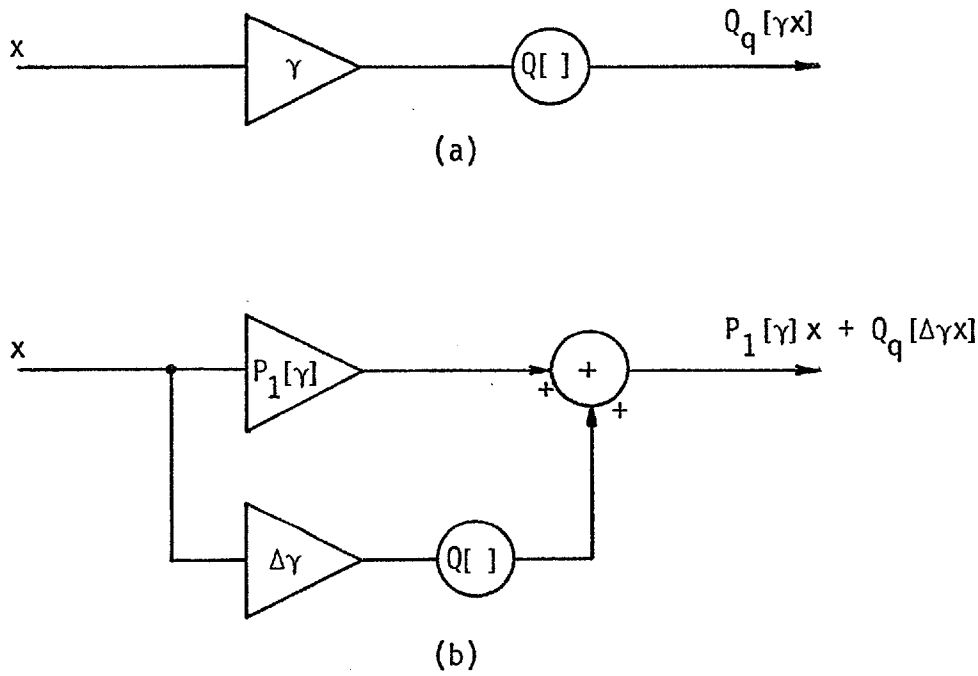


FIG. 4.1

The coefficient γ can be expressed in two parts,

i.e. $$\gamma = P_1[\gamma] + \Delta\gamma \quad , \tag{4.1}$$

where $P_1[\ ]$ is any quantization function with the quantization step of one. The output of the branch considered can now be written as:

$$Q_q[\gamma x] = Q_q[P_1[\gamma]\,x + \Delta\gamma x] \quad . \tag{4.2}$$

If $Q_q[\ ]$ represents rounding or value truncation, the properties of these quantization processes (see Section 1.4) can be used to express equation (4.2) in the following form:

$$Q_q[\gamma x] = P_1[\gamma]\,x + Q_q[\Delta\gamma x] \quad . \tag{4.3}$$

Equation (4.3) is a transformation that can be applied to any multiplier branch of a finite precision digital filter with single

precision arithmetic using rounding or value truncation. In this transformation, a coefficient of the filter is transformed into an integer part and a non-integer part, as shown in Fig. 4.1b.

The quantization process $P_1[\ ]$ can be any of, rounding R[ ], value truncation T2[ ] or magnitude truncation T1[ ]. For convenience in the analysis it is found desirable to choose the rounding process for $P_1[\ ]$; however, this choice does not limit the generality of the results.

Applying this transformation to each branch of a digital filter, a new structure (or a model) consisting of two parts can be formed. One part is a filter with integer coefficients, and therefore no quantization effects. This part will be called the "sub-filter". The other part provides a non-linear feedback to the sub-filter and will be called the "feedback section".

Each of the realizations for a second-order recursive filter will be considered separately, and modified as suggested above.

4.1.1 Type-A realization

The second-order filter of type-A realization, shown in Fig. 4.2, has its difference equations given by:

$$x_1(n) = -Q_q[\alpha x_1(n-1)] - Q_q[\beta x_2(n-1)] + u(n) \quad , \tag{4.4a}$$

and

$$x_2(n) = x_1(n-1) \quad . \tag{4.4b}$$

Let

$$\alpha = R_1[\alpha] + \Delta\alpha \quad , \tag{4.5a}$$

and

$$\beta = R_1[\beta] + \Delta\beta \quad . \tag{4.5b}$$

Using the transformation described by equation (4.3), the state equations

(4.4a and b) can be written as:

$$x_1(n) = -R_1[\alpha]\,x_1(n-1) - R_1[\beta]\,x_2(n-1) + v(n) \quad , \tag{4.6a}$$

and

$$x_2(n) = x_1(n-1) \quad , \tag{4.6b}$$

where

$$v(n) = -Q_q[\Delta\alpha x_1(n-1)] - Q_q[\Delta\beta x_2(n-1)] + u(n) \quad . \tag{4.7}$$

The structure of the digital filter described by equations (4.6a and b) and (4.7) is shown in Fig. 4.3. In this transformation we note that the filter is modelled as a sub-filter with the transfer function.

$$H_{0A}(z) = \frac{z^2}{z^2 + R_1[\alpha]\,z + R_1[\beta]} \quad , \tag{4.8}$$

and a non-linear feedback defined by equation (4.7).

4.1.2 Type-B realization

The second-order filter of type-B realization, as shown in Fig. 4.4, has its difference equations given by:

$$x_1(n) = -Q_q[ax_1(n-1)] - Q_q[bx_2(n-1)] + u(n) \quad , \tag{4.9a}$$

and

$$x_2(n) = x_1(n-1) + x_2(n-1) \quad . \tag{4.9b}$$

Let

$$a = R_1[a] + \Delta a \quad , \tag{4.10a}$$

and

$$b = R_1[b] + \Delta b \quad . \tag{4.10b}$$

As for the type-A realization, we can write the state equations (4.9a and b) in the following forms:

$$x_1(n) = -R_1[a]\,x_1(n-1) - R_1[b]\,x_2(n-1) + v(n) \quad , \tag{4.11a}$$

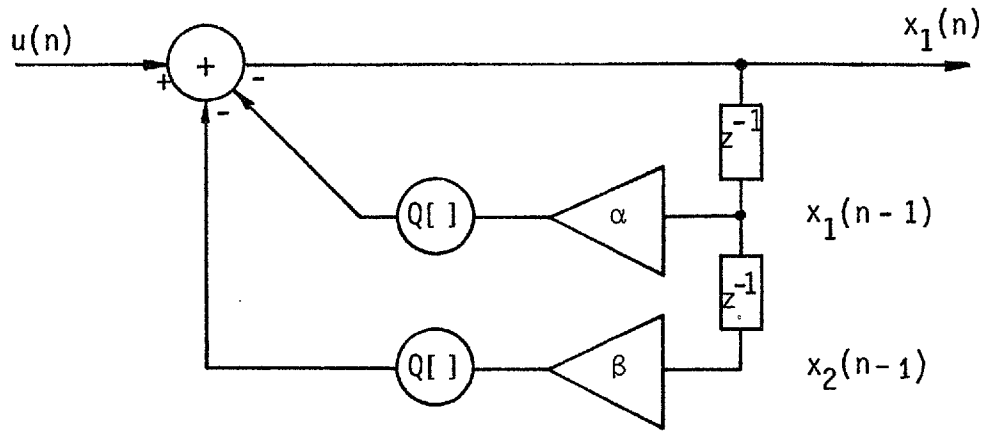


FIG. 4.2 A second-order single precision structure of type-A realization

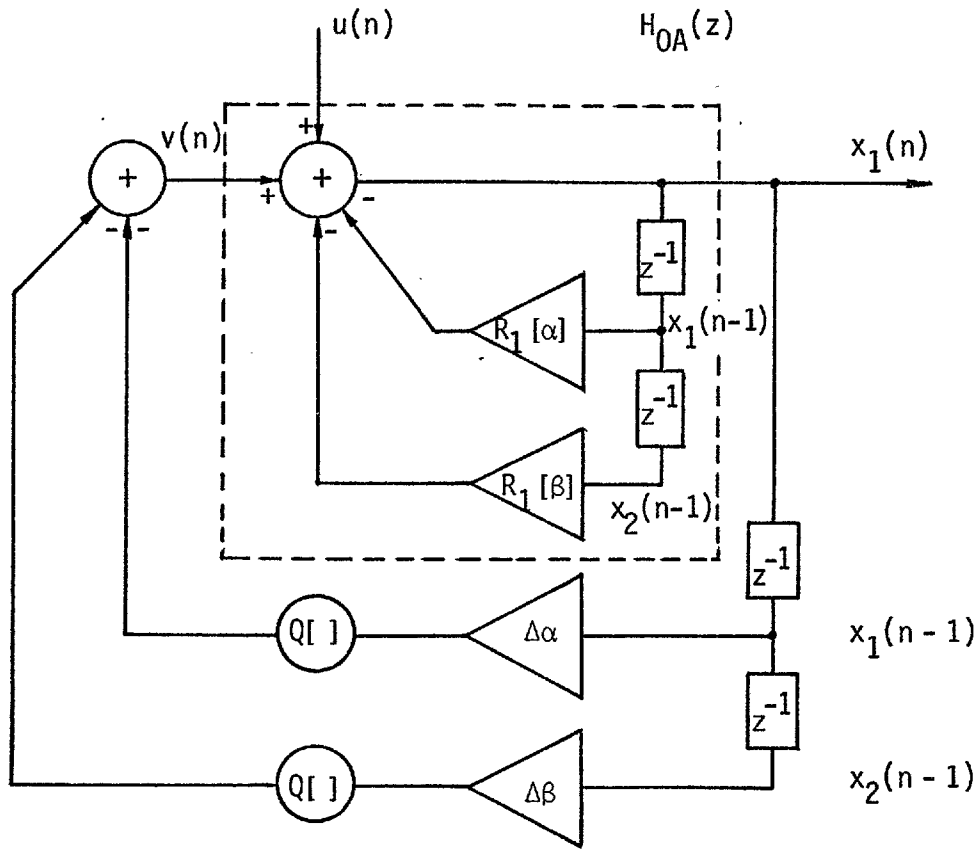


FIG. 4.3 A modified structure similar to Fig. 4.2

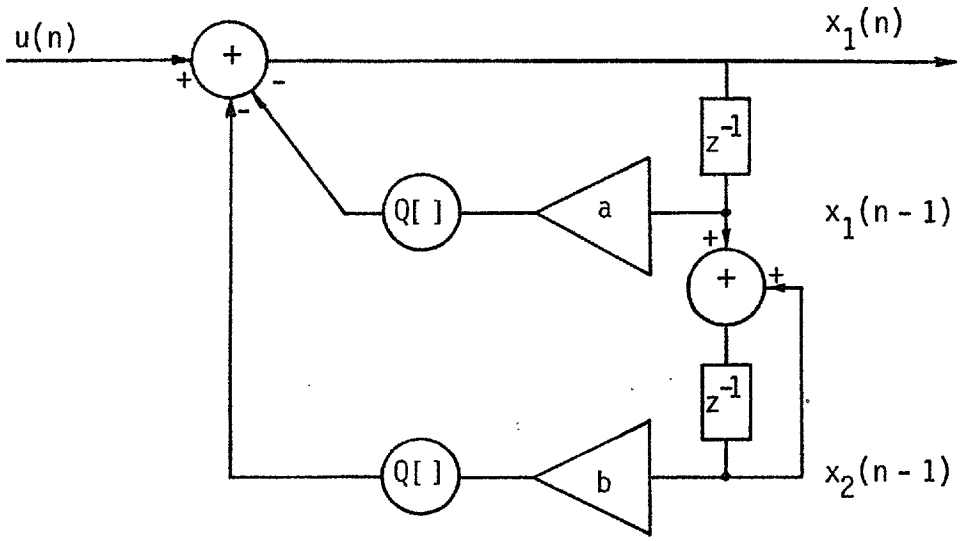


FIG. 4.4 A second-order single precision structure of type-B realization

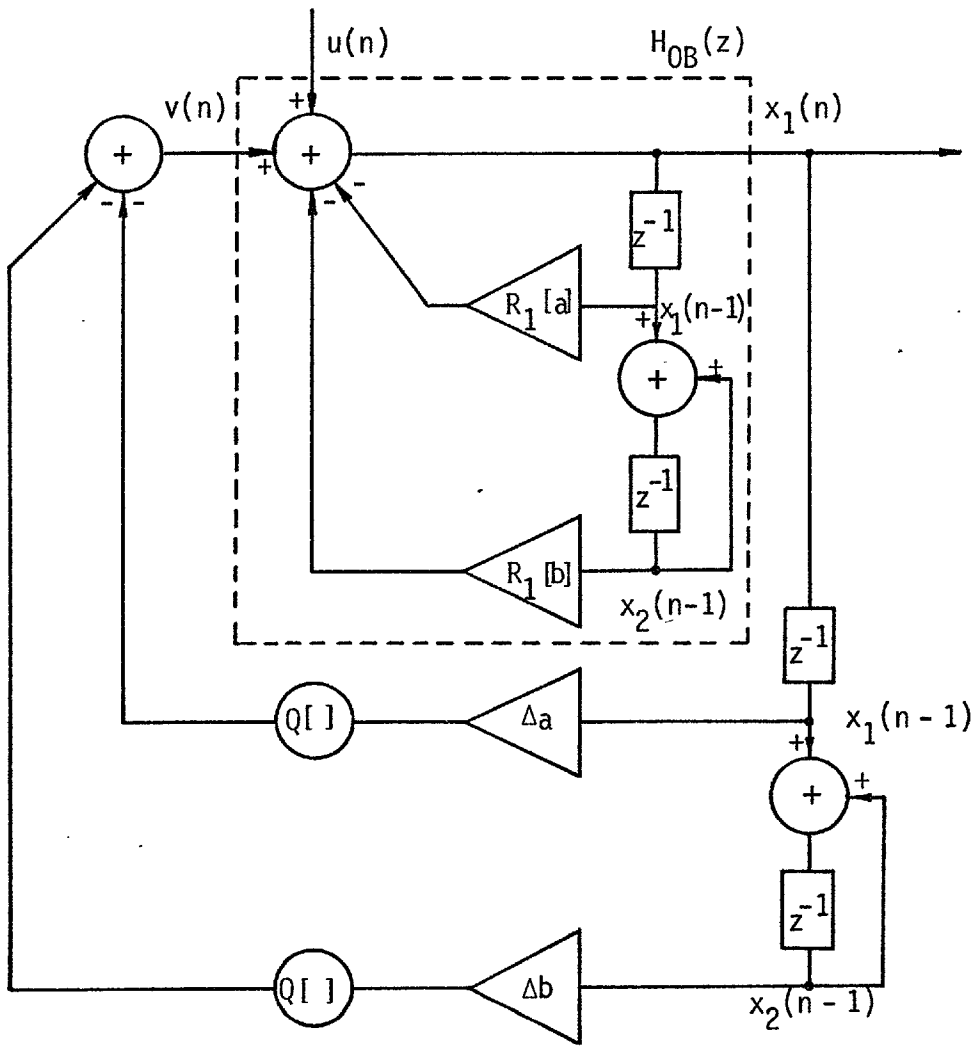


FIG. 4.5 A modified structure similar to Fig. 4.4

and

$$x_2(n) = x_1(n-1) + x_2(n-1) \quad , \tag{4.11b}$$

where

$$v(n) = -Q_q[\Delta a x_1(n-1)] - Q_q[\Delta b x_2(n-1)] + u(n) \quad . \tag{4.12}$$

The structure of the digital filter described by equations (4.11a and b) and (4.12) is shown in Fig. 4.5. The sub-filter for this realization has the transfer function given by

$$H_{OB}(z) = \frac{z(z-1)}{z^2 + (R_1[a] - 1)z + (R_1[b] - R_1[a])} \quad , \tag{4.13}$$

and the feedback section is defined by equation (4.12).

4.2 Studies of the model

The coefficients of the sub-filter, as described above, are of integer values, which means that the sub-filter is free from quantization errors and has a limited number of possible coefficients. Figures 4.6 and 4.7 show the positions for all the possible coefficient pairs of the sub-filter in the coefficient plane for the type-A and type-B realizations respectively.

In Figures 4.6 and 4.7 the boundaries for stability are indicated by the triangle ABC, and the corresponding impulse response of the state variable $x_1(n)$ for each coefficient pair of the sub-filter are also shown. We note that, except for the cases when

$$R_1[\alpha] = R_1[\beta] = 0 \quad ,$$

and

$$R_1[a] = R_1[b] = 1 \quad ,$$

the sub-filter has the transfer function of a digital oscillator or an unstable filter. It follows that the feedback section must have a stabilizing effect on the sub-filter. This indicates that a condition

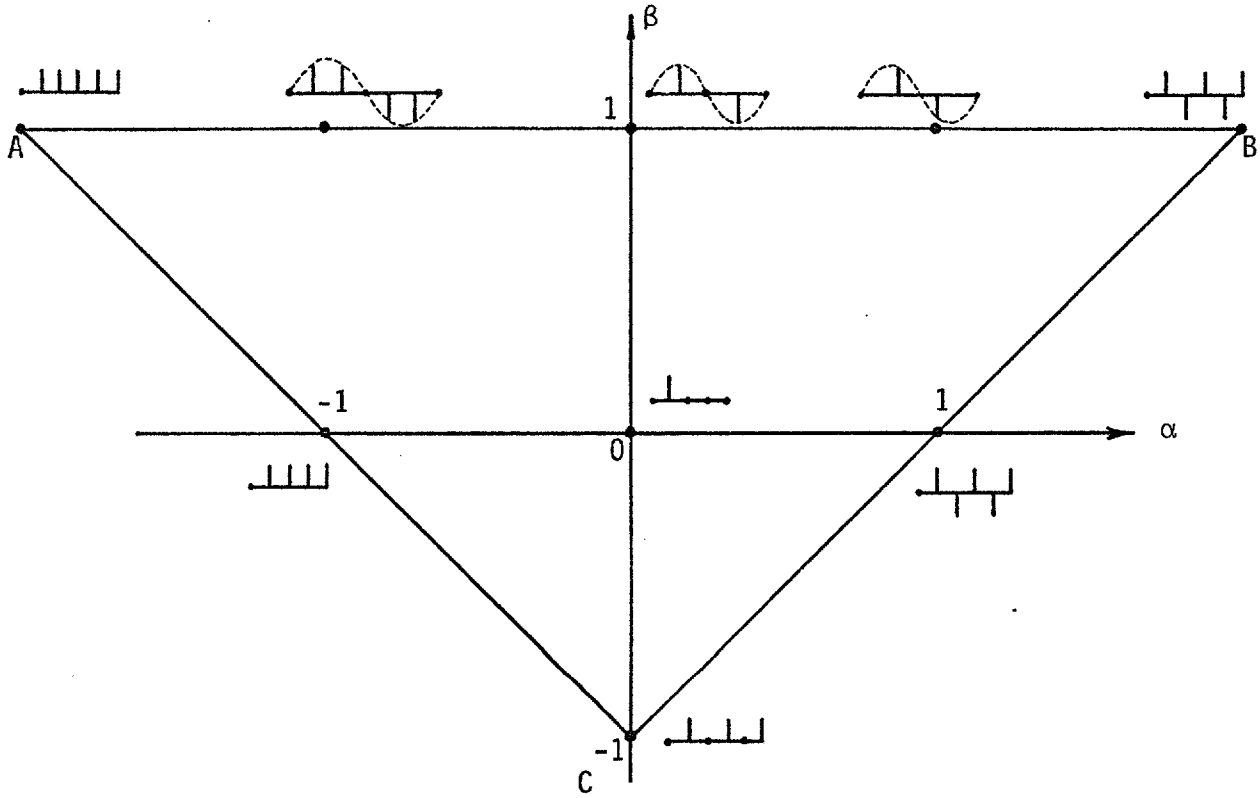


FIG. 4.6 Coefficient pairs, and $x_1(n)$ impulse response of $H_{0A}(z)$

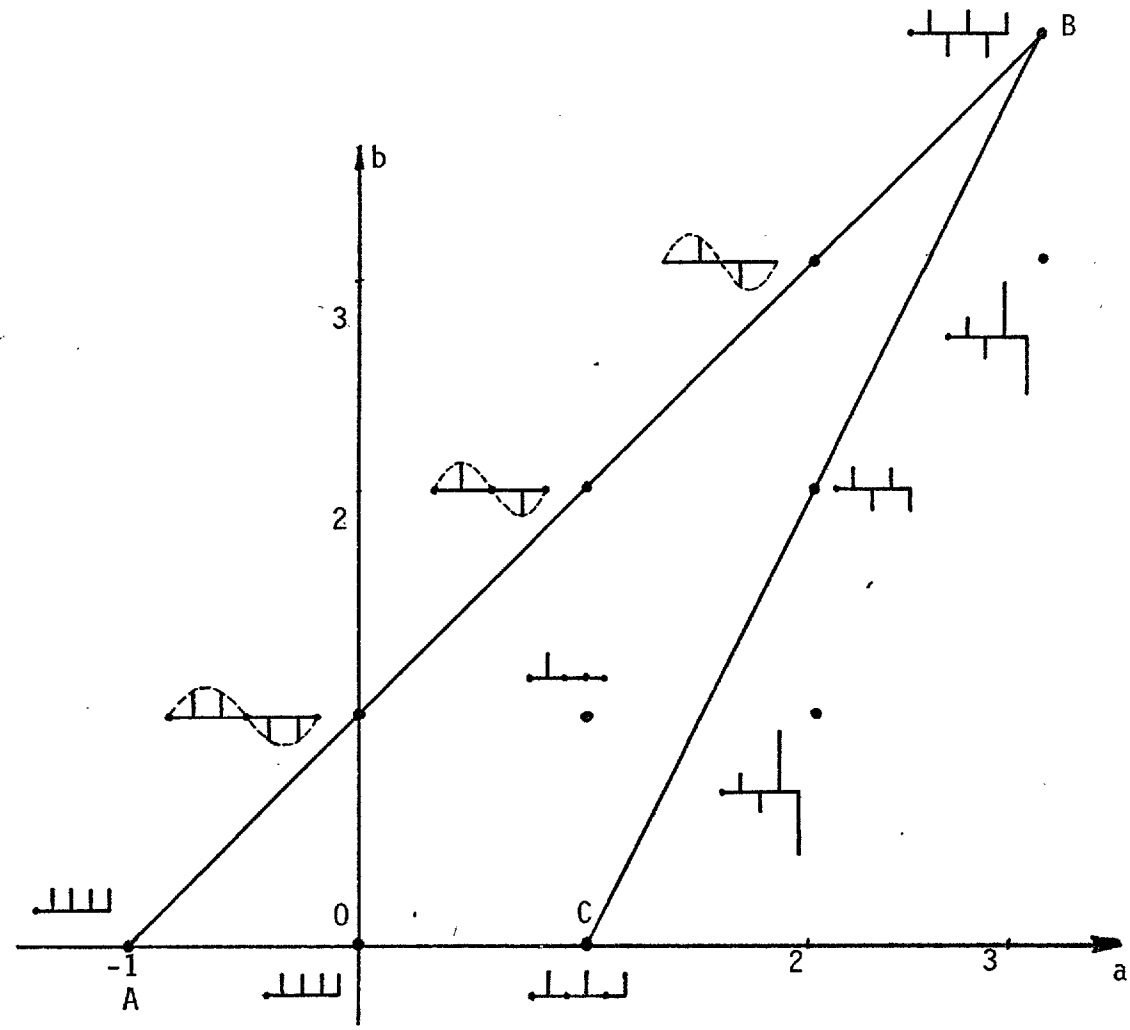


FIG. 4.7 Coefficient pairs, and $x_1(n)$ impulse response of $H_{0B}(z)$

for the model to go into oscillation occurs when the feedback to the sub-filter has the transfer of zero.

Using this idea we can suggest a condition for the model to go into oscillation. This condition would also be the condition which would guarantee the occurrence of limit cycle oscillations in the original filter.

Simulation results show that (see Chapter 6) this condition is satisfied for a large number of cases. However, other factors produce limit cycles which do not obey this rule. It appears that if a slight randomization (such as the "stabilization technique" discussed in Chapter 5) is introduced, then the results based on the arguments presented in this chapter are generally correct.

4.3 Limit cycle oscillations in the type-A realization

For the structure shown in Fig. 4.2 the limit cycle oscillations of the state variables $x_1(n)$ and $x_2(n)$ are similar, since $x_1(n) = x_2(n-1)$. The general conditions for limit cycle oscillations are therefore independent of whichever state variable is considered.

Using the model shown in Fig. 4.3 the conditions for limit cycle oscillations can be obtained by considering the different types of limit cycles. In Chapter 1 three types of limit cycles were introduced. These were:

i) limit cycles with period greater than two samples,

ii) limit cycles with period of two samples, and

iii) limit cycles with period of one sample.

To have limit cycle oscillations with period greater than two samples, the overall system must behave as a digital oscillator having complex poles. On the other hand, for one- or two-sample limit cycles the overall system must behave as a digital oscillator with one or two

real poles (Jackson [112]).

Considering the possibility of each type of limit cycles separately, and using the model discussed above, the results presented in the following sub-sections can be obtained for the different limit cycle oscillations.

4.3.1 Limit cycles with period greater than two samples

In the previous chapter (Section 3.4), it was shown that self-sustained oscillations of period greater than two samples will necessarily exist in a type-A realization provided the coefficient $\beta = 1$, and α satisfies the inequality

$$-2 < \alpha < 2 \quad . \tag{4.14}$$

Therefore, a condition which guarantees limit cycle oscillations of period greater than two samples is satisfied when $R_1[\beta] = 1$, and the feedback path corresponding to the coefficient $\Delta\beta$ has the transfer of zero for all the samples of oscillation. This idea is equivalent to the "effective value model" of Jackson [112]. However, it represents a new way of looking at the oscillation conditions, and it will assist us later to determine a suitable technique for the suppression of limit cycle oscillations.

If rounding is the quantization process, a sufficient condition to allow a limit cycle oscillation to exist occurs when

$$R_q[\Delta\beta x_2(n)] = 0 \quad , \tag{4.15}$$

where equation (4.15) is true for all the samples of the limit cycle oscillation. The condition expressed by equation (4.15) can also be written as

$$Kq \leq \frac{0.5q}{|\Delta\beta|} \quad . \tag{4.16}$$

Further, we can state that, if

$$Kq \leq \frac{0.5q}{|\Delta\alpha|} \quad , \tag{4.17}$$

the period, N (samples), of all limit cycle oscillations will be that of the sub-filter, which is given by

$$N = \frac{2\pi}{\cos^{-1}[-R_1[\alpha]/2]} \quad , \tag{4.18}$$

i.e. 1, 2, 3, 4, 6, or ∞. The values 1 and ∞ are equivalent and correspond to a constant steady state value.

For value truncation, the conditions for limit cycle oscillations can also be derived by similar arguments. For example, limit cycles with the period greater than two samples can occur when $R_1[\beta] = 1$, and the feedback corresponding to the coefficient $\Delta\beta$ has a signal transfer of zero for all the samples of that limit cycle,

i.e. $$T2_q[\Delta\beta x_2(n)] = mq \quad , \tag{4.19}$$

where m is an integer of value 0 or -1.

For equation (4.19) to hold, the state variable $x_2(n)$, or $x_1(n)$, must have no zero crossings. That is

$$-1 \leq \Delta\beta x_2(n) < 0 \quad \text{for} \quad m = -1 \quad ,$$

or

$$0 \leq \Delta\beta x_2(n) < 1 \quad \text{for} \quad m = 0 \quad .$$

Therefore, the amplitude Kq of the limit cycle is given by

$$Kq < \frac{q}{|\Delta\beta|} \quad . \tag{4.20}$$

Alternatively, the amplitudes of the limit cycles with no zero crossings normally satisfy

$$Kq \leq 2 \text{ ave}[x_2(n)] \quad , \tag{4.21}$$

where ave$[x_2(n)]$ represents the time average of the sequence $[x_2(n)]$ during the limit cycle oscillations (ie. the DC. component).

To estimate the value of ave$[x_2(n)]$, consider the model as described so far. Replace the quantizer in the branch having the coefficient α by an error source $e_\alpha(n)$, where for value truncation

$$\text{ave}[e_\alpha(n)] \approx -0.5q \quad ,$$

(see Fig. 4.8).

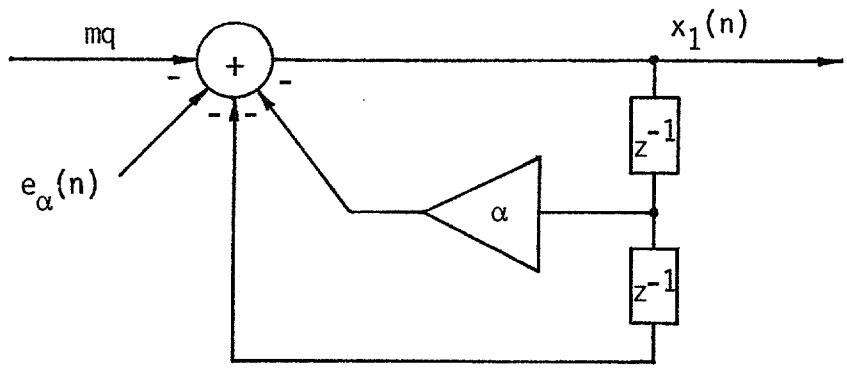


FIG. 4.8

From this description we can show that

$$\text{ave}[x_2(n)] \approx \left(\frac{0.5 - m}{2 + \alpha}\right)q \quad . \tag{4.22}$$

Inequality (4.21) can now be written as

$$Kq \leq 2\left(\frac{0.5 - m}{2 + \alpha}\right)q \quad . \tag{4.23}$$

From equation (4.19) we note that, if $m = 0$,

$$0 \leq \Delta\beta x_2(n) < 1 \quad .$$

However, for the case when $R_1'[\beta] = 1$, we have $\Delta\beta < 0$, which gives $x_2(n) < 0$. On the other hand, from equation (4.22) we have

$$\text{ave } x_2(n) \approx \left[\frac{0.5}{2+\alpha}\right]q \quad ,$$

i.e. $x_2(n) > 0$. Therefore, the case m = 0 does not satisfy both of equations (4.19) and (4.22). This indicates that m = -1.

Inequalities (4.20) and (4.23) can now be combined to give a rough estimate of the amplitude of the limit cycles with the period greater than two samples, for the case of value truncation,

i.e. $$Kq \leq \min\left[\frac{3q}{2+\alpha}, \frac{1}{|\Delta\beta|}\right] \quad . \qquad (4.24)$$

Experimental results presented in Chapter 6 (Fig. 6.13) indicate that the condition (4.24) is accurate for a majority of coefficients, and initial conditions. However, in some of the cases the amplitude of oscillation satisfies only condition (4.20). This indicates that, for value truncation, inequality (4.20) is an absolute limit for the amplitude of limit cycles with period greater than two samples. Experimental results also indicate that every limit cycle that satisfies condition (4.24) has no zero crossings, and its average is positive according to equation (4.22) when m = -1.

4.3.2 Limit cycles with period of one or two samples

In Chapter 2 (Section 2.4) it was shown that, for the study of the amplitude of limit cycles with period of one or two samples in the type-A realization, it is sufficient to consider the condition given by

$$Kq + d\, Q_q[\alpha Kq] + Q_q[\beta Kq] = 0 \quad , \qquad (4.25)$$

where $Q_q[\ ]$ represents either rounding, or value truncation; and d = +1 or -1 for one- or two-sample limit cycles respectively (see equations 2.25 and 2.37). In terms of the coefficients of the model, this condition can be expressed as

$$\{1 + R_1[\alpha] + R_1[\beta]\}Kq = -Q_q[\Delta\alpha Kq] - Q_q[\Delta\beta Kq] \quad , \tag{4.26a}$$

for one-sample limit cycles, and as

$$\{1 - R_1[\alpha] + R_1[\beta]\}Kq = Q_q[\Delta\alpha Kq] - Q_q[\Delta\beta Kq] \quad , \tag{4.26b}$$

for two-sample limit cycles.

Considering all the possible values for the coefficients of the sub-filter (i.e. $R_1[\alpha] = 0, \pm1, \pm2$, and $R_1[\beta] = 0, \pm1$), we note that there are two cases which can give solutions to equations (4.26a and b). These are the following:

i) If $1 + d\ R_1[\alpha] + R_1[\beta] = 0$, in which case the linear sub-filter has at least one real pole on the unit circle, then the condition for the limit cycle oscillations of the amplitude Kq is given by

$$d\ Q_q[\Delta\alpha Kq] + Q_q[\Delta\beta Kq] = 0 \quad , \tag{4.27}$$

where d is as defined above.

The solutions to equation (4.27) can be found by plotting the functions $d\ Q_q[\Delta\alpha Kq]$ and $-Q_1[\Delta\beta Kq]$, and finding the values of K which give equality between these two functions. For example, if $d\ \mathrm{sgn}(\Delta\alpha) \neq \mathrm{sgn}(\Delta\beta)$, the solutions to equation (4.27) for the case of rounding are given by the heavy lines of Fig. 4.9.

