

SUPPRESSION OF LIMIT CYCLES
IN
DIGITAL FILTERS

by

HOU-MING LIU

Submitted for the Degree of
Master of Philosophy
at the
University of Aston in Birmingham

July 1982

SUPPRESSION OF LIMIT CYCLES

IN DIGITAL FILTERS

Hou-Ming Liu

Submitted for the Degree of
Master of Philosophy
at the
University of Aston in Birmingham
1982

Summary

The injection of random dither to suppress limit cycle oscillations in second-order direct form digital filter sections is discussed. Three types of dither signal are used: uniformly distributed random dither, binary random dither and bandstop dither. All the limit cycles in the second-order filter sections can be suppressed by the injection of any one of three dither signals. No remaining noise appears in the output from the filter in the zero-input condition.

Experimental comparisons are made of the average time taken to suppress limit cycles and of the increase in output noise caused by the dither. With binary random dither, the time to suppress a limit cycle is comparable with the time for zero-input response of a linear filter to decay below the quantization threshold. Bandstop dither has the advantage to suppress limit cycles about as quickly as binary dither yet it causes an increase in output noise of less than 2 dB.

Key Words

LIMIT CYCLE, DIGITAL FILTER, DITHER.

ACKNOWLEDGEMENTS

I express my appreciation to the University of Aston in Birmingham for the opportunity to pursue this degree programme. I also wish to thank Chengdu Institute of Radio Engineering, People's Republic of China, for the permission of my study leave; British Council and Chinese Government for the financial support.

This work would not have been possible without the continuing aid and encouragement of my supervisor, Dr. M. H. Ackroyd, whom I thank for his support. I am also greatly indebted to Dr. G. K. Steel for his help. A special note of thanks is due to Mrs. H. Turner and Miss N. P. Freeman for their patient typing of the manuscript.

TABLE OF CONTENTS

	<u>Page</u>
 <u>CHAPTER 1</u>	
INTRODUCTION	1
1.1 Introductory Remarks	1
1.2 Methods for Limit Cycle Oscillation Suppression	4
1. Increasing the Wordlength of the Signal Representation in the Filter Sections	4
2. Using Wave Digital Filters	5
3. Using Controlled Quantization	6
4. Using Multirate Digital Filter with Periodically Varying Coefficients	6
5. Injection of Dither Signal	7
1.3 Preview of Results	8
 <u>CHAPTER 2</u>	
BASIC FILTER SECTIONS	14
2.1 Linear Basic Filter Sections	16
1. First-Order Filter Section	16
2. Second-Order Filter Section	23
2.2 Basic Filter Sections with Quantization	37
1. Quantization	41
(A) Rounding	42
(B) Truncation	42
2. Limit Cycles in the First-Order Filter Sections	48
(A) The First-Order Filter with Magnitude Truncation Quantizer	48

	<u>Page</u>
(B) The First-Order Filter with Rounding Quantizer	51
3. Limit Cycles in the Second-Order Filter Sections	54
(A) Limit Cycles in the Second-Order Filter Section with Rounding Quantizer	56
(B) Limit Cycles in the Second-Order Filter Section with Magnitude Truncation Quantizer	60
2.3 Summary	65
 <u>CHAPTER 3</u>	
ZERO-INPUT LIMIT CYCLES IN THE SECOND-ORDER DIGITAL FILTERS	
	69
3.1 Quantization Error in Presence of Roundoff	69
3.2 Classification of Limit Cycles	73
1. Classification of Limit Cycles Based on the Period	73
2. Classification of Limit Cycles Based on Accessibility Considerations	74
3. Classification of Limit Cycles Based on Symmetry Consideration	76
3.3 Successive-Value Phase-Plane Plot	81
3.4 Parameter Space	83
3.5 Different Properties of the Limit Cycles in the Second-Order Filters	85
3.6 Amplitude Bounds of Limit Cycles	87

	<u>Page</u>
1. Absolute Bound	87
2. RMS Bound	89
3. Approximate Bound	90
3.7 Frequency of Limit Cycles	92
3.8 Summary	94
 <u>CHAPTER 4</u>	
STABILIZATION BY THE INJECTION OF DITHER	96
4.1 The Proposed Method to Suppress Limit Cycles	96
4.2 Some Notes on the Proposed Method to Suppress Limit Cycles	99
4.3 Summary	102
 <u>CHAPTER 5</u>	
HOW DITHER AFFECTS THE LIMIT CYCLES	104
5.1 Verification of the Stabilization by the Use of Dither for a Particular Pair of Coefficient Values A and B	104
5.2 The Maximum Transition Time Needed to Move from any Limit Cycle to the Origin State	110
5.3 Verification of the Stabilization by the Injection of Dither for General Cases	115
5.4 Summary	121
 <u>CHAPTER 6</u>	
DITHER SIGNALS	123
6.1 Review and Discussion of the Previous Work on the Use of Dither for Limit Cycle Suppression	123

	<u>Page</u>
1. Randomised Quantization Method	123
2. The Method for Limit Cycle Suppression Proposed by Rashidi and Bogner	127
3. The Method for Limit Cycle Suppression Proposed by Büttner	131
4. Folding-Frequency Dither Method	132
6.2 Principal Considerations in the Dither Signal Design	134
6.3 Some Useful Dither Signals	136
1. Uniformly Distributed Dither Signal	136
2. Binary Random Dither	140
3. Bandstop Dither	143
6.4 A Note on the Dither Signals	148
6.5 Summary	152
 <u>CHAPTER 7</u>	
EXPERIMENTAL RESULTS	154
7.1 General Simulations on the Limit Cycles in the Second-Order Digital Filter Sections	155
1. Second-Order Filter Section with Quantization	155
2. Limit Cycles in the Second-Order Filter Sections	156
3. Limit Cycle Displaying	158
7.2 Generations of Three Types of Dither Signal in Simulation	158
1. Uniformly Distributed Random Dither	160
2. Binary Random Dither	160
3. Bandstop Dither	161

	<u>Page</u>
7.3 Simulation of Limit Cycle Suppression	176
7.4 Time for Stabilization	186
1. Results of the Use of Uniformly Distributed Random Dither	186
2. Results of the Use of Binary Random Dither	208
3. Results of the Use of Bandstop Dither	208
4. Summary About the Time for Stabilization	211
7.5 The Effect of Dither on the Output Noise	216
7.6 Summary	226
 <u>CHAPTER 8</u>	
CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH	229
 <u>APPENDICES</u>	
APPENDIX 1 Program Used with Computer PET to Display the Procedure of Limit Cycle Suppression on the State Plane	234
APPENDIX 2 Existence Conditions of Limit Cycles in the Second-Order Basic Filter Section with One Rounding Quantizer	236
APPENDIX 3 When the Coefficient Value B of the Second-Order Filter Section Satisfies $1 > B > 0.5$, the Origin State (0,0) on the State Plane becomes a Branch Point by the Use of Dither	242
APPENDIX 4 Uniformly Distributed Random Dither Tends to Linearize the Roundoff Quantization Characteristic	246

	<u>Page</u>	
APPENDIX 5	Program Used for the Generation of the Uniformly Distributed Random Dither	251
APPENDIX 6	Program Used with Computer PET to Display the Limit Cycles on the State Plane	253
APPENDIX 7	Program Used to Print Out the Limit Cycles on the State Plane	255
APPENDIX 8	Program Used for the Generation of a Normal (Gaussian) Distribution Sequence	259
APPENDIX 9	Program for the Calculation of Standard Deviation of Output from Bandstop Filter When a Gaussian Random Sequence $N(0,1^2)$ in Input	263
APPENDIX 10	Program for the Calculation of Time needed for Stabilization by the Use of Dither	266
APPENDIX 11	Program for the Calculation of the Increase in Output Noise from the Basic Filter Section by the Dither	271
<u>REFERENCES</u>		278

LIST OF TABLES

	<u>Page</u>
Table 1 The Example Shows that when a Periodic Limit Cycle Exists the Quantization Error is Also a Periodic Sequence	72
Table 2 The Relation Between the Limit Cycles of a Second-Order Filter with the Coefficients (-A, B) and a Second-Order Filter that has the Same Structure but Coefficients (A,B)	78
Table 3 The Transition Time to the Origin for Various Filter Sections with Uniform and Binary Dither Signals	189
Table 4 The Median Values of the Transition Time to the Origin State Corresponding to Various Filter Sections and Dithers	210
Table 5 The Median Transition Times to the Origin with Various Stop Band Widths of the Bandstop Filter in Bandstop Dither 1 Generation	212
Table 6 The Mean Time for Stabilization and the Increase in Output Noise with Sinusoidal Input for Various Filter Sections with Dithers	222

LIST OF FIGURES

	<u>Page</u>	
Fig. 1	Block diagram of a first-order digital filter	17
Fig. 2	Frequency responses of several first-order filters	19
Fig. 3	Pole and zero locations for first-order filters	20
Fig. 4	The impulse response of a stable ($ A < 1$) and unstable ($ A > 1$) first-order filter	22
Fig. 5	Block diagram of a second-order basic filter section	24
Fig. 6	Frequency responses of several second-order filters	26
Fig. 7	Pole and zero locations for a second-order filter	28
Fig. 8	The stable zone (triangle) of the linear second-order filter section	35
Fig. 9	Quantizer characteristic with rounding	43
Fig. 10	Quantizer characteristic with magnitude-truncation	44
Fig. 11	Quantizer characteristic with value-truncation	46

- Fig. 12 Probability density functions; (a) for roundoff error (b) for magnitude-truncation error and (c) for value-truncation error 47
- Fig. 13 Block diagram of a first-order digital filter with quantizer 49
- Fig. 14 Two ways of implementing the quantizations in the second-order filter sections (a) one quantizer version, (b) two quantizer version 55
- Fig. 15 Stability diagram for the second-order digital filter with one magnitude-truncation quantizer 61
- Fig. 16 Part of the state plane of the second-order filter ($A=-1.640625$, $B=0.953125$) with one magnitude-truncation, with initial conditions indicated from which a limit cycle will result 63
- Fig. 17 Stability diagram for the second-order digital filter with two magnitude-truncation quantizers 66
- Fig. 18 Successive-value phase plane plot for the second-order section with coefficient $A=-1.74$, $B=0.95833$ 82
- Fig. 19 Region of asymptotic stability for one rounding quantizer and regions where limit cycles can occur 84

	<u>Page</u>	
Fig. 20	Region of stability for two rounding quantizers and regions where limit cycles can occur	86
Fig. 21	The "Distribution Diagram" of the limit cycles in the example shows to which limit cycle each state belongs	107
Fig. 22	Transition diagram for the second-order section ($A=-1.74$, $B=0.95833$) with uniformly distributed random dither	111
Fig. 23	Equivalent quantizer	116
Fig. 24	Characteristic of equivalent quantizer Q_e	117
Fig. 25	The characteristic of equivalent quantizer in the randomized quantization method	125
Fig. 26	Block diagram of the method for limit cycle suppression proposed by Rashidi and Bogner	128
Fig. 27	The characteristic of equivalent quantizer in Fig. 26	129
Fig. 28	The probability distribution function of the uniformly distributed random dither signal	137
Fig. 29	The relative frequency histogram of the uniformly distributed random dither	139

	<u>Page</u>	
Fig. 30	Probability distribution function of the binary dither	142
Fig. 31	Block diagram of the bandstop dither generation	147
Fig. 32	The state plane plot shows which limit cycle each state belongs to for the filter section (A=-1.74, B=0.95833) with two quantizers	159
Fig. 33	The power density spectrum of Gaussian random sequence used in the research. The FFT with Hamming Window was repeated 127 times. The average values are shown here	163
Fig. 34	The probability density function of the Gaussian random sequence generated by program	164
Fig. 35	The characteristic of nonlinear memoryless network in the bandstop dither 1	169
Fig. 36	The relative frequency histogram of the bandstop dither 1. The nonlinear network whose characteristic is $\frac{1}{2}\text{erf}(Y_1)$ has been used	172
Fig. 37	The relative frequency histogram of the bandstop dither 2. The nonlinear network whose characteristic is $\frac{1}{2}\text{erf}\left(\frac{Y_1}{\sqrt{2}}\right)$ has been used	173

	<u>Page</u>	
Fig. 38	The power density spectrum of the bandstop dither 1 whose stop band width was equal to (0.08~0.14)Fs. The nonlinear function $\frac{1}{2}\text{erf}(Y_1)$ was used	174
Fig. 39	The power density spectrum of the bandstop dither 2 whose stop band width was equal to (0.08~0.14)Fs. The nonlinear function $\frac{1}{2}\text{erf}\left(\frac{Y_1}{\sqrt{2}}\right)$ was used	175
Fig. 40	An example of limit cycle and its suppression (a) without dither (b) with uniform dither	177
Fig. 41	An example of limit cycle suppression by the injection of binary dither	180
Fig. 42	The trajectory corresponding to Fig. 41 on the state plane	181
Fig. 43	The experimental block diagram of the limit cycle suppression in the second-order filter section with bandstop dither	182
Fig. 44-	The cumulative distribution function of the	
Fig. 61	transition time to the origin state for various filter sections with different dithers	190-207
Fig. 62	Frequency response of a Hamming Window	220

	<u>Page</u>
Fig. 63 If \bar{X} lies within the limits $\mu - 1.96\sigma_{\bar{X}}$ and $\mu + 1.96\sigma_{\bar{X}}$ then μ must lie within the limits $\bar{X} - 1.96\sigma_{\bar{X}}$ and $\bar{X} + 1.96\sigma_{\bar{X}}$	225
Fig. A4.1 (a) Cumulative probability distribution of the uniform dither	
(b) Cumulative probability distribution of the composite signal (dither and quantization error)	247
Fig. A4.2 Equivalent quantizer	249
Fig. A4.3 (a) Roundoff characteristic with dither	
(b) Statistical characteristic of the equivalent quantizer	250
Fig. A7.1 An example of the print by the use of the program shown in Appendix 7	258

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTORY REMARKS

When a digital filter is implemented either by software or by hardware, numbers are ultimately stored in finite-length registers. Consequently, coefficients and signal values must be quantized so that they can be stored. In this circumstance, errors due to finite precision in the representation of numbers are unavoidable. The quantization characteristic is a nonlinearity which gives rise to nonlinear effects such as limit cycle oscillation as well as approximation in a filter realization^(1,2). Limit cycle oscillations are undesirable, except in the digital oscillator applications.

The suppression of limit cycle oscillation has been discussed by many authors^(1,3,4,5). Many methods of limit cycle suppression by the use of dither have been proposed. Although in most cases, the methods reported are effective for limit cycle suppression they have some disadvantages. In some cases, not all limit cycles are suppressed. In other cases, even the zero-input limit cycle oscillation has been suppressed there is still some noise at the filter output. This research is mainly concerned with the limit cycle suppressions by the injection of somewhat different

dither signals which have no disadvantages mentioned above.

This dissertation first contains a brief review of relevant information on digital filter limit cycles and discusses the previous work on the use of dither for their suppression. Secondly, it shows how a random dither signal can be used to suppress limit cycles and yet, in the steady state, a zero valued output results when the input signal is zero. Then, the principal considerations in the dither signal design are introduced. Three types of dither signal are proposed. Extensive simulations have verified that the limit cycles in the second-order sections can be suppressed by the use of any one of the three types of dither and once the zero-input limit cycle has been suppressed the output of the filter remains zero. Two specifications, the time needed to stabilise the filter and the increase in output noise by the dither, are investigated.

At least four factors have to be considered when implementing a filter. They are:-

- (1) Selection of a specific configuration for the filter.
- (2) Choice of the arithmetic mode, i.e., the number system to be used.
- (3) Choice of the type of quantization, and

- (4) Specification of the number of significant digits.

Limit cycle oscillations may occur in fixed-point implementations of recursive digital filters. A recursive digital filter is defined as a filter in which the present output depends on the present input and past inputs and outputs, while for a nonrecursive filter the output depends on past and present inputs only. Most digital filters are of the fixed-point variety because floating-point arithmetic involves more hardware. Also, most filters are recursive because for the same degree of approximation, recursive filters are generally simpler than nonrecursive forms. A cascade or parallel form composed of first- and second-order subfilters is preferable over any direct realization of a higher order digital filter. Thus, in practice, a higher order filter is obtained by combining second-order sections. For these reasons, in this dissertation only the fixed-point implementations of second-order recursive digital filter sections are considered. These filter sections normally include feedforward coefficients, as well as feedback coefficients. The feedforward coefficients are not considered in this dissertation because they are not relevant to the limit cycle suppression problem. Throughout the paper, it is assumed that quantization is performed by rounding.

Many detailed studies of limit cycles in digital filters have been made. A comprehensive summary of this

work is given in Reference 6. Some authors have developed bounds on the amplitude of the limit cycle oscillations in terms of the filter section feedback coefficient values A and B^(7,8,9). Other work has determined the range of values of A and B for which various types of limit cycle may exist⁽¹⁰⁾. Some important research about limit cycle oscillations is also published in two special issues of the IEEE Transactions^(11,12).

1.2 METHODS FOR LIMIT CYCLE OSCILLATION SUPPRESSION

In addition to the injection of the dither there are several other methods for limit cycle suppression. In order to understand the application of the proposed methods in this dissertation it is necessary to introduce other methods of limit cycle suppression. In this section the methods for limit cycle suppression will be discussed briefly.

1. Increasing the Wordlength of the Signal Representation in the Filter Sections

Because the amplitude of limit cycle oscillation in the digital filter is proportional to the quantization step q it will be decreased by the reduction of the quantization step. For a fixed signal dynamic range, each bit increase in the wordlength of the signal representation will make the quantization step be half smaller. In other

words, it will suppress the limit cycle oscillation by 6 dB. If the wordlength in the filters can be increased sufficiently the limit cycle oscillations can be ignored but, of course, still cannot be eliminated totally. As will be seen later, in some cases, the amplitude of limit cycle may be much bigger (one hundred q , for example) than the quantization step, thus a big extra bit is needed. This extra bit requirement in wordlength will increase the complexity and the cost of the digital filters very much. It is worth noting that because the zero-input limit cycle oscillation is a correlated noise, it is even more harmful than normal noise. In some applications, the zero-input limit cycles may be not tolerable. Therefore, a more efficient method of limit cycle suppression is needed in practice.