Alternatively equation (4.27) may be solved using the method given in Appendix A. These solutions for the case of rounding can be expressed in the following forms:

$$|K| \leq \frac{0.5}{P} \quad , \tag{4.28}$$

and if $d\ \mathrm{sgn}(\Delta\alpha) \neq \mathrm{sgn}(\Delta\beta)$ additional solutions are given by

$$\frac{i - 0.5}{S} \leq |Kq| < \frac{i + 0.5}{P} \quad , \tag{4.29}$$

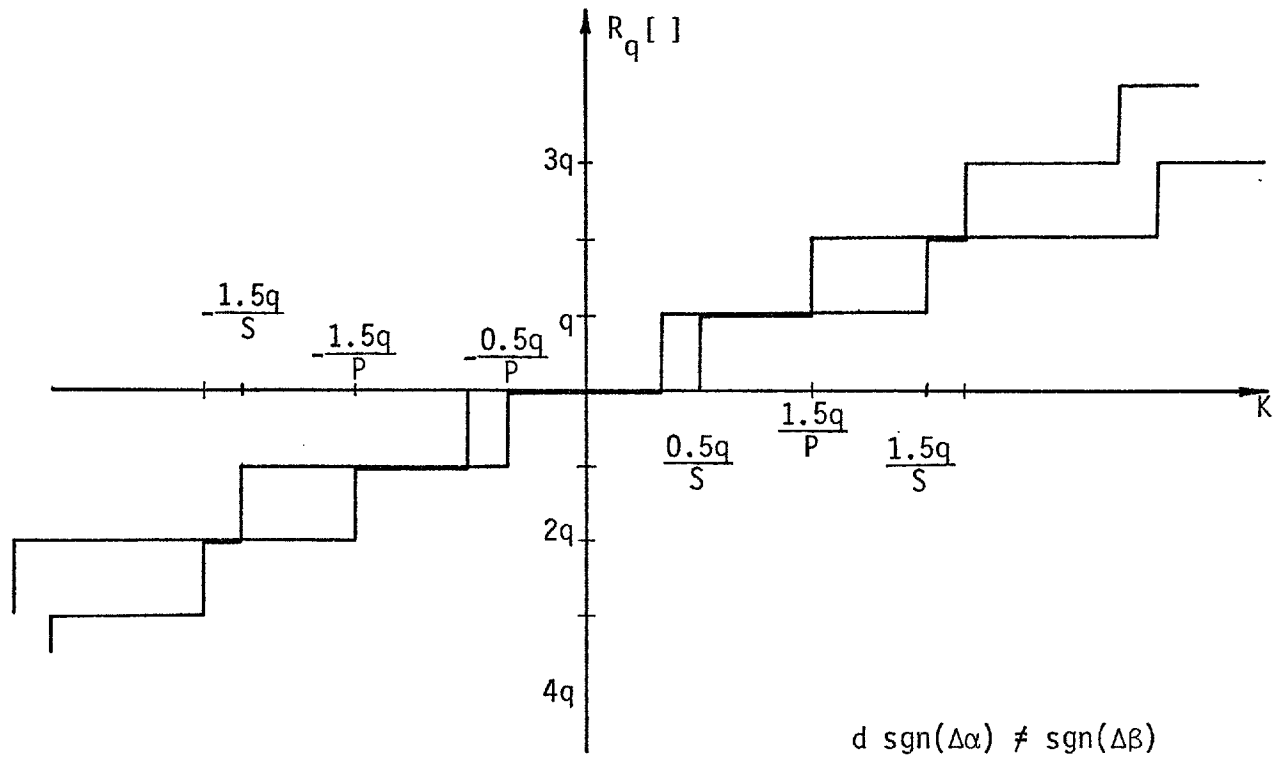


FIG. 4.9 Solutions to equation (4.27) for rounding

where

$$i = 1, 2, 3, \ldots I \quad , \tag{4.30a}$$

$$P = \max\ [|\Delta\alpha|, |\Delta\beta|] \quad , \tag{4.30b}$$

$$S = \min\ [|\Delta\alpha|, |\Delta\beta|] \quad , \tag{4.30c}$$

and I is the largest integer such that

$$\frac{I - 0.5}{S} < \frac{I + 0.5}{P} \quad ,$$

i.e.

$$I < \frac{P + S}{2(P - S)} \quad . \tag{4.31}$$

In the case of value truncation the solutions to equation

(4.27) can be found by similar method to the above. However, the results cannot be generalized in a simple form, and each special case has to be considered in its own right. For example, if $d = +1$, $\Delta\alpha > 0$ and $\Delta\beta < 0$, equation (4.27) in the case of value truncation can be written as

$$T2_q[|\Delta\alpha|Kq] = -T2_q[-|\Delta\beta|Kq] \quad . \tag{4.32}$$

The solutions to equation (4.32) have to satisfy

$$i \leq |\Delta\alpha|K < i + 1 \quad , \tag{4.33a}$$

and

$$-i \leq -|\Delta\beta|K < -i + 1 \quad , \tag{4.33b}$$

i.e.

$$\max\left[\frac{i}{|\Delta\alpha|}, \frac{i-1}{|\Delta\beta|}\right] < K < \min\left[\frac{i+1}{|\Delta\alpha|}, \frac{i}{|\Delta\beta|}\right] \quad , \tag{4.34}$$

where i is an integer such that the left hand side of expression (4.34) is less than the right hand side.

ii) If $d\, R_1[\alpha] + R_1[\beta] = 0$, then the condition for the limit cycle oscillation of amplitude Kq is given by

$$d\, Q_q[\Delta\alpha Kq] + Q_q[\Delta\beta Kq] + Kq = 0 \quad . \tag{4.35}$$

For rounding the condition represented by equation (4.35) can be simplified because, $|\Delta\alpha| \leq 0.5$ and $|\Delta\beta| \leq 0.5$. This means that

$$|R_q[\Delta\alpha Kq]| \leq 0.5K \quad ,$$

and

$$|R_q[\Delta\beta Kq]| \leq 0.5K \quad .$$

Therefore, equation (4.35) can be expressed as

$$d\, R_q[\Delta\alpha Kq] = R_q[\Delta\beta Kq] = -0.5Kq \quad . \tag{4.36}$$

To have non-zero solutions for equation (4.36) we require, $d\,\Delta\alpha < 0$, $\Delta\beta < 0$, and K an even integer. If these conditions are satisfied, then the solutions to equation (4.36), and thus equation (4.35),

are the following:

$$K = 0, 2, 4, \ldots 2L \quad , \tag{4.37}$$

where L is the largest integer that satisfies

$$L < \frac{0.5}{1-2S} \quad , \tag{4.38}$$

and S is defined by equation (4.30c).

For value truncation, equation (4.35) can be written as

$$T2_q[(1+\Delta\beta)Kq] = -d\ T2_q[\Delta\alpha Kq] \quad . \tag{4.39}$$

This equation may be solved by the graphical method described above or the method of Appendix A. However, as before, there are a number of cases for the solutions of equation (4.39), and these cannot be generalized.

The above results for the amplitude of limit cycle oscillations in the type-A realization, for rounding and value truncation, are summarized in Tables 4.1 and 4.2 respectively.

4.4 Limit cycle oscillations in the type-B realization

The amplitude of limit cycle oscillations in a second-order digital filter with the type-B realization can also be analysed using similar arguments as those in the previous section. However, because the samples and the amplitude of the limit cycles of the state variables $x_1(n)$ and $x_2(n)$ are not necessarily equal for this structure, the results based on this method will not be as simple. Therefore, in this study we can only apply this idea to determine the sufficient conditions for some of the limit cycle oscillations.

Table 4.1 Summary of the conditions and solutions for the amplitude of limit cycles in the type-A realization for rounding.

Type of limit cycle	Sufficient condition for oscillation	Solutions for the amplitude of oscillation
Period of one sample	If $1 + R_1[\alpha] + R_1[\beta] = 0$, then $R_q[\Delta\alpha Kq] + R_q[\Delta\beta Kq] = 0$	$K < \frac{0.5}{P}$ or if $\mathrm{sgn}(\Delta\alpha) \neq \mathrm{sgn}(\Delta\beta)$ $\frac{i-0.5}{S} \leq K < \frac{i+0.5}{P}$
	If $R_1[\alpha] + R_1[\beta] = 0$, and $\Delta\alpha < 0$, $\Delta\beta < 0$, then $R_q[\Delta\alpha Kq] = R_q[\Delta\beta Kq] = -0.5Kq$	$K = 0, 2, 4, \ldots 2L$, where $L < \frac{0.5}{1-2S}$
Period of two samples	If $1 - R_1[\alpha] + R_1[\beta] = 0$, then $-R_q[\Delta\alpha Kq] + R_1[\Delta\beta Kq] = 0$	$K < \frac{0.5}{P}$ or if $\mathrm{sgn}(\Delta\alpha) = \mathrm{sgn}(\Delta\beta)$ $\frac{i-0.5}{S} \leq K < \frac{i+0.5}{P}$
	If $-R_1[\alpha] + R_1[\beta] = 0$, and $\Delta\alpha > 0$, $\Delta\beta < 0$, then $-R_q[\Delta\alpha Kq] = R_q[\Delta\beta Kq] = -0.5Kq$	$K = 0, 2, 4, \ldots 2L$, where $L < \frac{0.5}{1-2S}$
Period of greater than two samples	$R_1[\beta] = 1$, and $R_q[\Delta\beta Kq] = 0$	$K \leq \frac{0.5}{\lvert\Delta\beta\rvert}$

Note: Kq = amplitude of oscillation

i, P, and S are defined in equations (4.30a, b and c)

Table 4.2 Summary of the conditions for limit cycles in the type-A realization for value truncation.

Type of limit cycle	Sufficient condition for oscillation
Period of one sample	If $1 + R_1[\alpha] + R_1[\beta] = 0$, then $T2_q[\Delta\alpha Kq] + T2_q[\Delta\beta Kq] = 0$.
	If $R_1[\alpha] + R_1[\beta] = 0$, then $T2_q[(1+\Delta\beta)Kq] + T2_q[\Delta\beta Kq] = 0$.
Period of two samples	If $1 - R_1[\alpha] + R_1[\beta] = 0$, then $T2_q[\Delta\alpha Kq] = T2_q[\Delta\beta Kq]$.
	If $-R_1[\alpha] + R_1[\beta] = 0$, then $T2_q[(1+\Delta\beta)Kq] = T2_q[\Delta\alpha Kq]$.
Period of greater than two samples	$R_1[\beta] = 1$, and $T2_q[\Delta\beta Kq] = -1$.

Note: Kq = amplitude of oscillation.

4.4.1 Limit cycles with period greater than two samples

To have limit cycle oscillations with period greater than two samples, the overall system must behave as a digital oscillator with complex poles. One way this can be possible is if we have complex poles for the sub-filter, with the effect of the feedback being zero.

For the sub-filter to have complex poles we require

$$R_1[b] - R_1[a] = 1 \quad . \tag{4.40}$$

If the above equation is satisfied, then a condition for the limit cycle oscillation is given by

$$Q_q[\Delta a x_1(n)] + Q_q[\Delta b x_2(n)] = 0 \quad . \tag{4.41}$$

One simple case which guarantees a limit cycle oscillation of this type is given by

$$Q_q[\Delta a x_1(n)] = Q_q[\Delta b x_2(n)] = 0 \quad . \tag{4.42}$$

For example, in the case of rounding, equation (4.42) has solutions given by

$$A_1 \leq \frac{0.5q}{|\Delta a|} \quad , \tag{4.43a}$$

and

$$A_2 \leq \frac{0.5q}{|\Delta b|} \quad , \tag{4.43b}$$

where A_1 and A_2 are the amplitudes of the limit cycles of the state variables $x_1(n)$ and $x_2(n)$ respectively.

4.4.2 Limit cycles with the period of one or two samples

In Chapter 2 (Section 2.4.1) limit cycle oscillations in the type-B realization having period of one sample was discussed and suitable conditions for oscillation and their solutions were suggested (see equation 2.28 and inequalities 2.29a, b and c). These results are

comprehensive and will not be discussed any further.

Limit cycle oscillations with period of two samples were also considered. A simplified equation was suggested for obtaining the amplitude of these limit cycle oscillations, equation (2.49). Using this result we can apply the transformation method of Section 4.2 to find their solutions.

Expanding equation (2.49) we have

$$\{R_1[a] - 1 - R_1\left[\frac{b}{2}\right]\}\, Kq = Q_q\left[\Delta\left(\frac{b}{2}\right)Kq\right] - Q_q[\Delta a Kq] \quad , \tag{4.44}$$

where

$$\Delta\left(\frac{b}{2}\right) = \frac{b}{2} - R_1\left[\frac{b}{2}\right] \quad , \tag{4.45a}$$

$$\Delta a = a - R_1[a] \quad , \tag{4.45b}$$

Kq is the amplitude of the limit cycle, and $Q_q[\ ]$ represents either rounding or value truncation.

For all the possible values of $R_1[a]$ and $R_1\left[\frac{b}{2}\right]$ (i.e. $R_1[a] = 0, \pm 1, 2, 3$, and $R_1\left[\frac{b}{2}\right] = 0, 1, 2$), there are two cases which can give solutions to equation (4.44). These are the following:

i) If $R_1[a] - R_1\left[\frac{b}{2}\right] = 1$, then equation (4.44) can be expressed as

$$Q_q\left[\Delta\left(\frac{b}{2}\right)Kq\right] = Q_q[\Delta a Kq] \quad . \tag{4.46}$$

For rounding, using the graphical method suggested in Section 4.3.2 or the method of Appendix A, the solutions to equation (4.46) are found to be

$$|K| \leq \frac{0.5}{P} \quad , \tag{4.47}$$

and if $\operatorname{sgn}\left(\Delta\left(\frac{b}{2}\right)\right) = \operatorname{sgn}(\Delta a)$, additional solutions are given by

$$\frac{i - 0.5}{S} \leq |K| < \frac{i + 0.5}{P} \quad , \tag{4.48}$$

where

$$i = 1, 2, \ldots I \quad , \tag{4.49a}$$

$$P = \max \left[|\Delta a|, \left|\Delta\left(\frac{b}{2}\right)\right| \right] \quad , \tag{4.49b}$$

$$S = \min \left[|\Delta a|, \left|\Delta\left(\frac{b}{2}\right)\right| \right] \quad , \tag{4.49c}$$

and I is given as the largest integer such that

$$I < \frac{P+S}{2(P-S)} \quad . \tag{4.50}$$

In the case of value truncation the same method can be used to find the solutions, however, the results cannot be generalized as above.

ii) If $R_1[a] - R_1\left[\frac{b}{2}\right] = 0$, then equation (4.44) can be expressed as

$$Q_q\left[\Delta\left(\frac{b}{2}\right)Kq\right] - Q_q[\Delta aKq] + Kq = 0 \quad . \tag{4.51}$$

But we have $|\Delta a| \leq 0.5$ and $|\Delta b| \leq 0.5$, therefore for the case of rounding, equation (4.51) can be written as

$$R_q[\Delta aKq] = -R_q\left[\Delta\left(\frac{b}{2}\right)Kq\right] = 0.5Kq \quad . \tag{4.52}$$

To have non-zero solutions, we require $\Delta a > 0$, $\Delta\left(\frac{b}{2}\right) < 0$, and K an even integer. If these conditions are satisfied, then the solutions are the following:

$$K = 0, 2, 4, \ldots 2L \quad , \tag{4.53}$$

where L is the largest integer that satisfies

$$L < \frac{0.5}{1-2S} \quad . \tag{4.54}$$

For value truncation equation (4.51) can be written as

$$T2_q\left[\left(1+\Delta\left(\frac{b}{2}\right)\right)Kq\right] = T2_q[\Delta aKq] \quad . \tag{4.55}$$

Table 4.3 Summary of the conditions and solutions for the amplitude of limit cycles in the type-B realization for rounding.

Type of limit cycle	Sufficient conditions for oscillation	Solutions for the amplitude of oscillation
Period of one sample	$R_q[bMq] = 0$	$M \leq \frac{0.5}{\lvert b \rvert}$
Period of two samples	If, $R_1[a] - R_1\left[\frac{b}{2}\right] = 1$, then $R_q\left[\Delta\left(\frac{b}{2}\right)Kq\right] = R_q[\Delta aKq]$.	$K \leq \frac{0.5}{P}$, and if $\operatorname{sgn}\left(\Delta\left(\frac{b}{2}\right)\right) = \operatorname{sgn}(\Delta a)$ $\frac{i-0.5}{S} \leq K < \frac{i+0.5}{P}$
	If $R_1[a] - R_1\left[\frac{b}{2}\right] = 0$, and $\Delta a > 0, \Delta\left(\frac{b}{2}\right) < 0$, then $R_q[\Delta aKq] = -R_q\left[\Delta\left(\frac{b}{2}\right)Kq\right] = 0.5Kq.$	$K = 0, 2, 4, \ldots 2L$, where $L < \frac{0.5}{1-2S}$.
Period of greater than two samples	$R_1[a] + 1 = R_1[b]$, and $R_q[\Delta aKq] + R_q[\Delta bMq] = 0$	normal solutions were not found

Note: Kq = amplitude of oscillation for $x_1(n)$,

Mq = amplitude of oscillation for $x_2(n)$,

i, P, and S are defined in equations (4.49a, b and c).

Table 4.4 Summary of the conditions for limit cycles in the type-B realization for value truncation.

Type of limit cycle	Sufficient condition for oscillation
Period of one sample	$T2_q[bMq] = 0$
Period of two samples	If $R_1[a] - R_1\left[\frac{b}{2}\right] = 1$, then $T2_q\left[\Delta\left(\frac{b}{2}\right)Kq\right] = T2_q[\Delta aKq]$.
	If $R_1[a] - R_1\left[\frac{b}{2}\right] = 0$, then $T2_q\left[\left(1+\Delta\left(\frac{b}{2}\right)\right)Kq\right] = T2_q[\Delta aKq]$.
Period of greater than two samples	$R_1[a] + 1 = R_1[b]$, and $T2_q[\Delta aKq] + T2_q[\Delta bMq] = 0$.

Note: Kq = amplitude of oscillation for $x_1(n)$,
Mq = amplitude of oscillation for $x_2(n)$.

Solutions to equation (4.55) may be found by similar method to the above, but cannot be generalized in a simple form.

The results for the limit cycle oscillations in the type-B realization for rounding and value truncation are summarized in Tables 4.3 and 4.4 respectively.

4.5 Conclusions

In this chapter a new idea for modelling (or transforming) the structure of a second-order digital filter was introduced. This technique was based on representing a coefficient in two parts, an integer part, and a non-integer part. Using this transformation the structure of the filter was divided into two sections (Figures 4.3 and 4.5), one having integer coefficients and therefore no quantization errors (the sub-filter), and the other with coefficients less than 0.5 in magnitude which therefore had quantization effects (the feedback section).

By considering the oscillation modes of the model, the conditions for some of the limit cycle oscillations were derived. For limit cycles with period of one or two samples the transformation technique helped to find a comprehensive set of solutions for the amplitude of oscillation, see Tables 4.1 to 4.4.

The results show that limit cycles in second-order filters normally are due to a predictable behaviour of the branch of the model with the coefficient $\Delta\beta$ for the type-A and coefficient Δb for the type-B realizations, where the coefficients $\Delta\beta$ and Δb are defined in equations (4.5b) and (4.10b) respectively. This result is followed up in the next chapter to suggest a suitable method for suppressing limit cycle oscillations.

CHAPTER 5

SUPPRESSION OF LIMIT CYCLE OSCILLATIONS

5.1 Introduction

The discussion of the previous chapter shows that limit cycle oscillations normally occur as the result of a consistency in the behaviour of the quantized output of the feedback section of the model described. For example, in the case of the type-A realization with rounding, this consistency was found to be the following:

i) for limit cycles with period greater than two samples:

$$R_q[\Delta\beta Kq] = 0 \quad , \tag{5.1a}$$

ii) for the limit cycles with period of one or two samples:

$$d\ R_q[\Delta\alpha Kq] + R_q[\Delta\beta Kq] = 0 \quad , \tag{5.1b}$$

or

$$d\ R_q[\Delta\alpha Kq] = R_q[\Delta\beta x_2(n)] = -0.5Kq \quad , \tag{5.1c}$$

where Kq is the amplitude of oscillation.

In these cases, to suppress the limit cycle oscillations, it is sufficient to alter the quantity $R_q[\Delta\beta x_2(n)]$ such that only the limit cycles with a smaller amplitude, or no limit cycle, can satisfy the conditions for oscillation.

In this chapter the method of "signal stabilization" [138-140] is considered for modifying these conditions for oscillation. This method is based on the addition of an external signal to the system to make that system more linear, and therefore reduce non-linear instabilities.

This signal is referred to as the "stabilizing signal".

In the following discussion, the stabilizing signal is assumed to be random with a uniform probability density and a flat spectrum. In Section 5.6 the use of a stabilizing signal with other properties is discussed.

5.2 The method of signal stabilization

Consider a multiplier branch of a digital filter defined by the function $Q_q[\gamma x(n)]$, where γ is the coefficient of the multiplier, and $x(n)$ is the corresponding state variable of that branch. In the method of "signal stabilization" an external signal independent of $x(n)$ is added to this multiplier branch. The effect of this signal is to modify the quantization process by making its transition regions indeterminant (see Fig. 5.2).

The stabilizing signal can be added to the branch either by a signal $s(n)$ added to $x(n)$ before multiplication or by a signal $s'(n)$ added to the product $\gamma x(n)$. These methods are shown in Figures 5.1a and 5.1b respectively.

If the modified quantization processes in the two cases are represented by $P_q[\ ]$, and $P'_q[\ ]$ respectively, then

$$P_q[\gamma x(n)] = Q_q[\gamma(x(n) + s(n))] \quad , \tag{5.2}$$

and

$$P'_q[\gamma x(n)] = Q_q[\gamma x(n) + s'(n)] \quad . \tag{5.3}$$

For the two methods, the stabilizing effect of the signal is the same if $s'(n) = \gamma s(n)$ but the increase in the error can be different. This effect will be investigated in Section 5.4.

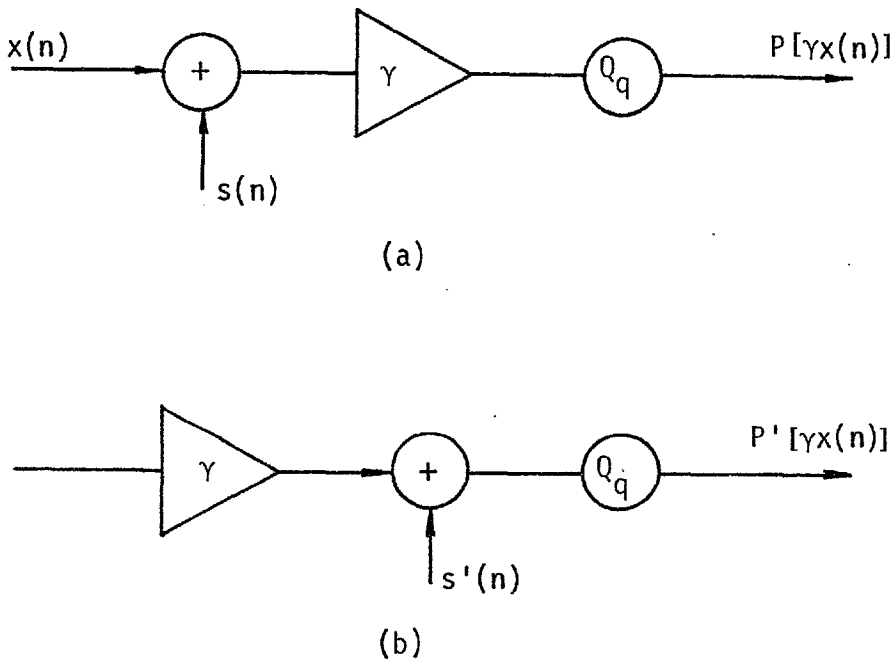


FIG. 5.1 Two ways of adding the stabilizing signal

If rounding is the quantization process, and the stabilizing signal is random and of amplitude A_s, the modified quantization process $P_q[\ ]$ may be represented as shown in Fig. 5.2. In this diagram the shaded areas represent the regions where the transition of the function can exist. The stabilizing signal therefore has the effect of jittering the transition regions of the staircase non-linearity. We note that by this process we effectively reduce the region where the transfer of the non-linearity is zero. The modified quantization process $P'_q[\ ]$ can also be described in the same way.

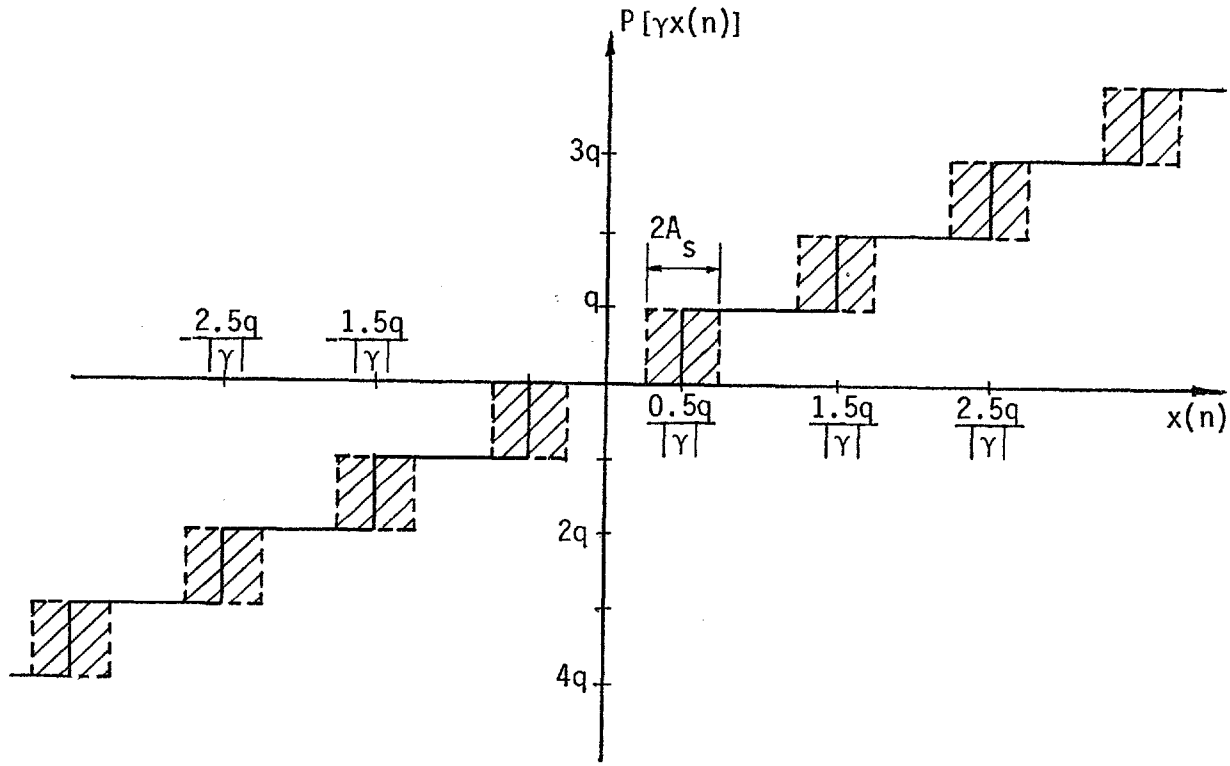


FIG. 5.2 The modified quantization function (case of rounding)

Consider the sufficient condition for the existence of the limit cycles in the type-A realization, i.e. Tables 4.1 and 4.2. To apply the method of signal stabilization, it is sufficient to modify only the quantized variable $Q_q[\Delta\beta x_2(n)]$. Therefore a suitable structure for applying the method of signal stabilization can be suggested to be that shown in Fig. 5.3.

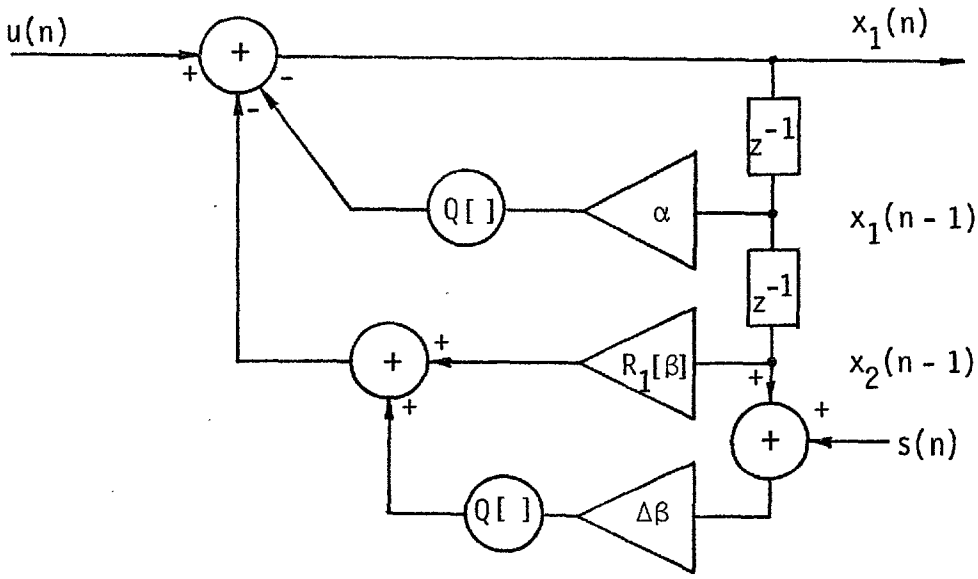


FIG. 5.3 Type-A realization with signal stabilization

The stabilizing signal s(n) must be of a suitable amplitude to suppress the limit cycles effectively, and must be of sufficiently low amplitude such that the excess noise added is negligible. In the following two sections we shall analyse the suppression effect of the signal stabilization in the type-A realization and the increase in the noise level due to this technique. In Section 5.5 the stabilization of the type-B realization will be considered.

5.3 The amplitude of the modified limit cycles for the type-A realization

The sufficient conditions for limit cycle oscillations in the structure of Fig. 5.3 can be investigated using similar methods as those discussed in Section 4.3. The results for each type of limit cycle

oscillation are presented in the following sub-sections.

5.3.1 Limit cycles with period greater than two samples

For the occurrence of the limit cycles with period greater than two samples, in the case of rounding, it is sufficient that

$$R_q[\Delta\beta(x_2(n) + s(n))] = 0 \quad . \tag{5.4}$$

The amplitude K'q of the limit cycle oscillations which satisfy equation (5.4) must therefore satisfy

$$K'q \leq \frac{0.5q - |\Delta\beta|A_s}{|\Delta\beta|} \quad . \tag{5.5}$$

To suppress all of the limit cycles satisfying equation (5.4) A_s should be chosen such that K' < 1, i.e.

$$A_s \geq \left[\frac{0.5 - |\Delta\beta|}{|\Delta\beta|}\right]q \quad . \tag{5.6}$$

On the other hand, to minimize the increase in the overall noise due to s(n), A_s must be small. Thus, from inequality (5.6), it follows that a suitable value for A_s sufficient to suppress the limit cycles with period greater than two samples and cause the least increase in the noise level is given by

$$A_s = T1_q\left[\frac{0.5q}{|\Delta\beta|}\right] \quad . \tag{5.7}$$

Equation (5.7) represents the smallest integral multiple of q which satisfies inequality (5.6). For example, let $\Delta\beta = 0.125$ (i.e. 0.001 in binary representation), to have no limit cycle of period greater than two samples, $A_s = 4q$.

For value truncation, the condition for the limit cycles of period greater than two samples in the type-A realization using signal stabilization can be expressed by

$$T2_q[\Delta\beta(x_2(n) + s(n))] = mq \quad , \tag{5.8}$$

where, as in equation (4.19), m is an integer of value -1. Therefore, to suppress all of the limit cycles which satisfy equation (5.8), a suitable signal is defined as

$$|s(n)| \leq A_s \quad , \tag{5.9a}$$

and

$$\text{sgn}(s(n)) = \text{sgn}(\Delta\beta) \quad , \tag{5.9b}$$

where the amplitude A_s of the signal, sufficient to suppress the limit cycle oscillations of this type is given by

$$A_s \geq \left(\frac{1 - |\Delta\beta|}{|\Delta\beta|}\right)q \quad . \tag{5.10}$$

5.3.2 Limit cycles with period of one or two samples

When signal stabilization is used, limit cycles with period of one or two samples can occur if the stabilizing signal is not of sufficiently large amplitude. The conditions for the limit cycles of these types are given by

$$\{1 + d\ R_1[\alpha] + R_1[\beta]\}K'q = -d\ Q_q[\Delta\alpha K'q] - Q_q[\Delta\beta(K'q + s(n))] \quad , \tag{5.11}$$

where K'q is the amplitude of the modified limit cycle, and d = +1 or -1, for one- or two-sample limit cycles respectively. The proof of the conditions given in equation (5.11) follows on the same line as that discussed in Chapter 2 (Section 2.4). For details, in the case of two-sample limit cycles, see Appendix B.

There are two possible cases for equation (5.11). These are as follows:

i) If $1 + d\ R_1[\alpha] + R_1[\beta] = 0$, then the condition for the limit cycle oscillations of the amplitude K'q is given by

$$d\ Q_q[\Delta\alpha K'q] + Q_q[\Delta\beta(K'q + s(n)] = 0 \quad . \tag{5.12}$$

The solutions to equation (5.12) can be found, using the graphical method discussed in the previous chapter (Section 4.3.2), or using the method suggested in Appendix A. For example, in the case of rounding, if $|\Delta\alpha| > |\Delta\beta|$, and $d\ \mathrm{sgn}(\Delta\alpha) \neq \mathrm{sgn}(\Delta\beta)$, then the solutions to equation (5.12) are the heavy lines of Fig. 5.4a. However, if $|\Delta\alpha| < |\Delta\beta|$, and $d\ \mathrm{sgn}(\Delta\alpha) \neq \mathrm{sgn}(\Delta\beta)$, then the solutions to equation (5.12) are the heavy lines of Fig. 5.4b.

Alternatively, we note that if

$$Q_q[\Delta\beta(K'q + s(n)] = iq \quad , \tag{5.13}$$

then for rounding we can write

$$\frac{(i-0.5)q + |\Delta\beta|A_s}{|\Delta\beta|} \leq K'q < \frac{(i+0.5)q - |\Delta\beta|A_s}{|\Delta\beta|} \quad , \tag{5.14}$$

i.e. from Appendix A, the following results for the case of rounding can be derived:

$$K' \leq \min\left[\frac{0.5}{|\Delta\alpha|},\ \frac{0.5q - |\Delta\beta|A_s}{|\Delta\beta|q}\right] \quad , \tag{5.15}$$

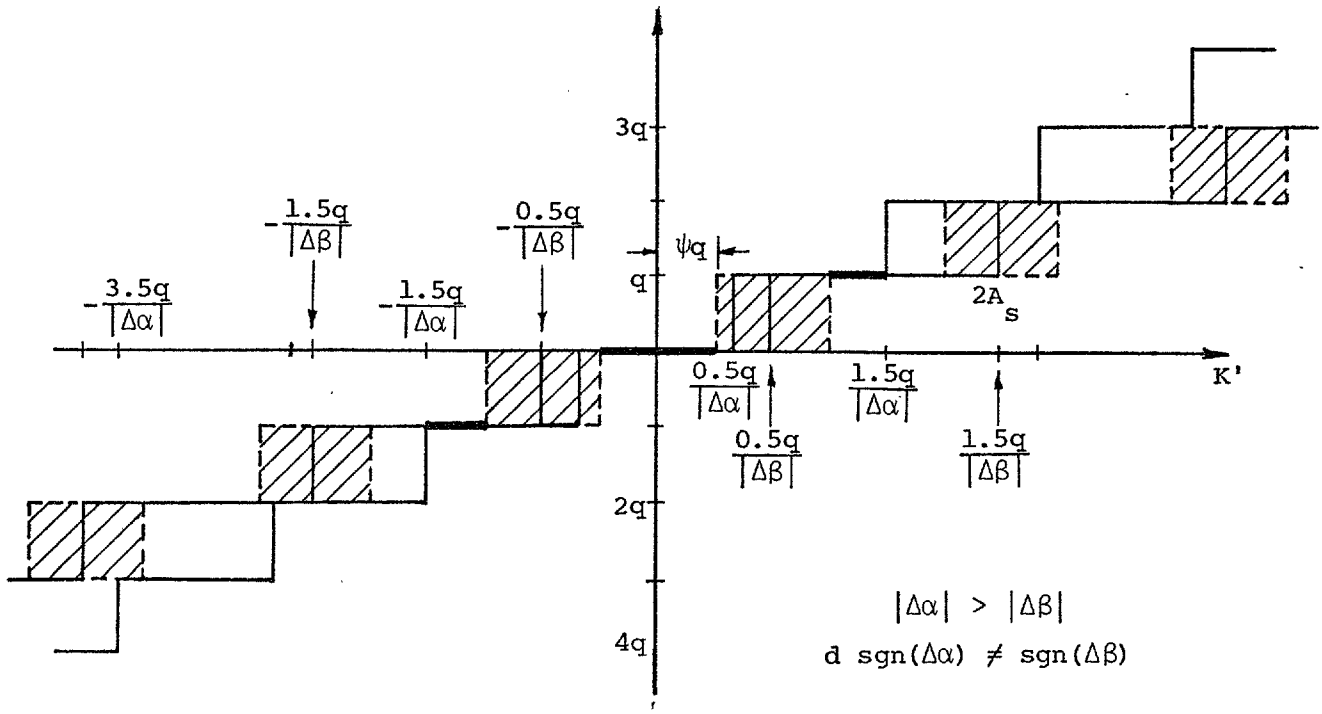
and, if $d\ \mathrm{sgn}(\Delta\alpha) \neq \mathrm{sgn}(\Delta\beta)$, additional solutions are given by

$$\max\left[\left\{\frac{i-0.5}{|\Delta\alpha|}\right\},\ \frac{(i-0.5)q + |\Delta\beta|A_s}{|\Delta\beta|q}\right] \leq K' \quad , \tag{5.16a}$$

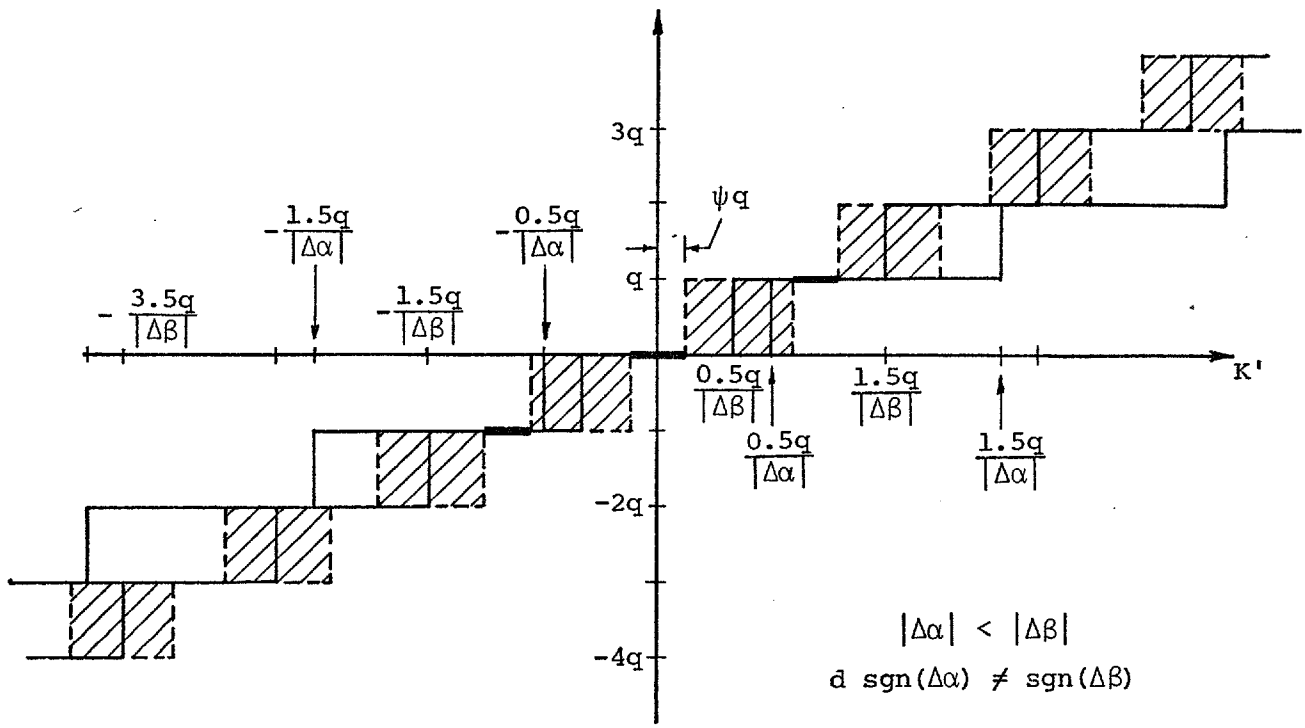
$$K' < \min\left[\left\{\frac{i+0.5}{|\Delta\alpha|}\right\},\ \frac{(i+0.5)q - |\Delta\beta|A_s}{|\Delta\beta|q}\right] \quad , \tag{5.16b}$$

where $i = 1, 2, 3, \ldots I$, for I defined as the largest integer whereby the left hand side of inequality (5.16a) is less than the right hand side of inequality (5.16b), i.e.

$$I < \frac{0.5(|\Delta\alpha| + |\Delta\beta|)q - A_s|\Delta\alpha|\ |\Delta\beta|}{|(|\Delta\alpha| - |\Delta\beta|)|} \quad . \tag{5.17}$$



(a)



(b)

FIG. 5.4 Solutions to equation (5.12)

Conditions (5.15) to (5.17) can be simplified by defining

$$\psi = \frac{0.5q - |\Delta\beta|A_s}{|\Delta\beta|q} \quad , \tag{5.18}$$

where the quantity ψ is also shown in Fig. (5.4). Now we can write

$$K' \leq \min \left[\left(\frac{0.5}{|\Delta\alpha|} \right), \psi \right] \quad , \tag{5.19}$$

and, if d $\text{sgn}(\Delta\alpha) \neq \text{sgn}(\Delta\beta)$, additional solutions are given by

$$\max \left[\left(\frac{i - 0.5}{|\Delta\alpha|} \right), \left(\frac{i}{|\Delta\beta|} - \psi \right) \right] \leq K' < \min \left[\left(\frac{i + 0.5}{|\Delta\alpha|} \right), \left(\frac{i}{|\Delta\beta|} + \psi \right) \right] \quad , \tag{5.20}$$

where $i = 1, 2, 3, \ldots I$, for I defined as the largest integer which satisfies

$$I < \frac{(0.5 + |\Delta\alpha|\psi)\ |\Delta\beta|}{|(|\Delta\alpha| - |\Delta\beta|)|} \quad . \tag{5.21}$$

From these results a suitable amplitude for s(n) which can suppress all of the limit cycles of period one or two samples can be shown to be given by

$$A_s \geq \left(\frac{0.5 - |\Delta\beta|}{|\Delta\beta|} \right) q \quad . \tag{5.22}$$

A comparison of the conditions (5.22) and (5.6) shows that these results are equivalent. Therefore, equation (5.7), the formula for choosing a suitable value for A_s to suppress all limit cycles, also applies for the limit cycles with period of one or two samples, when $1 + d\ R_1[\alpha] + R_1[\beta] = 0$.

For the case of value truncation, the solutions to equation (5.12) can be found following the same procedure as above. However, the results cannot be generalized in a similar way, but a suitable form for the stabilizing signal can be suggested to be as defined by the conditions (5.9a and b) and (5.10).

ii) If $d\ R_1[\alpha] + R_1[\beta] = 0$, then the condition for the limit cycle oscillation of amplitude K'q and period of one or two samples is given by

$$d\ Q_q[\Delta\alpha K'q] + Q_q[\Delta\beta(K'q + s(n))] + K'q = 0 \quad . \tag{5.23}$$

For rounding, as in Section 4.3, the above conditions can be simplified to

$$d\ R_q[\Delta\alpha K'q] = R_q[\Delta\beta(K'q + s(n))] = -0.5K'q \quad . \tag{5.24}$$

Non-zero solutions require: $d\ \Delta\alpha < 0$, $\Delta\beta < 0$, and K to be an even integer. If these conditions are satisfied then the solutions to equation (5.24) are given by

$$K' = 0,\ 2,\ 4,\ \ldots\ 2L \quad , \tag{5.25}$$

where

$$L < \min\left[\left(\frac{0.5}{1 - 2|\Delta\alpha|}\right),\ \left[\frac{0.5q - |\Delta\beta|A_s}{(1 - 2|\Delta\beta|)q}\right]\right] \quad . \tag{5.26}$$

From conditions (5.25) and (5.26) we note that, for the cases when $d\ R_1[\alpha] + R_1[\beta] = 0$, to suppress all of the limit cycles having the period of one or two samples, we require:

$$A_s > \left(\frac{2|\Delta\beta| - 0.5}{|\Delta\beta|}\right)q \quad . \tag{5.27}$$

In fact, if $|\Delta\beta| \geq 0.25$, then $A_s = q$, otherwise $A_s = 0$.

In the case of value truncation equation (5.23) may be written as

$$d\ T2_q[\Delta\alpha K'q] + K'q = -T2_q[\Delta\beta(K'q + s(n))] \quad . \tag{5.28}$$

The solutions to equation (5.28) cannot be generalized in the same way as for the case of rounding. However, a suitable stabilizing signal could be suggested to be defined by the conditions (5.9a and b) and (5.10).

Summarising the results derived in this section, we have the following:

i) For rounding the suitable amplitude of the stabilizing signal may be chosen according to the summary given in Table 5.1. If the amplitude of the stabilizing signal is less than the limits suggested the amplitude of the expected limit cycle oscillation will be given by the results summarised in Table 5.2.

ii) For value truncation the suitable stabilizing signal is defined as:

$$0 \leqslant s(n) \leqslant T1_q\left[\frac{q}{\Delta\beta}\right] \quad , \tag{5.29}$$

for $\Delta\beta > 0$, and as

$$T1_q\left[\frac{q}{\Delta\beta}\right] \leqslant s(n) \leqslant 0 \quad , \tag{5.30}$$

for $\Delta\beta < 0$.

Table 5.1 Amplitude of the stabilizing signal sufficient to suppress the limit cycles in the type-A realization for the case of rounding.

Conditions for the coefficients	Suitable amplitude of the stabilizing signal
$R_1\lvert\beta\rvert = 1$, or $1 \pm R_1[\alpha] + R_1[\beta] = 0$	$A_s = T1_q\left[\frac{0.5q}{\lvert\Delta\beta\rvert}\right]$
$\pm R_1[\alpha] + R_1[\beta] = 0$ with $\lvert\Delta\beta\rvert > 0.25$	$A_s = q$

Note: A_s = amplitude of the stabilizing signal in the structure of Fig. 5.3.

Table 5.2 Amplitude of the modified limit cycle oscillation in the type-A realization for the case of rounding.

Type of limit cycle	Amplitude of the modified oscillation
Period of one or two samples	If $1 + d\,R_1[\alpha] + R_1[\beta] = 0$, then $K' \leq \min\left[\frac{0.5}{\lvert\Delta\alpha\rvert}, \psi\right]$ and if $d\,\mathrm{sgn}(\Delta\alpha) \neq \mathrm{sgn}(\Delta\beta)$ $\max\left[\left(\frac{i-0.5}{\lvert\Delta\alpha\rvert}\right), \left(\frac{i}{\lvert\Delta\beta\rvert} - \psi\right)\right] \leq K'$ and $K' < \min\left[\left(\frac{i+0.5}{\lvert\Delta\alpha\rvert}\right), \left(\frac{i}{\lvert\Delta\beta\rvert} + \psi\right)\right]$
	If $d\,R_1[\alpha] + R_1[\beta] = 0$, then $K' = 0, 2, 4, \ldots 2L$ where $L < \min\left[\left(\frac{0.5}{1-2\lvert\alpha\rvert}\right), \left(\frac{0.5q - \lvert\Delta\beta\rvert A_s}{(1-2\lvert\Delta\beta\rvert)q}\right)\right]$
Period of greater than two samples	$K' \leq \psi$

Note: $K'q$ = amplitude of the modified oscillation.

$$\psi = \frac{0.5q - |\Delta\beta|A_s}{|\Delta\beta|q} \quad .$$

$i = 1, 2, 3, \ldots I$, for I defined by inequality (5.21).

$d = +1$ or -1 for one- or two-sample limit cycles respectively.

5.4 Increase in the noise level

We now consider the increase in the noise level of the system caused by addition of the stabilizing signal. For comparison purposes we only evaluate the increase in the noise level in the multiplier branch in which the stabilizing signal is added. The increase in the noise level of the state variables can then be evaluated by simple realizations (1.21) and (1.22). This increase for the case of rounding can be shown to be the same as the increase in the equivalent noise level at the input stage of the filter.

Consider the realization of equation (5.2), i.e. Fig. 5.1a. To add the stabilizing signal s(n) to the state variable x(n), without having to increase the wordlength, the signal s(n) has to be quantized and must have the quantization level given by

$$q_s = r_s q \tag{5.31}$$

where r_s is any positive integer. Therefore the possible samples of s(n) are given by

$$\{0, \pm q_s, \pm 2q_s, \ldots \pm N_s q_s\} \quad ,$$

where $N_s q_s$ is the amplitude A_s of the signal s(n). The probability density of s(n), assuming it to be uniform, is therefore as shown in Fig. 5.5.

The modified quantization process $P_q[\gamma x(n)]$ can be written as

$$P_q[\gamma x(n)] = \gamma x(n) + \gamma s(n) + e(n) \quad , \tag{5.32}$$

where e(n) is the error generated by the quantization of $[\gamma(x(n) + s(n))]$. As discussed in Section 2.3, e(n) is also a quantized variable, having a quantization level of qq_c, where q_c is the quantization level of the coefficient γ as defined by equations (2.14) and (2.15).

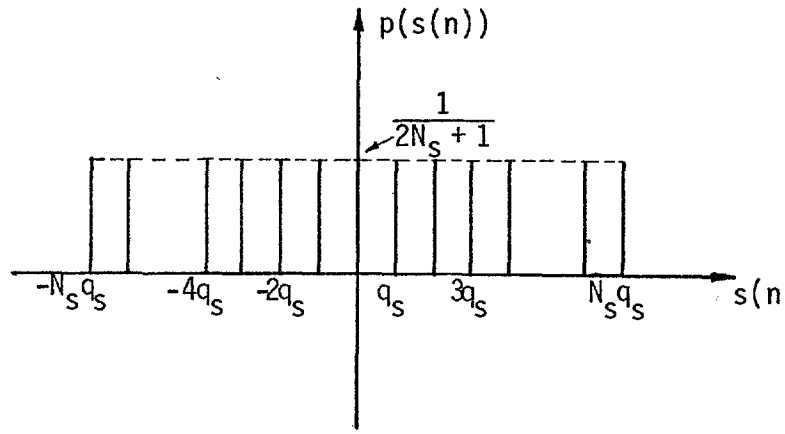


FIG. 5.5

Define $e_s(n)$ to be the total error due to the quantization and the added signal, i.e.

$$e_s(n) = \gamma s(n) + e(n) \quad . \tag{5.33}$$

Therefore we have

$$\rho_{e_s}^2 = \gamma^2 \rho_s^2 + \rho_e^2 \quad , \tag{5.34}$$

where ρ_e, ρ_s and ρ_{e_s} are the root mean square (rms) values of e(n), s(n) and $e_s(n)$ respectively.

From the property of s(n), as defined above, we have

$$\rho_s = \left[\sqrt{\frac{N_s(N_s+1)}{3}}\right] q_s \quad . \tag{5.35}$$

For rounding, the rms. of e(n) is given by

$$\rho_e = \left[\sqrt{\frac{1+2q_c^2}{12}}\right] q \quad , \tag{5.36}$$

i.e.

$$\rho_{e_s} = \left[\sqrt{\gamma^2\left[\frac{N_s(N_s+1)}{3}\right]r_s^2 + \left[\frac{1+2q_c^2}{12}\right]}\right]q \quad , \tag{5.37}$$

where r_s is defined by equation (5.31).

The increase in the rms. of the error in the branch considered is given by

$$\delta\rho_e = 10\ \log_{10}\left[1 + 4\gamma^2 r_s^2\left[\frac{N_s(N_s+1)}{1+2q_c^2}\right]\right] \text{dB} \quad . \tag{5.38}$$

For example, in Fig. 5.3, if the multiplier $\Delta\beta$ has the coefficient of value $\frac{1}{32}$, then $q_c = \frac{1}{32}$, and if the amplitude A_s of the stabilizing signal s(n) satisfies equation (5.7), i.e. $N_s = 16$, then $\delta\rho_e$ for the branch considered, is 3.14 dB .

In a digital filter with single precision arithmetic and M multiplier branches, the increase in the rms. of the equivalent error source for the case of rounding is given by

$$\delta\rho_e = 10\ \log_{10}\left[1 + \frac{4\gamma^2 r_s^2}{M}\left[\frac{N_s(N_s+1)}{1+2q_c^2}\right]\right] \text{dB} \quad . \tag{5.39}$$

For example, using the parameters defined above for the second-order filter of Fig. 5.3, $\delta\rho_e$ can be shown to be 1.84 dB.

For the realization of equation (5.3) (i.e. Fig. 5.1b), the stabilizing signal s'(n) has to be quantized to the quantization level given by

$$q_s' = r_s' q_c q \quad , \tag{5.40}$$

where r' is any positive integer. Analysis similar to that given above yields the rms. of the equivalent error source, due to M multiplier branches and the stabilizing signal s'(n), to be

$$\rho_{e_s'} = \left[\sqrt{q_c^2\left[\frac{N_s(N_s+1)}{3}\right]r_s'^2 + M\left[\frac{1+2q_c}{12}\right]}\right]q \quad . \tag{5.41}$$

Thus, the increase in the rms. of the equivalent error is given by

$$\delta\rho_{e'} = 10\log_{10}\left[1 + \frac{4q^2r_s'^2}{M}\left[\frac{N_s(N_s+1)}{1+2q_c^2}\right]\right]\text{dB} \quad . \tag{5.42}$$

Comparing equations (5.42) and (5.39), we note that the increase in the rms. of the overall error, for the structure shown in Fig. 5.1a, depends on the coefficient γ; however, for the structure shown in Fig. 5.1b, this depends on q_c. Therefore, the error in the former method will be larger when $|\alpha| > q_c$. However, considering hardware implementation, the method of Fig. 5.1a has a simpler hardware when compared to the method of Fig. 5.1b, since in the latter case the implementation requires the full multiplication of the product $\gamma x(n)$ and a double precision adder for adding the stabilizing signal. It is therefore more economical to use the signal addition method shown in Fig. 5.1a. The fact that it may result in a larger increase in the noise level can be ignored since normally this difference is very small.

5.5 Suppression of limit cycle oscillations in the type-B realization

Considering the conditions of limit cycle oscillations in the type-B realization (i.e. Tables 4.3 and 4.4) we note that the output of the multiplier branch with the coefficient Δb is a common factor for the build-up of all types of limit cycle oscillations. To suppress these oscillations, it is sufficient to add the stabilizing signal s(n) to the multiplier branch with the coefficient Δb. A suitable structure for this purpose is shown in Fig. 5.6.

The condition for limit cycle oscillations in this structure can

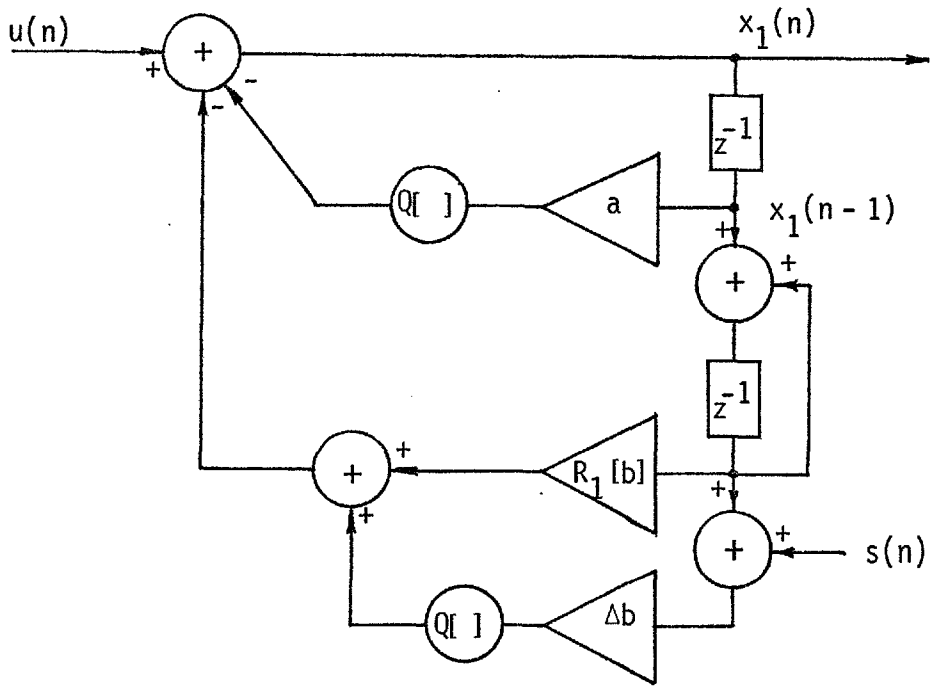


FIG. 5.6 Type-B realization with signal stabilization

be formulated using methods similar to those discussed in Sections 2.4 and 4.4. These conditions are the following:

i) For limit cycles with period of greater than two samples and satisfying equation (4.40), the condition for oscillation is given by

$$Q_q[\Delta a x_1(n)] + Q_q[\Delta b x_2(n) + \Delta b s(n)] = 0 \quad . \tag{5.43}$$

ii) For limit cycles with period of two samples if $R_1[a] - R_1\left[\frac{b}{2}\right] = 1$, then

$$Q_q\left[\Delta\left(\frac{b}{2}\right)K'q + \Delta b s(n)\right] = Q_q[\Delta a K'q] \quad , \tag{5.44}$$

and if $R_1[a] - R_1\left[\frac{b}{2}\right] = 0$, then

$$Q_q\left[\Delta\left(\frac{b}{2}\right)K'q + \Delta bs(n)\right] - Q_q[\Delta aK'q] + K'q = 0 \quad , \tag{5.45}$$

where K'q is the amplitude of the modified limit cycles for the state variable $x_1(n)$.

iii) For limit cycles with period of one sample, we have $x_1(n) = 0$, and

$$Q_q[bx_2(n) + \Delta bs(n)] = 0 \quad . \tag{5.46}$$

The proof of the above conditions is similar to the discussion of Sections 2.4 and 4.4. For example, the case of two-sample limit cycles is discussed in Appendix B.

The solutions to equations (5.44) to (5.46) can be determined using the same procedures as those in Section 5.3. Consider, for example, equations (5.44) and (5.12). The solutions to equation (5.44) for the case of rounding can be expressed by

$$K' \leq \min\left[\frac{0.5}{|\Delta a|}, \psi_B\right] \quad , \tag{5.47}$$

and if $\text{sgn}(\Delta a) = \text{sgn}(\Delta\left(\frac{b}{2}\right))$, additional solutions are given by

$$\max\left[\left(\frac{i - 0.5}{|\Delta a|}\right), \left(\frac{i}{\Delta\left(\frac{b}{2}\right)} - \psi_B\right)\right] \leq K' < \min\left[\left(\frac{i + 0.5}{|\Delta a|}\right), \left(\frac{i}{|\Delta\left(\frac{b}{2}\right)|} + \psi_B\right)\right] \quad , \tag{5.48}$$

where

$$\psi_B = \frac{0.5q - |\Delta b|A_s}{|\Delta\left(\frac{b}{2}\right)|q} \quad , \tag{5.49}$$

and $i = 1, 2, 3, \ldots I$, for I defined as the largest integer which satisfies

$$I < \frac{(0.5 + |\Delta a|\psi_B)|\Delta\left(\frac{b}{2}\right)|}{|(|\Delta a| - |\Delta\left(\frac{b}{2}\right)|)|} \quad . \tag{5.50}$$

Comparing equation (5.45) and (5.23), the solutions to equation (5.45) for the case of rounding can be expressed as

$$K' = 0, 2, 4, \ldots 2L' \tag{5.51}$$

where

$$L' < \min \left[\left(\frac{0.5}{1 - 2|\Delta a|} \right), \left[\frac{0.5q - |\Delta b| A_s}{(1 - 2|\Delta \left(\frac{b}{2} \right)|)q} \right] \right] \quad . \tag{5.52}$$

For non-zero solutions to exist, we require $\Delta a > 0$, and $\Delta \left(\frac{b}{2} \right) < 0$.

The solution to equation (5.46) in the case of rounding is given by

$$x_2(n) \leq \frac{0.5q - |\Delta b| A_s}{|b|} \quad . \tag{5.53}$$

From the above results a suitable amplitude for the stabilizing signal to suppress the limit cycle oscillations can be suggested. For the case of rounding these are as follows:

i) If $R_1[b] = 0$, then

$$A_s \geq \frac{0.5 - |b|}{|\Delta b|} q \quad , \tag{5.54}$$

ii) If $R_1[a] - R_1\left[\frac{b}{2}\right] = 1$, then

$$A_s \geq \frac{0.5 - |\Delta \left(\frac{b}{2} \right)|}{|\Delta b|} q \quad , \tag{5.55}$$

iii) If $R_1[a] - R_1\left[\frac{b}{2}\right] = 0$, then

$$A_s \geq \frac{2|\Delta \left(\frac{b}{2} \right)| - 0.5}{|\Delta b|} q \quad , \tag{5.56}$$

iv) If $R_1[a] - R_1[b] = -1$, then

$$A_s \geq \left(\frac{0.5 - |\Delta b|}{|\Delta b|} \right) q \quad . \tag{5.57}$$

5.6 Remarks on the method of signal stabilization

So far we have considered the method of signal stabilization with a random stabilizing signal of uniform probability density. This particular choice is an arbitrary one and may not be the most suitable. For example a signal with a different probability density or a special spectral distribution may result in a better performance. In this section a brief discussion of the basis for the choice of the different stabilizing signals will be considered.

The suitability of the stabilizing signal may be based on the following criteria:

i) the increase in the overall noise level caused by the signals,

ii) the stabilizing ability of the signal, i.e. the speed with which a particular limit cycle may be effected,

iii) the extra hardware necessary to implement the stabilization method.

Considering the increase in the noise level on the output of the filter or in the branch concerned, we note that a suitable signal which causes the least increase should have a small variance σ_s^2. For example, a signal with triangular probability density causes less increase in the noise level than a signal with a uniform probability density.

On the other hand the stabilizing effect of the signal with the triangular probability density is less, because the probability of having a sample of magnitude A_s (i.e. a magnitude sufficient to suppress most limit cycles) at a suitable time relative to the period of the limit cycle is less. In fact the requirements for a signal to satisfy the first two criteria are opposite to each other.

The other criterion for the choice of the stabilizing signal was pointed out to be the hardware necessary for the implementation of the method of signal stabilization. For example, if a suitable signal is found which provides the best performance, but if it requires a complex hardware for its implementation, then it may work out cheaper to simply increase the wordlength of the filter by several bits and therefore reduce the effects of both quantization noise and limit cycle oscillations.

Generally it appears that the implementation of a signal which can give a good performance would most likely involve the use of a complex hardware. Therefore from this discussion we can suggest that the suitability of a stabilizing signal should be first based on the complexity of the hardware for its implementation and then other factors could be considered.

From the results derived in Section 5.3 and 5.5 we note that the signal, with the amplitude recommended, has the effect of altering only the Least Significant Bit (LSB) of the quantized output of the multiplier branch. Therefore it appears that, for a simple hardware realization of the method of signal stabilization, we can randomly jitter the LSB of the quantized output of the multiplier branch. This method has been reported by Kieburtz [132].

To compare the performance of this simple technique with that described by Fig. 5.1a, we express the effective quantization, when the LSB is randomly jittered and rounding quantization used, by the function shown in Fig. 5.7. The broken lines in this description represent the limits of the effective function.

This representation shows that the method is equivalent to adding a three level signal of amplitude A_s, twice that recommended by equation (5.7). The equivalent signal has only the values of $\pm A_s$ and zero, with the probability of occurrence of 0.25 for the values $+A_s$ and $-A_s$ each.

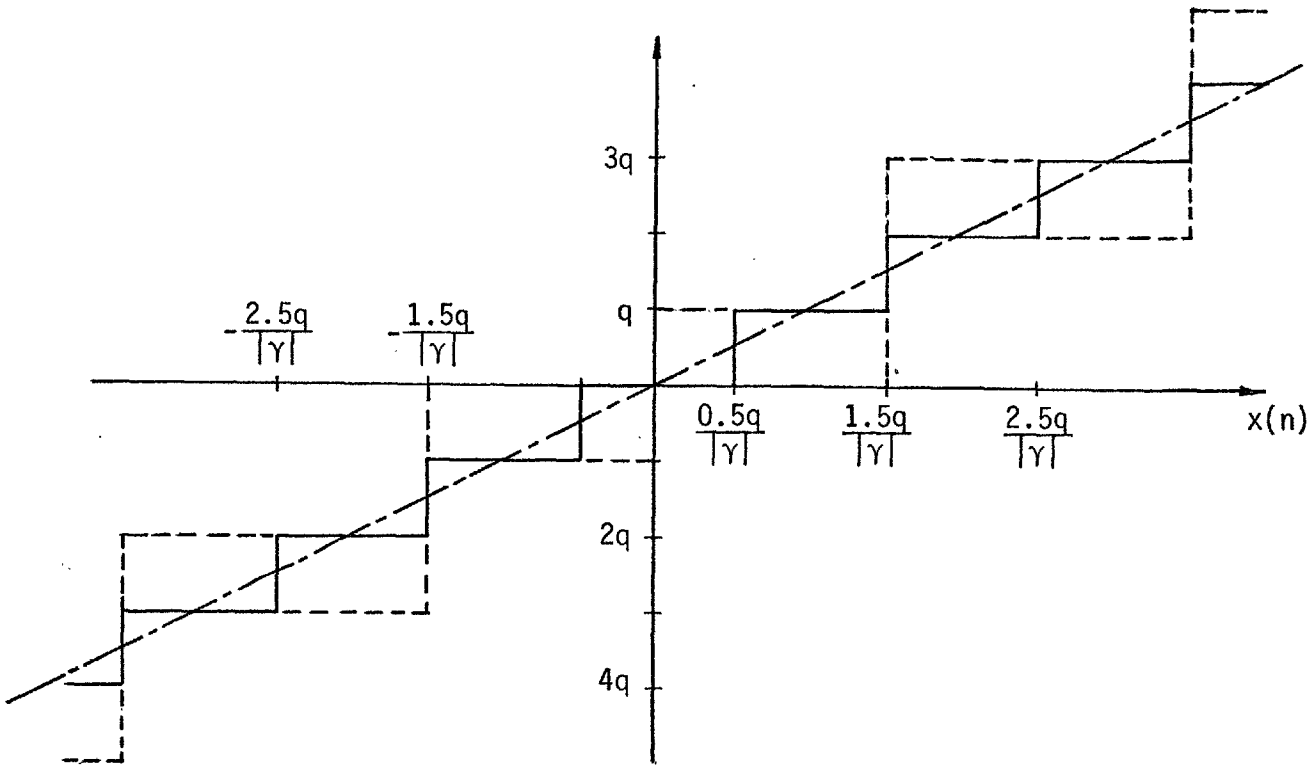


FIG. 5.7 The modified quantization function when jittering the LSB

Thus the increase in the noise level in the branch where this technique is used can be shown to be about 8.45 dB for the cases when $q_c \ll 1$. Therefore, compared to the increase of about 3.01 dB for the method of signal addition suggested in this chapter, this simple technique has a poor noise performance.

Based on this method and the signal addition method of Fig. 5.1a, the dithering of the suitable bits of the input to the multiplier may be suggested. One example of this is shown in Fig. 5.8. Here a random signal is generated by a shift register sequence generator [22]. An effective signal addition is obtained by the exclusive-OR (XOR) operation

of the suitable bits of the input to the multiplier and the random sequence. The bits of the input to the multiplier which are dithered depend on the required amplitude of the stabilizing signal.

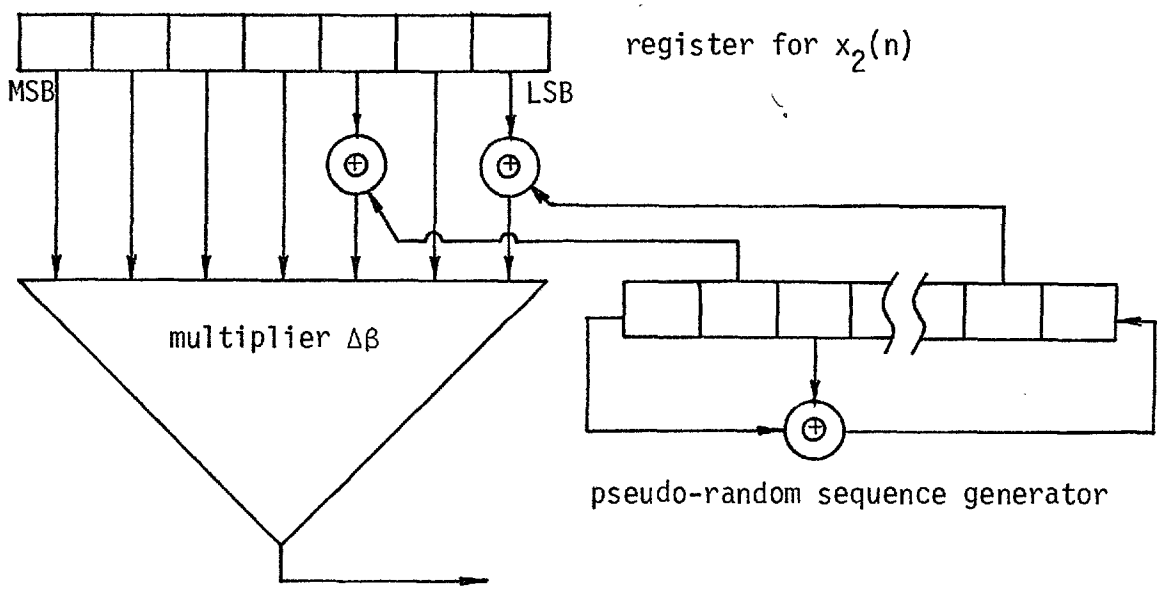


FIG. 5.8 A simple hardware to implement the signal stabilization method

Other alternative gates, for example AND, OR, etc. may also be used; however, the probability distribution of the effective signal may not be symmetrical and therefore result in a shift in the average value of the output.

This technique,when implemented in the form of the structures shown in Figures 5.3 and 5.6, involves one extra single precision addition, a random sequence generator and the gates necessary to implement the addition process.

5.7 Conclusions

Based on the results of Chapter 4, for suppressing limit cycle oscillations it was found sufficient to only interfere with the multiplier branch corresponding to the coefficient $\Delta\beta$ for the type-A and Δb for the type-B realizations. The structures shown in Figures 5.3 and 5.6 were suggested for the implementation of signal stabilization technique in the second-order filter structures.

Extending the analysis technique presented in Chapter 4, it was possible to estimate the amplitude of the limit cycle oscillations when insufficient signal stabilization is used (Table 5.2). From these results a suitable amplitude for the stabilizing signal to suppress limit cycles was suggested, see Table 5.1 for the type-A realization, and conditions (5.29), (5.30) and (5.54) to (5.57) for the type-B realization.

In Section 5.6 the criteria for the choice of the stabilizing signal were considered. It was concluded that the suitability of a signal stabilization technique should be based on the extra hardware necessary for its implementation.

A brief consideration of the simplest method for signal stabilization (i.e. the random dither of the LSB of the output of the multiplier) showed that this method had a poor noise performance. The method shown by Fig. 5.8 was therefore suggested. This used the concepts described in this chapter and required a simple hardware.

CHAPTER 6

SIMULATION STUDIES

6.1 Introduction

To develop further properties of the limit cycle oscillations, and to confirm the results from the previous chapters, second-order recursive filters of both the type-A and type-B realizations were simulated on a digital computer. The results of these simulations, which are related to the works of the previous chapters have all been presented in this chapter for convenience.

These simulation results relate to some of the studies of Chapters 3 to 5 and also present further investigations which arise from this work. The results in the previous chapters which were developed in detail, for example the behaviour of one- or two-sample limit cycles, were not investigated in the simulation studies.

For the simulation studies three different classes of programs were developed. These were written in Fortran IV language and were executed on the computers of Imperial College and Adelaide University (Cyber 173/CDC 6400).

The first class of programs was developed for investigating the zero-input Limit Cycle (LC.) behaviour of a filter when it is operated at different Initial Conditions (ICS.). These programs provided a comprehensive search through all initial conditions in a specific search area for a number of different filter coefficients. The general flow diagram for these programs is shown in Fig. 6.1 and

the results are discussed in Section 6.3.

The second class of programs was developed to investigate the suppression ability of the stabilization method suggested in chapter 5. The zero-input responses of both type-A and type-B structures were studied for different amplitudes of the stabilizing signal and the coefficients of the filter. The flow diagram for these programs is shown in Fig. 6.2 and the results are discussed in Section 6.4.

The third class of programs was developed to examine the noise performance of the stabilizing method. In these tests the filter was excited by a random signal with uniform probability and spectral distributions. The flow diagram for these programs is shown in Fig. 6.3, and the results are investigated in Section 6.4.

6.2 The programming techniques

Several factors had to be considered for simulating the digital filter algorithm on the computer. These were

i) the choice of the quantization steps for the signal and the coefficients of the filter,

ii) the routines for the quantization technique,

iii) the detection of limit cycle oscillations and their identification,

iv) the generation of the random sequences for the various signals in the stabilization tests.

Each of the above factors are discussed in the following sub-sections.