2. Using Wave Digital Filters

Fettweis proposed some digital filter structures related to classical filter networks called wave digital filters (WDF)⁽¹³⁾. Fettweis and Meerkötter have been able to prove that the absence of zero-input limit cycles can be guaranteed in WDF if the ideal linear counterpart is pseudopassive and if the nonlinear modifications required by the finite arithmetic are carried out in such a way that the absolute values of nonlinear component output is less than or equal to that of the linear counterpart^(14,15). This condition is satisfied by the characteristic of a

magnitud^e truncation quantizer. This means that it is possible to design a wave digital filter of arbitrary order, without limit cycles.

Although WDF may be free of limit cycle oscillations, second-order direct form sections are likely to remain in use for some time. This is partly because of the investment that has been made in implementing such filter sections as integrated circuits⁽¹⁶⁾ and also because filters using a cascade of second-order sections are easy to design.

3. Using Controlled Quantization

Controlled quantization has been proposed⁽¹⁷⁾ whereby the signal is quantized to a larger or a smaller value depending on the state variables in the filter. With a proper design of such a "controlled rounding" arithmetic, the most relevant limit cycles in digital filters can be suppressed. An algorithm has been given which guarantees the absence of limit cycles of periods larger than two. The disadvantages of this method are, first, some constant or alternating limit cycles still cannot be suppressed, second, it seems a bit complicated to be implemented.

4. Using Multirate Digital Filter with Periodically Varying Coefficients

Wong and King⁽¹⁸⁾ have shown that a multirate digital filter with periodically varying coefficients is capable of suppressing limit cycle oscillation in the output

completely, provided that the coefficients are suitably chosen.

The disadvantage of this method is that it is complicated for ^{the} complementing a simple second-order filter section and before the method is used, one has to do a lot of experiments so can choose the coefficients of multirate filter suitably. Because a different second-order filter section has different coefficients of the multirate filter which has no limit cycle. Perhaps that is why this method has not been applied widely yet.

5. Injection of Dither Signal

For many years, it has been known that limit cycle oscillations in continuous-time nonlinear feedback systems can often be suppressed by the injection of a dither signal^(19,20). A digital filter is a discrete-time, nonlinear system and some attention has been paid to the possibility of suppressing limit cycle oscillations in digital filters by the use of dither. Several methods for limit cycle suppression have been proposed^(1,3,4,5).

These methods will be reviewed and discussed in Chapter 6. In this research, somewhat different dither signals have been used. The methods used have no disadvantages of the methods proposed before.

The essential disadvantage of limit cycle suppression by the injection of a small random dither signal is that

it introduces a new random noise in the output. But as will be seen later, by the use of the bandstop dither which is a new type of dither used in our research the increase in output noise is very small.

1.3 PREVIEW OF RESULTS

From the remarks of the preceding section it follows that the limit cycle oscillations occurring in fixed-point implementation of recursive second-order digital filters can be suppressed by the injection of dither signal. In this section, the major results of the following chapters are previewed.

In Chapter 2, the main properties of the basic sections are given. This information is necessary for understanding the following chapters. It is shown that the direct form is inferior to both the cascade and parallel form when the effect of coefficient quantization errors and roundoff noise after arithmetic operations are considered. The first- and second-order filters are basic building blocks from which all higher order systems can be synthesized. The zeros of the digital filters do not change the nature of the limit cycle but influence the magnitude of the limit cycle amplitude. Therefore, the basic section which has two zeros at the origin on the z -plane is considered as a basic configuration in this dissertation.

In section 2.1, the stable region of the linear second-order filter section (without quantization) is derived which is bounded by a triangle on the parameter plane.

In section 2.2, two types of quantization, magnitude-truncation and roundoff are discussed. It shows in this section that in the first-order filter with magnitude-truncation quantizer, no limit cycle can be sustained, but with rounding quantizer, the constant or alternating limit cycles can exist. With rounding quantizer, variety limit cycles may exist in the second-order section. In the two quantizer version with magnitude-truncation quantization only limit cycles of periods 1 and 2 can be sustained. In the one quantizer version with magnitude-truncation quantization limit cycles will be possible only for very few values of A and B on the parameter plane. The magnitude-truncation quantization has a certain advantage over roundoff with respect to the occurrence of limit cycles but its quantization error is bigger than that with roundoff quantization.

In Chapter 3, the main properties of zero-input limit cycles in the second-order filter sections are discussed. It shows that the second-order filter sections with multiplication coefficients B for which $|B| > 0.5$ will always exhibit limit cycles. The existing conditions of limit cycles in the second-order filter section with one

rounding quantizer are derived in the appendix. Three different types of amplitude bound for limit cycles in the second-order filter sections are introduced and discussed. The frequency expression of the impulse response in linear second-order filter can be used as an approximation of the frequency of limit cycles. Especially, when the poles of the filter ^{are} close to the unit circle in the z-plane the frequency estimate becomes more accurate.

The proposed method to suppress the limit cycles in the second-order filter sections is described in Chapter 4. It shows that the use of the dither may cause the filter to leave the limit cycles and make the origin state (0,0) on the state plane be a branch point. Once the origin state has been reached the output of the filter remains zero as long as the input signal is zero. These properties of the proposed method support us to speculate that the dither will suppress all limit cycles in any second-order filter sections eventually.

In Chapter 5, the necessity of the limit cycle suppression in the second-order filter sections is proved, though partly, on the experimental basis. Because the quantization nonlinearities occurring in digital filters are highly discontinuous functions, it is difficult to prove strictly the stabilisation. But for a specified pair of coefficient values A and B, it is possible to verify strictly whether or not the dither will stabilise the filter. In this

chapter, the verification procedures are described with an example. By the transition matrix, the maximum transition time needed for transition from any limit cycle to the origin state can be calculated. The result is verified by simulation.

The previous work on the use of dither for limit cycle suppression is first reviewed and discussed in Chapter 6. Then, the principal considerations in the dither signal design are described. The dither signal should be a random signal distributed in the open range $(-\frac{q}{2}, \frac{q}{2})$. Three types of dither signal are derived from the principal considerations of the dither signal design. They are uniformly distributed random dither, binary random dither and bandstop dither.

In Chapter 7, the experimental results are introduced. Extensive simulations have verified that all the limit cycles in the second-order filter sections can be suppressed by the use of any one of the three types of dither. There is no remaining noise at the output of filter with zero-input signal.

When dither signal is used to stabilise a digital filter, two specifications are of particular interest. One of these is the length of time taken for the filter, with zero input, to reach the state plane origin from a limit cycle. The other specification of interest is the increase in the output noise from the filter above the

quantization noise which is present when nonzero input signals are applied without dither.

As far as the transition time is concerned, in the three types of dither, the preferred order is the binary random dither, bandstop dither and uniform dither. The mean times for the former two types of dither to effect stabilisation are comparable with the decay time for the strictly linear filter.

As far as the increase in output noise is concerned, in the three types of dither, the order of preference is the bandstop dither, the uniform random dither and the binary random dither. The increase in output noise by the bandstop dither when an input signal is present is small - equivalent to only a small fraction of one bit of the filter wordlength.

There have been a number of integrated circuit implementation of second-order direct form digital filter sections, in which large investments have been made. The results in this paper should be useful when these integrated circuits needed to be used for applications where limit cycles are not tolerable.

The main conclusions about the research are given in Chapter 8. This chapter concludes with an indication of those problems which remain subject to further research.

The major results of this research have been presented in Saraga Memorial Colloquium on Electronic Filters⁽⁴⁵⁾. One paper about this research has been prepared for publication⁽⁴⁶⁾.

CHAPTER 2

BASIC FILTER SECTIONS

This chapter provides a general discussion of the first- and second-order digital filters which are basic building blocks in constructing higher-order digital filters. The information in this chapter is necessary for understanding the following chapters. The conclusions in this chapter will be used later.

Kaiser⁽²¹⁾ has shown that the high order direct form digital filter should be avoided because of coefficient sensitivity, i.e., the effect of change of numerical coefficients of the filter causes large variations in the filter response. Also, Knowles and Edwards⁽²²⁾ have concluded that the direct form is inferior to both the cascade and parallel form when the effect of roundoff errors after arithmetic operations is considered. In a paper by Edwards, Bradley and Knowles⁽²³⁾, the above mentioned conclusions have been verified using the 11th order elliptical bandstop filter. Taking scaling into account to assure the proper dynamic range for the filter, the ratio of the rms noise level due to roundoff after multiplication for the direct form, to the rms noise of the parallel or the cascade form, was about $10^{12}:1$.

The second-order section has been chosen as a basic

block because this is the minimum order for realising a pair of complex conjugate roots such that the polynomials of the numerator and denominator of the transfer function have real coefficients. Real roots can be realised in pairs also, except for the case where the order of the filter is odd, in which case use of a first-order section becomes necessary. Therefore, the first- and second-order filters are basic building blocks from which all higher order systems can be synthesized. In addition, Hess⁽²⁴⁾ has shown that the zeros of a filter are not to change the nature of the limit cycle, but to influence the magnitude of the limit cycle amplitude. It is for these reasons that the study of limit cycles and their suppression in discrete-time systems will be restricted to the second-order basic section which has two zeros at the origin on the z-plane. The first-order filter can be considered as a degenerated case of the second-order filter. In this thesis, the first- and the second-order filter which has two zeros at the origin on the z-plane are called basic filter sections or basic sections.

First, we assume that both the values of sequences and the coefficients of linear filter have infinite bit precision. Later on, we shall discuss the finite word length effects in digital filters where the amplitudes are quantized to some specified accuracy.

2.1 LINEAR BASIC FILTER SECTIONS

As mentioned above, in this section both the sample values of sequences and the coefficients of linear filter are assumed to have infinite bit precision.

1. The First-Order Filter Section

Fig. 1 shows the first-order filter section. The corresponding difference equation is

$$Y(n) = X(n) - AY(n-1) \quad (1)$$

where $X(n)$ represents the input sequence and $Y(n)$ denotes the filter output signal after n sampling intervals each with a duration of sampling period T_s . A is the coefficient value of the filter.

Its transfer function is

$$H(z) = \frac{1}{1+Az^{-1}} \quad (2)$$

The impulse response, $h(n)$, is readily obtained as

$$h(n) = \begin{cases} (-A)^n & n \geq 0 \\ 0 & n < 0 \end{cases} \quad (3)$$

Substituting the equation

$$z = e^{j\omega}$$

into Eqn. (2) the frequency response of the first-order filter can be obtained as

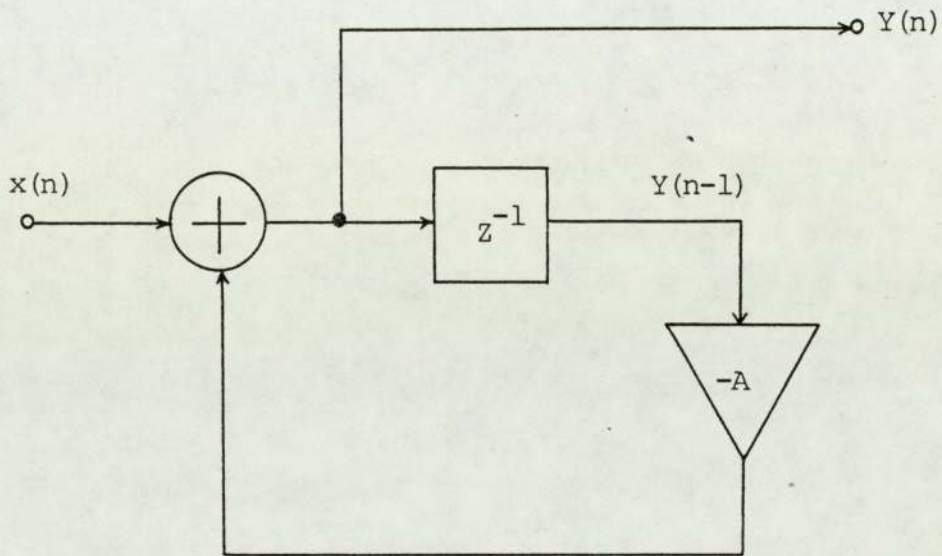


Fig. 1 Block diagram of a first-order digital filter

$$H(e^{j\omega}) = \frac{1}{1+ Ae^{-j\omega}} \quad (4)$$

Representing $H(e^{j\omega})$ as

$$H(e^{j\omega}) = |H(e^{j\omega})| e^{j\angle H(e^{j\omega})}$$

gives

$$|H(e^{j\omega})| = \frac{1}{(1+A^2+2A \cos\omega)^{\frac{1}{2}}} \quad (5)$$

$$\begin{aligned} \angle H(e^{j\omega}) &= \tan^{-1} \frac{A \sin\omega}{1+A \cos\omega} \\ &= \omega - \tan^{-1} \left(\frac{\sin\omega}{\cos\omega + A} \right) \end{aligned} \quad (6)$$

Fig. 2⁽²⁵⁾ shows plots of $\log|H(e^{j\omega})|$ and $\angle H(e^{j\omega})$ for various values of A . As can be seen from this figure, the first-order filter has a lowpass characteristic.

The zero and pole of the transfer function can be obtained from Eqn (2). The zero is at the origin on the z -plane. And the pole can be determined by $p=-A$ which is on the real axis of the z -plane. Fig. 3 shows the positions of the zero and pole in the z -plane.

A linear, time-invariant system is said to be stable if every bounded input produces a bounded output. A necessary and sufficient condition on the impulse response for stability is

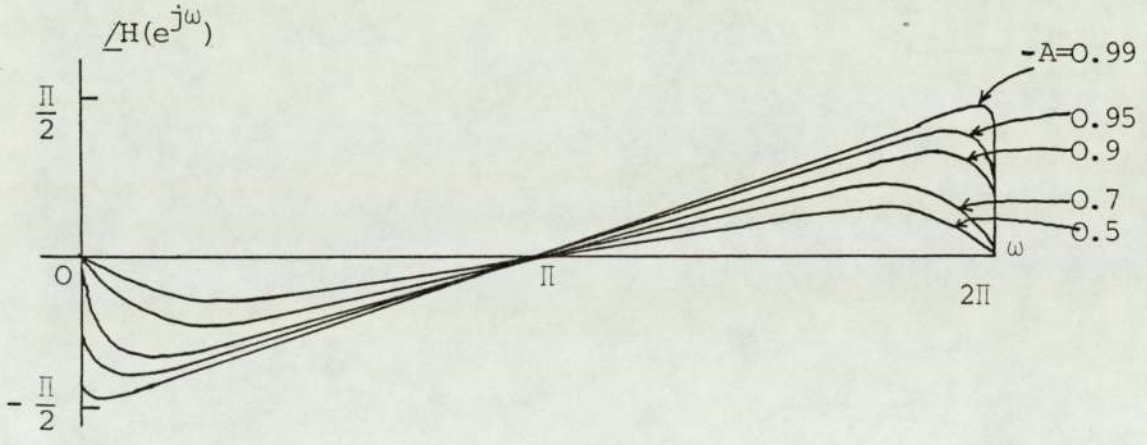
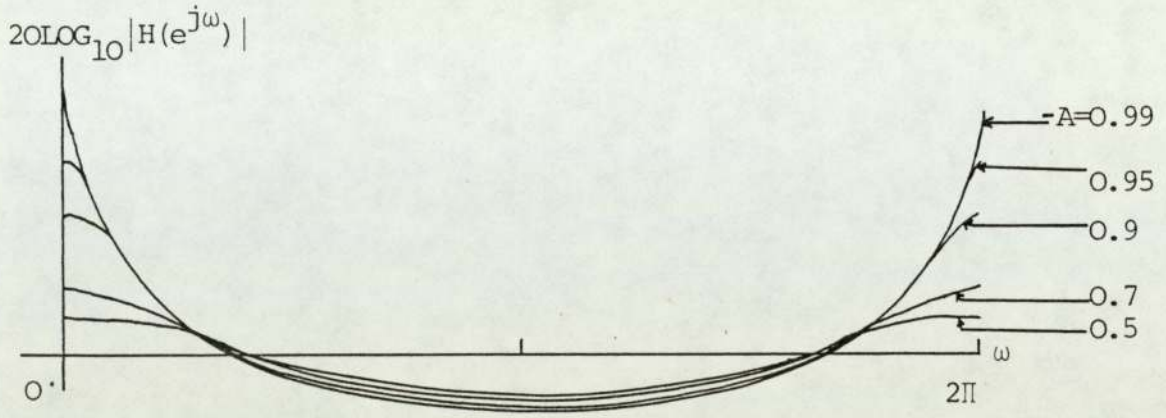
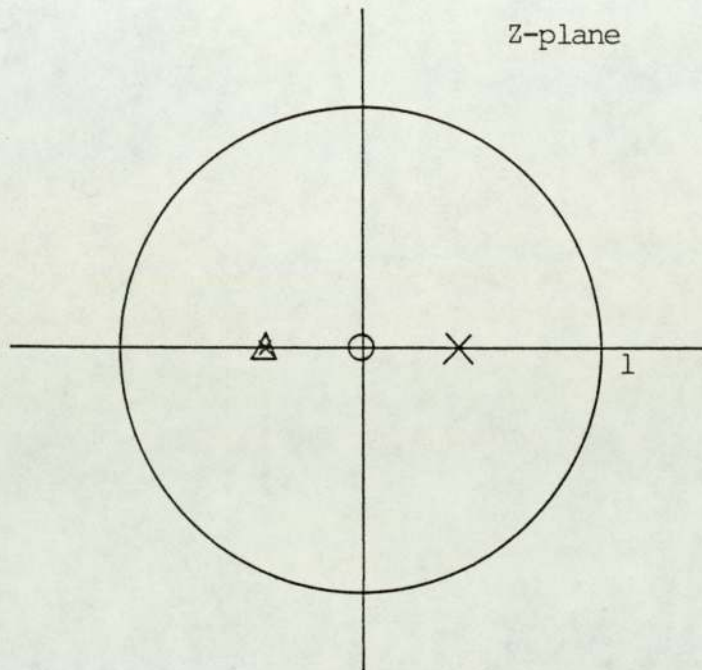


Fig.2 Frequency response of several first-order filters



- shows the position of zero.
- × shows the position of pole when $A < 0$
- Δ shows the position of pole when $A > 0$

Fig. 3 Pole and zero locations for first-order filters.