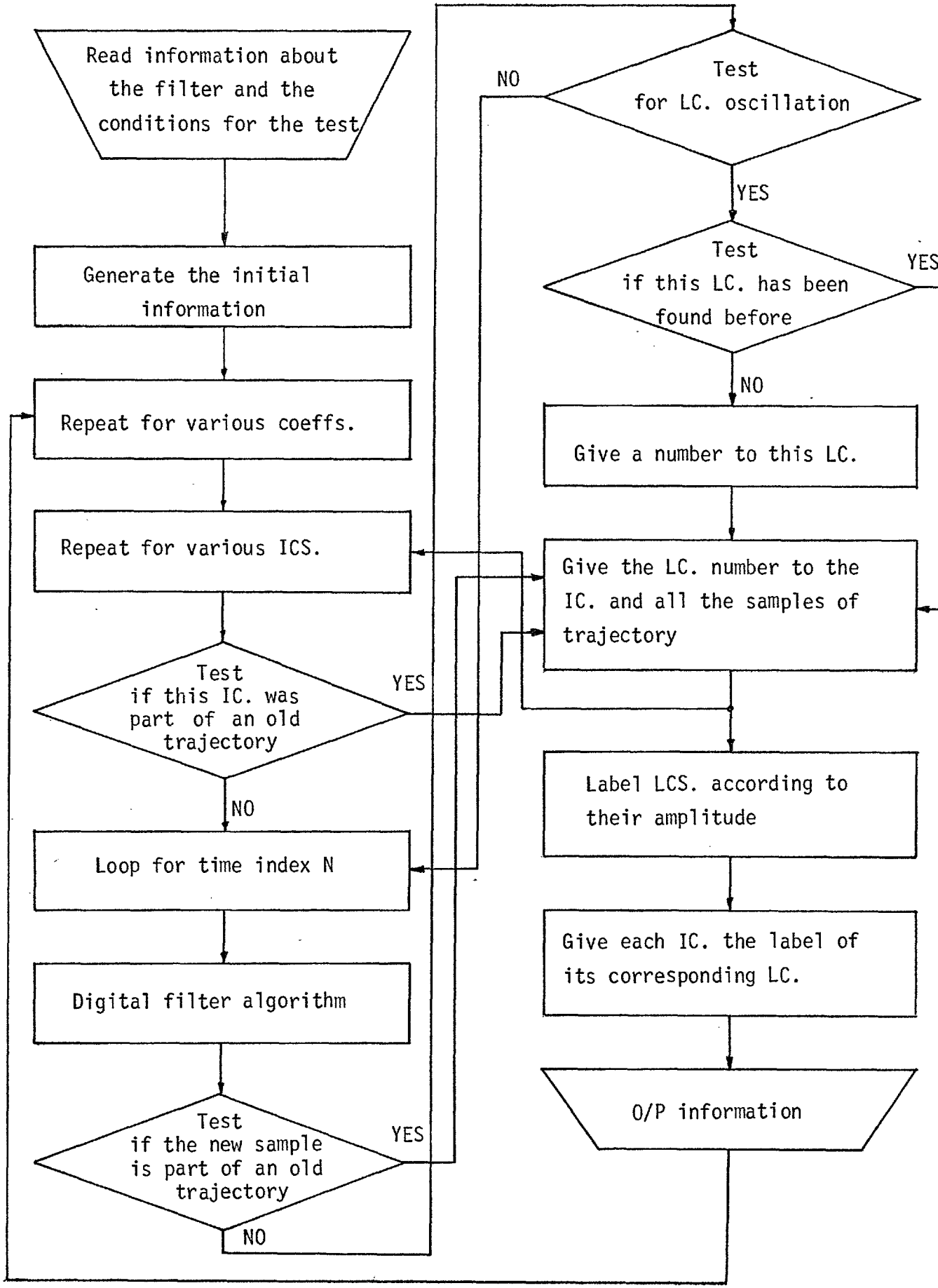


FIG. 6.1 The flow diagram for search through a range of initial conditions

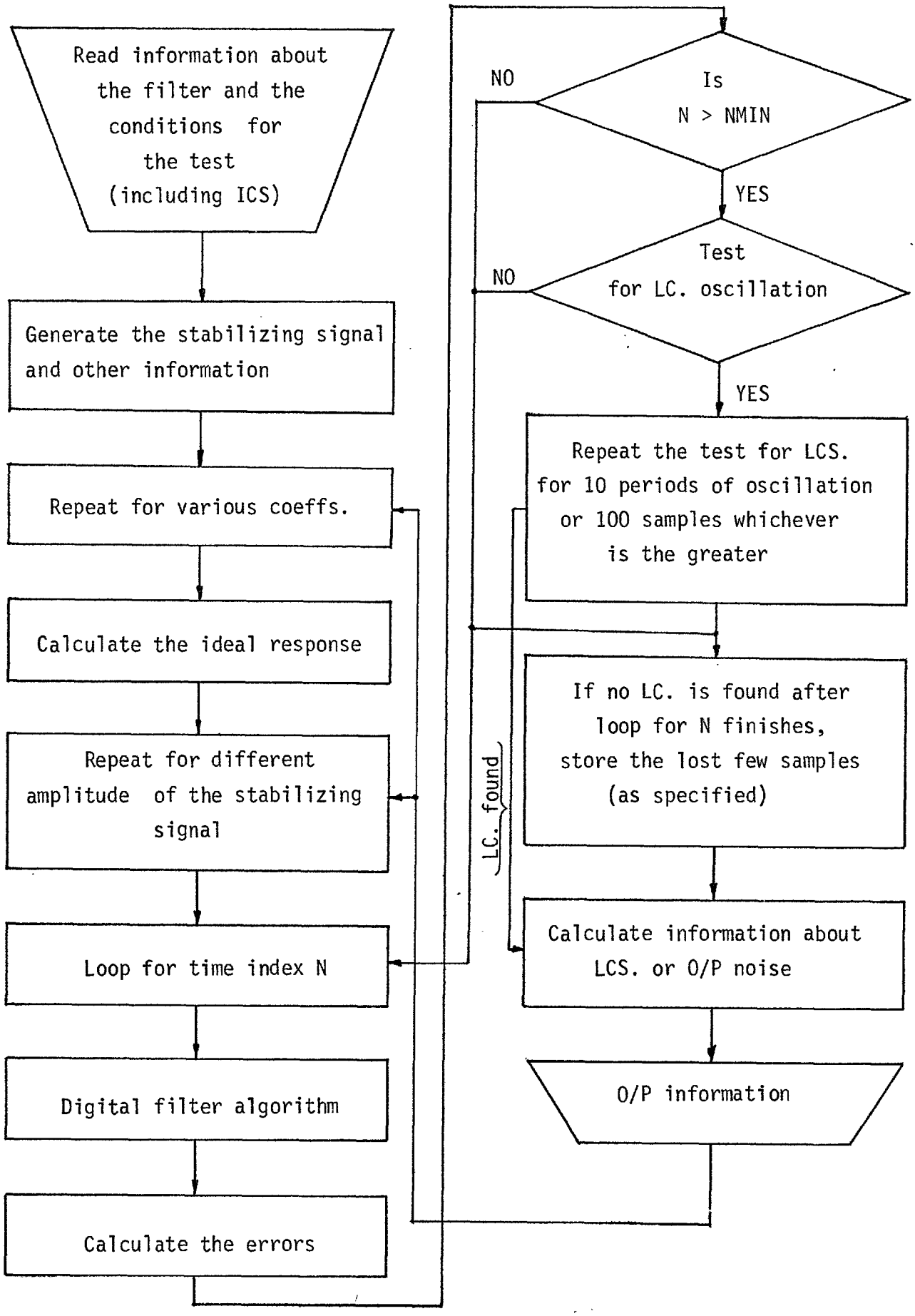


FIG. 6.2 Flow diagram for the test of signal stabilization

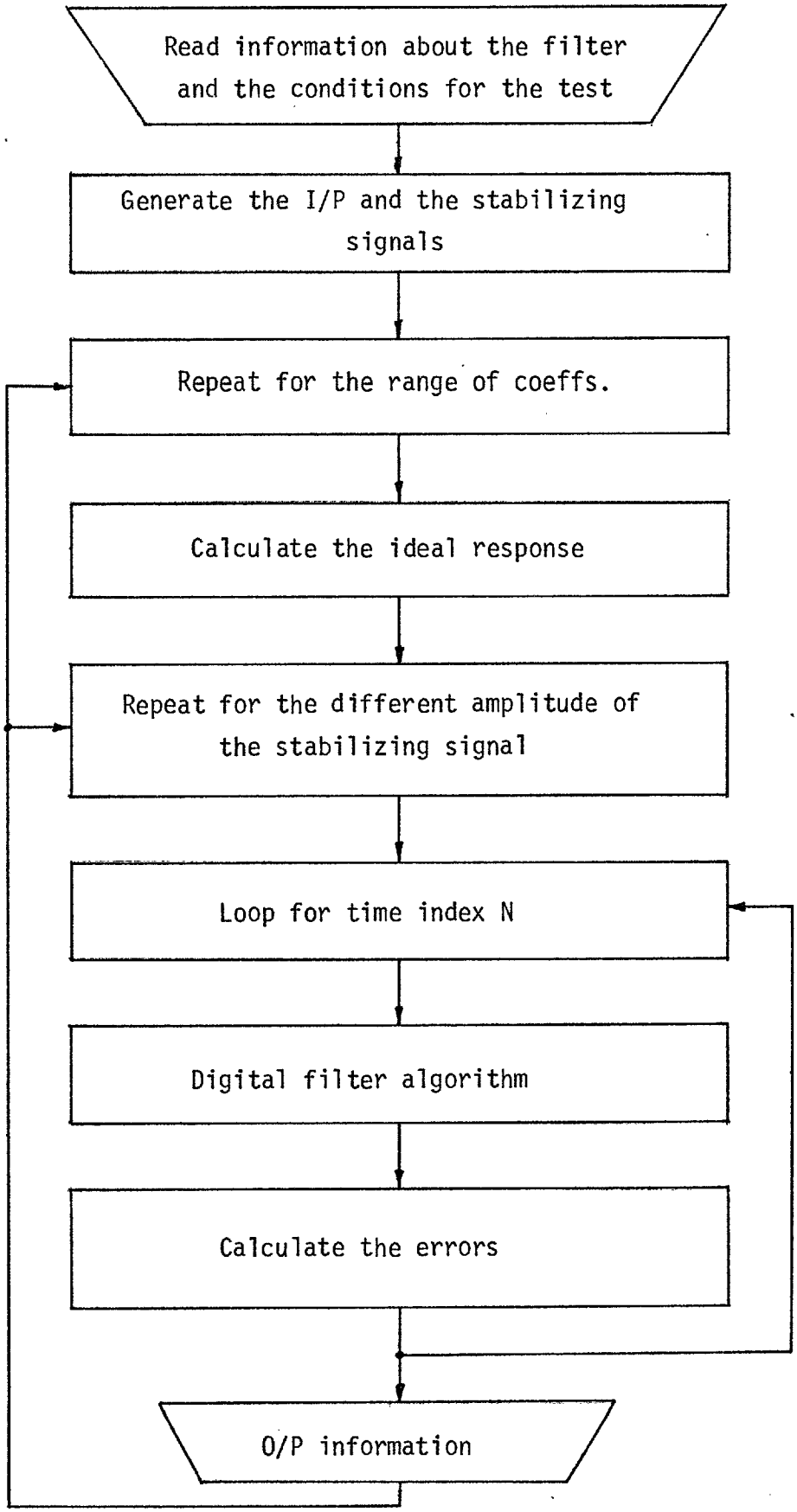


FIG. 6.3 The flow diagram for estimating increase in noise

6.2.1 Choice of quantization steps for the signal and the coefficients

Consider the properties of the quantization function as discussed in Section 1.4. If x is a quantized variable with the quantization step of q, and γ is any coefficient, we can write

$$Q_q[\gamma x] = Q_1[\gamma \frac{x}{q}]\, q \qquad (6.1)$$

where $Q_q[\ \ ]$ is any quantization function with the quantization step of q, and the quantity $(\frac{x}{q})$ is an integer variable. Therefore the simulation of the finite precision filter can be performed in integer arithmetic form, without loss of generality.

The sections of the programs used for evaluating the response of the filter are shown in Fig. 6.4 for type-A and type-B realizations. In this figure, the normal representation for defining integer variables in the Fortran language is used (i.e. variable names starting with letters I to N). The output of the filter, IY(N), was assumed to be the state variable $x_1(n)$. The quantized output of the multipliers were specified by the state variables IY1 and IY2, and the coefficient values were represented by IB1/MF and IB2/MF, where IB1 and IB2 are integer variables and MF is an integer multiplication factor.

The output of the multiplier branches were quantized by the subroutine QTZ, as given in Fig. 6.5, for the various quantization techniques.

6.2.2 Detection and identification of limit cycle oscillations

The limit cycle oscillations were detected by finding an equality between two samples in the phase plane. The details are given in Fig. 6.6.

When signal stabilization was used, the check for a limit cycle was repeated for 10 periods of oscillation or 100 samples, whichever

was the greater number (see Fig. 6.2). This was to ensure that a detected limit cycle was in fact the steady state oscillation for the filter. For example, in the cases of one or two sample limit cycles, the check was repeated for one hundred samples.

When searching for all the possible limit cycle oscillations in a specified search area in the phase plane, the amplitudes of the limit cycles were calculated and all limit cycles were categorised according to their amplitude (the largest limit cycle was given the label A, the next largest one was given the label B, etc.). Limit cycles of the same amplitude were categorised depending on which one was found first (i.e. first one found had a higher ranking label than the next).

To limit the computation time for search through a large number of initial conditions a method had to be developed to increase the speed of the programs. This was based on remembering all the samples which a transient trajectory passes through before it reaches the steady state value (see Fig. 6.1).

6.2.3 Random number generation

From the discussion of Section 5.6, the suitability of the various random number generator routines was not considered, and therefore the simplest form of these (i.e. RANF function) was chosen arbitrarily.

The RANF function for random number generation is a single congruential generator of nearly uniform random sequences with numbers in the range of zero to one. Fig. 6.7 shows how RANF was used to generate the various random sequences for the tests on the signal stabilization method.

The required random sequences were of three types. These were:

```
C   THE BASIC LOOP FOR THE EVALUATION OF THE NEXT
C   SAMPLE OF THE TYPE-A REALIZATION
      N1 = 2
      N2 = 1
      DO 1 N = 3,M
      IY1 = IB1*IY(N1)
      IY2 = IB2*IY(N2)
      CALL QTZ(IY1)
      CALL QTZ(IY2)
      IY(N) = - IY1 - IY2
      N2 = N1
      N1 = N
 1    CONTINUE


C   THE BASIC LOOP FOR THE EVALUATION OF THE NEXT
C   SAMPLE OF THE TYPE-B REALIZATION
      N1 = 1
      DO 1 N = 2,M
      IY1 = IB1*IX1(N1)
      IY2 = IB2*IX2(N1)
      CALL QTZ(IY1)
      CALL QTZ(IY2)
      IX1(N) = - IY1 - IY2
      IX2(N) = IX1(N1) + IX2(N1)
      N1 = N
 1    CONTINUE
```

FIG. 6.4 The digital filter algorithm

```
      SUBROUTINE QTZ(IA)
C   SUBROUTINE FOR ROUNDING THE QUANTITY FLOAT(IA)/FLOAT(MF)
      COMMON /Q/ MF,MF2
C   MF2 EQUALS MF/2
      IF(IA.LT.0)GO TO 1
      IA = (IA + MF2)/MF
      RETURN
 1    IA = (IA - MF2)/MF
      RETURN
      END


      SUBROUTINE QTZ(IA)
C   SUBROUTINE FOR VALUE TRUNCATION OF FLOAT(IA)/FLOAT(MF)
      COMMON /Q/ MF, MF1
C   MF1 EQUALS MF - 1
      IF(IA.LT.0)GO TO 1
      IA = IA/MF
      RETURN
 1    IA = (IA - MF1)/MF
      RETURN
      END
```

FIG. 6.5 The quantization routines

```
C     TEST FOR LIMIT CYCLE OSCILLATIONS , TYPE-A
        I1 = NMIN - 1
        DO 1 I = NMIN,N1
        IF(IY(I).EQ.IY(N).AND.IY(I1).EQ.IY(N1))GO TO 2
        I1 = 1
 1      CONTINUE


C     TEST FOR LIMIT CYCLE OSCILLATIONS , TYPE-B
        DO 1 I = NMIN,N1
        IF(IX1(I).EQ.IX1(N).AND.IX2(I)EQ.IX2(N))GO TO 2
 1      CONTINUE
```

NOTES:
1) LOOP FOR I IS INSIDE THE LOOP FOR THE TIME INDEX N
2) IF THIS LOOP FINISHES THEN LC. IS NOT REACHED
3) STATEMENT 2 IS REACHED WHEN A MATCHED CONDITION IS FOUND
   IF SIGNAL STABILIZATION IS USED THEN A COUNTER IS STARTED
   TO CHECK THIS CONDITION FOR SEVERAL SAMPLES

FIG. 6.6 The procedure for detecting limit cycle oscillations

```
C     TO GENERATE A SEQUENCE OF RANDOM NUMBERS
        DO 1 N = 1,NPTS
        S(N) = RANF(1.0)
        JS1(N) = INT(S(N)*2.9999)-1
        KS(N) = INT(S(N)*1.0E 14)
 1      CONTINUE


        AS = FLOAT(2*IAS) + 0.9999


C     TO GENERATE A RANDOM SEQUENCE OF INTEGERS
        IS = INT( S(N)*AS) - IAS


C     TO GENERATE A RANDOM SEQUENCE OF VALUES +IAS, -IAS, AND ZERO
        IS1 = JS1(N)*IAS


C     TO STIMULATE THE HARDWARE SHOWN IN FIG. 5.8
        IS = KS(N).AND.IAS
        IX = XOR(IY(N2),IS)
        IY2 = IB2*IX
```

NOTE: (IS) IS GENERATED IN THE LOOP FOR THE TIME INDEX N

FIG. 6.7 Random signal generators

i) random sequences of amplitude A_s and of integer quantities with nearly uniform probability distribution,

ii) random sequences of values $+A_s$, $-A_s$ and zero, each value having equal probability,

iii) one sided random numbers for use with tests on value truncation.

For tests on the performance of the hardware suggested in Section 5.6 (i.e. Fig. 5.8) large amplitude random sequences of integer values were generated and used in the form shown in Fig. 6.7.

6.3 Scan of initial conditions

The purpose of these tests was to verify some of the results of Chapter 3 and Chapter 4, and to investigate other properties of limit cycle oscillations. These tests were based on a thorough search of initial conditions for a large number of coefficients.

To limit the computation time, the tests were divided into two parts.

i) The first part of these tests was designed to search for the complete family of limit cycle oscillations corresponding to a limited number of coefficient pairs. This was achieved by the analysis of every initial condition in a suitably large area in the phase plane enclosing the origin. This area covered all possible limit cycle trajectories. Some examples of the results obtained are given in Figures 6.8 to 6.11.

In these results the initial conditions in the search area are represented by the label of their subsequent limit cycle oscillations. For each test the complete pattern of the initial conditions is presented. In the same presentation, for ease of comparison, the

pattern for all initial conditions which correspond to some of the individual limit cycles are also shown. The limit cycles selected are all of the accessible limit cycles and a few of the larger inaccessible limit cycles.

In the separated patterns of initial conditions, the samples of the limit cycles are represented by the asterisk symbol (*). The sequence for these samples is indicated by the solid lines joining these asterisks. In the figures corresponding to the type-A realization, the boundaries for the p-points in the search area are also shown (the broken lines enclosing x_2 values of (+9, +10), and (-9, -10)). The conclusions for these results will be given later in this section.

ii) The other part of these tests was set up to obtain information about the behaviour of the most popular limit cycle oscillations (i.e. accessible limit cycles). This was performed for a large and comprehensive range of coefficient pairs. To limit the computation time for these results, the search area was limited to the boundaries for the p-points so that only the accessible limit cycles could be detected. Further, to ensure that all of the accessible oscillations were detected, the search was carried out within the first two positive boundaries covering the energy level Γ (as defined in Section 3.2) of at least five times that of the minimum energy in those bands. Referring to the definition for Γ, we note that by this choice we would find all accessible limit cycles (if any) of amplitude at least five times greater than the lower limit of the second boundary for the p-points. This was considered sufficient for detecting all accessible limit cycles. Examples of these results for the cases of rounding and value truncation are shown in Figures 6.12 and 6.13 respectively.

To restrict the number of possible cases for these tests, they were mainly carried out for the coefficients corresponding to complex

conjugate poles with angles of less than 90° (i.e. $4\beta > \alpha^2$, and $\alpha < 0$), and 7 bit wordlength. In the case of the type-A realization with rounding, it should be noted that the family of the limit cycles for the coefficient pair (α,β) are of the same amplitude as the family of limit cycles with the coefficient pair $(-\alpha, \beta)$. Therefore, for this example it is not necessary to cover both positive and negative values of α.

From the results of Fig. 6.8 to 6.13 and other trial tests, the following conclusions could be made about the behaviour of limit cycle oscillations in second-order filters.

i) The results discussed in Section 3.6, concerning accessibility and inaccessibility of limit cycle oscillations in the type-A realization, were found to agree with the simulation results, i.e. an accessible limit cycle had at least one sample in the regions for the p-points, inaccessible limit cycles did not have any samples in these regions.

ii) The inaccessible limit cycles were normally smaller in amplitude than the lower limit of the first boundary for the p-points (defined by the left hand side of the inequality (3.31a), when $j = 1$). However, this was not always true, and in some examples such as the limit cycle labelled "A" in the results shown in Fig. 6.10, inaccessible limit cycles were observed which did not obey this rule (note that the limit cycles of this amplitude were not detected in the tests for the results of Fig. 6.12). For the cases of one or two sample limit cycles, large inaccessible limit cycles were found more readily.

iii) Normally for the case of rounding, the amplitude kq of the most popular limit cycle oscillations for a majority of the coefficients was bounded by the upper limit of the first boundary for the p-points, i.e.

$$Kq \leq \left[\frac{1.5 - |\beta|}{1 - |\beta|}\right] q \ . \qquad (6.2)$$

Therefore, the condition (6.2) can be considered a good estimate for the amplitude of limit cycles in the type-A realization (for the case of rounding).

The condition (4.16), specifying the amplitude of limit cycle oscillations with period greater than two samples, is shown in Fig. 6.12b. We note that this condition only holds for most cases when $|\alpha| < 1$. This proves that more complex phenomena than those discussed in Section 4.2 set up limit cycle oscillations. However, later in this chapter (Section 6.4) we note that the results based on this idea are quite accurate when slight randomisation, such as the signal stabilization technique is used.

iv) The boundary defined by the condition (6.2) was exceeded in a number of cases; see Figures 6.9 and 6.12. In general it was found that limit cycles exceeding this limit had a smaller probability of occurrence. This is illustrated by the pattern for the limit cycle "A" in Fig. 6.9a. In very rare cases (e.g. α = -201/128, β = 121/128 in Fig. 6.12a) limit cycles which almost reach the second boundary for the p-points were observed. Therefore, the absolute limit $K_{abs}q$, for the amplitude of limit cycle oscillations in the type-A realization for rounding may be suggested to be

$$K_{abs}q < \left[\frac{|\beta| + 0.5}{1 - |\beta|}\right] q \qquad (6.3)$$

(i.e. substituting j = 2, in the left hand side of the inequality (3.31a)).

v) From Figures 6.12 and 6.13, we note that the amplitude of limit cycle oscillations in the type-A realization appear to be related to the bit-pattern of the coefficient α. The exact nature of this

relation is not known.

vi) The boundaries for the p-points (condition (3.31)) were verified by noting that in every case (i.e. Figures 6.8 to 6.10), the initial conditions with the same limit cycle labels always appeared in pairs in these regions.

vii) For value truncation, the amplitudes of the accessible limit cycles in the type-A realization are shown in Fig. 6.13, for $\beta = 121/128$ and a wide range of the coefficient α. It appears that the amplitude of a majority of these limit cycle oscillations satisfy condition (4.24). However, the absolute limit for the amplitude was found to be the condition (4.20).

viii) For magnitude truncation, tests with conditions similar to those for the results of Figures 6.12 and 6.13 were conducted, but only accessible limit cycles with a magnitude of one were observed. By a thorough search through a wide range of initial conditions, all accessible and inaccessible limit cycles were found to have period of oscillation of one or two samples only. This result agrees with the results of Kao [126]. Accessible limit cycles with amplitude greater than one were possible, but these were restricted to only a small range of coefficients.

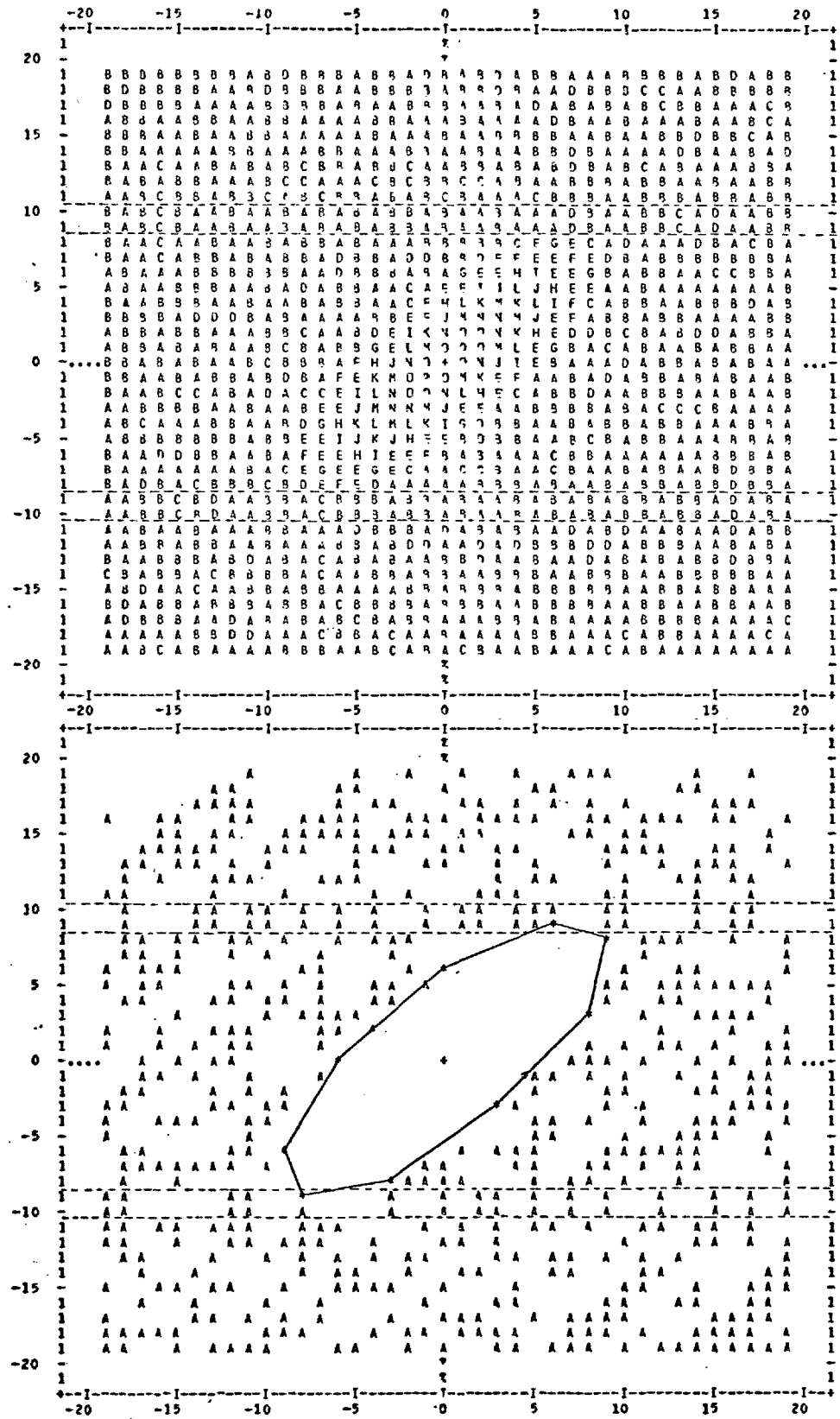
ix) For the type-B realization a search of all initial conditions in a suitable area of the phase plane gave the result of Fig. 6.11. From this result, three categories of limit cycles could be identified:

a) accessible limit cycles,

b) semi-accessible limit cycles,

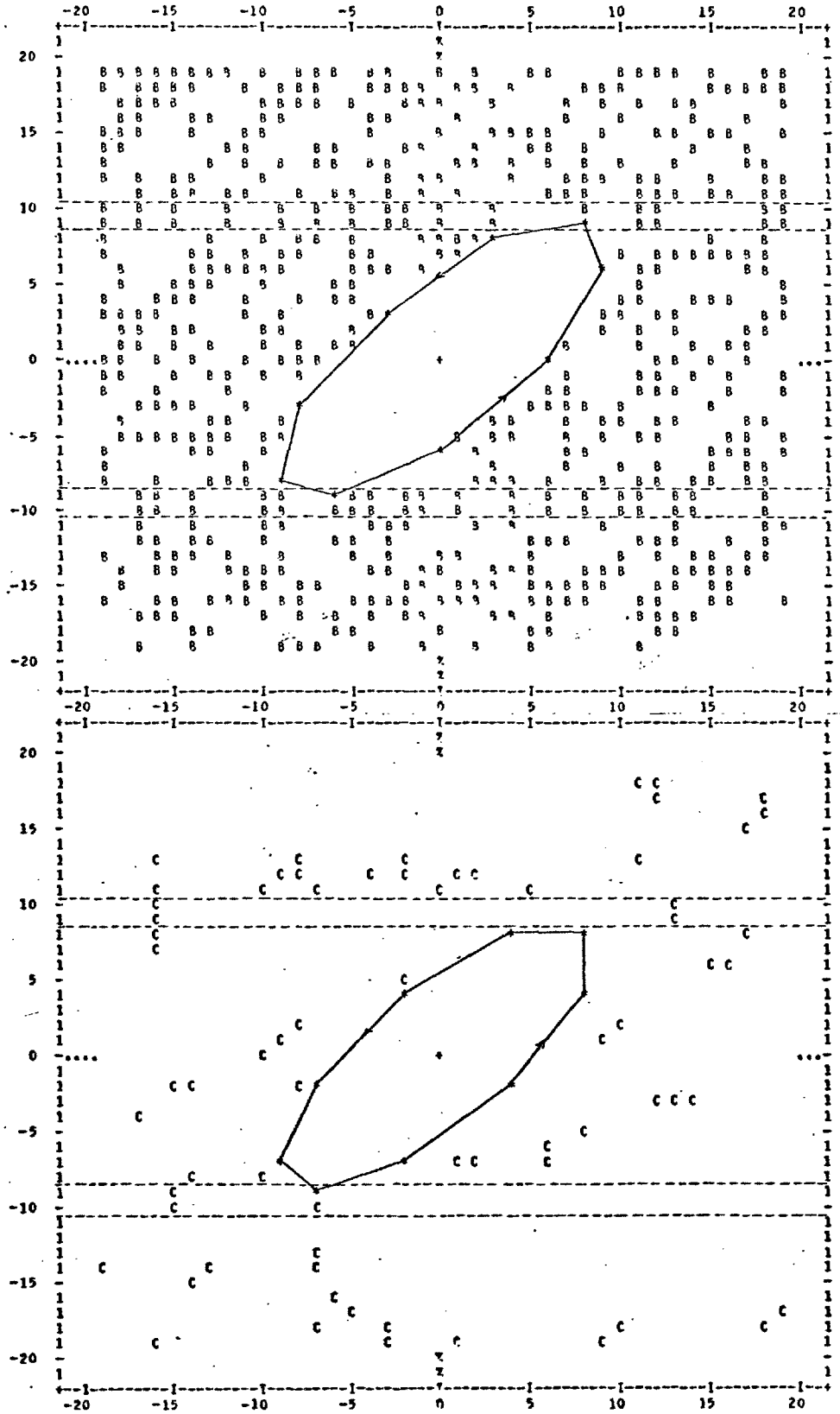
c) inaccessible limit cycles.