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty \quad (7)$$

According to Eqn. (7) from Eqn. (3) we know that the stable condition for the first-order filter section is

$$|A| < 1 \quad (8)$$

Fig. 4 shows the impulse responses when $|A| > 1$ and $|A| < 1$ respectively.

For the first-order filter with $|A| < 1$ under the zero-input condition from any initial state the response of the filter will eventually tend to zero. For example, suppose $A = -0.875$ and the initial condition $Y(-1) = 8$, its zero-input response is shown in the following table.

n	0	1	2	3	4	5	6
Y(n)	8	7	6.125	5.359	4.689	4.103	3.590
n	7	8	9	10	11	12	13
Y(n)	3.142	2.749	2.405	2.105	1.842	1.611	1.410
n	14	15	16	17	18	19	20
Y(n)	1.234	1.079	0.945	0.826	0.723	0.633	0.554
n	21	22	...	40	...		
Y(n)	0.484	0.424	...	0.038	...		

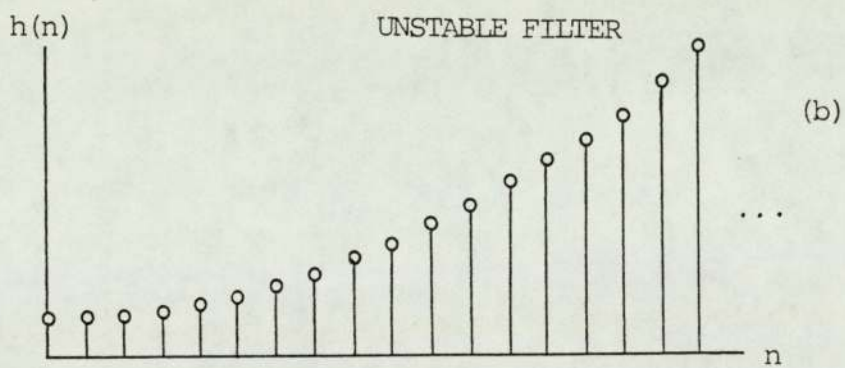
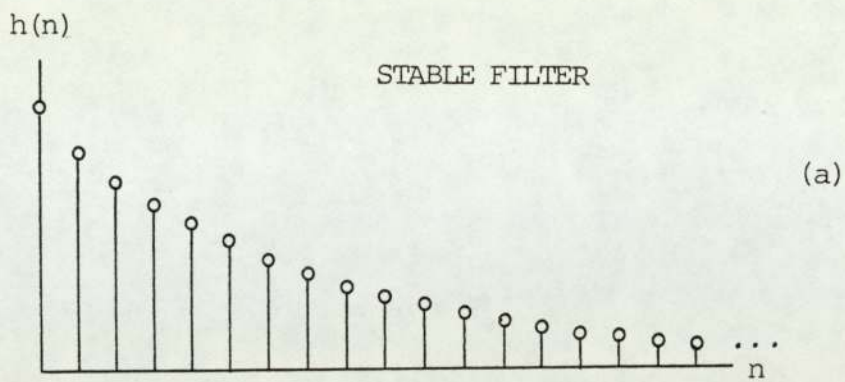


Fig. 4 The impulse response of a stable ($|A| < 1$) and unstable ($|A| > 1$) first-order filter.

2. Second-Order Filter Section

Fig. 5 shows the block diagram of the second-order basic filter section. The corresponding difference equation may be written in the form

$$Y(n) = X(n) - AY(n-1) - BY(n-2) \quad (9)$$

and its transfer function is

$$H(z) = \frac{1}{1 + Az^{-1} + Bz^{-2}} \quad (10)$$

If we assume the initial conditions $Y(-1)=0$ and $Y(-2)=0$, then the impulse response is readily shown to be one of two types.

TYPE 1

If the poles both are real (but not equal to each other), then

$$h(n) = \alpha_1 (p_1)^n + \alpha_2 (p_2)^n \quad (11)$$

where p_1 and p_2 are real poles and α_1, α_2 are constant.

TYPE 2

If the poles are conjugate complex, then

$$h(n) = \frac{r^n}{\sin\theta} \sin[(n+1)\theta] \quad (12)$$

where $r = \sqrt{B}$ and $\theta = \arccos \frac{-A}{2\sqrt{B}}$

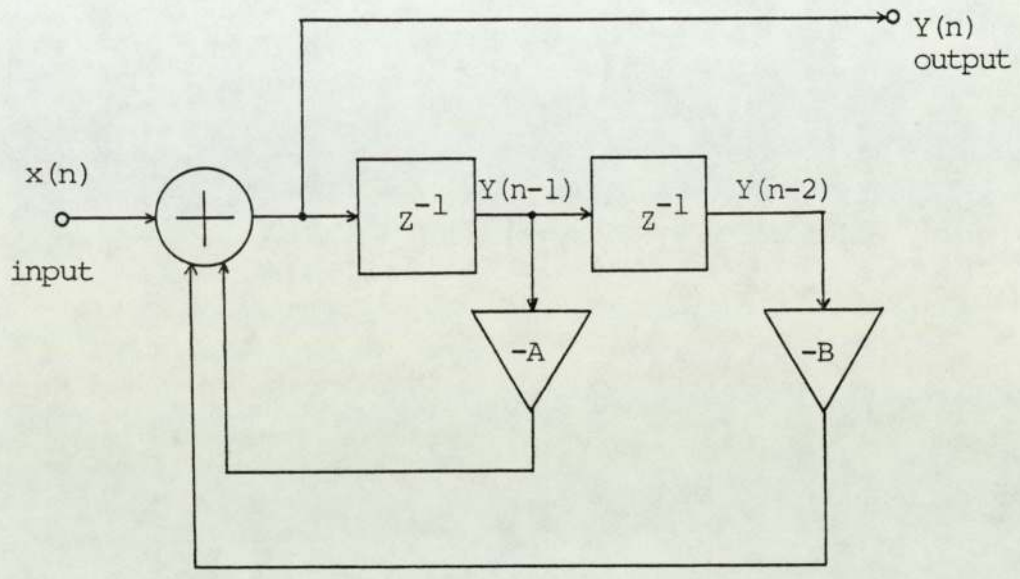


Fig. 5 Block diagram of a second-order basic filter section.

The frequency expression can be written as

$$\omega = \frac{1}{T_s} \arccos\left(-\frac{A}{2\sqrt{B}}\right) \quad (13)$$

For $B=1$, the impulse response is a sinusoid with constant amplitude and frequency

$$\omega = \frac{1}{T_s} \arccos\left(\frac{-A}{2}\right)$$

Type 1 represents two first-order systems, and the frequency response of the first-order section has been considered before.

The frequency response corresponding to Type 2 can be written as

$$H(e^{j\omega}) = \frac{1}{1 - 2rcos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \quad (14)$$

The log magnitude and phase response of second-order systems corresponding to a fixed value of $\theta (\frac{\pi}{4})$ and varying r , are shown in Fig. 6⁽²⁵⁾. From these plots it is clear that a second-order system represents a simple digital resonator.

The complex poles are readily obtained from Eqn. (10)

$$P_{1,2} = \frac{-A \pm \sqrt{A^2 - 4B}}{2} \quad (15)$$

The existing condition of complex poles is

$$A^2 - 4B < 0 \quad \text{or} \quad B > \frac{A^2}{4}$$

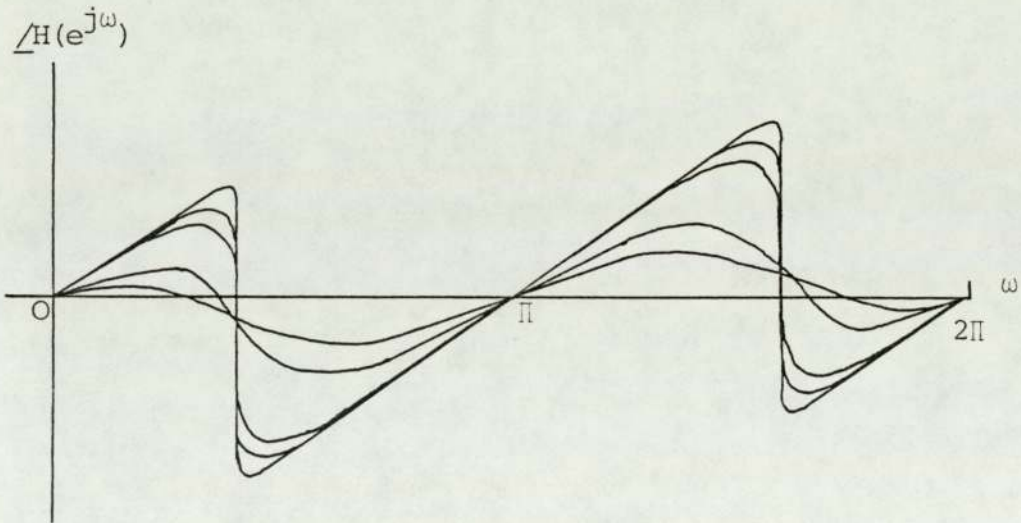
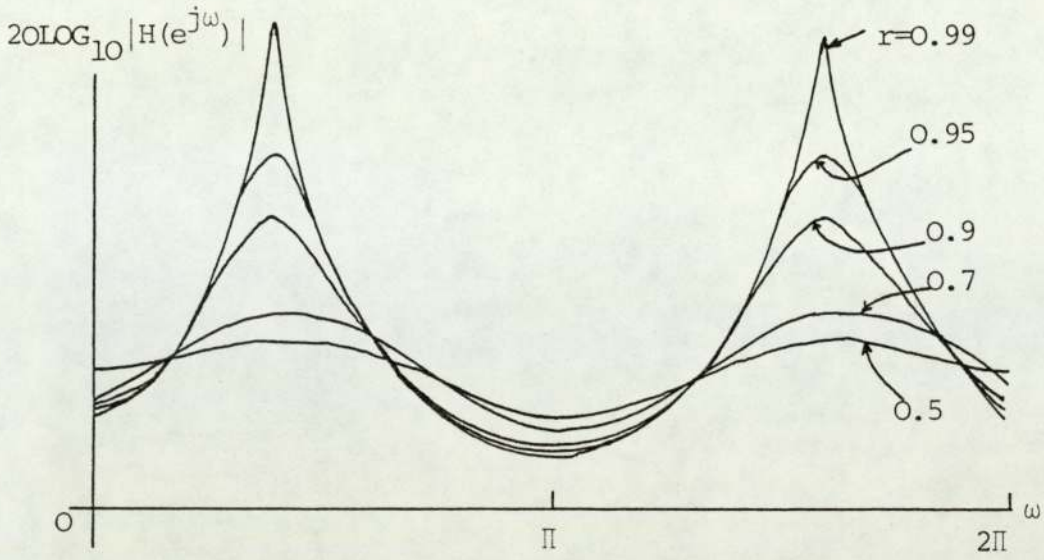


Fig. 6 Frequency response of several second-order filters

which implies

$$B > 0. \tag{16}$$

Fig. 7 shows the positions of the zeros and poles of the second-order filter section.

For the comparison of the time needed for limit cycle suppression, in a later chapter, the zero-input decay response of the linear second-order filter with specified initial conditions is needed. This response can be obtained by using state-space techniques⁽²⁶⁾. Consider a filter F_0 in which

$$Y(n) = \sum_{i=0}^N a_i X(n-i) - \sum_{i=1}^N b_i Y(n-i) \tag{17}$$

Then the Nth-order filter can be represented by the system

$$\overline{q}(n+1) = \overline{A} \overline{q}(n) + \overline{B}X(n) \tag{18}$$

$$Y(n) = \overline{C} \overline{q}(n) + \overline{D}X(n) \tag{19}$$

where \overline{A} , \overline{B} , \overline{C} and \overline{D} are the matrices defined as follows:

$$\overline{A} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ -b_N & -b_{N-1} & \dots & \dots & \dots & -b_1 \end{bmatrix} \tag{20}$$

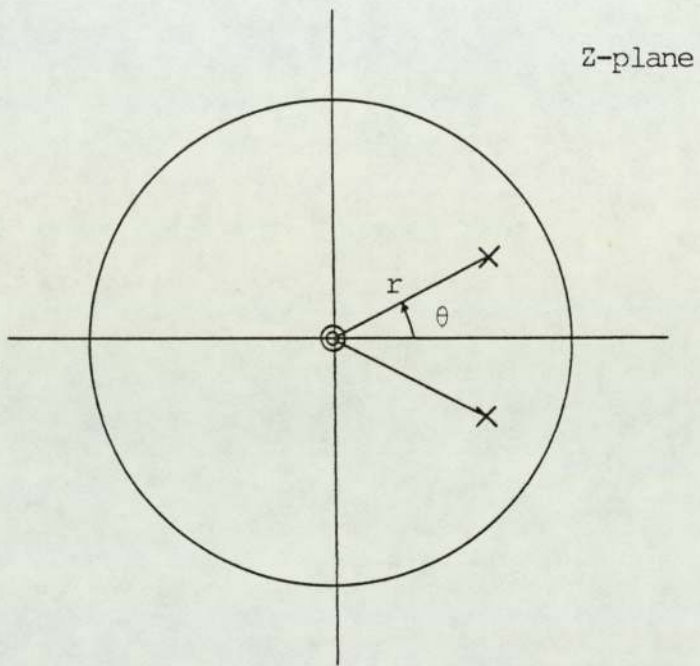


Fig. 7 Pole and zero locations for a second-order filter.

$$\bar{B} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ 0 \\ 1 \end{bmatrix} \quad (21)$$

$$\bar{C} = [c_1 \quad c_2 \quad \dots \quad c_N] \quad (22)$$

where

$$c_1 = a_N - a_0 b_N$$

$$c_2 = a_{N-1} - a_0 b_{N-1}$$

.....

$$c_N = a_1 - a_0 b_1$$

and

$$\bar{D} = [a_0] \quad (23)$$

The state matrix, $\overline{q(n+1)}$, consists of N auxiliary variables $q_1(n), q_2(n), \dots, q_N(n)$ which are called state variables.

$$\overline{q(n+1)} = \begin{bmatrix} q_1(n+1) \\ q_2(n+1) \\ \cdot \\ q_{N-1}(n+1) \\ q_N(n+1) \end{bmatrix} \quad (24)$$

A second-order basic filter section is characterised by

$$Y(n) = X(n) - AY(n-1) - BY(n-2)$$

It can be verified that in the second-order section case the matrices are

$$\bar{A} = \begin{bmatrix} 0 & 1 \\ -B & -A \end{bmatrix} \quad (25)$$

$$\bar{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (26)$$

$$\bar{C} = [-B \quad -A] \quad (27)$$

and

$$\bar{D} = [1] \quad (28)$$

For a zero-input second-order section, let the initial state be $Y(-1)$ and $Y(-2)$.

Then

$$\overline{q(0)} = \begin{bmatrix} Y(-2) \\ Y(-1) \end{bmatrix} \quad (29)$$

and the input

$$X(K) = 0 \quad K=0,1,2,\dots,n-1$$

For $n = 0, 1, 2, \dots, n-1$ Eqn. (18) gives

$$\overline{q(1)} = \bar{A} \overline{q(0)}$$

$$\overline{q(2)} = \bar{A} \overline{q(1)}$$

$$\overline{q(3)} = \bar{A} \overline{q(2)}$$

... ..

Hence

$$\overline{q(2)} = \bar{A}^2 \overline{q(0)}$$

$$\overline{q(3)} = \bar{A}^3 \overline{q(0)}$$

and in general

$$\overline{q(n)} = \bar{A}^n \overline{q(0)} \tag{30}$$

From Eqn. (19) it is clear that

$$Y(n) = \bar{C} \bar{A}^n \overline{q(0)}$$

or

$$Y(n) = \begin{bmatrix} -B & -A \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -B & -A \end{bmatrix}^n \begin{bmatrix} Y(-2) \\ Y(-1) \end{bmatrix} \tag{31}$$

If the coefficient values of the second-order section A , B and the initial condition $Y(-1)$, $Y(-2)$ are known, for any n , the response of the filter can be obtained from the above equation. But as can be seen when n becomes large the calculating of $Y(n)$ is rather troublesome. The convenient way to get $Y(n)$, the zero-input response of the second-order

filter section, is to simulate the section by the use of a small program.

Now let us derive the stable condition of a linear second-order section which is relative to the existing condition of the limit cycles in the nonlinear counterpart. The second-order filter section as shown in Fig. 5 is a closed-loop system. As mentioned before, its closed-loop transfer function $H(z)$ can be written as

$$H(z) = \frac{1}{1+G(z)} = \frac{1}{1+Az^{-1}+Bz^{-2}} \quad (32)$$

where $G(z) = Az^{-1}+Bz^{-2}$ is its open-loop transfer function.

According to the Nyquist criterion, the closed-loop system will be stable if the point $(-1,0)$ is not encircled by the polar plot of $G(j\omega)$ for $-\infty < \omega < \infty$.

The frequency response of the open-loop transfer function $G(j\omega)$ can be achieved by letting $z = \exp(j2\pi fT_s)$ where T_s is the sampling period.