The definitions of the accessible limit cycles (limit cycle



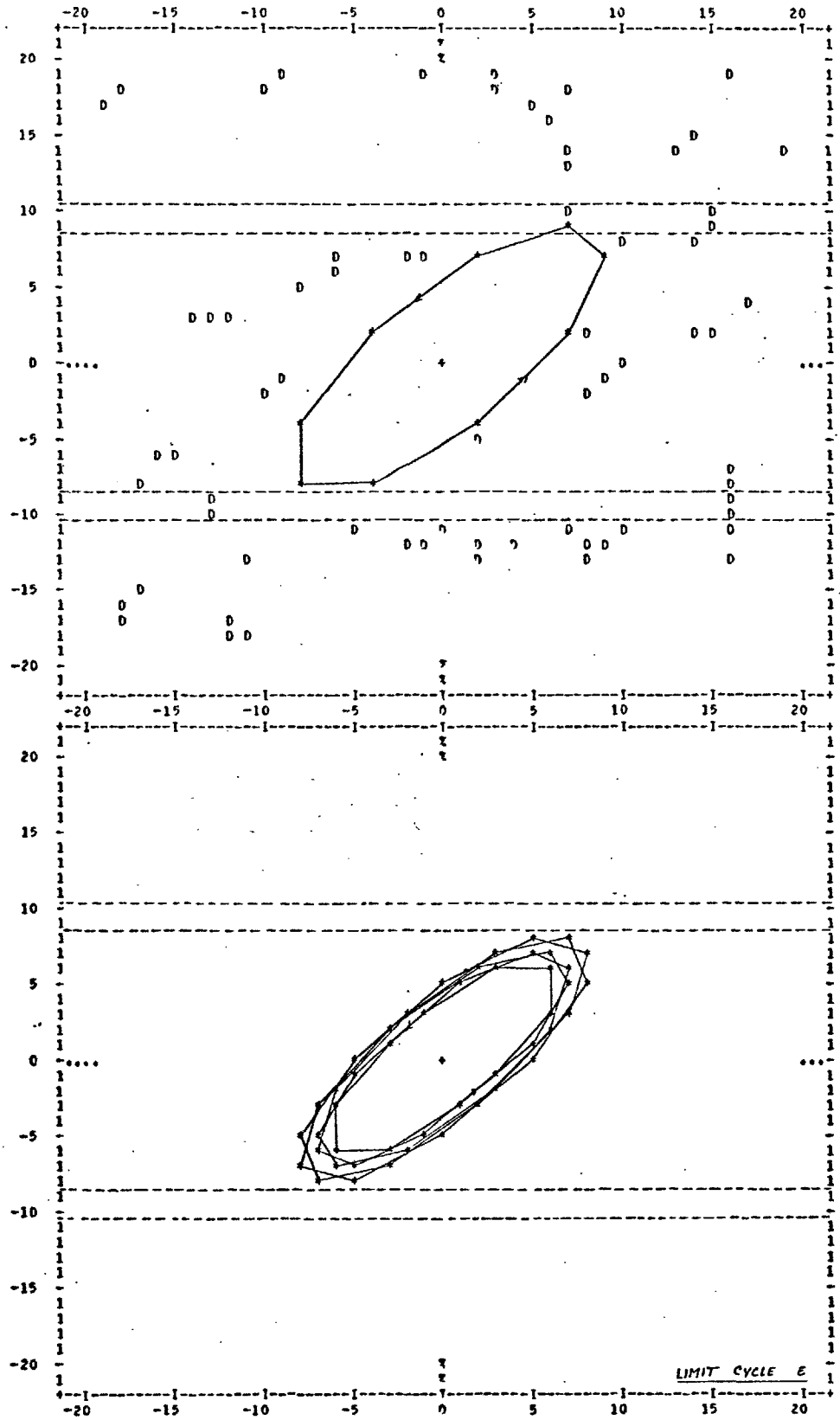
(a)

FIG. 6.8 Pattern of initial conditions in the phase plane labelled by their corresponding limit cycle label for the type-A realization with rounding, and coefficients α = -192/128 β = 121/128



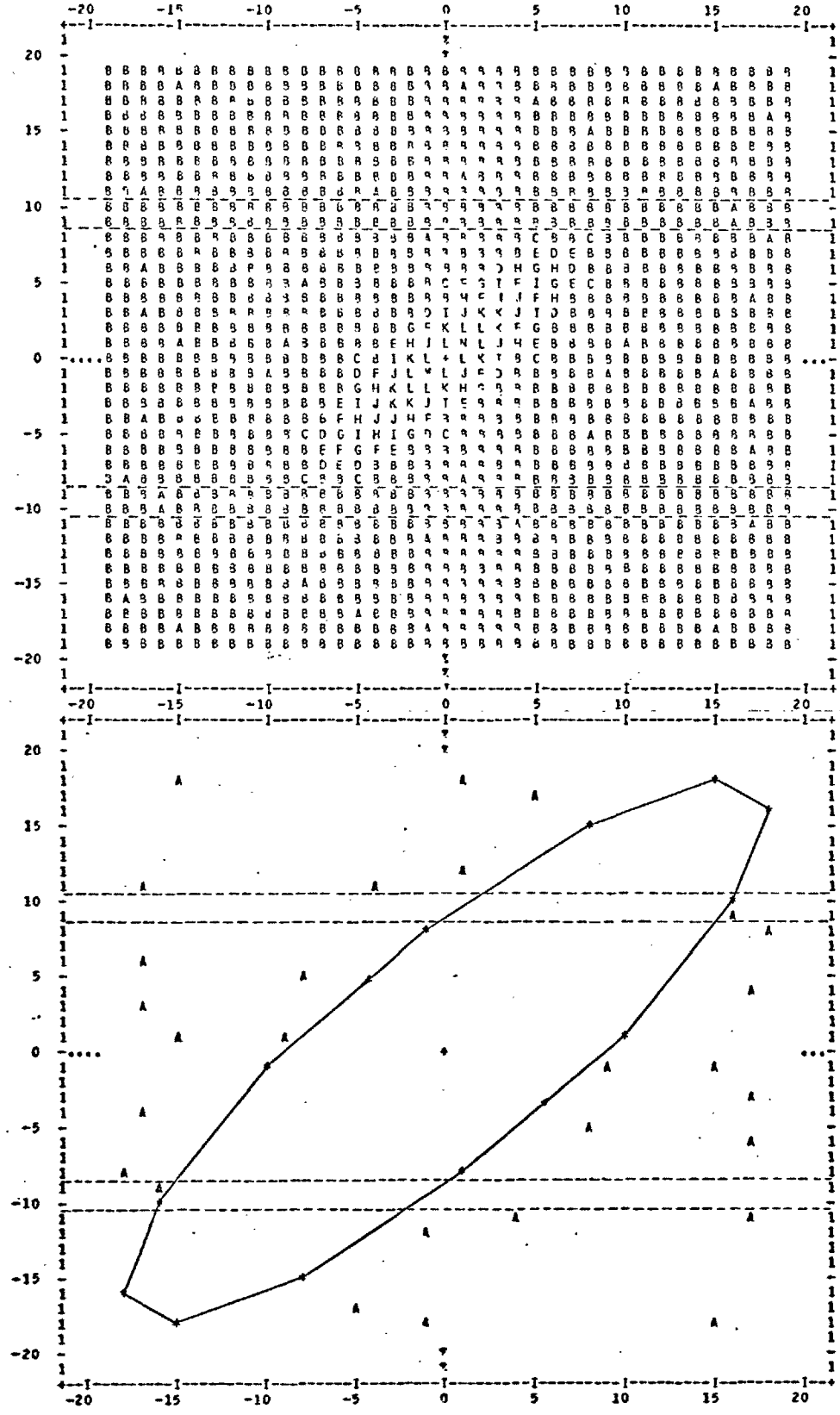
(b)

FIG. 6.8



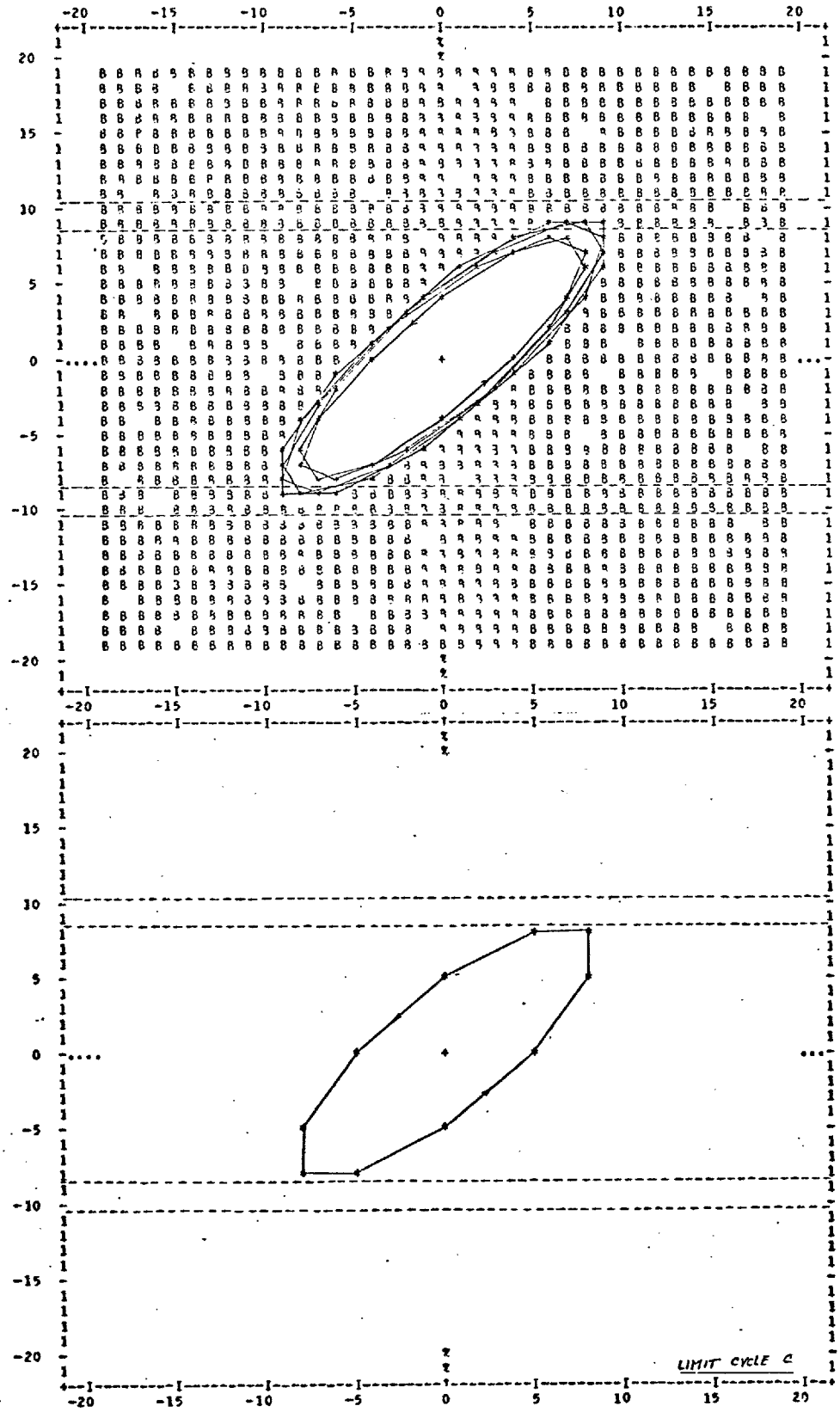
(c)

FIG. 6.8



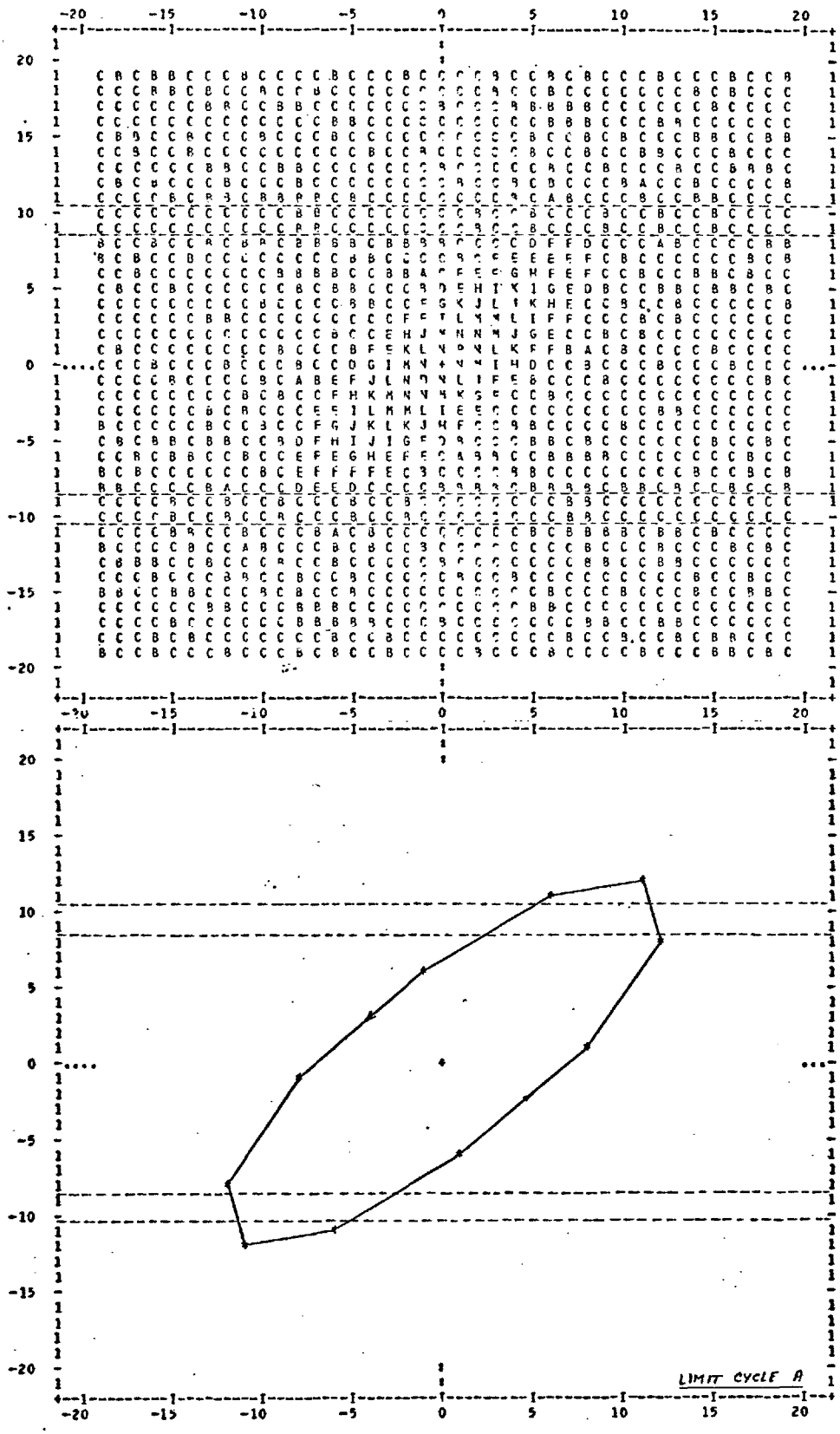
(a)

FIG. 6.9 Pattern of initial conditions in the phase plane labelled by their corresponding limit cycle label for the type-A realization with rounding, and coefficients α = -214/128, β = 121/128



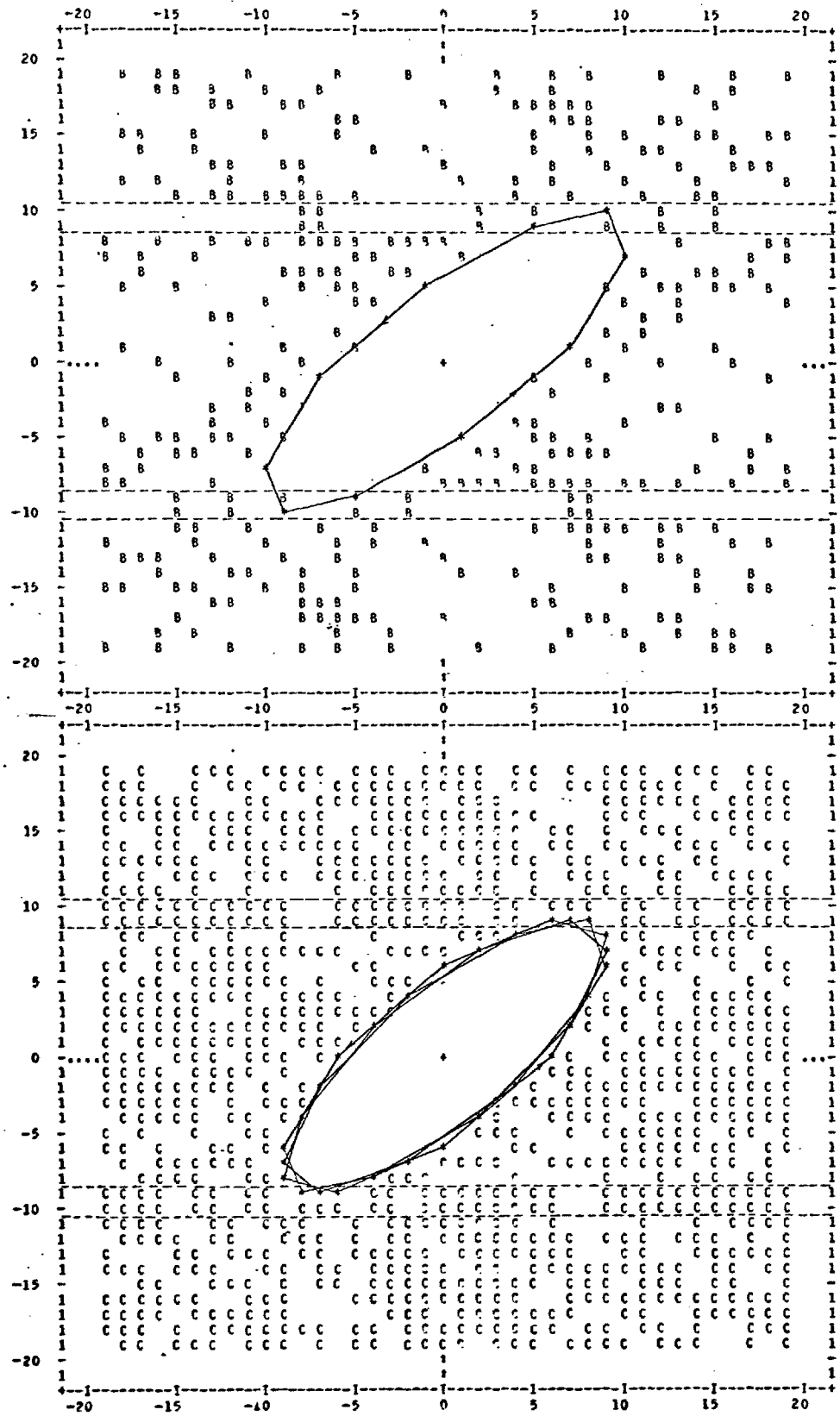
(b)

FIG. 6.9



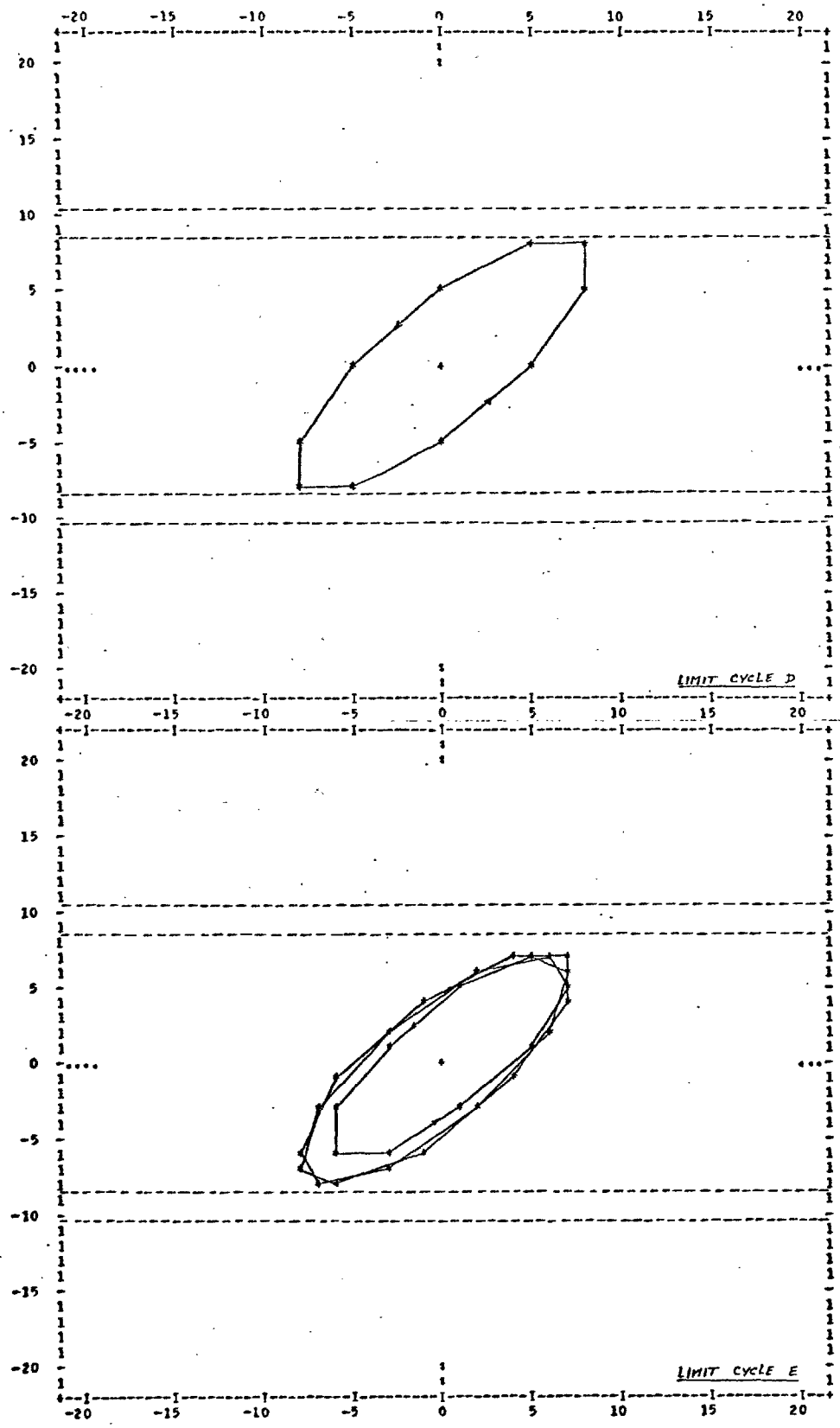
(a)

FIG. 6.10 Pattern of initial conditions in the phase plane labelled by their corresponding limit cycle label for the type-A realization with rounding, and coefficients $\alpha = -202/128$ $\beta = 121/128$.



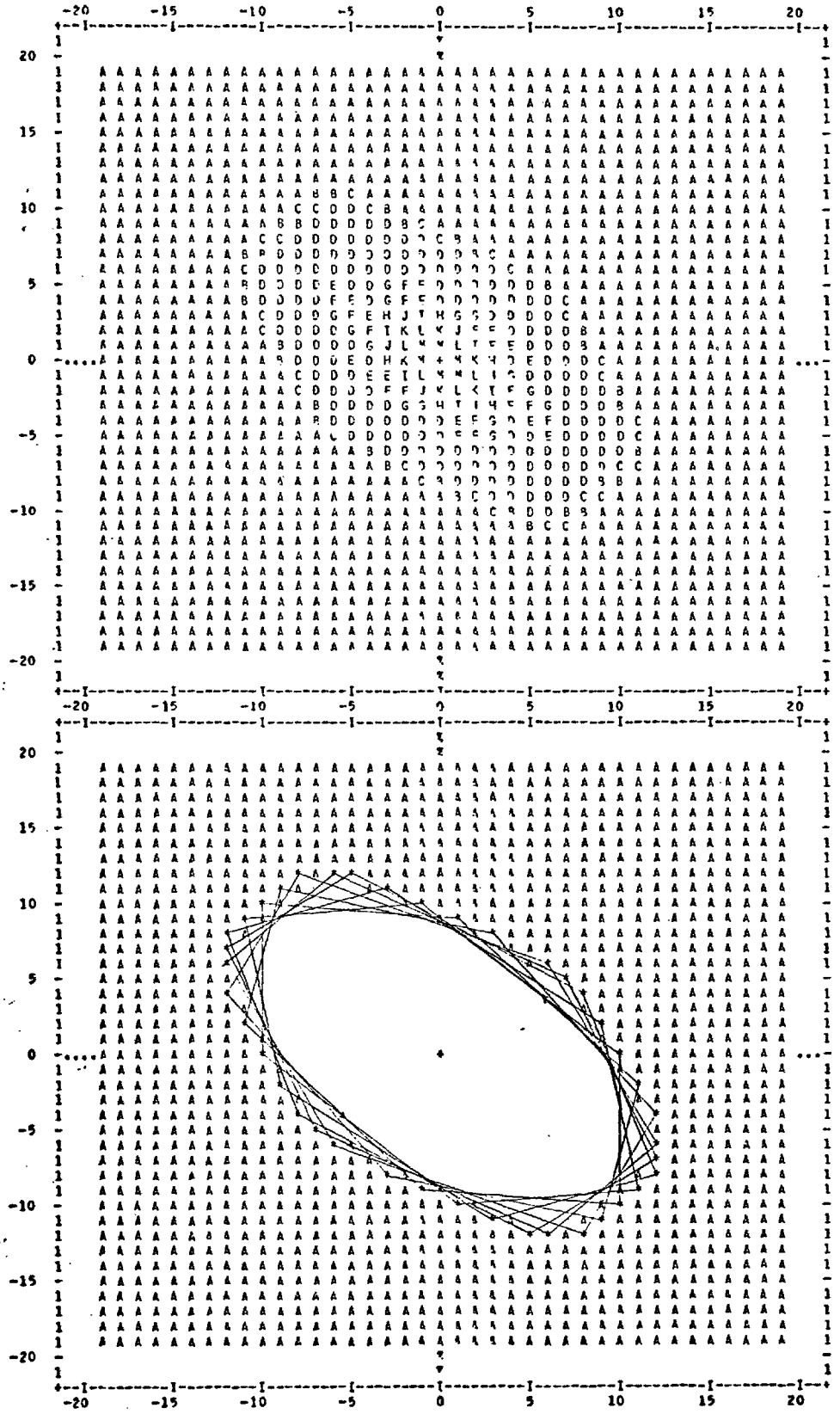
(b)

FIG. 6.10



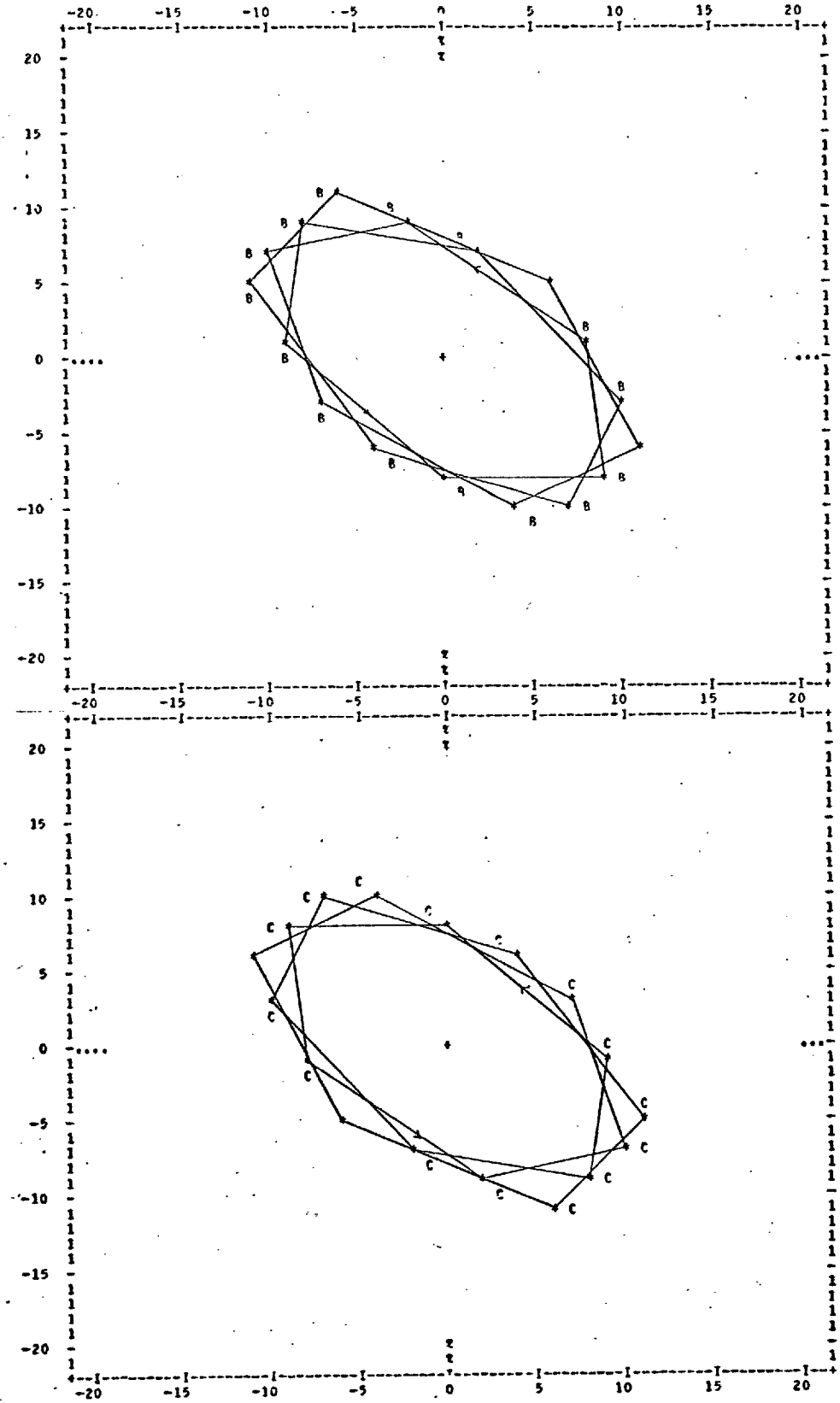
(c)

FIG. 6.10



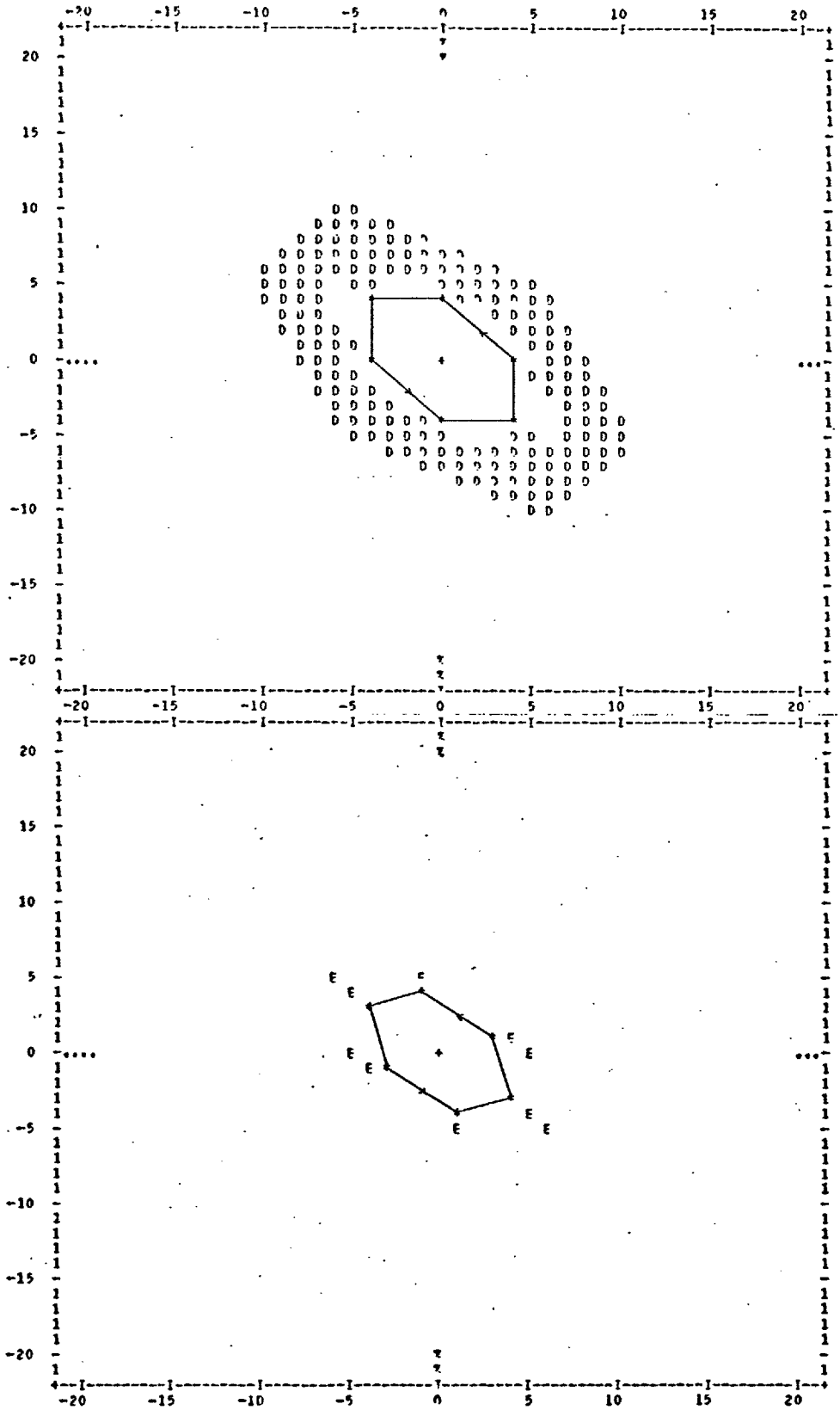
(a)

FIG. 6.11 Pattern of initial conditions in the phase plane labelled by their corresponding limit cycle label for the type-B realization with rounding, and coefficients a = 15/128, b = 139/128



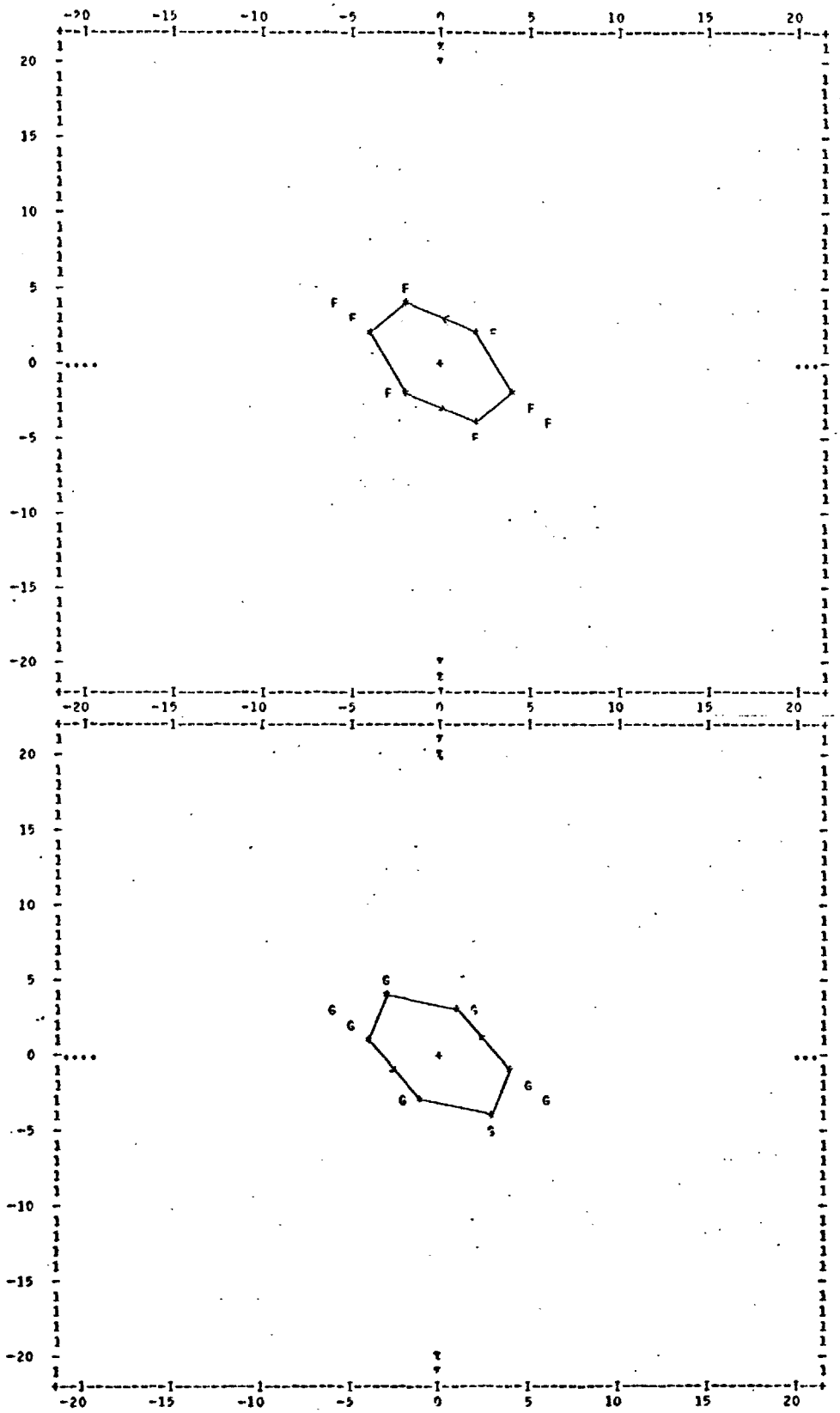
(b)

FIG. 6.11



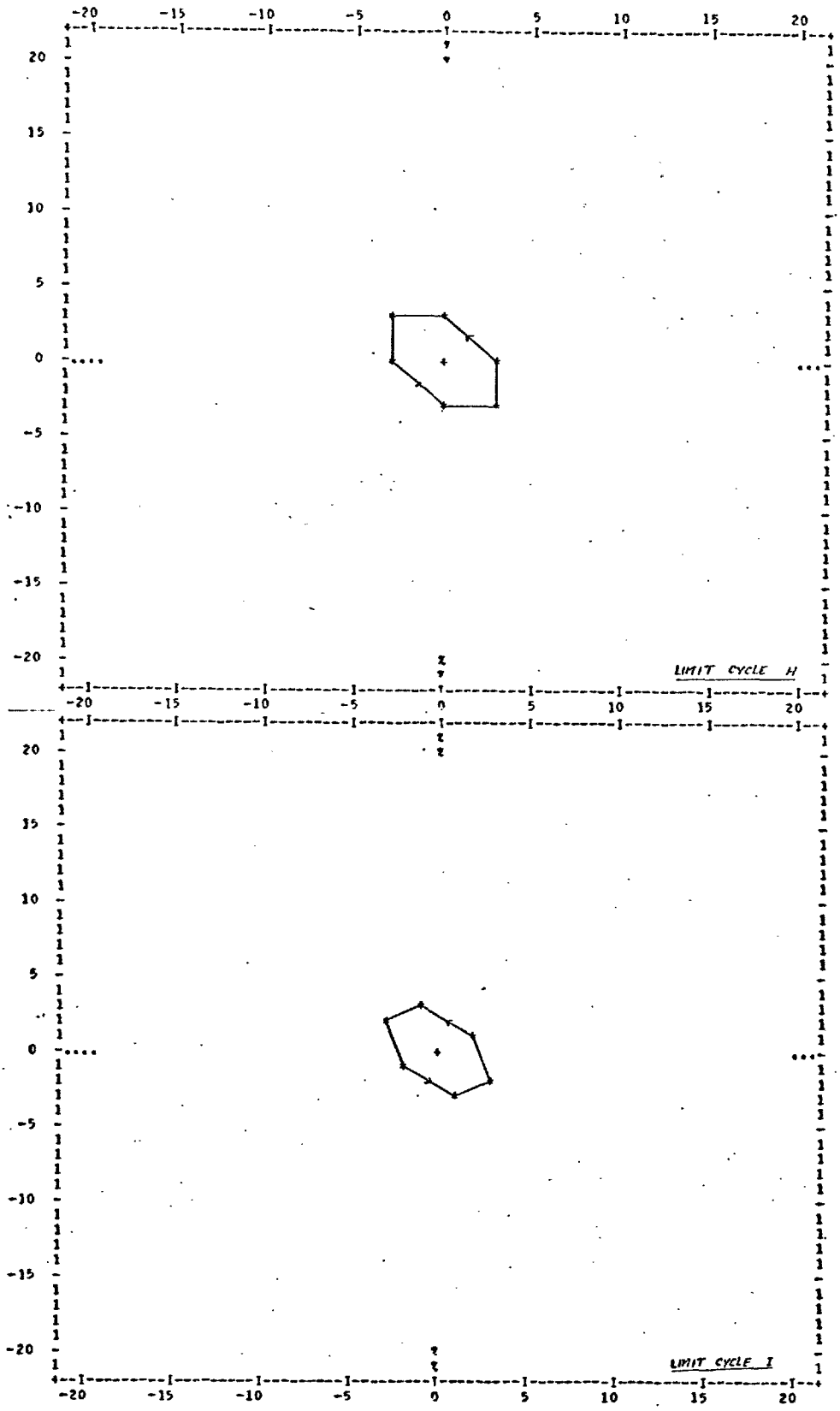
(c)

FIG. 6.11



(d)

FIG. 6.11



(e)

FIG. 6.11

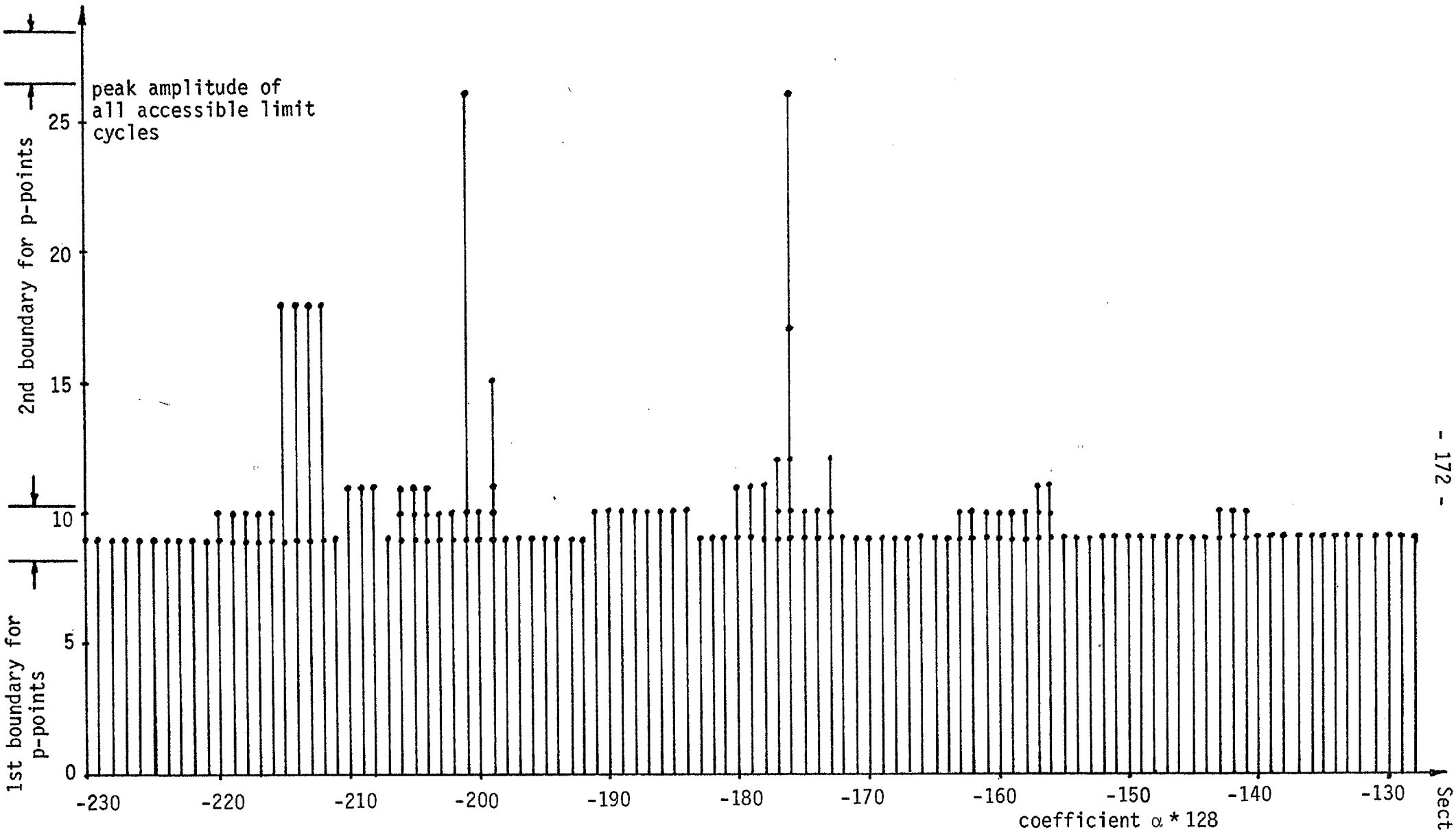


FIG. 6.12a Amplitude of all accessible limit cycle oscillations for type-A realization with rounding: α = -230/128 to -128/128, β = 121/128

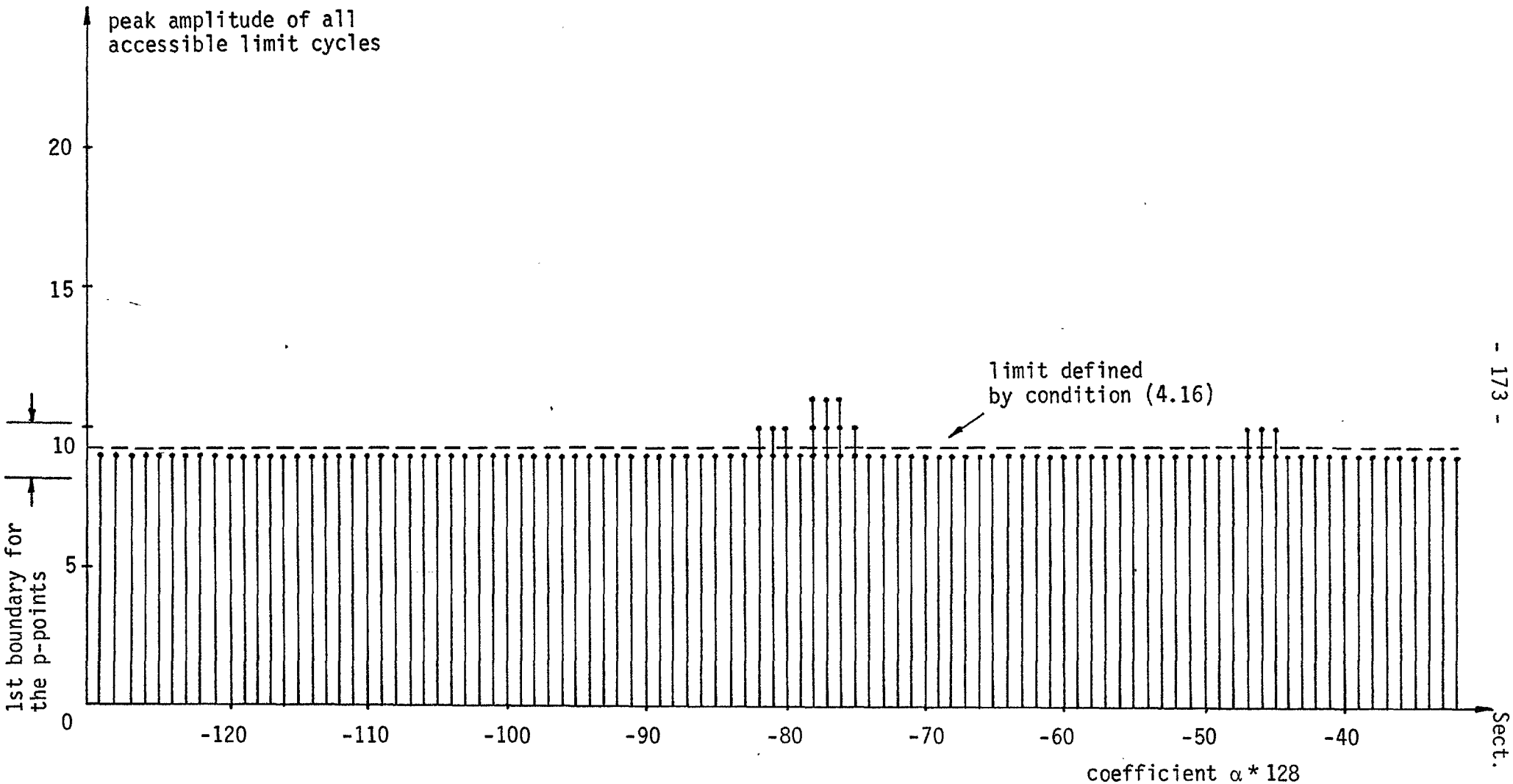


FIG. 6.12b Amplitude of all accessible limit cycle oscillations for type-A realization with rounding: $\alpha = -130/128$ to $-32/128$, $\beta = 121/128$

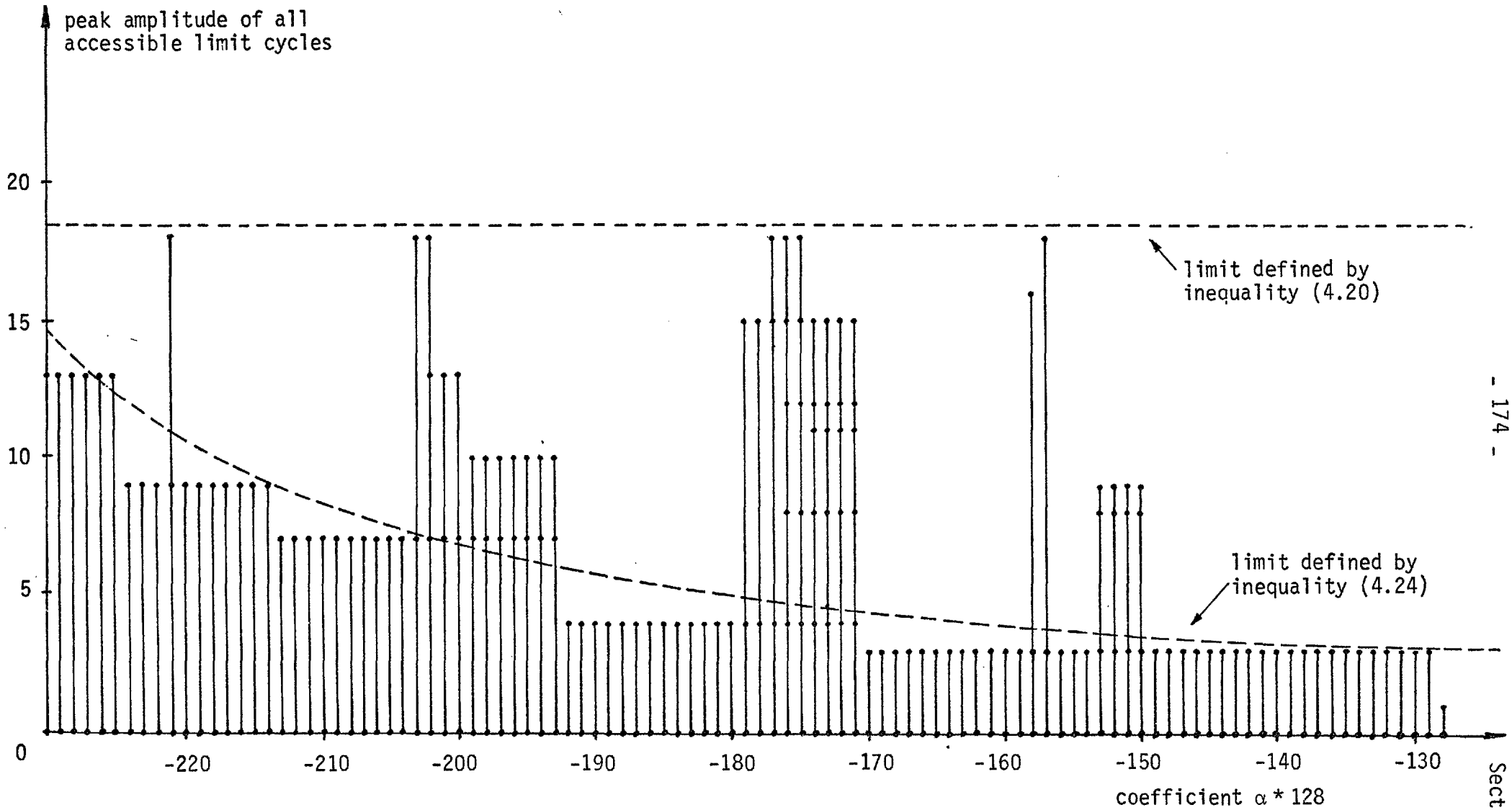


FIG. 6.13a Amplitude of all accessible limit cycle oscillations for type-A realization with value truncation: α = -230/128 to -128/128, β = 121/128.

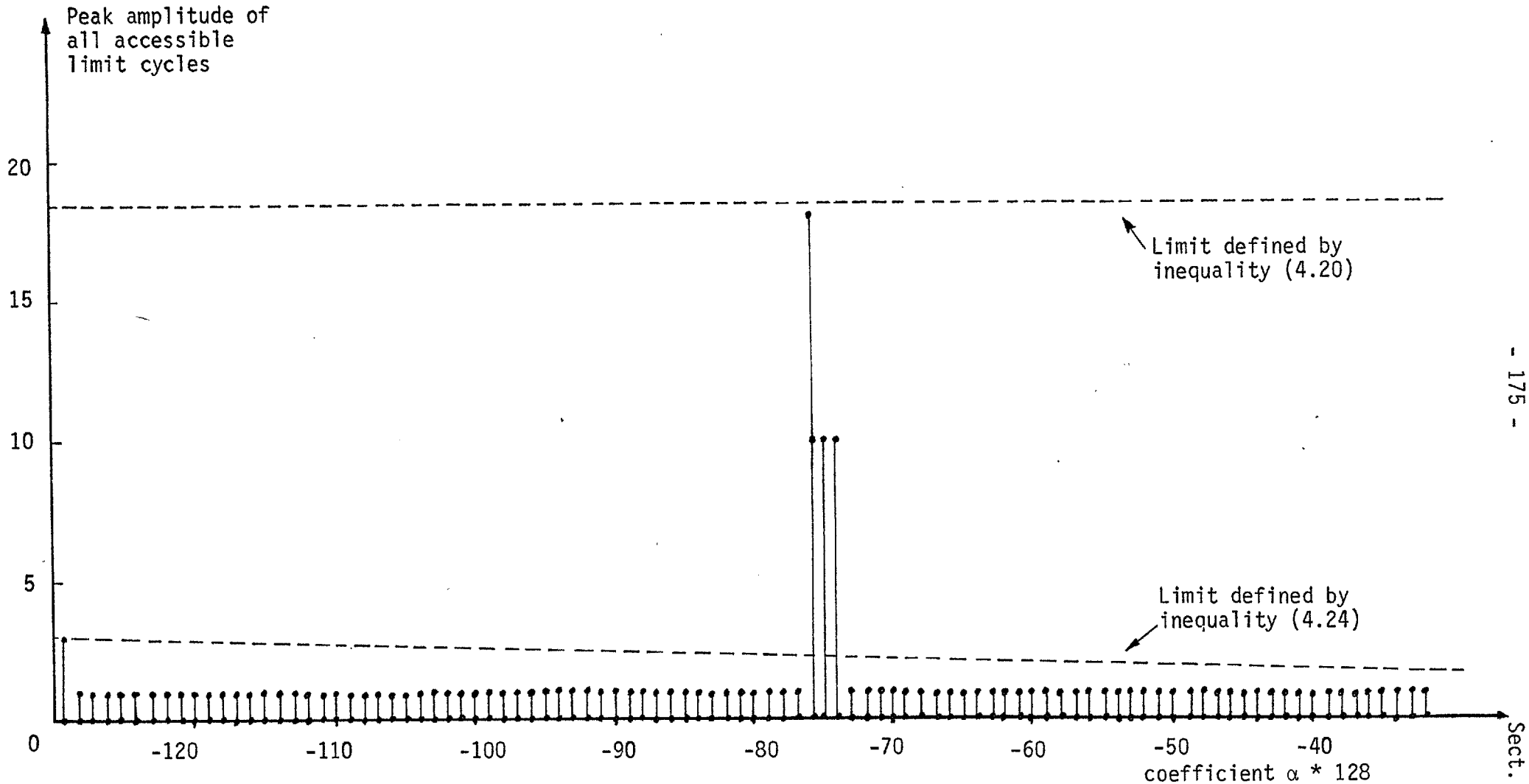


FIG. 6.13b Amplitude of all accessible limit cycle oscillations for type-A realization with value truncation; $\alpha = -130/128$ to $-32/128$, $\beta = 121/128$

labelled "A") are the same as the definitions given in Section 3.6. The "semi-accessible limit cycle" (limit cycles labelled "B" to "G") may be defined as: limit cycles which result from initial conditions that are not part of the trajectory of the limit cycle, and have a limited number of these conditions which can produce them (see Figures 6.11 b, c and d). The behaviour pattern of these conditions is not understood yet.

6.4 Suppression of limit cycle oscillations

Simulations were carried out to study the performance of the method of signal stabilization as presented in Chapter 5. The purposes of these tests were the following:

i) to experimentally verify the results of Section 5.3, i.e. the amplitude of the limit cycle oscillations when insufficient signal stabilization is used, and to verify the amplitude for the suitable stabilizing signal (see Tables 5.1 and 5.2),

ii) to verify the increase in the noise level due to the stabilization method, i.e. equation (5.39).

iii) to experimentally determine the performance of the simple techniques for implementing the signal stabilization method.

The results of these tests (i.e. Figures 6.14 to 6.24) were obtained and presented using the following methods:

i) The performance parameters (for example, amplitude of the limit cycle oscillations or the amplitude of the output noise) were determined for a range of amplitudes of the stabilizing signal.

ii) The performance parameters were presented by averaging over a large number of coefficient pairs which were expected to give the same performance. For each set of results the maximum and the

minimum limits for the performance parameters were also presented for comparison.

The discussions of the individual results are given in the following sub-sections:

6.4.1 Amplitude of the suppressed limit cycle oscillations

Tests were performed to determine the steady state response of the second-order filter for an arbitrary initial condition that is expected to give an accessible limit cycle oscillation. The results for these tests are shown in Figures 6.14 to 6.16.

Figures 6.14 and 6.15 show the results for the type-A realization in the cases of rounding and value truncation respectively. These correspond to a comprehensive range of filters with the coefficients $\beta = 121/128$ and all possible values of α in the range -230/128 to zero at steps of 1/128 (i.e. all possible 7 bit coefficients in that range).

In the tests giving the results illustrated in Fig. 6.14, it was found that, for $0 \leq A_s < 9$ the steady state response was always a limit cycle oscillation, and for $A_s > 9$, the steady state response was noise. For $A_s = 9$, the steady state response was always suppressed to zero which agrees with the results of Section 5.3 (i.e. Table 5.1).

Further, it was observed that for $0 < A_s < 9$, a large majority of the test cases resulted in a limit cycle which corresponded to the lower limit shown in Fig. 6.14. This can be noted from the closeness of the average line to the lower limit. This limit is also the estimate for the amplitude of the suppressed limit cycles, as derived in Section 5.3 (see Table 5.2).

In a small number of cases a larger limit cycle was observed (the upper limit in Fig. 6.14). However, using a different random

sequence it was possible to suppress these to the amplitude corresponding to the lower limit. A similar effect was also obtained by extending the check for the limit cycle oscillation. For the results shown, a period of only 100 samples or 10 periods of oscillation was used. The extended check for limit cycle oscillation resulted in a large increase in the computation time, and therefore could not be used for all of the test cases.

The close agreement between the simulation results and the results derived in Chapter 5, shows that the arguments for the development of limit cycles with period greater than two samples (see Sections 4.3 and 5.3) are valid for all cases when a slight randomization such as the signal stabilization is introduced.

The suitable signal for suppressing the limit cycle oscillations in the case of value truncation (see Fig. 6.15) was given by the conditions (5.29) and (5.30). However, the result of Fig. 6.15 shows that similar effects could also be achieved with a signal of half that amplitude. In fact this should be preferred since the resulting increase in the noise level would be less.

For the type-B realization, the suppression ability of the signal stabilization was observed for every case tested. Some examples of these results are presented in Fig. 6.16. In some of the tests, which are not illustrated in Fig. 6.16, the complete suppression of the limit cycle oscillations was not possible. The exact reason for this is not understood.

6.4.2 Increase in the noise level due to signal stabilization

The noise performance of the signal stabilization techniques was studied by estimating the increase in the noise level on the state variable $x_1(n)$, when the input to the filter was an arbitrary random

signal with an arbitrary amplitude. The results for a large number of coefficient pairs, which were expected to give the same performance, were compiled and presented in the form shown in Figures 6.17 to 6.22. In these results three curves are shown:

i) the solid line curve representing the average result for all the cases which are expected to have the same performance.

ii) the broken line curves representing the maximum and the minimum result for all the coefficients tested. The theoretical results derived in Section 5.4 for the case of rounding are also presented for comparison.

In selecting the values for the coefficients in each test it was necessary to ensure that the base quantization noise (i.e. noise when $s(n) = 0$) generated by each multiplier branch was the same. This was to ensure that the increase of noise level for filters with higher base noise would not be less.

In the discussion of Section 2.3 it was mentioned that the coefficients with same q_c, as defined in equations (2.14) and (2.15), would have the same noise level. Therefore, the values of α considered were defined by $(2i + 1)/128$, where i is any integer. This choice gives $q_c = 1/128$.

For each coefficient, the results correspond to tests on 3000 samples. The average for all the coefficients tested would then be equivalent to the result for one coefficient when a much higher number of samples are used.

The results from these tests are as follows:

i) Fig. 6.17 shows the results for the increase in the noise level in the case of rounding for the type-A realization. The results calculated from equation (5.39) are also indicated for comparison.

ii) Fig. 6.18 shows results with test conditions similar to those of Fig. 6.17, but with an input signal of a different amplitude. We note that the input signal level did not alter the noise performance.

iii) Fig. 6.19 shows the increase in the noise level for the case of value truncation with a stabilizing signal defined by the conditions (5.29) and (5.30). In this particular case, we note that the method of signal stabilization caused a decrease in the noise level in the system. This phenomenon is due to the use of a one sided signal which caused a reduction in the DC offset inherent in value truncation.

iv) Three level signals, defined by $\pm A_s$ and zero, with equal probability were also used as a stabilizing signal. The results for noise increase in the cases of rounding and value truncation are given in Figures 6.20 and 6.21 respectively. In general the increase in the noise level for those types of signals was greater than for the signals quantized to the nearest quantization step (for example, for $A_s = 9$, the difference was about 1.1 dB). Comparison of the simulation results and the theoretical results as shown in Fig. 6.20, show a good agreement.

v) The increase in the noise level for the type-B realization was also tested. The results are shown in Figure 6.22. We note that for this type, the theoretical and the simulation results are in close agreement.

6.4.3 Simulation of the simple implementation of signal stabilization

Simulation tests were also conducted to determine the performance of the signal stabilization methods discussed in Section 5.6. For comparison purposes the test conditions and the result presentation were chosen similar to those for the cases described in the previous sub-sections.

The first test was to determine the performance of the method based on jittering the LSB of the quantized output of the multiplier. The test was for 100 coefficient pairs with β = 121/128 and α from -221/128 to -23/128 at steps of 2/128. The average increase in the noise level was about 6.13 dB. However, the results of Fig. 6.17 show an increase of about 2dB in the noise level, for the optimum signal of amplitude 9. Comparing these two results we note that the noise performance of the first method is poor.

On simulating the stabilization technique shown in Fig. 5.8, the results of Figures 6.23 and 6.24 were obtained. Fig. 6.23 shows the amplitude of the suppressed limit cycles for a large range of the filters coefficients. These results indicate that by this method it is possible to suppress the limit cycles in most cases. However, the suppression of these oscillations to zero steady state was not possible. Also in some cases large amplitude limit cycles could exist.

It was noted that for the cases when large amplitude oscillations could exist there was a relation between the bit pattern of the samples of the limit cycle, the coefficient of the multiplier and the affected bits of the input to the multiplier. The exact form of this is not understood.

On the average, the hardware suggested in Fig. 5.8 was found to be successful in suppressing limit cycles.

The noise increase due to this method of signal stabilization is shown in Fig. 6.24. Comparing this with the results of Fig. 6.17, we note that the noise performance of this method is comparable to that of the signal addition method. If should also be noted that in the cases when A_s is of value 2, 4, 8, ... etc., the effect of this technique would be equivalent to the use of a three level signal.

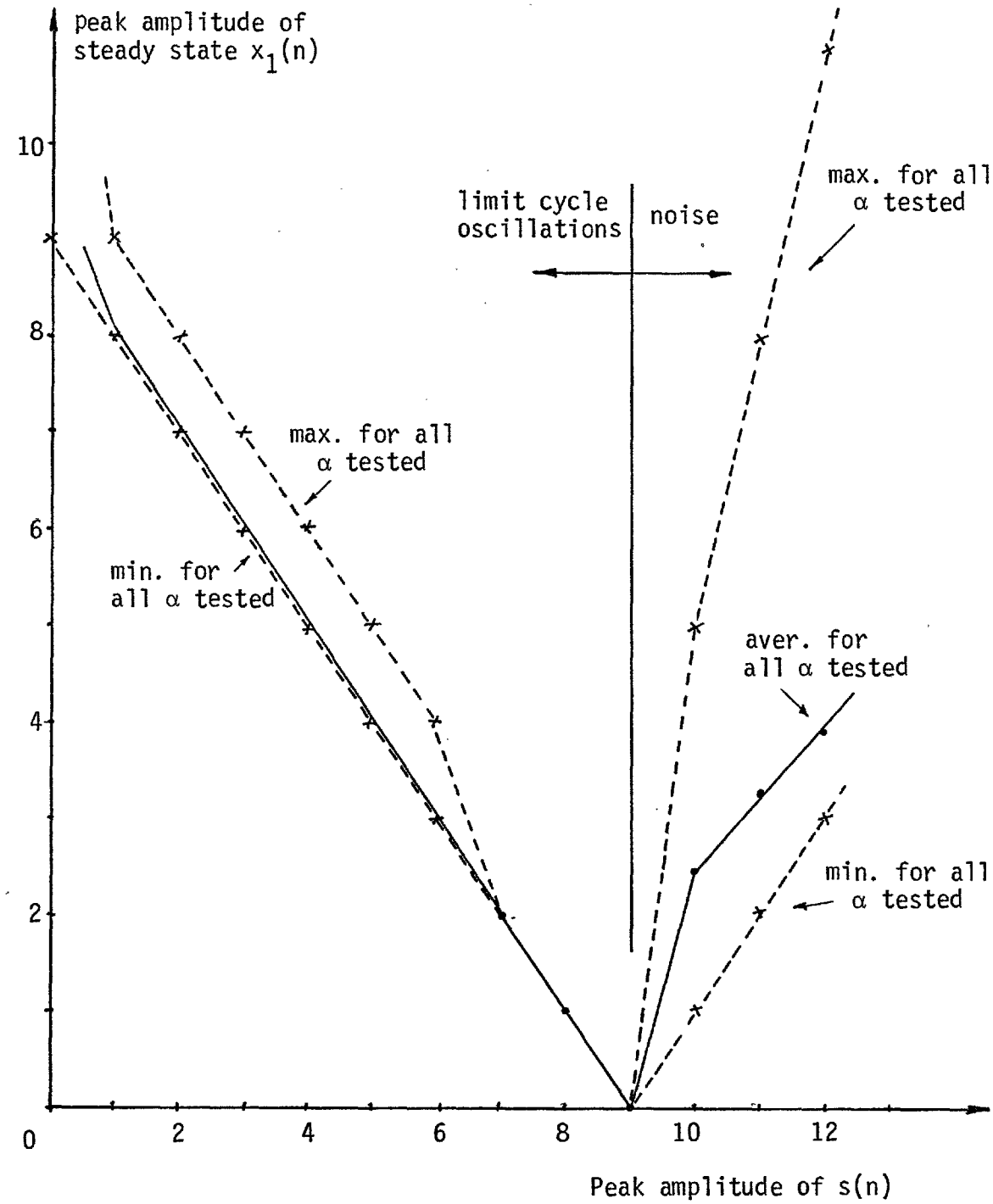


FIG. 6.14 Amplitude of the suppressed limit cycles or steady state noise in type-A realization with rounding:
α from -230/128 to zero at steps of 1/128, β = 121/128

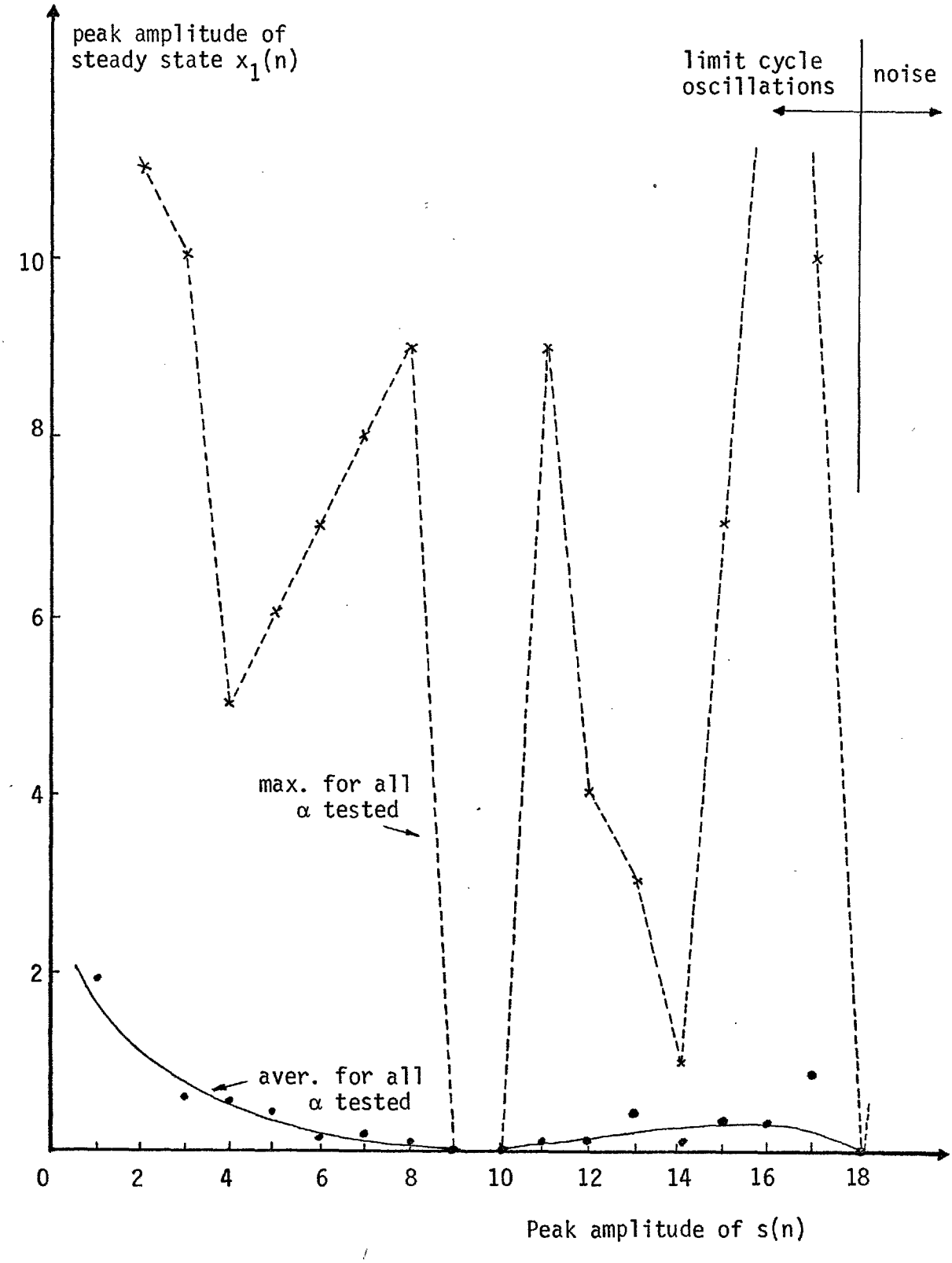


FIG. 6.15 Amplitude of the suppressed limit cycles or steady state noise in type-A realization with value truncation: α from -230/128 to zero at steps of 1/128, β = 121/128