Hence,

$$\begin{aligned} G(j\omega) &= G(z) \Big|_{z=e^{j2\pi fT_s}} \\ &= A \exp(-j2\pi fT_s) + B \exp(-j4\pi fT_s) \end{aligned} \quad (33)$$

Let $2\pi fT_s = \gamma$, Eqn. (33) becomes

$$\begin{aligned} G(\gamma) &= A\exp(-j\gamma) + B\exp(-j2\gamma) \\ &= A\cos\gamma + B\cos 2\gamma - j(A\sin\gamma + B\sin 2\gamma) \end{aligned} \quad (34)$$

In order to check whether the $G(j\omega)$ encircles the point $(-1,0)$, it is enough to check whether the intersections of $G(j\omega)$ with the real axis lie on the right of the point $(-1,0)$. If all the intersections are on the right of the point $(-1,0)$ the closed-loop system will be stable. These intersections can be found by letting the imaginary part of Eqn. (34) be equal to zero.

$$\begin{aligned} \text{Im}G(\gamma) &= -(A\sin\gamma + B\sin 2\gamma) \\ &= -\sin\gamma(A + 2B\cos\gamma) \\ &= 0 \end{aligned} \quad (35)$$

Eqn. (35) leads to two equations

$$\sin\gamma = 0 \quad (36)$$

$$A + 2B\cos\gamma = 0 \quad (37)$$

There may be three intersections:

(A) $\gamma = 0$, corresponding to $f=0$ condition.

The real intersection is

$$G(\gamma) \Big|_{\gamma=0} = A+B \quad (38)$$

This real intersection will be on the right of the

point $(-1,0)$,

$$\text{if } A+B > -1 \quad (38)$$

$$\text{or } 1+A+B > 0 \quad (39)$$

(B) $\gamma = \pi$, corresponding to $f=(2T_s)^{-1}$

The real intersection is

$$G(\gamma) \Big|_{\gamma=\pi} = -A+B \quad (40)$$

This real intersection will be on the right of the point $(-1,0)$,

$$\text{if } -A+B > -1$$

$$\text{or } 1-A+B > 0 \quad (41)$$

(C) $\cos \gamma = -\frac{A}{2B}$ corresponding to periodic oscillation

The real intersection is

$$G(\gamma) \Big|_{\gamma = \arccos(-\frac{A}{2B})} = -B \quad (42)$$

This real intersection will be on the right of the point $(-1,0)$,

$$\text{if } -B > -1$$

$$\text{or } B < 1 \quad (43)$$

A filter will be linear stable, only if all real intersections lie on the right of the point $(-1,0)$. Eqns(39), (41) and (43) define a triangle in the A,B parameter plane as shown in Fig. 8. Any linear second-order digital filters are stable only if its coefficients are within this triangle

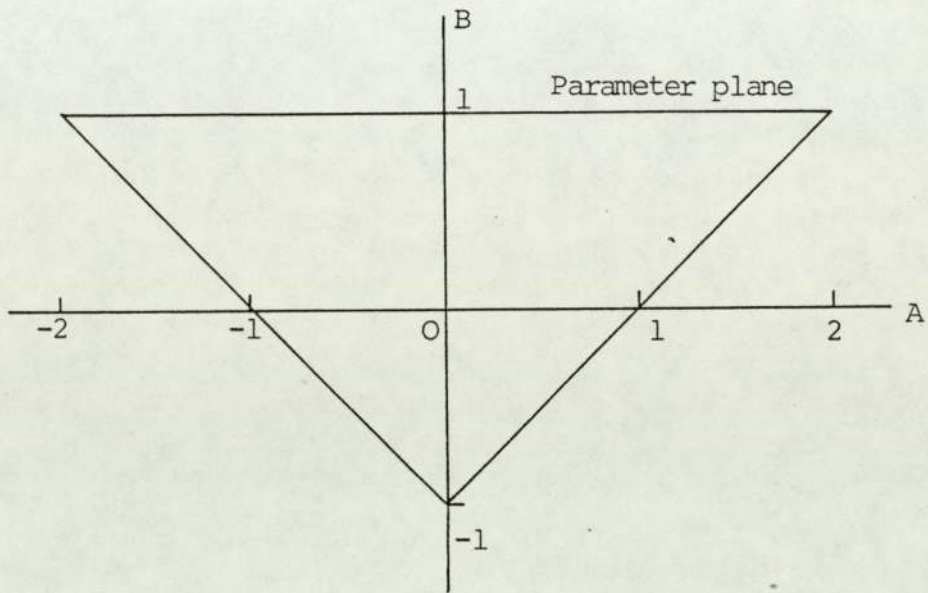


Fig. 8 The stable zone (triangle) of the linear second-order filter section.

in the parameter plane. In zero-input conditions, for all initial states, the filter whose coefficients are within the triangle in the parameter plane will tend to the zero state. For example, suppose that two second-order basic sections whose coefficients are $A_1 = -1.25$, $B_1 = 0.625$, and $A_2 = -1.74$, $B_2 = 0.95833$ respectively have initial states (11,11) then their response are as follows:

$$A_1 = -1.25, \quad B_1 = 0.625.$$

n	1	2	3	4	5	6	7	8
Y(n)	6.875	4.297	2.868	1.678	1.049	0.656	0.410	0.256
n	9	10	11	12	13	14	15	...
Y(n)	0.160	0.100	0.063	0.039	0.024	0.015	0.010	...

$$A_2 = -1.74, \quad B_2 = 0.95833$$

n	1	2	3	4	5	6
Y(n)	8.60	4.42	-0.55	-5.19	-8.51	-9.83
n	7	8	9	10	...	100
Y(n)	-8.95	-6.15	-2.13	2.19	...	-1.02
n	...	145	146	...	200	...
Y(n)	...	0.51	0.41	...	0.06	...

As can be seen from above tables, only if the coefficients of a second-order filter are within the triangle in the parameter plane whatever the Q-value of the filter is the filter must tend to the zero state from any initial states. The higher the Q-value, the longer the time needed to tend to the zero state. The time corresponding to the zero-input response of the filter from the initial state to the state after which the *interest* absolute values of response are less than 0.5 is interested. Later on this time will be used as a reference time when we compare with the time needed for suppressing limit cycles. For above two examples the reference times are respectively equal to $7T_s$ and $146T_s$ where T_s is the sampling period in the digital filters.

2.2 BASIC FILTER SECTIONS WITH QUANTIZATION

In software as well as hardware digital filter implementations numbers are ultimately stored in finite-length registers. Consequently, coefficients and signal values must be quantized before they can be stored.

The effects of quantization after arithmetic operations can be demonstrated with the example of a first-order digital filter described by the following difference equation:

$$\hat{Y}(n) = \hat{X}(n) - A\hat{Y}(n-1) \quad (44)$$

where $\hat{X}(n)$ is the input sequence.

Throughout this thesis, the circumflex is used to designate the results of finite precision arithmetic, i.e., quantized numbers. Suppose that all numbers $\hat{Y}(n)$, A , $\hat{X}(n)$ are expressed initially with k significant digits and that fixed-point arithmetic is employed for the implementation of the difference equation. Calculation of the filter response shows that after n iterations $\hat{Y}(n)$ is expressed by numbers with $(n+1)k$ significant digits.

This example indicates that the number of significant digits needed to compute the filter response precisely, increase linearly with each iteration. As long as the number of operations performed on a signal remains finite, for example, in a nonrecursive digital filter, the increasing wordlength is also finite. But as can be seen from Fig. 1 and Fig. 5, the basic filter sections are recursive filters. In a recursive filter, a wordlength reduction is necessary to prevent the signals from acquiring an ever-increasing wordlength. Any practical filter, realised with k significant digits, has to include quantization after each arithmetic operation so as to keep the results at a specified finite precision. Quantization introduces inherent nonlinearity which tend to make the original linear stable system zero-input unstable⁽²⁷⁾.

If the input to a system is identically zero, then starting from some initial condition the signals in the

system will either grow beyond any bound or will converge to one of the so called equilibrium solutions. As far as digital filter is concerned, since every state $\hat{Y}(n)$ in the absence of an input has a unique successor, and since the nonlinear digital filter is a finite state system, (quantized and bounded in amplitude), there are only two possibilities for its autonomous behaviour. Either the zero state is reached after a finite-time, or a periodic oscillation will result, which is referred to as a limit cycle or zero-input limit cycle⁽²⁷⁾.

To be able to analyse the nonlinear effects on the response of digital filters, it is necessary to consider the type of arithmetic used, and the type of nonlinearity introduced into the digital filter through finite precision arithmetic.

There are a variety of types of arithmetic that are used in the implementation of digital systems. Among the most common are fixed-point and floating-point. A hybrid between these arithmetic types was introduced called block floating point arithmetic.

Kaneko⁽²⁸⁾, while excluding the possibilities of overflow and underflow, proved that limit cycles of considerable amplitudes can be found with floating-point arithmetic. Lacroix⁽²⁹⁾ has studied the limit cycles that may result from underflow and he found regions for

the coefficients of a second-order digital filter for which such limit cycles can be found. Sandberg⁽³⁰⁾ has derived the asymptotically stable condition for floating-point arithmetic in the presence of roundoff and shown that this condition will be satisfied if the damping of the infinite precision counterpart of the digital filter is sufficiently "large" relative to the number of bits allotted to the mantissa of the data. Under these conditions limit cycle response to a zero-input or to an input sequence that approaches zero is also ruled out. If in case of underflow the signal is made zero then the stability region, which can be derived, is always approximately that of fixed-point with magnitude-truncation⁽¹⁰⁾. As will be seen later, this stability region of fixed-point with magnitude-truncation is very "large", in other words, the instability region is very small. Thus, generally speaking, limit cycle oscillations are not a problem when floating-point is used. In addition, most digital filters use fixed-point arithmetic because floating-point arithmetic involves more hardware. Therefore, in this thesis, only fixed-point arithmetic is considered.

In digital filters, two types of nonlinearities are connected with the adders and quantizers respectively.

If numbers are added whose sum exceeds the dynamic range of the adder "overflow" occurs. This "overflow" leads to a severe nonlinearity. Ebert et al.⁽³¹⁾ and

Jackson independently recognised the possibility of large amplitude limit cycles resulting from adder overflow with wrap-around arithmetic. Ebert et al. derived the conditions for the existence of overflow limit cycles and showed that by introducing saturation arithmetic the oscillations could be eliminated. In following chapters, it will be assumed that the adders in the digital filter are linear and overflow effects can be neglected.

The other type of nonlinearity is connected with the quantizers in the digital filters.

1. Quantization

It has been mentioned at the beginning of this section that wordlength reduction must be applied in the closed loop in digital filters. This can be done by affecting the least significant bits only, i.e., quantization. Quantization can be performed by substituting the nearest possible word that can be represented by the limited number of bits. There are two standard methods for eliminating the low-order bits; rounding and truncation.

Suppose that any number x with $E_{\min} \leq x \leq E_{\max}$ is represented by a fixed-point format and the quantization step size is q . In the following chapters we assume q equal to one.

(A) Rounding

Rounding x to the nearest integral multiple of q is a familiar method. The rounded number is designated by $[x]_R$. The relationship between $[x]_R$ and x is shown in Fig. 9. The difference of the signals, $\delta(n) = x - [x]_R$ is called quantization error or quantization noise. It is clear from Fig. 9, that the error signal satisfies the relation

$$-\frac{q}{2} \leq \delta \leq \frac{q}{2} \quad (45)$$

Under certain, not overly restrictive, assumptions it can be shown that if the number x can be treated as a random sequence then the quantization error is uniformly distributed in the closed range $[-\frac{q}{2}, \frac{q}{2}]$. The probability distribution of the quantization error for rounding is shown in Fig. 12(a).

(B) Truncation

Depending on the negative number representation used, there exist two types of truncation; magnitude truncation and value truncation.

(a) Magnitude Truncation

In a representation of the signals by sign and magnitude this leads to magnitude truncation quantization with a characteristic as shown in Fig. 10. It is clear

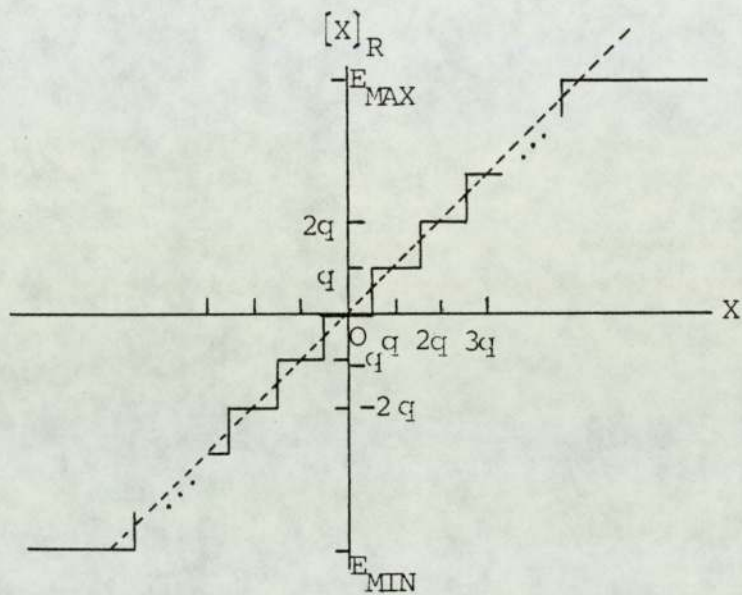


Fig. 9 Quantizer characteristic with rounding.

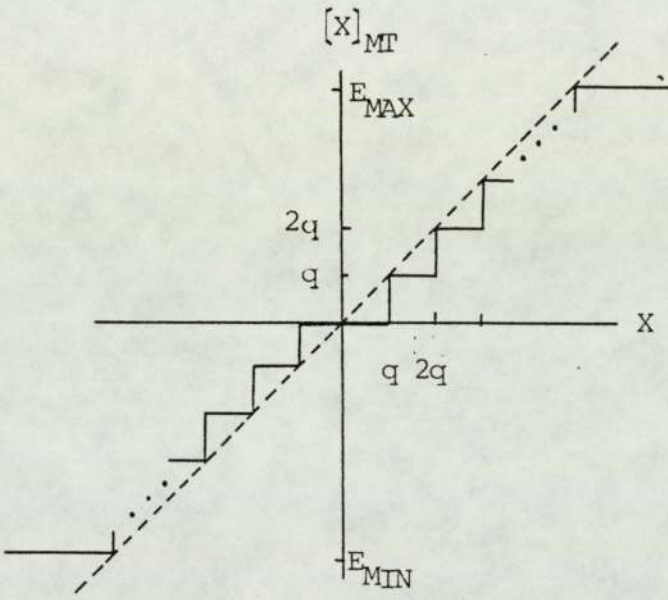


Fig. 10 Quantizer characteristic with magnitude-truncation.

from Fig. 10 that the quantization error satisfies the relation

$$|\delta| < q \quad (46)$$

Similarly, if the number x can be treated as a random sequence the probability distribution of the quantization noise can be shown as in Fig. 12(b).

(b) Value Truncation

Value truncation results when a two's complement number representation is used. Fig. 11 and Fig. 12(c) show its characteristic and probability distribution of the quantization error respectively.

The value truncation is not considered in detail because the results are similar to the ones for rounding with a constant input added as value truncation introduces only a bias of $\frac{1}{2}q$ for every quantizer.

As can be seen from Fig. 12, the variance of the quantization error for magnitude-truncation is four times bigger than that for roundoff. Claasen et al.⁽⁶⁾ have shown that under nonzero input condition, the digital filters with magnitude-truncation quantizer have quantization noise power (5-10) times bigger than that when roundoff is used. Therefore, in practice a roundoff quantizer is preferable to a truncation quantizer. But as will be

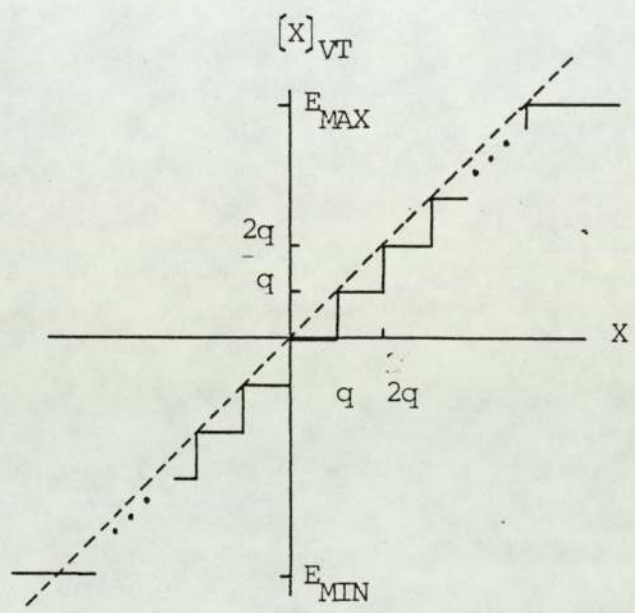


Fig. 11 Quantizer characteristic with value-truncation.

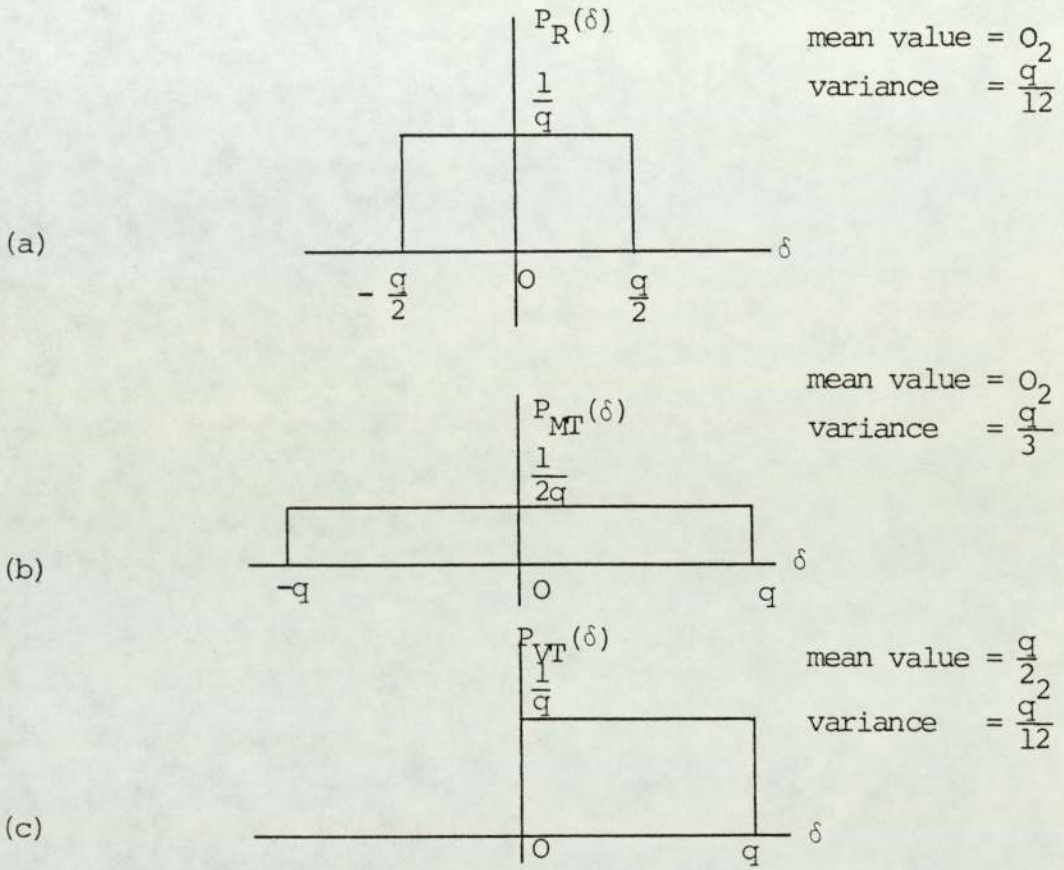


Fig. 12 Probability density functions: (a) for roundoff error
(b) for magnitude-truncation error and (c) for value-truncation error.

pointed out later, from the view point of limit cycle appearance the truncation quantization has its own advantage.