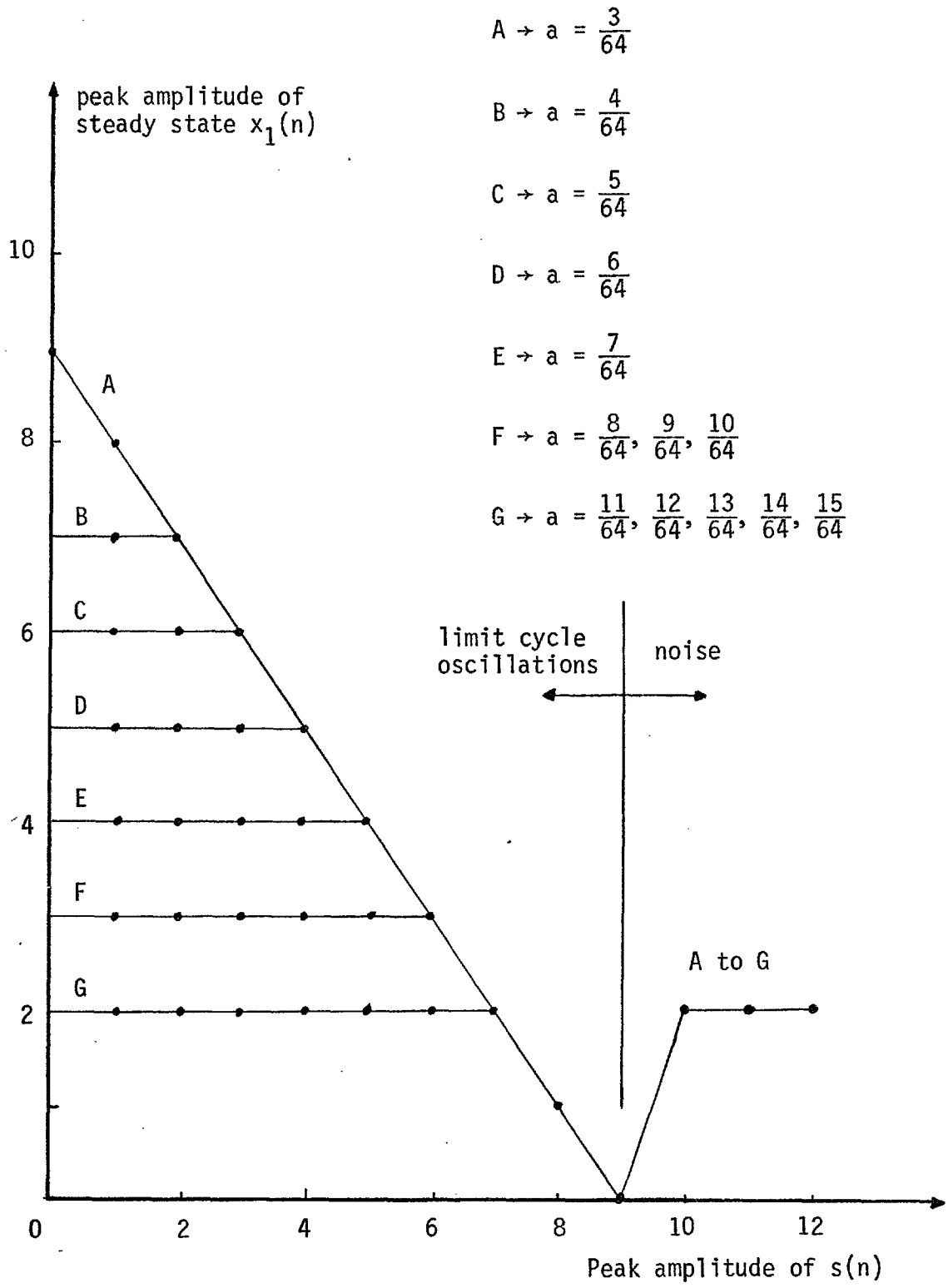


FIG. 6.16 Amplitude of the suppressed limit cycles or steady state noise in type-B realization with rounding:
a = values shown, b = 121/128

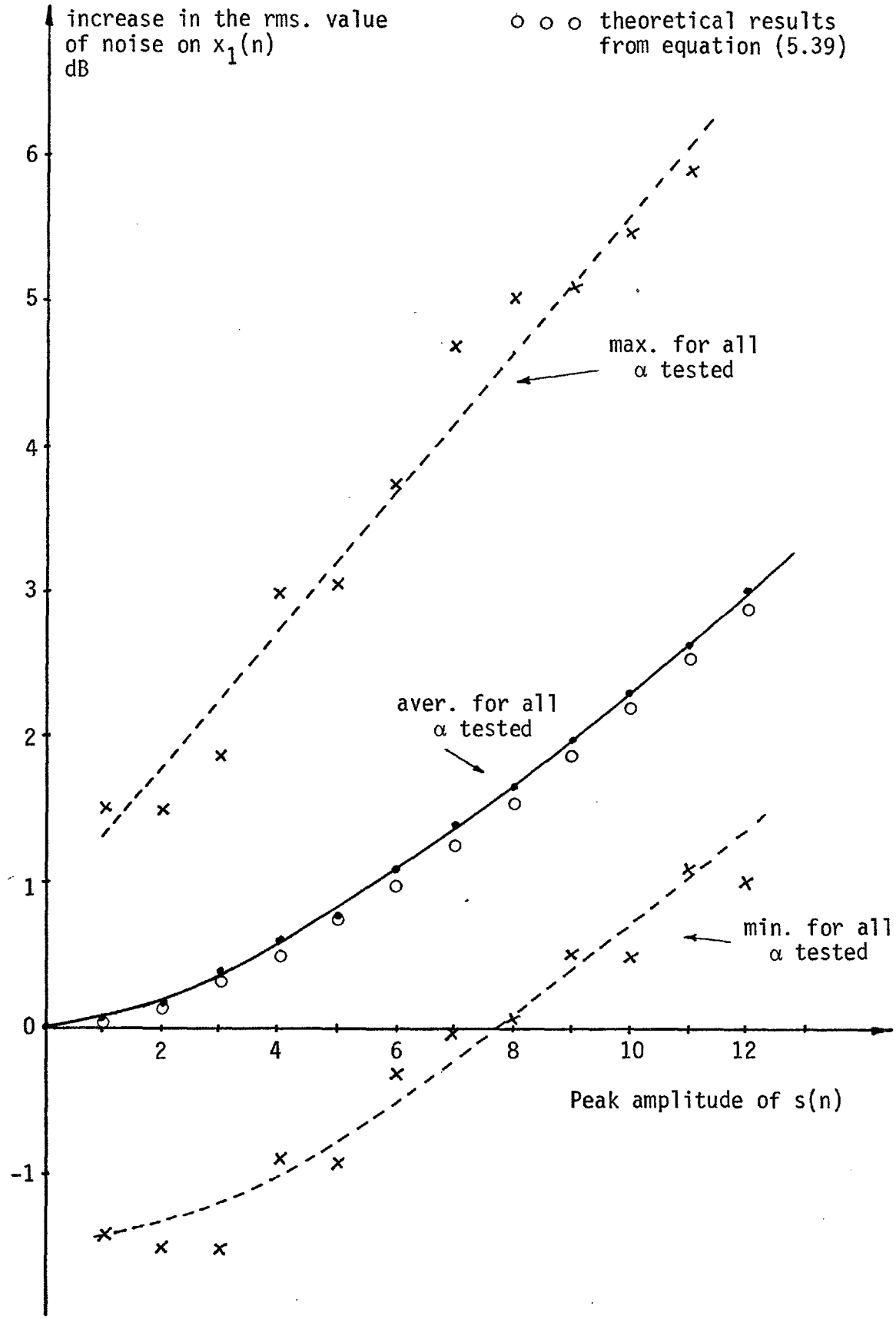


FIG. 6.17 Increase in noise level for type-A realization with rounding: α from -221/128 to -23/128 at steps of 2/128, β = 121/128, amp. of I/P = 100, s(n) quantized to nearest integer

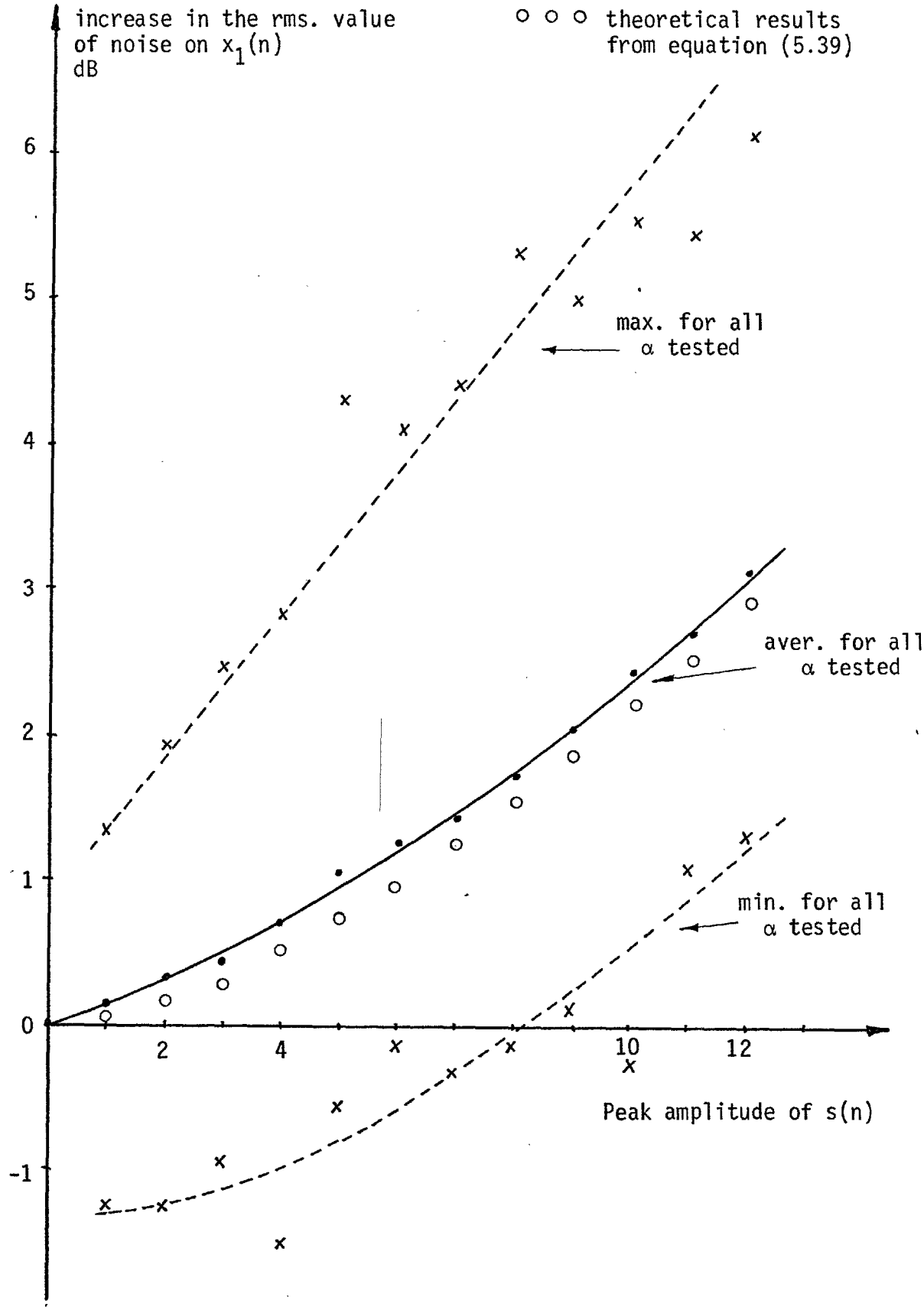


FIG. 6.18 Increase in noise level for type-A realization with rounding: α from -221/128 to -23/128 at steps of 2/128, β = 121/128, amp. of I/P = 1000, s(n) quantized to nearest integer

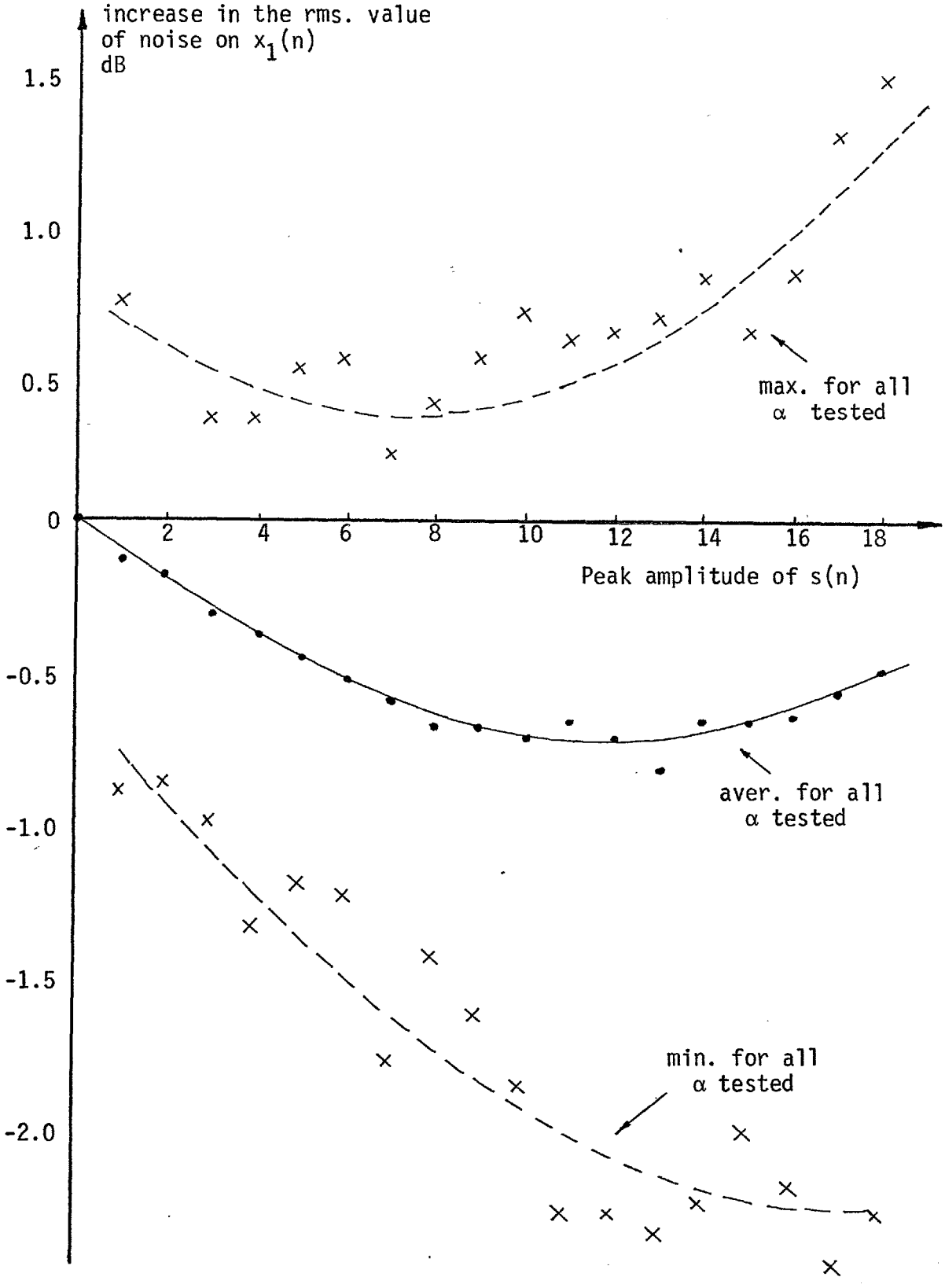


FIG. 6.19 Increase in noise level for type-A realization with value truncation: α from -221/128 to -23/128 at steps of 2/128, β = 121/128. Amp . of I/P = 100, s(n) quantized to nearest integer

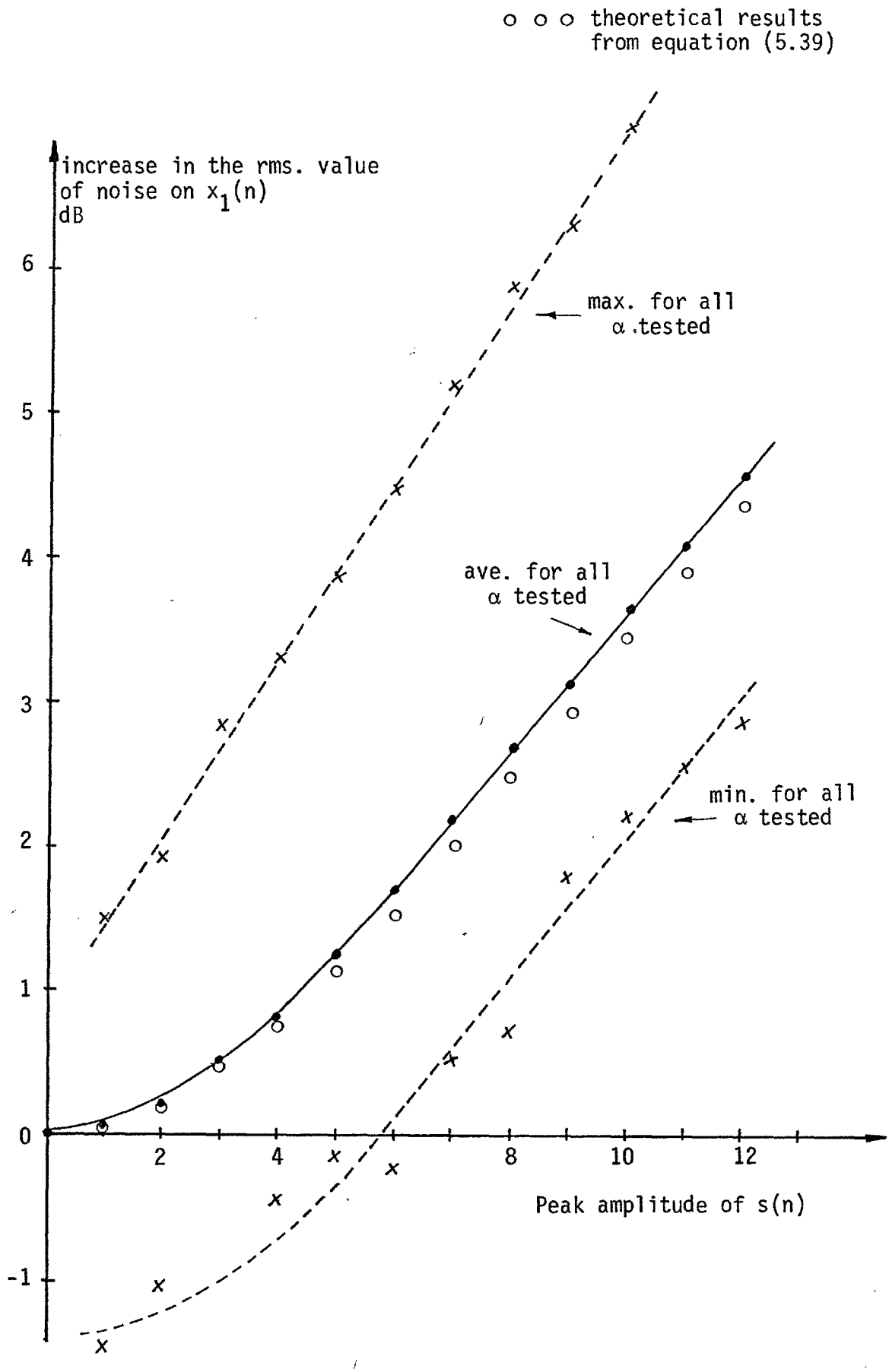


FIG. 6.20 Increase in noise level for type-A realization with rounding: α from -221/128 to -23/128 at steps of 2/128, β = 121/128, amp. of I/P = 100, s(n) quantized to 3 levels

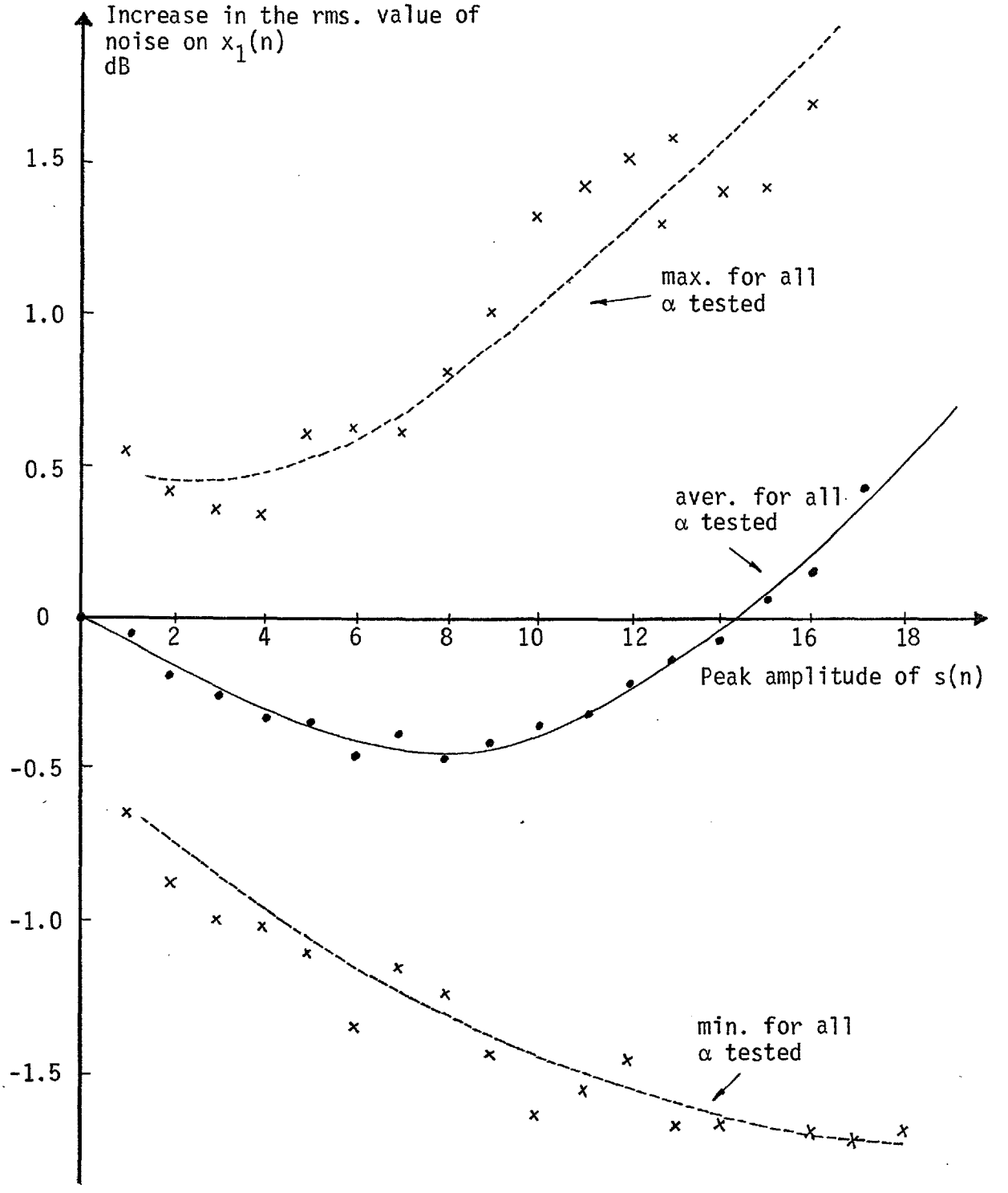


FIG. 6.21 Increase in noise level for type-A realization with value truncation; α = -221/128 to -123/128 at steps of 2/128, β = 121/128, amp. of I/P = 100, s(n) quantized to 3 levels

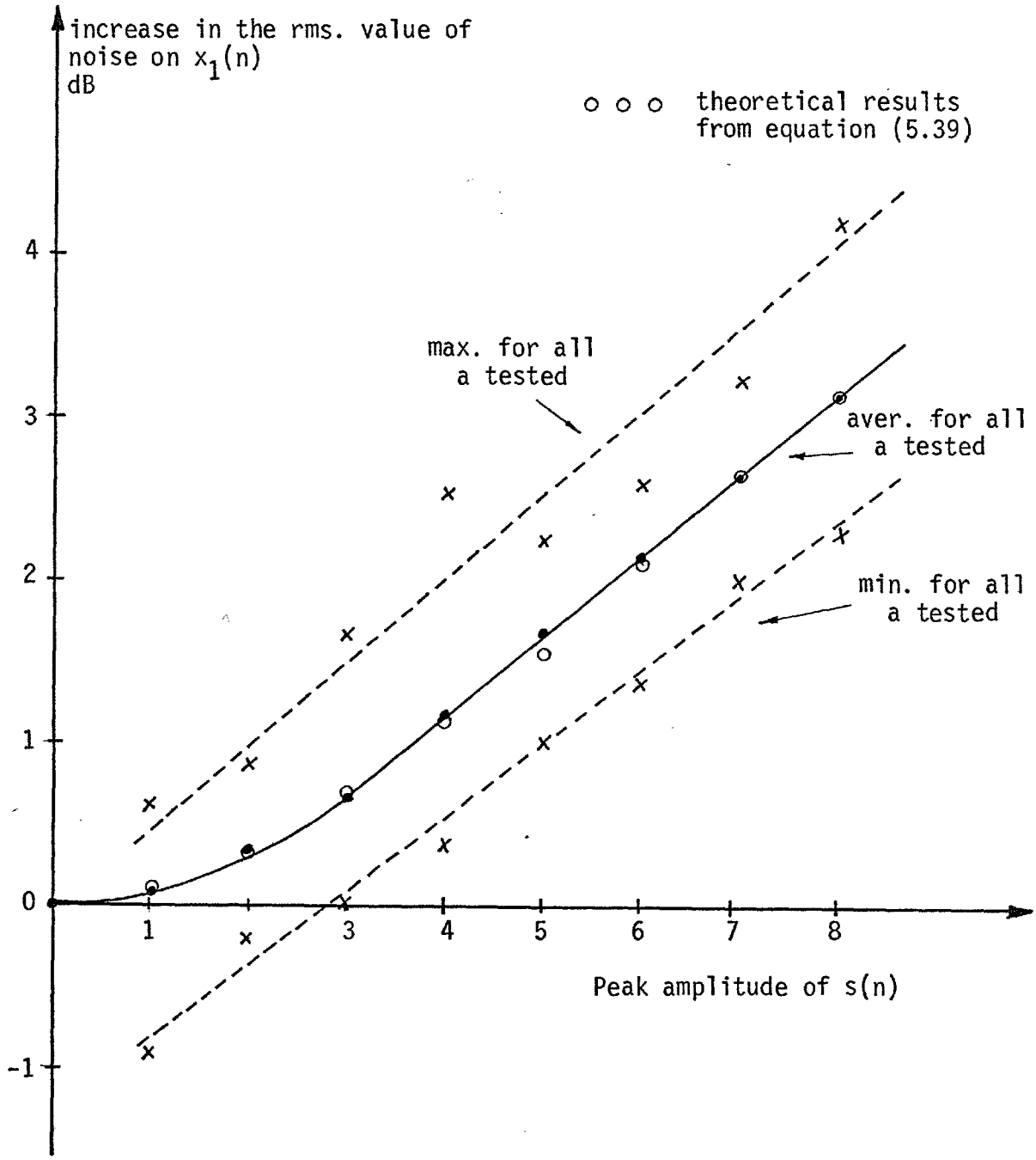


FIG. 6.22 Increase in noise level for type-B realization with rounding: a from 15/128 to 73/128 at steps of 2/128, b = 139/128, amp. of I/P = 100, s(n) quantized to nearest integer

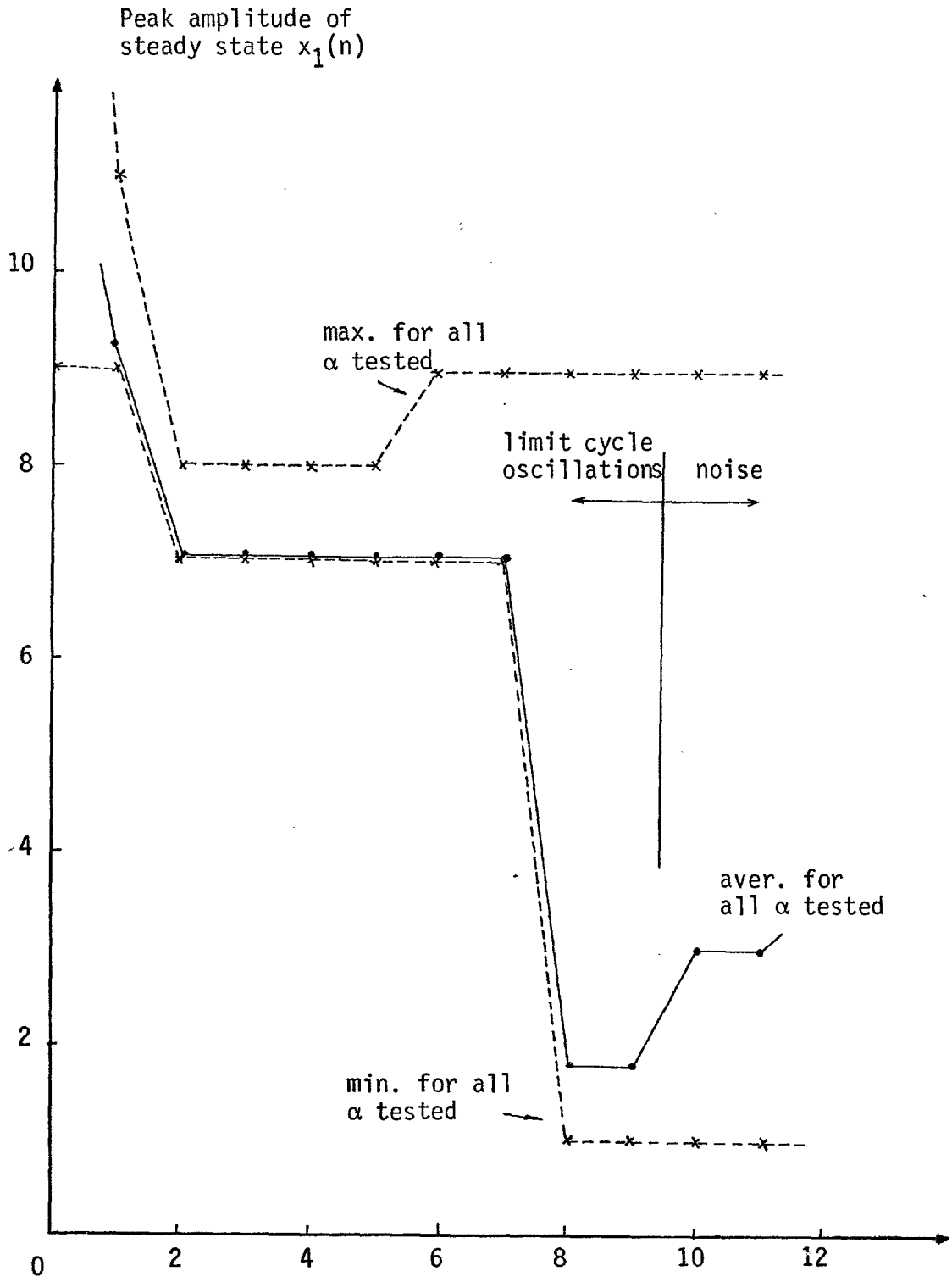


FIG. 6.23 Amplitude of the suppressed limit cycles or steady state noise in type-A realization with rounding using signal stabilization method of Fig. 5.8: α = -230/128 to zero at steps of 1/128, β = 121/128

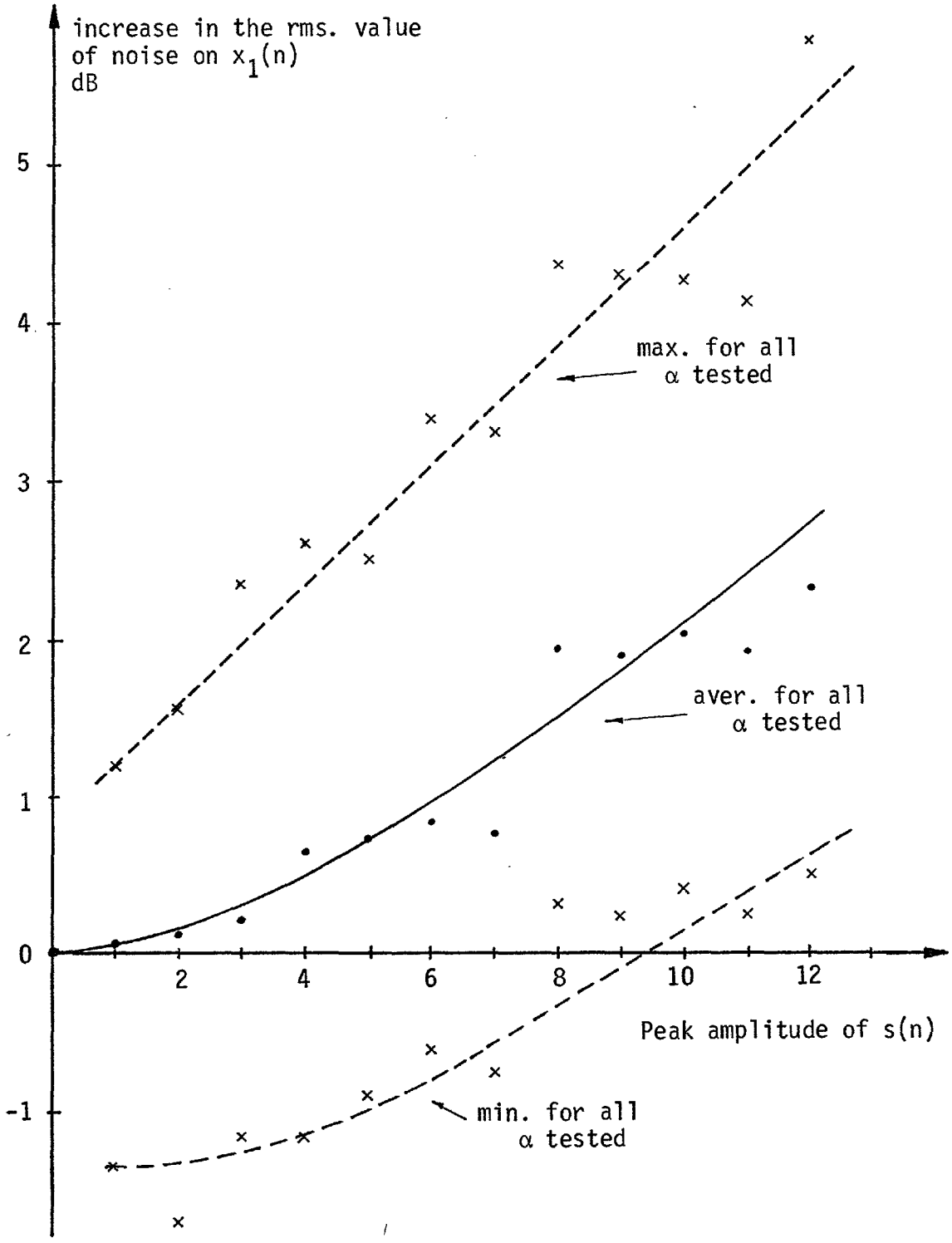


FIG. 6.24 Increase in noise level for type-A realization with rounding using signal stabilization method of Fig. 5.8: α = -221/128 to -23/128 at steps of 2/128, β = 121/128, amp, of I/P = 100

6.5 Conclusions

The simulation results presented in this chapter confirmed several of the theoretical results of the previous chapters. They also presented further results concerning the behaviour of limit cycle oscillations.

From the simulation studies it was possible to suggest and confirm absolute limits for the amplitude of limit cycles in the cases of rounding and value truncation.

The type-B realization was found to have three classes of limit cycles. These were referred to as: accessible, semi-accessible and inaccessible limit cycles.

On tests of signal stabilization method discussed in Chapter 5, good agreement was found with the theoretical results. The simple technique suggested in Fig. 5.8 was found, on the average, to suppress limit cycle oscillations.

CHAPTER 7

CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDIES

In this thesis several aspects of the behaviour of finite word-length recursive second-order filters have been investigated. The conclusions from these studies and suggestions on the possible areas of further research will be discussed in this chapter.

The studies of limit cycle oscillations were based on the investigation of limit cycles of the filter's state variables. Using the method of Jackson [25] and Hess [26] it was shown that there are only two simple structures for the realization of the state variables of the second-order filter. These structures were called the type-A and the type-B realizations (see Section 2.2).

The studies in this thesis are mainly on limit cycle oscillation of these structures. The areas of investigation may be divided into the following groups:

i) the phase plane study of zero input response of the type-A realization.

ii) the study of the occurrence of limit cycle oscillations resulting from a transformation of the filter's structure.

iii) the synthesis of a suitable structure for the suppression of limit cycle oscillations using the method of signal stabilization.

iv) the studies of the behaviour of the finite wordlength digital filter using comprehensive simulation tests.

Examining the zero input response of the finite wordlength filters in the phase plane, it was found that the trajectories of the transient response can go through a process called merging. For $|\beta| < 1$, two or more trajectories can merge at a single point. The condition for the maximum number of trajectories merging at one point relates to the value of β and were specified by theorem 3.1.