2. Limit Cycles in the First-Order Filter Section

(A) The First-Order Filter With Magnitude Truncation Quantizer

The conclusion is that in zero-input conditions no nonzero limit cycle can be sustained with this kind of system. Fig. 13 shows the block diagram of the first-order filter with magnitude truncation quantizer. Its difference equation with zero-input can be written as

$$\hat{Y}(n) = [-A \hat{Y}(n-1)]_{MT} \pm \delta_T(n) \quad (47)$$

where from the magnitude quantization characteristic one knows that $\delta_T(n) < 1$ and $|A| < 1$ from the linear stable condition Eqn. (8).

In the first-order system, the limit cycles occurring can be of only two forms: constant magnitude and sign for A negative, or constant magnitude with alternating signs for A positive.

The conclusion about no limit cycle existing in the first-order filter with magnitude-truncation quantizer can be proved by the contrary method.

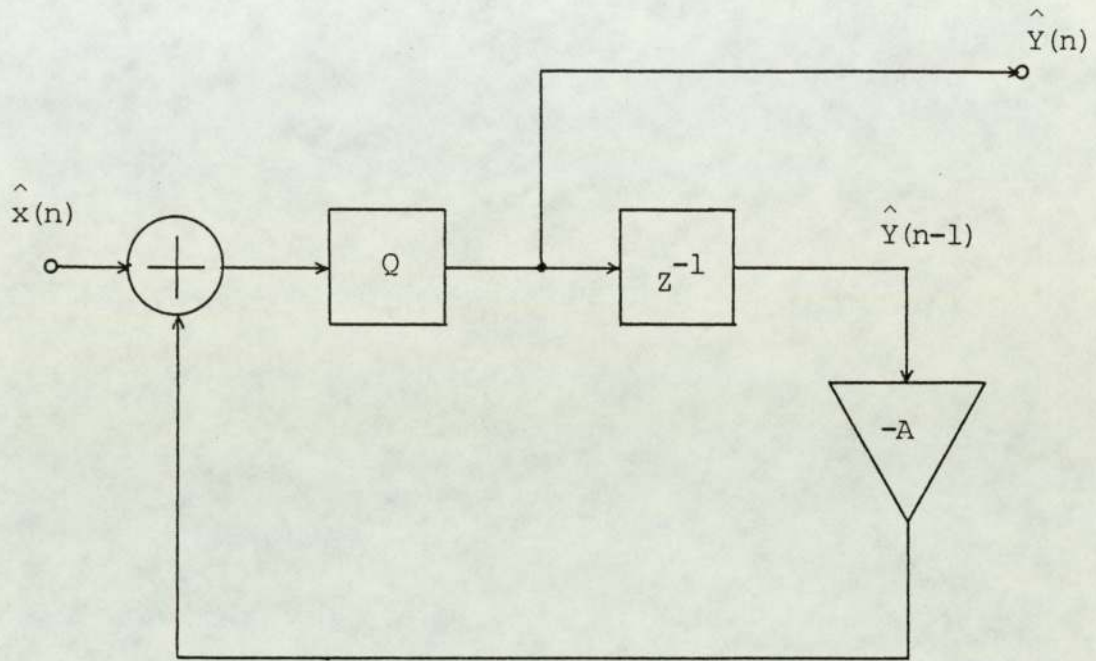


Fig.13 Block diagram of a first-order digital filter with quantizer.

In the first case, assume that $-1 < A < 0$ suppose a positive steady-state limit cycle exists, then from the exiting condition

$$\hat{Y}(n) = \hat{Y}(n-1) \quad (48)$$

Because $A < 0$ and $\hat{Y}(n) > 0$

$$-A\hat{Y}(n-1) > 0 \quad (49)$$

In this case, from the characteristic of the magnitude-truncation one knows that

$$\hat{Y}(n) = -A \hat{Y}(n-1) - \delta_T(n) \quad (50)$$

where

$$0 \leq \delta_T(n) < 1$$

Because

$$\hat{Y}(n) = \hat{Y}(n-1)$$

One obtains

$$\hat{Y}(n) = \frac{-\delta_T(n)}{1+A} \quad (51)$$

As can be seen from this equation, when $\delta_T(n) = 0$, then $\hat{Y}(n) = 0$. That means that no limit cycle exists.

But for

$0 < \delta_T(n) < 1$, Eqn (51) shows that

$$\hat{Y}(n) < 0$$

This conclusion is contrary to the hypothesis made above. In other words, the limit cycle does not exist.

Therefore, for this case the only equilibrium is

$$\hat{Y}(n) = \hat{Y}(n-1) = 0$$

for $n > N_0$ where N_0 is a finite value.

For other cases, along the similar lines with above, the same conclusion can be obtained. The same example used in Section 2.1 ($A = -0.875$, $Y(0) = 8$) is chosen but here, a magnitude-truncation quantizer is included. It is easy to verify that the zero-input response is as follows:-

7,6,5,4,3,2,1,0,0,0,...

In summary, in the first-order filter with magnitude truncation quantizer under zero-input conditions the only equilibrium is zero state, i.e., no zero-input limit cycle exists.

(B) The First-Order Filter Section With Rounding Quantizer

The conclusion is that in the first-order filter only constant magnitude (with constant signs or alternating signs) limit cycles exist.

Refer to Fig. 13, the difference equation with rounding quantizer under zero-input condition can be written as

$$\begin{aligned}\hat{Y}(n) &= [-A\hat{Y}(n-1)]_R \\ &= -A\hat{Y}(n-1) - \delta(n)\end{aligned}\tag{52}$$

where from the roundoff quantization characteristic one knows that

$$-0.5 \leq \delta(n) \leq 0.5$$

and from the stable condition Eq. (8)

$$|A| < 1$$

here $[\cdot]_R$ represents the roundoff quantization.

As mentioned above for the first-order system the limit cycles occurring can be of only two forms; constant magnitude and sign for A negative, or constant magnitude with alternating signs for A positive.

First, suppose A is negative, then a steady-state limit cycle exists.

$$\hat{Y}(n) = \hat{Y}(n-1)$$

Substitute above equation into Eqn (52), we obtain

$$\hat{Y}(n) = \text{INT}\left(\frac{-\delta(n)}{1+A}\right)$$

where $\text{INT}(X)$ represents the integer part of X.

Because $\delta(n)$ satisfies the inequality

$$-0.5 \leq \delta(n) \leq 0.5$$

therefore the limit cycle $\hat{Y}(n)$ satisfies

$$\text{INT}\left(\frac{-0.5}{1+A}\right) \leq Y(n) \leq \text{INT}\left(\frac{0.5}{1+A}\right)$$

or the amplitude of $\hat{Y}(n)$, K , satisfies the inequality

$$K \leq \text{INT}\left(\frac{0.5}{1+A}\right) \quad (53)$$

Similarly, suppose A is positive then constant magnitude with alternating signs limit cycles exist. The amplitude, K , satisfies the inequality

$$K \leq \text{INT}\left(\frac{0.5}{1-A}\right) \quad (54)$$

Combining the above two cases, the amplitude of limit cycles in the first-order filters with roundoff quantization, K , satisfies

$$K \leq \text{INT}\left(\frac{0.5}{1-|A|}\right) \quad (55)$$

As an example, we chose the example used in Section 2.1, i.e., $A=-0.875$ $Y(0) = 8$ but this time the roundoff quantizer is included. It is readily verified that the zero-input response is as follows:

n	1	2	3	4	5	6	...
Y(n)	7	6	5	4	4	4	...

The amplitude of the constant limit cycle is 4.

According to the Eqn. (55) the bound

$$\begin{aligned} K &= \text{INT} \left(\frac{0.5}{1-|A|} \right) \\ &= \text{INT} \left(\frac{0.5}{1-0.875} \right) \\ &= 4 \end{aligned}$$

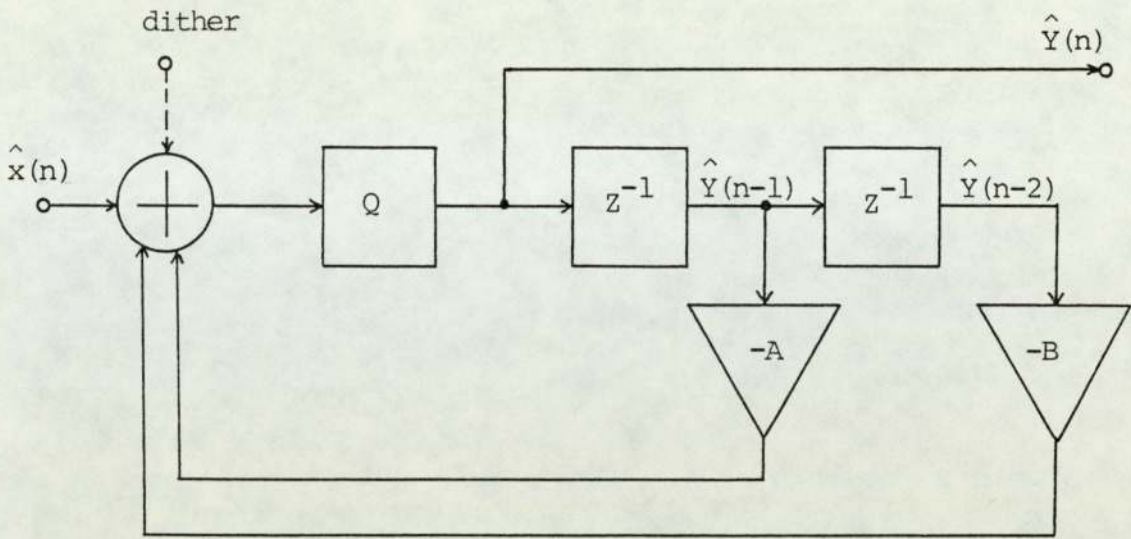
The experimental data is exactly consistent with the theoretical formula.

3. Limit Cycles in the Second-Order Filter Sections

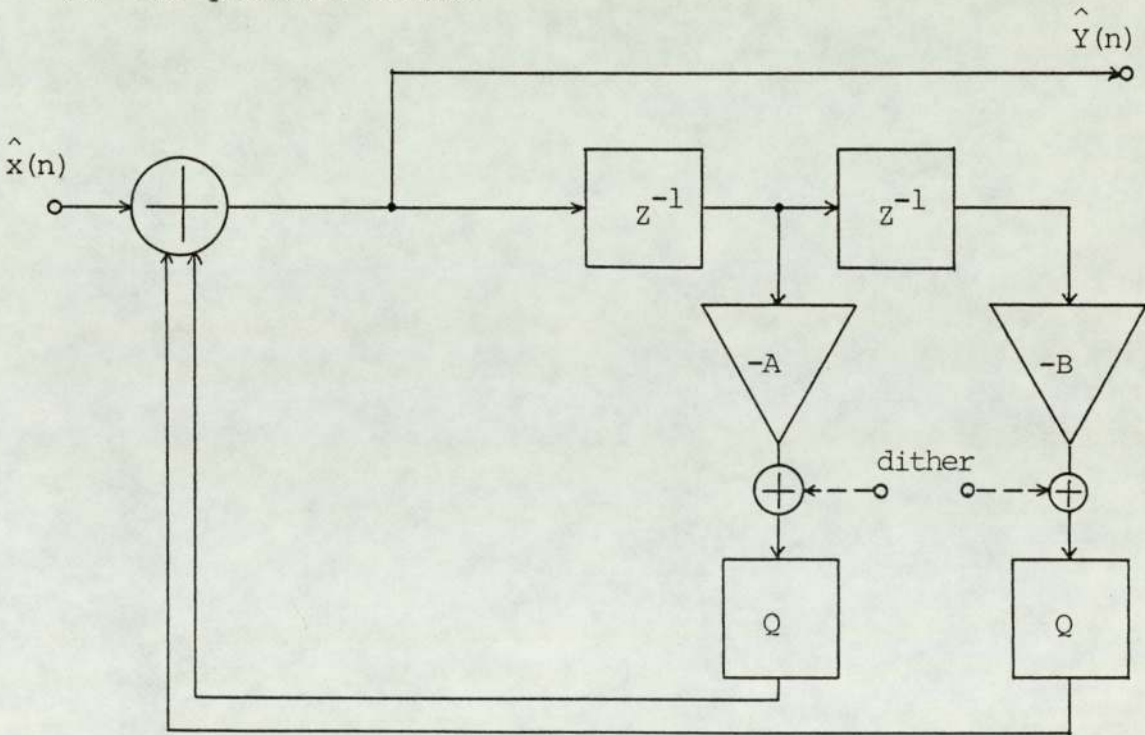
There are two different ways of implementing the quantizations in the second-order sections as shown in Fig. 14(a) and 14(b). Between these two ways there is the following difference: In Fig. 14(a), full precision is maintained in the feedback loop as long as possible and the necessary limitation of the wordlength occurs only once, whereas, in Fig. 14(b), the wordlength is limited immediately after the multipliers. The former structure is called one quantizer version and the later two quantizer version.

In the condition of zero-input, $\hat{x}(n)=0$, the sections given in Fig. 14(a) and (b) are described respectively by the difference equations

$$\hat{Y}(n) = [-A\hat{Y}(n-1) - B\hat{Y}(n-2)]_Q \quad (56)$$



(a) One quantizer version



(b) Two quantizer version

Fig. 14 Two ways of implementing the quantizations in the second-order filter sections (a) One quantizer version, (b) Two quantizer version.

and

$$\hat{Y}(n) = [-A\hat{Y}(n-1)]_Q + [-B\hat{Y}(n-2)]_Q \quad (57)$$

where $[\cdot]_Q$ represents the operation of quantization and $\hat{Y}(n)$ denotes the quantized output signal of the filter after n sampling intervals each with a duration of T_s . A and B are the coefficient values of the filter.

In this thesis, we have mainly concentrated our attention on the one quantizer version. But the principle of limit cycle suppression is also suitable to the two quantizer version. In the following sections we will discuss the limit cycles in the second-order filter with rounding- and truncation-quantizer, respectively.

(A) Limit Cycles in the Second-Order Filter Section with Rounding Quantizer

As mentioned before, quantization introduces inherent nonlinearity. Because of this nonlinearity limit cycles may appear in the digital filter.

For example, in one rounding quantizer version, suppose the coefficients, $A=-1.74$, $B=0.95833$. It can readily be verified by setting the filter to different initial states that the limit cycles which can appear in the filter are as follows, where one complete period of each limit cycle is shown.

Limit Cycle 1:

... 0, 5, 9, 11, 11, 9, 5, 0, -5, -9, -11, -11, -9, -5, ...

Limit Cycle 2:

... 1, 5, 8, 9, 8, 5, 1, -3, -6, -8, -8, -6, -3, ...

Limit Cycle 3:

... -1, -5, -8, -9, -8, -5, -1, 3, 6, 8, 8, 6, 3, ...

Limit Cycle 4:

... 0, 3, 5, 6, 6, 5, 3, 0, -3, -5, -6, -6, -5, -3, ...

Limit Cycle 5:

... 1, 3, 4, 4, 3, 1, -1, -3, -4, -4, -3, -1, ...

Limit Cycle 6:

... 0, 1, 2, 3, 3, 2, 1, 0, -1, -2, -3, -3, -2, -1, ...

Limit Cycle 7:

... 2, 2, 2, ...

Limit Cycle 8:

... 1, 1, 1, ...

Limit Cycle 9:

... -1, -1, -1, ...

Limit Cycle 10:

... -2, -2, -2, ...

This example will be frequently used in this thesis.

It is of interest to compare the limit cycles in the

two quantizer version with the same coefficients. The corresponding limit cycles are as follows:

Limit Cycle 1:

... 5, 12, 16, 17, 15, 10, 3, -5, -12, -16, -17, -15, -10, -3, ...

Limit Cycle 2:

... 3, 8, 11, 11, 8, 3, -3, -8, -11, -11, -8, -3, ...

Limit Cycle 3:

... 0, 6, 10, 11, 9, 5, 0, -5, -9, -11, -10, -6, ...

Limit Cycle 4:

... 0, -6, -10, -11, -9, -5, 0, 5, 9, 11, 10, 6, ...

Limit Cycle 5:

... 4, 8, 10, 9, 6, 1, -4, -8, -10, -9, -6, -1, ...

Limit Cycle 6:

... 1, 6, 9, 10, 8, 4, -1, -6, -9, -10, -8, -4, ...

Limit Cycle 7:

... 3, 7, 9, 9, 7, 3, -2, -6, -8, -8, -6, -2, ...

Limit Cycle 8:

... -3, -7, -9, -9, -7, -3, 2, 6, 8, 8, 6, 2, ...

Limit Cycle 9:

... 1, 5, 8, 9, 8, 5, 1, -3, -6, -7, -6, -3, ...

Limit Cycle 10:

..., -1, -5, -8, -9, -8, -5, -1, 3, 6, 7, 6, 3, ...

Limit Cycle 11:

... 0, 4, 7, 8, 7, 4, 0, -4, -7, -8, -7, -4, ...

Limit Cycle 12:

... 2, 5, 7, 7, 5, 2, -2, -5, -7, -7, -5, -2, ...

Limit Cycle 13:

... 1, 4, 6, 6, 4, 1, -2, -4, -5, -5, -4, -2, ...

Limit Cycle 14:

... -1, -4, -6, -6, -4, -1, 2, 4, 5, 5, 4, 2, ...

Limit Cycle 15:

... 0, 3, 5, 6, 5, 3, 0, -3, -5, -6, -5, -3, ...

Limit Cycle 16:

... 1, 3, 4, 4, 3, 1, -1, -3, -4, -4, -3, -1, ...

Limit Cycle 17:

... 0, 2, 3, 3, 2, 0, -2, -3, -3, -2, ...

Limit Cycle 18:

... 0, 1, 2, 2, 1, 0, -1, -2, -2, -1, ...

Limit Cycle 19:

... 1, 1, 1, ...

Limit Cycle 20:

... -1, -1, -1, ...

As can be seen from the above two examples, although with the same coefficients the limit cycles in two quantizer

version is quite different with that in one quantizer version. The properties of limit cycles in the second-order section with rounding quantizer will be discussed in detail later.

(B) Limit Cycles in the Second-Order Filter Section with Magnitude Truncation Quantizer

(a) One Quantizer Version

Claasen et al.^(32,33,34) have investigated the zero-input behaviour of second-order digital filters with one magnitude truncation quantizer. They proved that the area of absolute stability of the nonlinear filter is the shaded area in Fig. 15. This area, where no limit cycles can occur, is bounded by the left- and right-hand sides of the linear stable triangle, by a part of the ellipse $A^2 + 8B(B-1) = 0$ and by the two straight lines $|A| = 2 - B$. In the area remaining only in the small trapezoid area defined by the intervals $1 > B \geq 0.94$ and $1.42 \leq |A| < 2$ limit cycles have been found with simulations using fixed-point arithmetic. It is worth pointing out that in the above trapezoid area only small number of A, B coefficient combinations can lead to limit cycles. What is more, even though the limit cycles exist, there are a number of initial conditions from which the steady-state zero-input responses of the filter are zero. Only about (25-40)% of these limit cycles are

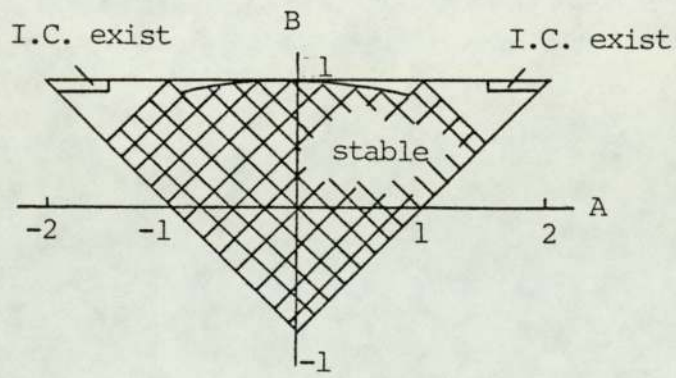


Fig. 15 Stability diagram for the second-order digital filter with one magnitude-truncation quantizer.

accessible. An accessible limit cycle means that it can be reached from initial conditions that do not pertain to that limit cycle. On the other hand, an inaccessible limit cycle only appears if the filter is started with initial conditions pertaining to that limit cycle. Apparently, in practice, only accessible limit cycles are troublesome, because the probability of occurrence of inaccessible limit cycles is very small. For the filter which has the coefficients $A=-1.74$, $B=0.95833$ but with one magnitude-truncation quantizer, the simulation showed that no limit cycles exist.

In the second-order section with the coefficients $A=-1.640625$, $B=0.953125$ and with one magnitude-truncation quantizer, only one limit cycle exists. The limit cycle sequence is as follows:

... 5, 9, 10, 7, 1, -5, -9, -10, -7, -1, ...

Fig. 16 shows the initial states in the area bounded by the amplitude bound of the limit cycle in the state plane from which the limit cycle can be reached. As can be seen, among the 442 states in the area there are only 76 states from which the limit cycle can be obtained. In other words, there are 82.8 percent of states from which the origin state can be reached.

Claasen et al. ⁽³⁴⁾ have shown the limit cycles for a filter with coefficients $A=-\left(\frac{214}{128}\right)$, $B = \left(\frac{126}{128}\right)$ and with

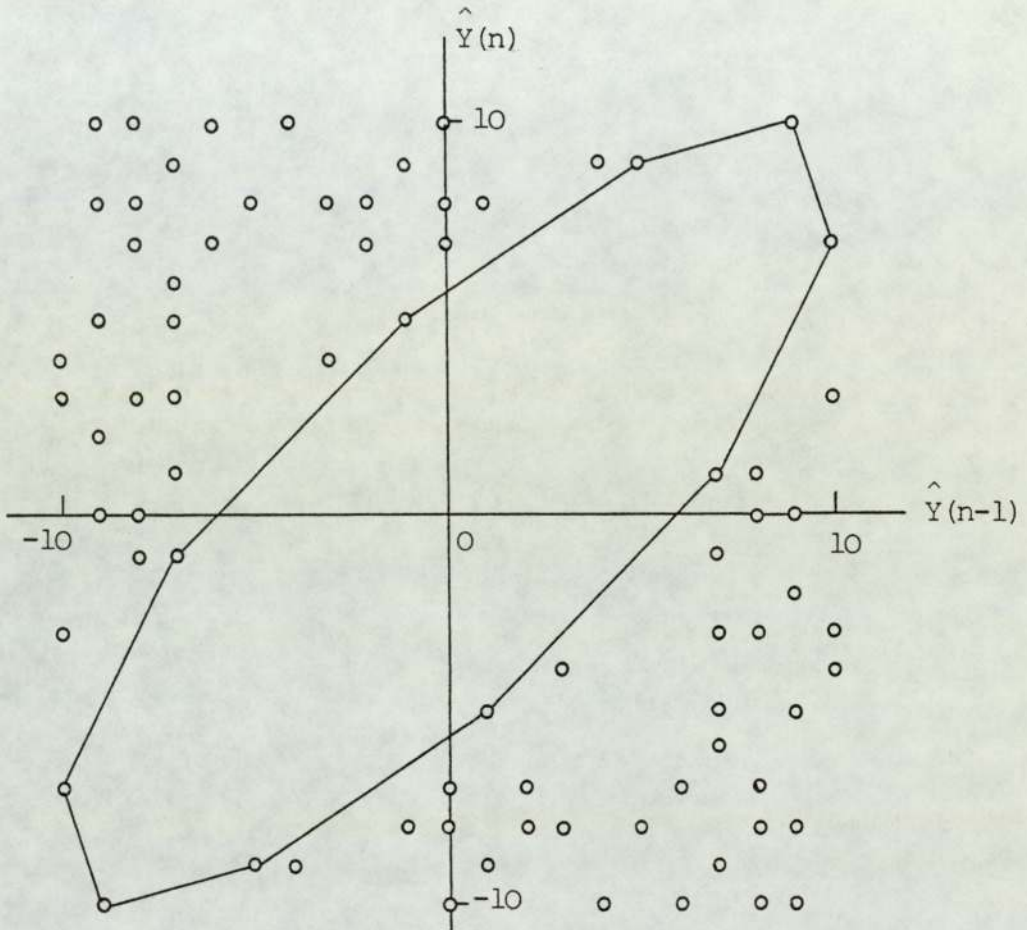


Fig. 16 Part of the state plane of the second-order filter ($A=-1.640625$, $B=0.953125$) with one magnitude-truncation, with initial conditions indicated from which a limit cycle will result.

one magnitude-truncation quantizer. In this filter, two limit cycles exist. One is accessible whereas another is inaccessible. In the area in the state plane which is bounded by ± 20 , there are 1681 states. But only 142 states exist from which the limit cycles will be obtained. There are more than 90 percent of states from which the origin state can be reached.

The conclusion is that in a filter with one magnitude-truncation quantizer limit cycles will be possible only for very few values of coefficients A and B. Moreover, for those values of A and B for which limit cycles are possible the probability of the occurrence of a limit cycle is small. Therefore, if the initial state of the second-order section is chosen randomly, it is unlikely that the limit cycle will be obtained, i.e., the origin state will be reached with large probability. This fact is helpful for understanding the principle of limit cycle suppression by the injection of random dither.

(b) Two Quantizer Version

The second-order filter with two magnitude-truncation quantizers has been analysed by Kao⁽³⁵⁾, who derived regions where limit cycles of periods 1 and 2 occur. These regions are defined by the linear stable triangle and $|A| > 1$. Claasen et al.⁽³⁶⁾ have derived the stability region with the frequency domain criteria. The stability diagram is

shown in Fig. 17. Only limit cycles of length 1 and 2 have been observed by simulations and they are found in the triangle for value $|A| > 1$ ⁽¹⁰⁾. For high-Q poles ($B \leq 1$) only limit cycles with amplitudes equal to one quantization step are accessible⁽³⁷⁾.

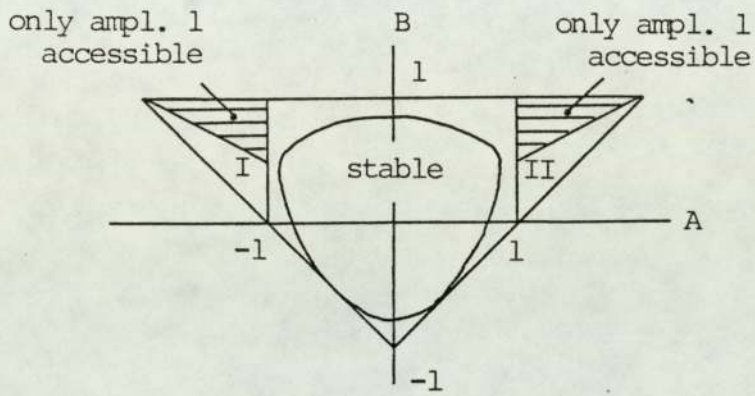
In some cases, limit cycles of periods 1 and 2 have no serious consequences in practical applications.

Comparing with its counterpart with rounding quantizers, with respect to the occurrence of limit cycles, the filter with truncation has its own advantage.

2.3 SUMMARY

The main properties of the basic sections have been given in this chapter.

The direct form is inferior to both the cascade and parallel form when the effect of coefficient quantization errors and roundoff noise after arithmetic operations are considered. The first- and second-order filters are basic building blocks from which all higher order systems can be synthesised. The zeros of the digital filters do not change the nature of the limit cycle but influence the magnitude of the limit cycle amplitude. Therefore the basic section which has two zeros at the origin on the Z-plane is used as a basic model in this research.



I : limit cycles of length 1 exist
 II: limit cycles of length 2 exist

Fig. 17 Stability diagram for the second-order digital filter with two magnitude-truncation quantizers.

The linear first-order filter has a lowpass characteristic. Its stability criterion is the absolute value of the filter coefficient less than unity.

The second-order filter represents a simple digital resonator. The frequency of the impulse response of the linear second-order section is shown in Eqn. (13). The stable region of the linear second-order filter section is bounded by a triangle in the parameter space.

In a recursive filter, quantization is necessary to prevent the signals from acquiring an ever-increasing wordlength. In this research, the fixed-point arithmetic has been used. Two types of quantization: magnitude-truncation and roundoff have been discussed.

In the first-order filter with magnitude-truncation quantizer, no limit cycle can be sustained. But with rounding quantizer, the constant amplitude limit cycles can exist.

There are two different ways of implementing the quantizations in the second-order sections. With rounding quantizer, variety limit cycle may exist in the second-order section. But in the two quantizer version with magnitude-truncation quantization only limit cycles of periods 1 and 2 can be sustained. In the one quantizer version with magnitude-truncation quantization limit cycles will be possible only for very few values of A and B in the



parameter space. Moreover, for those values of A and B for which limit cycles are possible the probability of the occurrence of a limit cycle is small.

Either the filters with one or two magnitude-truncation quantizers have certain advantages over roundoff with respect to the occurrence of limit cycles. But their quantization errors are bigger than that with roundoff quantization.

The properties of limit cycles in the second-order filter section with rounding quantizers will be discussed in more detail in the next chapter.

CHAPTER 3

ZERO-INPUT LIMIT CYCLES IN THE SECOND-ORDER

DIGITAL FILTERS

As mentioned in Chapter 2, although magnitude-truncation quantization has certain advantages over roundoff with respect to the occurrence of limit cycles, the digital filters with magnitude-truncation quantizations have much more quantization noise than that when roundoff is used. Therefore, in practice, a rounding quantizer is preferable to a truncation quantizer. This research only considers the filters with roundoff quantizers.

3.1 QUANTIZATION ERROR IN PRESENCE OF ROUND OFF

In the condition of zero input $\hat{X}(n)=0$, one quantizer and two quantizer version filters with roundoff (refer to Fig. 14(a) and 14(b)) can be described respectively by the difference equations

$$\hat{Y}(n) = [-A\hat{Y}(n-1) - B\hat{Y}(n-2)]_R \quad (58)$$

and

$$\hat{Y}(n) = [-A\hat{Y}(n-1)]_R + [-B\hat{Y}(n-2)]_R \quad (59)$$

where $[\cdot]_R$ represents the operation of roundoff quantization.

In this chapter, we first pay attention to the limit cycles in the one quantizer version, then describe the

features of limit cycles in the two quantizer version.

In the second-order basic filter section with one quantizer, the output of the rounding quantizer can be expressed as

$$[-A\hat{Y}(n-1) - B\hat{Y}(n-2)]_R = -A\hat{Y}(n-1) - B\hat{Y}(n-2) - \delta(n) \quad (60)$$

where

$-A\hat{Y}(n-1) - B\hat{Y}(n-2)$ and $\delta(n)$ are the exact products and quantization error respectively.

From the characteristic of roundoff quantization shown in Fig. 9, one knows that

$$|\delta(n)| \leq 0.5 \quad (61)$$

where the quantization step, q , has been assumed equal to unity.

If the input signal changes in its dynamic range and the signal levels throughout the filter are much larger than the quantization step q , the following reasonable assumptions can be made;

(1) $\delta_i(n)$ and $\delta_i(n+k)$ are statistically independent for any value of n ($k \neq 0$), and

(2) $\delta_i(n)$ and $\delta_j(n+k)$ are statistically independent for any value of n or k ($i \neq j$), here the subscripts i, j stand for different quantizers.

Once the above assumptions are valid, as far as the error signal $\delta(n)$ is concerned, the filter can be treated as a linear system. This results in a stochastic approach. The quantization error $\delta(n)$ is described by a uniform probability density function. The assumption leads to acceptably accurate results for most applications with high signal level and sufficient spectral content. As will be seen later, when we consider the quantization noise power with high level input signal, the stochastic approach will be used.

However, in the zero-input limit cycle study, above assumptions are not valid. Since the output of a quantizer is a single-valued function of the input, a given input yields a definite output and, consequently, a definite roundoff quantization error sequence $\delta(n)$. If a zero-input limit cycle exists, $\delta(n)$ can be a periodic or constant or alternating sign sequence depending on the limit cycle type.

The fact that there is correlation among $\delta(n)$ is a feature when limit cycles appear. For example, in Eqn (58) let $A=-1.74$, $B=0.95833$ and $Y(-1) = 5$, $Y(-2) = 9$. The signal values after rounding to the nearest integer and the roundoff quantization errors are shown in Table 1. As can be seen from the table, the filter with the specified initial state has a periodic limit cycle whose period is $14T_s$ and the roundoff quantization error is also a periodic sequence with


TABLE 1

The example shows that when a periodic limit cycle exists the quantization error is also a periodic sequence

$$\hat{Y}(n) = [-A\hat{Y}(n-1) - B\hat{Y}(n-2)]_R$$

$$\delta(n) = -A\hat{Y}(n-1) - B\hat{Y}(n-2) - \hat{Y}(n)$$

where $A=-1.74$, $B=0.95833$, $Y(0)=5$, $Y(-1)=9$

n	$\hat{Y}(n)$	$\delta(n)$	
1	0	-0.0750299999	 one period
2	-5	-0.20835	
3	-9	-0.3	
4	-11	-0.1316500001	
5	-11	-0.4849700001	
6	-9	-0.4016300001	
7	-5	0.11837	
8	0	0.0750299999	
9	5	0.20835	
10	9	0.3	
11	11	0.13165	
12	11	0.4849700001	
13	9	0.4016300001	
14	5	-0.11837	
15	0	-0.0750299999	
16	-5	-0.20835	
.	.	.	
.	.	.	
.	.	.	

the same period.

Therefore, in the zero-input limit cycle study, especially in the theory of limit cycle generation the stochastic approach cannot be applied. Parker and Hess⁽²⁾ by a deterministic approach have analyzed the limit cycle oscillations in fixed-point implementations of recursive digital filters due to roundoff and truncation quantization after multiplication.

3.2 CLASSIFICATION OF LIMIT CYCLES

There are several different ways to classify the limit cycles in the digital filters. Each way has its own feature and from it some important properties about limit cycles can be obtained.

1. Classification of Limit Cycles Based on the Period

Three different types of limit cycle may be distinguished; constant, alternating and periodic. In a constant limit cycle, the output is the same at each sampling instant e.g. (... ,2,2,2,...). In an alternating limit cycle, the output alternates between values of opposite polarity, e.g. (... ,4,-4,4,-4,...). Although, strictly speaking, constant and alternating limit cycles are also periodic, the term periodic limit cycle is reserved here for limit cycles whose period is greater than two clock instant, e.g.