For $0.5 \leq |\beta| < 1$, merging could only take place in particular regions of the phase plane. These regions were specified for the points (p-points) immediately prior to the merge points, and were presented by Figures 3.10, 3.13 and 3.15 for the cases of rounding, value truncation and magnitude truncation respectively. For $|\beta| < 0.5$, every point in the phase plane was found to be a merge point.

The idea of merging trajectories in the phase plane and the related results were found useful to categorise limit cycle oscillation into two types; the accessible and the inaccessible limit cycle oscillations.

The accessible limit cycles were defined as the oscillations resulting from initial conditions that are not part of the trajectory of the limit cycle. Inaccessible limit cycles are only possible if the initial conditions are part of the trajectory of that limit cycle. Because accessible limit cycles are produced by a large number of initial conditions, they are more prevalent. Inaccessible limit cycles are not as common.

The amplitudes of the accessible limit cycles in the type-A realization, when rounding is used and $|\beta| \geq 0.5$, are found to have a lower limit. This phenomenon was described by Theorem 3.3.

The second area of study was to develop a deterministic method to predict the occurrence of limit cycle oscillations. This was partially achieved by a transformation of the filter's structure into

two sections; one having integer coefficients and therefore no quantization effects, the other having coefficients less than 0.5 and therefore exhibiting the effects of quantization.

Using this idea, the conditions for the occurrence of limit cycle oscillations and some of the solutions for their amplitudes were developed (Tables 4.1 to 4.4). It was found that in all of these conditions the characteristics of a specific multiplier branch of the transformed structure was a common factor. This phenomenon was used to suggest a new structure in which limit cycle oscillations were suppressed by the method of signal stabilization.

The process of signal stabilization was related to a modified form of the quantization nonlinearity. Using this technique new results for the amplitude of oscillation in the case of insufficient stabilization, and a suitable amplitude for the stabilizing signal, were derived (see Chapter 5). For the type-A realization these results agreed closely with those from simulation studies, implying that the conditions for limit cycle oscillations from the transformation technique are correct when a slight randomization of the filter's algorithm is introduced.

In Section 5.6 a brief discussion of the choices for the stabilizing signal was given. It was suggested that the method of signal stabilization should use simple hardware to make it economical. A relatively simple implementation of this technique was suggested (Fig. 5.8) which was useful for suppressing limit cycle oscillation in most cases.

From simulation studies of the type-B realization it was noted that three types of limit cycles could exist. These were the accessible, the inaccessible and the semi-accessible limit cycles. The full reason for the occurrence of these different types of limit

cycle is not known.

Bounds for the amplitude of the limit cycle oscillations for the type-A realization were suggested using the results from a comprehensive search for accessible limit cycles and from tests for a wide range of coefficients. For limit cycles with period greater than two samples in the cases of rounding and value truncation, two possible bounds were suggested for each case. One was a pessimistic bound which gave an absolute upper limit for the amplitude of limit cycles for all possible coefficients. The other was a good estimate for a majority of coefficients, however it was not exact.

Further, these results showed that there appears to be a relation between the amplitude of the limit cycles and the bit pattern of the coefficient α for a given β. The exact nature of this relation is not known.

From the above conclusion several possible areas for further investigation are suggested.

i) The simulation study of limit cycle oscillations in the type-A realization has shown that there is a relation between the bit pattern of the coefficients and the amplitude of the limit cycle. The development of this relation would give better understanding of the limit cycle behaviour, and hence more efficient techniques for suppression of limit cycles could be found.

ii) In a number of cases the hardware for an inexpensive implementation of the method of signal stabilization (as shown in Fig. 5.8) was found not to suppress all limit cycle oscillation effectively. This is believed to be due to the correlation which could exist between the bit patterns of the input to the multiplier branch, the coefficients, and the bits of the input to the multiplier affected by the stabilizing sequence. Further work could investigate more

effective simple hardware.

iii) The cost versus performance of the signal stabilization technique compared to other techniques for suppressing limit cycle oscillations [141 - 146] could be investigated.

iv) The various degrees of the accessibility of limit cycles in the type-B realization is not yet understood and therefore could be pursued.

v) In the case of type-B realizations, for some conditions, the signal stabilization method could not suppress the limit cycles completely. This may be due to the existence of semi-accessible limit cycles. Further investigations could pursue this direction of study.

APPENDIX A

Solutions of an equation with quantization nonlinearity

We have to solve the general equation given by

$$Q_q[\alpha Kq] = Q_q[\beta Kq], \tag{A.1}$$

where α and β are any coefficients, K is an integer and Q [ ] is a quantization function. In the analysis we assume a rounding function for Q [ ].

Let $$R_q[\alpha Kq] = iq = R_q[\beta Kq]. \tag{A.2}$$

Then $$\frac{i - 0.5}{\alpha} \leqslant K < \frac{i + 0.5}{\alpha}, \tag{A.3}$$

and $$\frac{i - 0.5}{\beta} \leqslant K < \frac{i + 0.5}{\beta}. \tag{A.4}$$

Let $P = \max(|\alpha|, |\beta|)$,

and $S = \min(|\alpha|, |\beta|)$.

For K to satisfy both equations (A.3) and (A.4) we have to consider two cases:

(a) sgn (α) $\neq$ sgn (β). Then equation (A.2) implies i = 0 and hence

$$K < \frac{0.5}{P}, \tag{A.5}$$

(b) sgn (α) = sgn (β). Then

$$\frac{i - 0.5}{S} \leqslant K < \frac{i - 0.5}{P} \tag{A.6}$$

where i can have any positive integer value which allows K to be an integer. From equation (A.6) limit I for the value of i is given by

$$I < \frac{0.5(P + S)}{(P - S)} \tag{A.7}$$

APPENDIX B

PROOF OF THE CONDITIONS FOR THE MODIFIED LIMIT CYCLES WITH THE PERIOD OF TWO SAMPLES

Type-A realization

Let the sequence for the samples of the modified limit cycle of period two samples in the structure of Fig. 5.3 be {K'q, -L'q, K'q, -Lq, . . . etc.}. Then we can write

$$K'q + Q_q[-\alpha L'q] + Q_q[\beta K'q + \Delta\beta s(n)] = 0 \quad , \tag{B.1}$$

and

$$-L'q + Q_q[\alpha K'q] + Q_q[-\beta L'q + \Delta\beta s(n+1)] = 0 \quad , \tag{B.2}$$

For equations (B.1) and (B.2) to be true for all the possible values of $s(n)$ and $s(n+1)$, we require:

$$Q_q[\beta K'q + \Delta\beta s(n)] = Q_q[\beta K'q] \quad , \tag{B.3}$$

and

$$Q_q[-\beta L'q + \Delta\beta s(n+1)] = Q_q[-\beta L'q] \quad . \tag{B.4}$$

Therefore, we can write:

$$K'q + Q_q[-\alpha L'q] + Q_q[\beta K'q] = 0 \quad , \tag{B.5}$$

and

$$-L'q + Q_q[\alpha K'q] + Q_q[-\beta L'q] = 0 \quad . \tag{B.6}$$

Using the method of Section 2.4.2, we can show that the condition for oscillations of amplitude K'q in the structure of Fig. 5.3, is

$$K'q + Q_q[\beta K'q + \Delta\beta s(n)] = Q_q[\alpha K'q] \quad . \tag{B.7}$$

Type-B realization

Let the sequence for the samples of the modified limit cycle of period two samples in the structure of Fig. 5.6 be {K'q, -K'q, K'q, -K'q, . . . etc.} and {-L'q, M'q, -Lq, M'q, . . . etc.} for the state variables $x_1(n)$ and $x_2(n)$ respectively, where K' = L' + M'. Therefore, we can write

$$K'q + Q_q[-aK'q] + Q_q[bM'q + \Delta bs(n)] = 0, \tag{B.8}$$

and

$$-K'q + Q_q[aK'q] + Q_q[-bL'q + \Delta bs(n+1)] = 0 \quad . \tag{B.9}$$

For equation (B.8) and (B.9) to be true for all the possible values of s(n) and s(n+1), we need:

$$Q_q[bM'q + bs(n)] = Q_q[bM'q] \quad , \tag{B.10}$$

and

$$Q_q[-bL'q + \Delta bs(n+1)] = Q_q[-bL'q] \tag{B.11}$$

Therefore, from the results of Section 2.4.2 we can show that

$$Q_q\left|\left(\frac{b}{2}\right)K'q + \Delta bs(n)\right| + K'q = Q_q[aK'q] \tag{B.12}$$

REFERENCES

1. KUO, B.C.

"Analysis and synthesis of sampled-data control systems", Prentice-Hall Inc., Englewood Cliffs, N.J., 1963.

2. FREEMAN, H.

"Discrete-time systems; an introduction to the theory", John Wiley and Sons, N.Y., 1965.

3. BEAUCHAMP, K.G.

"Signal processing, using analog and digital techniques", George Allen and Unwin, London, 1973.

4. WIENER, N.

"The extrapolation, interpolation and smoothing of stationary time series", John Wiley and Sons Inc., N.Y., 1949.

5. BLACKMAN, R.B.

"Linear data smoothing and prediction in theory and practice", Addison-Wesley, Massachusetts, 1965.

6. KAISER, J.F.

"Digital filters", in Syst. Analysis by Digital Computers, by F. Kuo and J.F. Kaiser, McGraw-Hill, N.Y., 1959, Chapter 7, pp. 218-285.

7. GOLD, B. and RADER, C.M.

"Digital processing of signals", McGraw-Hill, N.Y., 1969.

8. OTNES, R.K. and ENOCHSON, L.

"Digital time series analysis", John Wiley and Sons, N.Y., 1972.

9. BOGNER, R.E. and CONSTANTINIDES, A.G.

"Introduction to digital filtering", John Wiley and Sons, London, 1975.

10. OPPENHEIM, A.V. and SCHAFER, R.W.
"Digital signal processing", Prentice-Hall, Englewood Cliffs, N.J., 1975.

11. RABINER, L.R. and GOLD, B.
"Theory and application of digital signal processing", Prentice-Hall, Englewood Cliffs, N.J., 1975.

12. CHILDERS, D. and DURLING, A.
"Digital filtering and signal processing", West Publishing Co., N.Y., 1975.

13. RABINER, L.R. and RADER, C.M.
"Papers in digital signal processing", IEEE Press Inc., N.Y., 1972.

14. IEEE DIGITAL SIGNAL PROCESSING COMMITTEE
"Selected papers in digital signal processing, II", IEEE Press, N.Y., 1976.

15. JURY, E.I.
"Theory and application of the z-transform method", John Wiley and Sons Inc., N.Y., 1964.

16. VIDAL, P.
"Non-linear sampled-data systems", Gordon and Breach, N.Y., 1969.

17. SCHWARTZ, M. and SHAW, L.
"Signal processing: discrete spectral analysis, detection, and estimation", McGraw-Hill, N.Y., 1975.

18. FRANKS, L.
"Signal theory", Prentice-Hall Inc., Englewood Cliffs, N.J., 1969.

19. WILKINSON, J.H.
"Rounding errors in algebraic processes", Prentice-Hall Inc., Englewood Cliffs, N.J., 1963.

20. BLACKMAN, R.B. and TUKEY, J.W.

"The measurement of power spectra", Dover Publications Inc., N.Y., 1959.

21. TIMOTHY, L.K. and BONA, B.E.

"State space analysis: an introduction", McGraw-Hill, N.Y., 1968.

22. GOLOMB, S.W.

"Shift register sequences", Holden-Day Inc., San Francisco, 1967.

23. FLORES, I.

"The logic of computer arithmetic", Prentice-Hall, Englewood Cliffs, N.J., 1963.

24. HILL, F.J. and PETERSON, G.R.

"Digital systems: hardware organisation and design", John Wiley and Sons, N.Y., 1973.

25. JACKSON, L.B.

"An analysis of round off noise in digital filters", D.Sc. Thesis, Stevens Institute of Technology, Hoboken, N.J., 1970.

26. HESS, S.

"A deterministic analysis of limit cycle oscillations in recursive digital filters due to quantization", Ph.D. Thesis, United State Naval Postgraduate School, Monterey, California, 1970.

27. LEE, W.N.

"Minimization of round off noise in digital filters", Ph.D. Thesis, Stanford University, 1971.

28. BESSEYRE, J.A.

"Quantization effects in digital filters", Ph.D. Thesis, North Carolina State University at Raleigh, 1971.

29. OJONGBEDE, H.A.

"Characterisation of the non-linear behaviour of recursive digital filters", Ph.D. Thesis, Univ. College Swansea, Wales, 1971.

30. FONG, M.H.K.

"Truncation and round off error bounds in digital computation", Ph.D. Thesis, Auburn University, 1971.

31. KYLE, Jr. J.C.

"An RMS bound for digital simulation", Ph.D. Thesis, Syracuse University, 1971.

32. CHAN, D.S.K.

"Round off noise in cascade realization of finite impulse response digital filters", S.B. and S.M. Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1972.

33. RABINER, L.R. et al.

"Terminology in digital signal processing", IEEE Trans. Audio and Electroacoust., Vol. AU-20, No. 5, pp. 322-337, Dec. 1972.

34. JURY, E.I. and TSYPKIN, Y.Z.

"Theory of discrete automatic systems (review)", Automation and Remote Control, No. 6, pp. 915-936, June 1970.

35. KAISER, J.F.

"Design methods for digital filters", Proc. First Annual Allerton Conf. Circuits and Syst. Theory, pp. 221-236, Nov. 1963.

36. GIBBS, A.J.

"An introduction to digital filters", Aust. Telecommun. Res., Vol. 3, No. 2, pp. 3-14, Nov. 1969.

37. GIBBS, A.J.

"The design of digital filters", Aust. Telecommun. Res., Vol. 4, No. 1, pp. 29-34, May 1970.

38. OTNES, R.K.

"An elementary design procedure for digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-16, No. 3, pp. 330-335, Sept. 1968.

39. CROCHIERE, R.E. and OPPENHEIM, A.V.

"Analysis of linear digital networks", Proc. IEEE, Vol. 63, No. 4, pp. 581-594, April 1975.

40. ANTONIOU, A.

"Realization of digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-20, No. 1, pp. 95-97, Mar. 1972.

41. MITRA, S.K. and SHERWOOD, R.J.

"Canonic realization of digital filters using the continued fraction expansion", IEEE Trans. Audio and Electroacoust., Vol. AU-20, No. 3, pp. 185-194, Aug. 1972.

42. MITRA, S.K. and SHERWOOD, R.J.

"Digital ladder networks", IEEE Trans. Audio and Electroacoust., Vol. AU-21, No. 1, pp. 30-36, Feb. 1973.

43. MITRA, S.K. and HIRANO, K.

"Multiplier extraction approach - a new method of digital filter realization", IEEE Arden House workshop on digital signal processing, Jan. 1974.

44. MITRA, S.K. and HIRANO, K.

"Digital all-pass networks", IEEE Trans. Circuits and Syst., Vol. CAS-21, No. 5, pp. 688-700, Sept. 1974.

45. SZCZUPAK, J. and MITRA, S.K.

"Digital filter realization using successive multiplier-extraction approach", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-23, No. 2, pp. 235-239, April 1975.

46. FETTWEIS, A.

"Some principles of designing filters imitating classical filter structures", IEEE Trans. Circuit Theory, Vol. CT-18, No. 2, pp. 314-316, Mar. 1971.

47. PARKER, S.R. and HESS, S.

"Canonic realization of second-order digital filters due to finite precision arithmetic", IEEE Trans. Circuit Theory, Vol. CT-19, No. 4, pp. 410-413, July 1972.

48 JACKSON, L.B., KAISER, J.F. and McDONALD, H.S.

"An approach to the implementation of digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-16, No. 3, pp. 413-421, Sept. 1968.

49. OPPENHEIM, A.V.

"Realization of digital filters using block floating point arithmetic", IEEE Trans. Audio and Electroacoust., Vol. AU-18, No. 2, pp. 130-136, June 1970.

50. ZOHAR, S.

"New hardware realization of nonrecursive digital filters", IEEE Trans. Computers, Vol. C-22, No. 4, pp. 328-338, Apr. 1973.

51. ZOHAR, S.

"The counting recursive digital filter", IEEE Trans. Computers, Vol. C-22, No. 4, pp. 338-347, Apr. 1973.

52. PELED, A. and LIU, B.

"A new approach to the realization of nonrecursive digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-21, No. 6, pp. 477-484, Dec. 1973.

53. PELED, A. and LIU, B.

"A new hardware realization of digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-22, No. 6, pp. 456-462, Dec. 1974.

54. FREENY, S.L.

"Special purpose hardware for digital filtering", Proc. IEEE, Vol. 63, No. 4, pp. 633-648, Apr. 1975.

55. ALLEN, J.

"Computer architecture for signal processing", Proc. IEEE, Vol. 63, No. 4, pp. 624-633, Apr. 1975.

56. KAISER, J.F.

"Some practical considerations in the realization of linear digital filters", Proc. 3rd Annual Allerton Conf. Circuits and Syst. Theory, pp. 621-633, Oct. 1965.

57. KNOWLES, J.B. and OLCAYTO, E.M.

"Coefficient accuracy and digital filter response", IEEE Trans. Circuit Theory, Vol. CT-15, No. 1, pp. 31-41, Mar. 1968.

58. MITRA, S. and SHERWOOD, R.J.

"Estimation of pole-zero displacement of a digital filter due to coefficient quantization", IEEE Trans. Circuits and Syst., Vol. CAS-21, No. 1, pp. 116-123, Jan. 1974.

59. STEIGLITZ, K.

"Designing short-word recursive digital filters", Proc. 9th Annual Allerton Conf. Circuit and Syst. Theory, pp. 778-788, Oct. 1971.

60. AVENHAUS, E.

"An optimization procedure to minimize the wordlength of digital filter coefficients", London Conf. on Digital Filtering, Imperial College, Aug. 1971.

61. AVENHAUS, E.

"On the design of digital filters with coefficients of limited wordlength", IEEE Trans. Audio and Electroacoust., Vol. AU-20, No. 3, pp. 206-212, Aug. 1972.

62. SUK, M. and MITRA, S.K.
"Computer-aided design of digital filters with finite wordlength", IEEE Trans. Audio and Electroacoust., Vol. AU-20, No. 5, pp. 356-363, Dec. 1972.

63. CHAO, K.S. and LU, K.S.
"On sequential refinement schemes for recursive digital filter design", IEEE Trans. Circuit Theory, Vol. CT-20, No. 4, pp. 153-156, July 1973.

64. MERON, P., SEKEY, A.A. and ZEHEB, E.
"Design method for stable second-order digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-22, No. 3, pp. 196-202, June 1974.

65. CHARALAMBOUS, C. and BEST, M.J.
"Optimization of recursive digital filters with finite wordlength", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-22, No. 6, pp. 424-431, Dec. 1974.

66. CROCHIERE, R.E.
"A new statistical approach to the coefficient wordlength problem for digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 3, pp. 190-196, Mar. 1975.

67. HADJIFOTIOU, A. and APPLEBY, D.G.
"Design of digital filters with severely quantized coefficients", Proc. Radio and Electron. Engineer, Vol. 46, No. 1, pp. 23-28, Jan. 1976.

68. EBERT, P.M., MAZO, J.E. and TAYLOR, M.G.
"Overflow oscillations in digital filters", The Bell Syst. Tech. J., Vol. 48, No. 9, pp. 2999-3020, Nov. 1969.

69. SANDBERG, I.W.

"A theorem concerning limit cycles in digital filters", Proc. 7th Annual Allerton Conf. on Circuits and Syst. Theory, pp. 63-68, Oct. 1969.

70. WILLSON, Jr. A.N.

"Some effects of quantization and adder overflow on the forced response of digital filters", Bell Syst. Tech. J., Vol. 51, No. 4, pp. 863-887, Apr. 1972.

71. WILLSON, Jr. A.N.

"Limit cycles due to adder overflow in digital filters", IEEE Trans. Circuit Theory, Vol. CT-19, No. 4, pp. 342-346, July 1972.

72. WILLSON, Jr. A.N.

"Computation of the periods of forced overflow oscillations in digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-24, No. 1, pp. 89-97, Feb. 1976.

73. CLAASEN, T.A.C.M. and KRISTIANSSON, L.O.G.

"Necessary and sufficient conditions for the absence of overflow phenomena in a second-order recursive digital filter", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-23, No. 6, pp. 509-515, Dec. 1975.

74. KRISTIANSSON, L.O.G.

"Jump phenomenon in digital filters", Electron. Lett., Vol. 10, No. 2, pp. 14-15, Jan. 1974.

75. LANDAURO, A. and LIENARD, J.

"On overflow detection and correction in digital filters", IEEE Trans. Computers, Vol. C-24, No. 12, pp. 1226-1228, Dec. 1975.

76. BENNETT, W.R.

"Spectra of quantized signals", Bell Syst. Tech. J., Vol. 27, No. 3, pp. 446-472, July 1948.

77. BERTRAM, J.E.

"The effect of quantization in sampled feedback systems", AIEE Trans., Vol. 77, Pt. II (Application and Industry), No. 38, pp. 177-182, Sept. 1958.

78. SLAUGHTER, J.B.

"Quantization errors in digital control systems", IEEE Trans. Autom. Control, Vol. AC-9, No. 1, pp. 70-74, Jan. 1964.

79 KNOWLES, J.B. and EDWARDS, R.

"Effect of a finite word length computer in a sampled-data feedback system", Proc. IEE , Vol. 112, No. 6, pp. 1197-1207, June 1965.

80. KNOWLES, J.B.

"Word-length limitation on digital filter performance", Proc. Seminar Digital Filters, Inst. Tech. Physik, Zurich, pp. 5.1-5.40, Jan. 1969.

81. JACKSON, L.B.

"On the interaction of roundoff noise and dynamic range in digital filters", Bell Syst. Tech. J., Vol. 49, No. 2, pp. 159-184, Feb. 1970.

82. JACKSON, L.B.

"Roundoff noise analysis for fixed point digital filters realized in cascade or parallel form", IEEE Trans. Audio and Electroacoust., Vol. AU-18, No. 2, pp. 107-122, June 1970.

83. EDWARDS, R., BRADLEY, T. and KNOWLES, J.B.

"Comparison of noise performance or programming methods in the realization of digital filters", Proc. Symp. Computer Processing in Commun., Polytechnique Institute Brooklyn, pp. 295-311, Apr. 1969.

84. KAN, E.P.F. and AGGARWAL, J.K.

"Minimum dead-band design of digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-19, No. 4, pp. 292-296, Dec. 1971.

85. SANDBERG, I.W. and KAISER, J.F.

"A bound on limit cycles in fixed point implementations of digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-20, No. 2, pp. 110-111, June 1972.

86. LONG, J.L. and TRICK, T.N.

"An absolute bound on limit cycles due to roundoff errors in digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-21, No. 1, pp. 27-30, Feb. 1973.

87. LONG, J.L. and TRICK, T.N.

"A note on absolute bounds on quantization errors in fixed-point implementation of digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 6, pp. 567-570, June 1975.

88. YAKOWITZ, S. and PARKER, S.R.

"Computation of bounds for digital filter quantization errors", IEEE Trans. Circuit Theory, Vol. CT-20, No. 4, pp. 391-396, July 1973.

89. ÜNVER, Z. and ABDULLAH, K.

"A tighter practical bound on quantization errors in second-order digital filters with complex conjugate poles", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 7, pp. 632-633, July 1975.

90. FINK, A.M.

"A bound on quantization errors in second-order digital filters with complex poles that is tighter for small θ", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 5, pp. 325-326, May 1975.

91. BELAL, A.A.

"On the quantization error bounds in second-order digital filters with complex conjugate poles", IEEE Trans. Circuits and Syst., Vol. CAS-24, No. 1, p. 45, Jan. 1977.

92. PARKER, S.R. and YAKOWITZ, S.

"Statistical error bounds due to finite precision arithmetic in digital filters", 16th Midwest Symp. on Circuit Theory, University of Waterloo, Canada, Apr. 1973.

93. PARKER, S.R. and YAKOWITZ, S.

"A general method for calculating quantization error bounds due to roundoff in multivariable digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 6, pp. 570-572, June 1975.

94. HWANG, S.Y.

"Roundoff noise in state-space digital filtering: a general analysis", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-23, No. 3, pp. 256-262, June 1976.

95. MITRA, S.K., HIRANO, K. and SAKAGUCHI, H.

"A simple method of computing the input quantization and multiplication round-off errors in a digital filter", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-22, No. 5, pp. 326-329, Oct. 1974.

96. APPEL, U.

"Bounds on second-order digital filter limit cycles", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 6, pp. 630-632, July 1975.

97. CHANG, T.L.

"A note on upper bounds on limit cycles in digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-24, No. 1, pp. 99-100, Feb. 1976.

98. JENKINS, W.K. and LEON, B.J.

"An analysis of quantization error in digital filters based on interval algebras", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 3, pp. 223-232, Mar. 1975.

99. YARLAGADDA, R. and MASSAD, K.E.

"Digital filter roundoff noise calculation using the driving-point impedance", Electron. Lett., Vol. 11, No. 16, pp. 359-360, Aug. 1975.

100. JOHNSON, G.W.

"Upper bound on dynamic quantization error in digital control systems via the direct method of Liapunov", IEEE Trans. Autom. Control, Vol. AC-10, No. 4, pp. 439-448, Oct. 1965.

101. FETTWEIS, A.

"On the evaluation of roundoff noise in digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 11, p. 896, Nov. 1975.

102. HALYO, N. and McALPINE, G.A.

"A discrete model for product quantization errors in digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-19, No. 3, pp. 255-256, Sept. 1971.

103. THOMPSON, J.S.

"An improved model for quantization errors in digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-21, No. 2, pp. 125-127, Apr. 1973.

104. ECKHARDT, B. and SCHUSSLER, H.W.

"On the quantization error of a multiplier", IEEE Symp. Circuits and Syst., Zurich, 1976.

105. PARKER, S.R. and GIRARD, P.E.

"Correlated noise due to roundoff in fixed point digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-23, No. 4, pp. 204-211, Apr. 1976.

106. FETTWEIS, A.

"On the connection between multiplier word length limitation and roundoff noise in digital filters", IEEE Trans. Circuit Theory, Vol. 19, No. 5, pp. 486-491, Sept. 1972.

107. FETTWEIS, A.

"Roundoff noise and attenuation sensitivity in digital filters with fixed point arithmetic", IEEE Trans. Circuit Theory, Vol. CT-20, No. 2, pp. 174-175, Mar. 1973.

108. BONZANIGO, F.

"Comment on 'Roundoff noise and attenuation sensitivity in digital filters with fixed-point arithmetic'", IEEE Trans. Circuits and Syst., Vol. CAS-21, No. 6, pp. 809-811, Nov. 1974.

109. JACKSON, L.B.

"Roundoff noise bounds derived from coefficient sensitivities for digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-23, No. 8, pp. 481-485, Aug. 1976.

110. LIU, B. and VAN-VALKENBURG

"On roundoff error of fixed-point digital filters using sign-magnitude truncation", IEEE Trans. Circuit Theory, Vol. CT-19, No. 5, pp. 536-537, Sept. 1972.

111. CLAASEN, T.A.C.M., MECKLENBRÄUKER, W.F.G. and PECK, J.B.H.

"Quantization noise analysis for fixed-point digital filters using magnitude truncation for quantization", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 11, pp. 887-895, Nov. 1975.

112. JACKSON, L.B.

"An analysis of limit cycles due to multiplication rounding in recursive digital (sub) filters", Proc. 7th Annual Allerton Conf. Circuit and Syst. Theory, pp. 69-78, Oct. 1969.

113. PARKER, S.R. and HESS, S.F.

"Limit cycle oscillations in digital filters", IEEE Trans. Circuit Theory, Vol. CT-18, No. 6, pp. 687-697, Nov. 1971.

114. WHITE, S.A.

"Quantizer-induced digital controller limit cycles", IEEE Trans. Autom. Control, Vol. AC-14, No. 4, pp. 430-432, Aug. 1969.

115. OJONGBEDE, H.A.

"Limit cycle constraints for recursive digital filter design", Electron. Lett., Vol. 6, No. 22, pp. 698-700, Oct. 1970.

116. OJONGBEDE, H.A.

"Comparison of dead band effects in constant-input digital filter systems", Electron. Lett., Vol. 7, No. 11, pp. 309-310, Jan. 1971.

117. HESS. S.F. and PARKER, S.R.

"Heuristic bounds for the frequency of digital oscillators due to quantization noise", Electron. Lett., Vol. 8, No. 4, pp. 86-87, Feb. 1972.

118. CLAASEN T.A.C.M., MECKLENBRÄUKER, W.F.G. and PEEK, J.B.H.

"Some remarks on the classification of limit cycles in digital filters", Philips Res. Rep., Vol. 28, No. 4, pp. 297-305, Aug. 1973.

119. CLAASEN, T.A.C.M., MECKLENBRÄUKER, W.F.G. and PEEK, J.B.H.

"Frequency domain criteria for the absence of zero-input limit cycles in nonlinear discrete-time systems, with applications to digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 3, pp. 232-239, Mar. 1975.

120. JATNIEKS, G.U. and SHENOI, B.A.

"Zero-input limit cycles in coupled digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-22, No. 2, pp. 146-149, Apr. 1974.

121. STORZBACH, W.H.

"Forced oscillations in recursive digital filters", Proc. Intl. Symp. Circuit Theory, pp. 232-236, Apr. 1972.

122. KAN, E.P.F. and AGGARWAL, J.K.

"Minimum-deadband design of digital filters", IEEE Trans. Audio and Electroacoust., Vol. AU-19, No. 4, pp. 292-296, Dec. 1971.

123. BRUBAKER, T.A. and GOWDY, J.N.

"Limit cycles in digital filters", IEEE Trans. Autom. Control, Vol. AC-17, No. 5, pp. 675-677, Oct. 1972.

124. MARIA, G.A. and FAHMY, M.M.

"Limit cycle oscillations in a cascade of first-and second-order digital filter sections, IEEE Trans. Circuits and Syst., Vol. CAS 22, No. 2, pp. 131-134, Feb. 1975.

125. BONZANIGO, F.

"Constant-input behaviour of recursive digital filters", IEEE Arden House Workshop on Digital Filtering, N.Y., Jan. 1970.

126. KAO, C.Y.

"An analysis of limit cycles due to sign magnitude truncation in multiplication in recursive digital filters", Proc. 5th Asilomar Conf. on Circuits and Syst., pp. 349-353, Nov. 1971.

127. CLAASEN, T.A.C.M., MECKLENBRÄUKER, W.F.G. and PEEK, J.B.H.

"Second order digital filters with only one magnitude-truncation quantizer and having practically no limit cycles", Electron. Lett., Vol. 9, No. 22, pp. 531-532, Nov. 1973.

128. CLAASEN, T.A.C.M., MECKLENBRÄUKER, W.F.G. and PEEK, J.B.H.
"A comparison between the stability of second-order digital filters with various arithmetics", IEE European Conf. on Circuit Theory and Design, Conf. Pub. No. 116, pp. 354-358, July 1974.

129. CLAASEN, T.A.C.M., MECKLENBRÄUKER, W.F.G. and PEEK, J.B.H.
"Remarks on the zero-input behaviour of second order digital filters designed with one magnitude truncation quantizer", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-23, No. 2, pp. 240-242, Apr. 1975.

130. THONG, T. and LIU, B.
"Limit cycles in the combinational implementation of digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-24, No. 3, pp. 248-256, June 1976.

131. BÜTTNER, M.
"Some experimental results concerning randomlike noise and limit cycles in recursive digital filters", NTZ, Vol. 28, No. 11, pp. 402-406, 1975.

132. KIEBURTZ, R.B.
"An experimental study of roundoff effects in a tenth order recursive digital filter", IEEE Trans. Commun., Vol. COM-21, No. 6, pp. 757-763, June 1973.

133. HWANG, S.Y.
"On optimization of cascade fixed-point digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-21, No. 1, pp. 163-166, Jan. 1974.

134. LEE, W.S.
"Optimization of digital filters for low roundoff noise", IEEE Trans. Circuits and Syst., Vol. CAS-21, No. 3, pp. 424-431, May 1974.

135. MITRA, S.K., MONDAL, K., HIRANO, K. and SAKAGUCH, H.
"A novel approach to recursive digital filter realization with low roundoff noise", Proc. IEEE Symp. Circuits and Syst., Zurich, 1976.

136. MULLIS, C.T. and ROBERTS, R.A.
"Synthesis of minimum roundoff noise fixed point digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-23, No. 9, pp. 551-562, Sept. 1976.

137. MULLIS, C.T. and ROBERTS, R.A.
"Roundoff noise in digital filters: frequency transformations and invariants", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-24, No. 6, pp. 538-550, Dec. 1976.

138. OLDENBURGER, R.
"Signal stabilization of a control system", Trans. American Society of Mechanical Engineers, Vol. 79, No. 8, pp. 1870-1872, Nov. 1957.

139. OLDENBURGER, R. and LIU, C.C.
"Signal stabilization of a control system", AIEE Trans., Vol. 78, Pt. II (Application and Industry), No. 42, pp. 96-100, May 1959.

140. OLDENBURGER, R. and SRIDHAR, R.
"Signal stabilization of control systems with random inputs", AIEE Trans., Vol. 80, Pt. II (Application and Industry), No. 57, pp. 260-268, Nov. 1961.

141. WONG, K.M. and KING, R.A.
"Method to suppress limit-cycle oscillations in digital filter", Electron. Lett., Vol. 10, No. 5, pp. 55-57, Mar. 1974.

142. BUTTERWECK, H.J.
"Suppression of parasitic oscillations in second-order digital filters by means of a controlled-rounding arithmetic", Arch. Elektron and Üebertragungstech, Vol. 29, No. 9, pp. 371-374, 1975.

143. FETTWEIS, A. and MEERKÖTTER, K.

"Suppression of parasitic oscillations in wave digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-22, No. 3, pp. 239-246, Mar. 1975.

144. FETTWEIS, A. and MEERKÖTTER, K.

"Suppression of parasitic oscillations in half-synchronic wave digital filters", IEEE Trans. Circuits and Syst., Vol. CAS-23, No. 2, pp. 125-126, Feb. 1976.

145. MEERKÖTTER, K. and WEGENER, W.

"A new second-order digital filter without parasitic oscillations", Arch. Elektron and Üebertragungstech (Electronics and Communication), Vol. 29, pp. 312-314, 1975.

146. MEERKÖTTER, K.

"Realization of limit cycle-free second-order digital filters", IEEE Intl. Symp. Circuits and Syst., pp. 295-298, 1976.

147. FEDOROV, S.M. and SHIDLOVSKII, V.I.

"Method for increasing accuracy and the design of an optimal structure for a digital filter", Automation and Remote Control, Vol. 37, No. 11, Pt. 1, pp. 1686-1691, Nov. 1976.

148. WEINSTEIN, C.J.

"Quantization effects in digital filters", Tech. Rep. No. 468, Lincoln Laboratory, MIT, Nov. 1969.

149. LIU, B.

"Effect of finite word length on the accuracy of digital filters - A review", IEEE Trans. Circuit Theory, Vol. CT-18, No. 6, pp. 670-677, Nov. 1971.

150. AGGARWAL, J.K.

"Input quantization and arithmetic roundoff in digital filters - A review", Tech. Rep. No. N130, Electron. Res. Centre, University of Texas, July 1972.

151. OPPENHEIM, A.V. and WEINSTEIN, C.J.

"Effects of finite register length in digital filtering and the fast fourier transform", IEEE Proc., Vol. 60, No. 8, pp. 957-976, Aug. 1972.

152. KAISER, J.F.

"On the limit cycle problem", Proc. IEEE Intl. Symp. Circuits and Syst., Zurich, 1976.

153. CLAASEN, T.A.C.M., MECKLENBRÄUKER, W.F.G. and PEEK, J.B.H.

"Effects of quantization and overflow in recursive digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-24, No. 6, pp. 517-529, Dec. 1976.