(... 1, 3, 4, 4, 3, 1, -1, -3, -4, -4, -3, -1, ...). The example $A=-1.74$, $B=0.95833$ shown in Chapter 2 has six periodic, four constant limit cycles and no alternating limit cycles. The constant and alternating limit cycles have zero- and Nyquist-frequency, respectively. Periodic limit cycle is a sort of thing that we should pay more attention to.

2. Classification of Limit Cycles Based on Accessibility Considerations (37)

The following two types of limit cycle can be distinguished.

(A) Inaccessible Limit Cycle

They only appear if the filter is started with initial condition pertaining to that limit cycle. Hence, if the filter is started with randomly chosen initial conditions it is unlikely that these will correspond to a point on a limit cycle, i.e., the probability of occurrence of inaccessible limit cycles is very small.

In our typical example (second-order section with one rounding quantizer, $A=-1.74$, $B=0.95833$), two periodic limit cycles (limit cycle 4 and 5) and all the constant limit cycles are inaccessible.

(B) Accessible Limit Cycle

They can be reached from initial conditions that do not pertain to that limit cycle. In this case, there has to be at least one state $\hat{Y}(n)$ of the filter corresponding to a point of the limit cycle, which state can be reached from at least two different states $\hat{Y}(n-1)$ and $\hat{Y}'(n-1)$, the predecessors of $\hat{Y}(n)$. Thus both states have, as their successor, the state $\hat{Y}(n)$. The point corresponding to such a state is called a branch point. Apparently, an accessible limit cycle has at least one branch point. But inaccessible limit cycles have no branch points. Accessible limit cycles are observed more frequently in the digital filter. In the example just mentioned above, the other four periodic limit cycles (limit cycles 1, 2, 3 and 6) are accessible. In the region bounded with $\hat{Y}(n-1) = \pm 11$ and $\hat{Y}(n) = \pm 11$ in the state plane which is defined by $\hat{Y}(n)$ and $\hat{Y}(n-1)$, there are 46 initial states except the states which pertain to that limit cycle from there the period limit cycle 1 will be reached eventually. For periodic limit cycles 2, 3 and 6 the corresponding numbers of the initial states are 48, 48 and 302 respectively.

As mentioned before, since every state $\hat{Y}(n)$ in the absence of an input has a unique successor, and since the nonlinear digital filter is a finite state system, there are only two possibilities for its autonomous behaviour. Either the zero-state is reached after a finite time, or

a limit cycle will result. In other words, if the zero state (0,0) is not a branch point it cannot be reached from other states and limit cycles must necessarily exist. Claasen et al⁽³⁷⁾ have proved that in case one or two rounding operations are used, the origin state (0,0) is a branch point only if $|B| < 0.5$. Therefore, such second-order filters with multiplication coefficient B for which $|B| \geq 0.5$ will always exhibit limit cycles. By using the same idea it is also shown that where one or two magnitude-truncation operations are used, (0,0) is always a branch point for $|B| < 1$, and limit cycles do not necessarily exist.

3. Classification of Limit Cycles Based on Symmetry Considerations⁽³⁷⁾

As regards symmetry, two types of limit cycle can be, in general, distinguished.

TYPE A Symmetric Limit Cycle

The length N of the limit cycle is even and the limit cycle has half-wave symmetry

$$\hat{Y}(n + \frac{N}{2}) = -\hat{Y}(n) \quad \text{for all } n \quad (62)$$

In our typical example, periodic limit cycle 1, 4, 5 and 6 pertain this type.

A further distinction is Type A1: $\frac{N}{2}$ is odd, and Type A2: $\frac{N}{2}$ is even.

As can be seen, in the example above, the limit cycle 1, 4 and 6 are Type A1 and limit cycle 5 is Type A2.

TYPE B Asymmetric Limit Cycle

All limit cycles which are not of Type A are called asymmetric limit cycles. These can be subdivided into Type B1: N is odd, and Type B2: N is even.

In the example above, both limit cycle 2 and 3 pertain Type B1.

With this classification method a relation has been given in Table 2⁽³⁷⁾ between the limit cycles of a second-order digital filter with the coefficients $(-A, B)$ and a second-order digital filter that has the same structure but coefficients (A, B) . It can be seen from this table that a limit cycle of Type A1 in the filter with $-A$ and B transforms into two limit cycles of Type B1 in the filter with A and B which have the same amplitudes as that of Type A1 but with lengths that have been halved. In addition, two constant limit cycles in the filter with $(-A, B)$ transform into an alternating limit cycle in the filter with (A, B) .

From these relations at least two useful conclusions can be made:

First, once one has found the limit cycles in the

TABLE 2

The relation between the limit cycles of a second-order filter with the coefficients $(-A, B)$ and a second-order filter that has the same structure but coefficients (A, B) ⁽³⁷⁾

Filter Coefficients	Type of Limit Cycle	Length of Limit Cycle	Number of Limit Cycle	Type of Limit Cycle	Length of Limit Cycle	Number of Limit Cycle	Type of Limit Cycle	Length of Limit Cycle	Number of Limit Cycle	Type of Limit Cycle	Length of Limit Cycle	Number of Limit Cycle
$-A, B$	A1	N	1	A2	N	1	B1	N	2	B2	N	2
A, B	B1	N/2	2	A2	N	1	A1	2N	1	B2	N	2

second-order filter with coefficients $(-A,B)$ one also knows the structure of the limit cycles in the filter with coefficients (A,B) . For example, by the relation just mentioned, we can derive the limit cycles in the second-order basic filter section with the coefficients $A=1.74$, $B=0.95833$ from the typical example used before ($A=-1.74$, $B=0.95833$).

According to Table 2, one Type A1 limit cycle will be transformed into two Type B1 limit cycles. Type A2 will be still transformed into Type A2. Two constant limit cycles will be transformed into one alternating limit cycle. In the procedure of transformation the feature is that the sign of the number is changed alternatively.

The transformed results are as follows, where the combination numbers which designate the limit cycle show the way of the transformation. For example, Limit Cycle

1A and 1B indicate that both limit cycles come from the original Limit Cycle 1, and Limit Cycle 23 indicates that this limit cycle comes from the original Limit Cycles 2 and 3.

Limit Cycle 1A:

... 0, -5, 9, -11, 11, -9, 5, ...

Limit Cycle 1B:

... 0, 5, -9, 11, -11, 9, -5, ...

Limit Cycle 23:

... -1, 5, -8, 9, -8, 5, -1, -3, 6, -8, 8, -6, 3, 1, -5, 8, -9, 8, -5,
1, 3, -6, 8, -8, 6, -3, ...

Limit Cycle 4A:

... 0, 3, -5, 6, -6, 5, -3, ...

Limit Cycle 4B:

... 0, -3, 5, -6, 6, -5, 3, ...

Limit Cycle 5:

... 1, 1, -3, 4, -4, 3, -1, -1, 3, -4, 4, -3, ...

Limit Cycle 6A:

... 0, -1, 2, -3, 3, -2, 1, ...

Limit Cycle 6B:

... 0, 1, -2, 3, -3, 2, -1, ...

Limit Cycle 89:

... 1, -1, 1, -1, ...

Limit Cycle 70:

... 2, -2, 2, -2, ...

Above results have been verified by the simulation.

The second conclusion is that any bounds on the magnitude of a limit cycle response evaluated with the assumption that $A < 0$, $B > 0$ are equally valid for $A > 0$ and $B > 0$.

3.3 SUCCESSIVE-VALUE PHASE-PLANE PLOT⁽²⁾

The successive-value phase-plane or sometimes called state plane for second-order digital filters is defined in a cartesian coordinate system with the Y axis representing $\hat{Y}(n)$ and the X axis representing $\hat{Y}(n-1)$. The successive-value phase-plane plot results of a second-order filter recorded on this state plane. For a given point or state in this plane, the successive state is uniquely determined for a digital filter with zero input. A limit cycle exists where a sequence of successive-value points in the phase-plane forms a closed curve when they are jointed by straight lines. Fig. 18 illustrates the successive-value phase-plane plot for the digital filter frequently used as an example. It can be seen that with a constant limit cycle, each successive state of the filter lies at the same point of the state plane. The plot corresponding to an alternating limit cycle includes only two state points, and the plot of a periodic limit cycle which has a period of NT_s includes N points.

Successive-value phase-plane plots provide a useful means for displaying the nature of the limit cycles of a digital filter. By the use of the program shown in Appendix 1, the successive-value phase-plane plot can be displayed on the screen of computer PET.

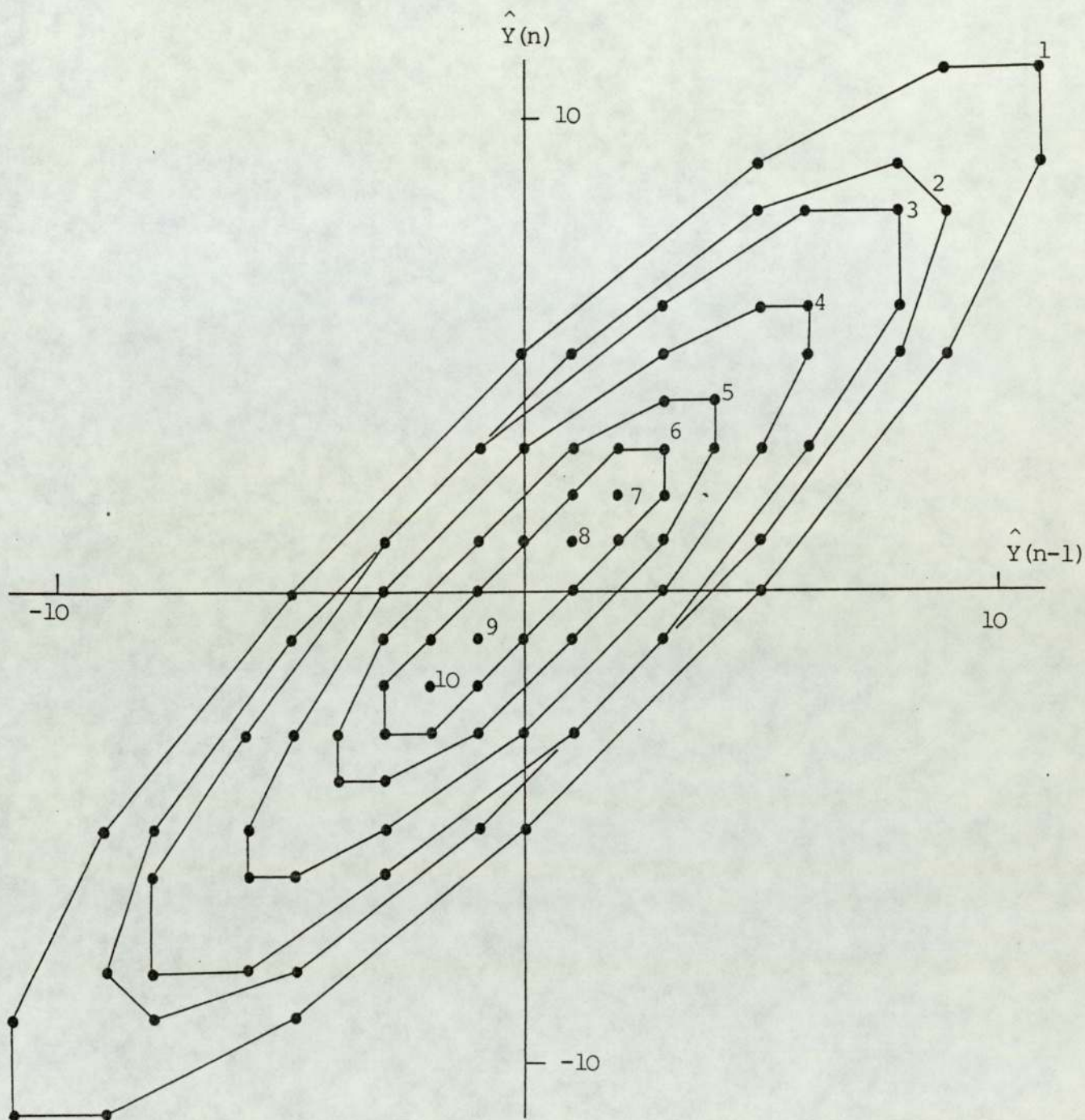
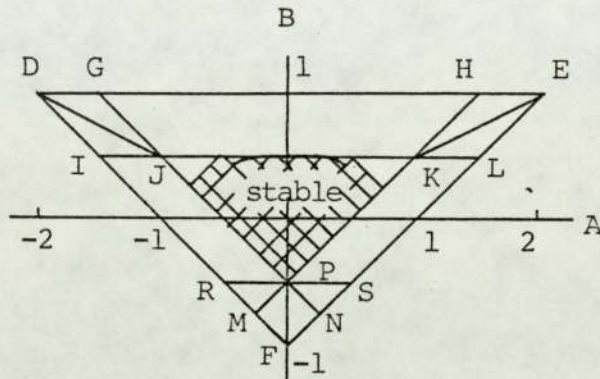


Fig. 18 The successive-value phase plane plot for the second-order section with coefficients $A=-1.74$, $B=0.95833$.

3.4 PARAMETER SPACE

The existing conditions of limit cycles in the second-order digital filter depend on the coefficient values A and B. Therefore, it is convenient to show these conditions on a coordinate system with the X axis representing coefficient A and the Y axis representing coefficient B. The coefficients A, B define a plane called parameter space.

The existing conditions of various limit cycles in the second-order filter with one rounding quantizer have been derived by a simple way in the Appendix 2. The regions in the parameter space where various limit cycles exist is shown in Fig. 19. The linear stable region bounded by a triangle mentioned earlier is also shown in the same figure. Values of A and B for which this filter is stable have been obtained by applying the frequency domain criterion⁽¹⁰⁾. This asymptotic stable region is also shown with shaded area in Fig. 19. Fig. 19 is the same with that shown in the reference (10), but here gives two extra bound lines which give more information. As can be seen from Fig. 19, the area which bounded by the triangle and $B > 0.5$ can be divided further into several subregions. In some subregions only constant or alternating limit cycles are possible. This information has not been found in the published literatures.



- In region GHKJG periodic limit cycles exist
- In region DGJD and RPMR both periodic and constant limit cycles exist
- In region EKHE and SPNS both periodic and alternating limit cycles exist
- In region DJPRD only constant limit cycles exist
- In region EKPSE only alternating limit cycles exist
- In region PMFNP three types of limit cycles exist

Fig. 19 Region of asymptotic stability for one rounding quantizer and regions where limit cycles can occur.

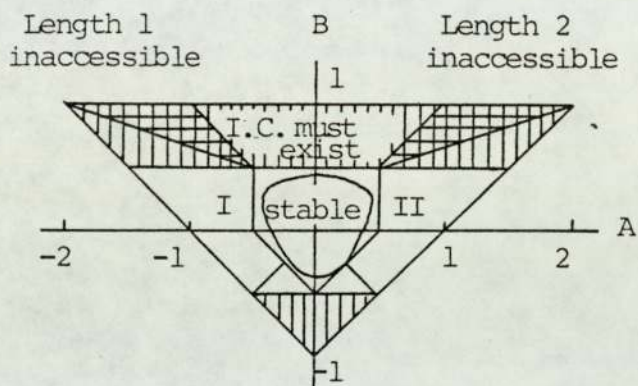
As references, Fig. 20, Fig. 15 and Fig. 17 show the parameter space plots with two rounding, one magnitude-truncation and two magnitude truncation quantizers respectively⁽¹⁰⁾.

3.5 DIFFERENT PROPERTIES OF THE LIMIT CYCLES IN SECOND-ORDER FILTER SECTIONS WITH ROUND OFF BETWEEN ONE- AND TWO-QUANTIZER VERSIONS

At least two differences between the limit cycles in the second-order filters with one- and two-rounding quantizers can be distinguished. These differences are as follows:

(1) As shown in Appendix 2, with one rounding quantizer all constant or alternating limit cycles are successive in unit of q in amplitude, i.e., they are $\pm 1, \pm 2, \pm 3, \dots, \pm C$ where C is the maximum amplitude of the limit cycles. It has been found by experiments that with one rounding quantizer all periodic limit cycle trajectories surround all constant or alternating limit cycles in the state plane. From this fact, the bound lines DJ, KE in Fig. 19 have been derived. These bounds were verified by simulations. No exceptions have been found. No constant or alternating limit cycles lie among the periodic limit cycles in the state plane.

As contrasted with one quantizer case, with two rounding quantizers constant or alternating limit cycles



I : limit cycles of length 1 exist
 II : limit cycles of length 2 exist

Fig. 20 Region of stability for two rounding quantizers and regions where limit cycles can occur.

can lie among the periodic limit cycles.

(2) With two rounding quantizers in some region of the parameter space there is a special kind of periodic limit cycle called pinwheel limit cycle⁽³⁾. Its trajectory surrounds the origin state (0,0) in the state plane several circles. But in contrast with two rounding quantizer version, with one quantizer we have not found any pinwheel limit cycles. As can be seen later, although there are those differences they will not influence the suppression of limit cycles by the injection of random dither.

3.6 AMPLITUDE BOUNDS OF LIMIT CYCLES

Three different types of amplitude bound for limit cycles in the second-order digital filters have been given in the literature.

1. Absolute Bounds

Several authors have derived bounds on the maximum value of the limit cycles for general types of digital filter^(2,7,38). Application of these bounds for determining the internal wordlength of the filter will guarantee the neglect of zero-input limit cycles in the output. The difficulty with the absolute bounds is that they apply to the situation where all errors add up in the worst possible way. Thus, the absolute bounds are in general overly pessimistic compared with the other bounds.

These bounds are seldom used in practice but they provide the biggest bounds which can never be exceeded. Among the absolute bounds, we only list out the bounds of Long and Trick⁽⁷⁾.

$$K \leq \left\{ \begin{array}{ll} \frac{\delta}{1-|A|+B} & \text{for } B \leq 0, \text{ or } B > 0 \text{ and } 2\sqrt{B} \leq |A| \\ \frac{\delta}{(1-\sqrt{B})^2} & \text{for } B > 0 \text{ and } 2B\sqrt{\frac{2}{B}} - 1 \leq |A| \leq 2\sqrt{B} \quad (63) \\ \frac{\delta(1+\sqrt{B})}{(1-B)\sqrt{1-\frac{A^2}{4B}}} & \text{for } B > 0 \text{ and } |A| \leq 2B\sqrt{\frac{2}{B}} - 1 \end{array} \right.$$

where δ is a constant and $\delta=0.5$ for one quantizer or $\delta=1$ for two quantizers.

Using the typical example, $A=-1.74$, $B=0.95833$. The coefficients satisfy the inequality

$$|A| \leq 2B\sqrt{\frac{2}{B}} - 1 = 1.957$$

therefore the bound

$$K \leq \frac{\delta(1+\sqrt{B})}{(1-B)\sqrt{1-\frac{A^2}{4B}}} = 103.6\delta$$

The bounds for one and two rounding quantizer versions are 51 and 103 respectively.

As respected, for the example, the bounds of one- and two-quantizer version, 51 and 103 are much bigger than the actual maximum amplitudes of limit cycles, 11 and 17.

2. RMS Bounds

Sandberg and Kaiser⁽⁸⁾ have derived a bound on the rms value of the quantization error. As contrasted with the absolute bounds, this bound gives no information on the maximum amplitude of a limit cycle. In other words, this bound can be exceeded.

For the second-order section, the rms bounds are as follows:

$$\text{For the constant limit cycles } K_c \leq \frac{\delta}{(1+A+B)} \quad (64)$$

$$\text{For the alternating limit cycles } K_a \leq \frac{\delta}{1-|A|+B} \quad (65)$$

For the periodic limit cycles

$$K_p \leq \begin{cases} \frac{\delta}{(1-B) \left(1 - \frac{A^2}{4B}\right)^{\frac{1}{2}}} & \text{for } B > 0 \text{ and } |A| \leq \frac{4B}{1+B} \\ \frac{\delta}{1-|A|+B} & \text{for either } B \leq 0 \text{ or } B > 0 \text{ and } |A| \geq \frac{4B}{1+B} \end{cases} \quad (66)$$

where δ is a constant and $\delta=0.5$ for one rounding quantizer version or $\delta=1$ for two quantizer version.

Let us still go back to the typical example, $A=-1.74$, $B=0.95833$. In the example, the coefficients satisfy the inequality

$$|A| \leq \frac{4B}{1+B} = 1.957$$

Therefore, the bound of the periodic limit cycles

$$K_p \leq \frac{\delta}{(1-B) \left(1 - \frac{A^2}{4B}\right)^{\frac{1}{2}}} = 52.3\delta$$

The bounds for one and two rounding quantizer versions are 26 and 52 respectively.

Comparing with the absolute bound one knows that the absolute bound is $(1+\sqrt{B})$ times bigger than rms bound. When $B \rightarrow 1$, $(1+\sqrt{B}) \rightarrow 2$. The rms bound is simple to evaluate and the maximum value of the limit cycle will not exceed this bound by a factor of more than 2.

3. Approximate Bound

Jackson⁽⁹⁾ has derived an estimate of the limit cycle amplitude based on an effective value linear model. This bound can be written as

$$K_p = \text{INT} \left(\frac{0.5}{1-|B|} \right) \quad (67)$$

where $\text{INT}(X)$ denotes the integer part of X .

Since this bound is based on the assumption that the nonlinear system oscillates if B has an "effective value" $B'=1$, which is a carry-over from linear theory, there may exist exceptions. Parker and Hess⁽²⁾ have pointed out that Jackson's bound may be exceeded in some cases. Claassen et al.⁽³⁷⁾ have proved that where there are two roundoff quantizers the value derived by Jackson is in fact a lower bound for the maximum amplitude of the possible limit cycles. This means that there must exist at least one limit cycle with an amplitude larger than or equal to this bound.

It is easy to verify that for the typical example $A=-1.74$, $B=0.95833$, the $K_p=11$ which is a very good estimation for one quantizer version but for two quantizer version this bound is exceeded by 6.

Jackson's bound is a simplest one in the bound expressions proposed. In most cases, especially for one quantizer version, it is accurate enough. Hess⁽²⁴⁾ has shown on his simulation studies that for two quantizer version exceptions from the effective value linear model occur for $B \geq \frac{5}{6}$ and for values of A around ± 1.5 , ± 1.0 , ± 0.5 . For these cases, he suggested that the following bound can be an alternative.

$$K_p = \frac{1.5}{1-|B|} \quad (68)$$

which is three times bigger than the bound of Jackson.

3.7 FREQUENCY OF LIMIT CYCLE

It is not difficult to obtain the frequency expression of the impulse response in a linear (without quantizer) second-order filter section.

As mentioned earlier, for a linear second-order basic section, its difference equation can be written as

$$Y(n) = X(n) - AY(n-1) - BY(n-2)$$

If the poles are complex its impulse response is

$$h(n) = \left(\frac{r^n}{\sin\theta}\right) \sin [(n+1)\theta]$$

where $r = \sqrt{B}$ and $\theta = \arccos\left(-\frac{A}{2\sqrt{B}}\right)$

For $B=1$, the impulse response is a sinusoid with constant amplitude and frequency

$$f = \left(\frac{1}{2\pi T_s}\right) \arccos\left(-\frac{A}{2}\right)$$

It is difficult to obtain an accurate frequency expression of limit cycles in the second-order digital filter, because of the nonlinearity. Although limit cycles in a digital system with only one nonlinearity can also be studied with the describing function method, this method is an approximate one which only gives results if the occurring limit cycles are almost sinusoidal. However, in order to get an impression about the possibility of

frequency of these limit cycles, Claassen et al⁽³³⁾ with describing function method have derived a frequency expression

$$f = \frac{1}{2\pi T_s} \arccos\left(\frac{A^2+B^2-1}{2AB}\right) \quad (69)$$

or

$$T = 2\pi T_s \left[\arccos\left(\frac{A^2+B^2-1}{2AB}\right)\right]^{-1} \quad (70)$$

Using the typical example, substituting the coefficients $A=-1.74$, $B=0.95833$ into above equation we know that the period of limit cycle is $13T_s$, which is accurate enough. But apparently, when the coefficients $A \rightarrow 0$, $B \neq 1$ above frequency expression cannot be applied because the argument $\left(\frac{A^2+B^2-1}{2AB}\right)$ may be greater than unity.

It is worth pointing out that if the poles are close to the unit circle in the Z-plane the limit cycle is approximately sinusoidal with a frequency close to the value given by the frequency expression of the linear model. Still substituting the coefficients $A=-1.74$, $B=0.95833$ into the Eqn. (13) the period is $14T_s$ which is a good estimation. As respected, in this case the above two frequency expressions are similar because in high-Q cases, the limit cycles approximate sinusoid and the describing function method becomes more accurate. In fact when $B \rightarrow 1$, Eqn. (70) degenerates to Eqn. (13).

3.8 SUMMARY

In this chapter we have discussed the main properties of zero-input limit cycles in second-order filters. There are two different analytical ways of quantization error; stochastic and deterministic. Both methods have their own applications. It is important to understand the assumptions of these two methods. In this research both methods will be applied in different situations.

When limit cycles exist the quantization error sequences $\delta(n)$ can be either constant, alternating or periodic. The correlation among $\delta(n)$ is a feature when limit cycles exist.

The second-order digital filters with multiplication coefficient B for which $|B| > 0.5$ will always exhibit limit cycles. Three different types of limit cycle may be distinguished; constant, alternating and periodic.

Successive-value phase-plane plots provide a useful means for displaying the nature of the limit cycles of a digital filter. The existing conditions of limit cycles can be shown, with a convenient way, in the parameter space. By the use of these parameter space plots, one can choose the coefficient values A and B correctly, so as to obtain a certain kind of limit cycle. This is important in the simulations.

Three different types of amplitude bound for limit cycles in the second-order digital filters have been given. These bounds are important not only in the determining of the internal wordlength of filter but also in the searching limit cycles in the filter because these bounds give the regions in the state plane where the limit cycles occur. One must be careful in the applications because each bound has its own restriction.

It is easy to estimate the occurrence of constant or alternating limit cycles from the parameter space plot. But one can only estimate the frequency of periodic limit cycles approximately. When the poles of the filter close to the unit circle in the Z-plane the frequency estimate becomes accurate. Simulations have shown that limit cycles of very long periods are possible.

Now we are in the position to discuss the methods of limit cycle suppression.

CHAPTER 4

STABILIZATION BY THE INJECTION OF DITHER

It has been known for a long time that limit cycle oscillations in nonlinear, continuous-time feedback systems can be suppressed by the injection of a random dither signal^(19,20). A recursive digital filter with quantization is a nonlinear, discrete-time feedback system. As mentioned in Chapter 1, several authors have studied the use of added dither for suppressing the limit cycles in digital filters, but in this research somewhat different approaches have been used which have certain advantages over the methods proposed before.

4.1 THE PROPOSED METHOD TO SUPPRESS LIMIT CYCLES

We have known that if a limit cycle in the second-order basic section exists, $\hat{Y}(n)$ and the quantization error $\delta(n)$ both are periodic or constant or alternating sequence depending on the limit cycle type. The basic idea of this proposed approach of suppressing the limit cycles is to inject a minimum pseudo-random noise at the front of quantizer (see Fig. 14), so as to break the periodicity of the quantization errors. This pseudo-random noise is called dither⁽¹⁾.

For the sake of simplicity, in the following chapters, unless specifically stated, only one quantizer version is considered. But the proposed method is effective to two quantizer version as well. In the absence of dither, the nth output from the filter, $\hat{Y}(n)$, is given by

$$\begin{aligned}\hat{Y}(n) &= [-A\hat{Y}(n-1)-B\hat{Y}(n-2)]_R \\ &= -A\hat{Y}(n-1)-B\hat{Y}(n-2)-\delta(n)\end{aligned}\quad (71)$$

where $[.]_R$ represents the operation of roundoff quantization, and $|\delta(n)| \leq 0.5$.

The above equation can be rewritten as

$$-A\hat{Y}(n-1)-B\hat{Y}(n-2) = [-A\hat{Y}(n-1)-B\hat{Y}(n-2)]_R + \delta(n)\quad (72)$$

If a random dither distributed in the open range $(-\frac{q}{2}, \frac{q}{2})$ is added at the nth instant, the resulting nth output from the filter, $\hat{Y}'(n)$, is given by

$$\hat{Y}'(n) = [-A\hat{Y}(n-1)-B\hat{Y}(n-2)+d(n)]_R\quad (73)$$

where $d(n)$ is the random dither and $|d(n)| < 0.5$ because the quantization step q has been assumed equal to one.

The difference of $\hat{Y}'(n)$ and $\hat{Y}(n)$ can be written as

$$\hat{Y}'(n) - \hat{Y}(n) = [-A\hat{Y}(n-1)-B\hat{Y}(n-2)+d(n)]_R - [-A\hat{Y}(n-1)-B\hat{Y}(n-2)]_R\quad (74)$$

Substitute Eqn. (72) into the above equation and we obtain

$$\begin{aligned}
& \hat{Y}'(n) - \hat{Y}(n) \\
&= \left[\left[-A\hat{Y}(n-1) - B\hat{Y}(n-2) \right]_R + \delta(n) + d(n) \right]_R - \left[-A\hat{Y}(n-1) - B\hat{Y}(n-2) \right]_R \\
&= \left[\delta(n) + d(n) \right]_R \tag{75}
\end{aligned}$$

In the case where the filter is not at the origin state, the sum $-A\hat{Y}(n-1) - B\hat{Y}(n-2)$ is, in general, not an integer multiple of q , even though $\hat{Y}(n-1)$ and $\hat{Y}(n-2)$ are integer multiples of q . Specially, if a limit cycle exists it is impossible that $\delta(n)$ are equal to zero for all n . Because suppose that a limit cycle exists and $\delta(n)=0$ for all n then that means no roundoff exists and the quantizer has no influence to the filter. But in this case there must be no limit cycle in the filter. This conclusion conflicts with the initial assumption.

Thus, if a random dither distributed in $(-\frac{q}{2}, \frac{q}{2})$ is added at the n th instant, there is a nonzero probability that

$$\begin{aligned}
\hat{Y}'(n) - \hat{Y}(n) &= \left[\delta(n) + d(n) \right]_R \\
&= \pm q \tag{76}
\end{aligned}$$

In other words, if the present state of the filter is on a limit cycle, with the addition of dither, there is a nonzero probability that the next state will be off that limit cycle. Although this does not guarantee that dither will stabilise the filter, i.e., ensure that with zero

input, the filter eventually reaches the origin state, the effect of the dither in causing the filter to leave the limit cycle makes it reasonable to speculate that the dither will suppress limit cycles in the second-order filters. In the next chapter, we will discuss this problem in detail.

4.2 SOME NOTES ON THE PROPOSED METHOD TO SUPPRESS LIMIT CYCLES

It is necessary to point out the properties of the proposed method to suppress limit cycles. Some properties show the possibility of limit cycle suppression.

- (1) Once a limit cycle has been suppressed the output signal from the filter remains at zero as long as the input signal is zero. From then on, the dither has no influence on the output. This statement can be readily proved from the equation

$$\hat{Y}'(n) = [-A\hat{Y}'(n-1) - B\hat{Y}'(n-2) + d(n)]_R \quad (77)$$

When a limit cycle has been suppressed means

$$\hat{Y}'(n-1) = \hat{Y}'(n-2) = 0$$

and because

$$|d(n)| < 0.5$$

$$\hat{Y}'(n) = [d(n)]_R = 0 \quad (78)$$

In some cases, this property is important because it means no remaining noise is left after the limit cycle has been suppressed.

- (2) The injection of the dither makes the origin point $(0,0)$ in the state plane be branch point even the coefficient B of the second-order filter satisfies that $1 > |B| > 0.5$.

As mentioned earlier, without dither, in case one or two rounding operations are used, the origin $(0,0)$ is a branch point only if $|B| < 0.5$ and from this statement one asserts that the filters with multiplication coefficient B for which $|B| > 0.5$ will always exhibit limit cycles.

But after adding dither the situation is quite different. As shown in Appendix 3, in case one or two rounding quantizers are used even though $1 > |B| > 0.5$ the origin state $(0,0)$ still can be a branch point only if

- (a) in one quantizer case the dither is added at the front of quantizer, or
- (b) in two quantiser case the dithers are added respectively at the front of two quantizers, or
- (c) in two quantizer case the dither is only added at the front of B coefficient product quantizer.

It is also shown in the Appendix 3 that it is impossible

when two quantizers are used to suppress the limit cycles by the addition of the dither at the front of coefficient A product quantizer only. Because in this case the origin state is not a branch point any more.

These arguments show the intuitive reason of suppressing limit cycles by the use of the random dither.

- (3) The injection of a uniformly distributed random dither makes the statistical quantization characteristic (mean value output versus input) linearise.

It has been proved in Appendix 4 that with a concept of equivalent quantizer that when a dither uniformly distributed in $(-\frac{q}{2}, \frac{q}{2})$ is used the mean value output from the equivalent quantizer varies with the input with a linear manner. This statement at least shows the tendency of linearisation by the use of the uniformly distributed random dither.

- (4) In the two quantizer version, two dither signals can be taken from the same random sequence generator.

In the two rounding quantizer version with dither, the difference equation can be written as

$$\hat{Y}'(n) = [-A\hat{Y}'(n-1) + d_a(n)]_R + [-B\hat{Y}'(n-2) + d_b(n)]_R \quad (79)$$

Suppose that

$$[-\hat{A}\hat{Y}'(n-1)]_R = -\hat{A}\hat{Y}'(n-1) - \delta_a(n) \quad (80)$$

and

$$[-\hat{B}\hat{Y}'(n-2)]_R = -\hat{B}\hat{Y}'(n-2) - \delta_b(n) \quad (81)$$

where

$$|\delta_a(n)| \leq 0.5 \quad \text{and} \quad |\delta_b(n)| \leq 0.5$$

Therefore,

$$\hat{Y}'(n) = [-\hat{A}\hat{Y}'(n-1)]_R + [\delta_a(n) + d_a(n)]_R + [-\hat{B}\hat{Y}'(n-2)]_R + [\delta_b(n) + d_b(n)]_R \quad (82)$$

Because, in general, $\delta_a(n)$ and $\delta_b(n)$ are different and their correlation is small, even $d_a(n) = d_b(n)$, the values of $[\delta_a(n) + d_a(n)]_R$ and $[\delta_b(n) + d_b(n)]_R$ still can be different and their correlation is also small.

It is expected that by the use of the same dither, the purpose of limit cycle suppression can still be carried out, and the simulation has verified this expectation.

4.3 SUMMARY

The proposed method to suppress the limit cycles in the second-order digital filters has been described. The use of the dither may cause the filter to leave the limit cycles and make the origin state (0,0) be branch point.

Once the origin state has been reached the state of filter will stay there as long as the input signal is zero. These properties of the proposed method support us to speculate that the dither will suppress limit cycles in the second-order filters eventually.

One of the advantages of this method is that in the zero-input condition, once the limit cycle has been suppressed the output signal remains at zero, i.e., no remaining noise left.

In the next chapter, the effect of the dither on the limit cycle suppression will be discussed further.

CHAPTER 5

HOW DITHER AFFECTS THE LIMIT CYCLES

Although two properties of the proposed method to suppress limit cycles in the second-order digital filters mentioned in Chapter 4 make it reasonable to speculate that the dither will suppress limit cycles, those have not proved that the dither will stabilize the filters yet. In this chapter, we will first verify the stabilization for a particular pair of coefficient values A and B, then prove it for general cases though it is not strict in the mathematical sense.

5.1 VERIFICATION OF THE STABILIZATION BY THE USE OF DITHER FOR A PARTICULAR PAIR OF COEFFICIENT VALUES A AND B

For a particular pair of coefficient values A and B, it is possible to verify that dither will stabilise the filter. This can be done as follows:

First, all the limit cycles are identified, for example, by simulating the digital filter using all possible initial states in the zone of the state plane where limit cycles may exist. This zone can be determined by use of the known bounds on limit cycle amplitudes.

Second, each state is examined to determine which limit

cycle it belongs to and is labelled to indicate this. The initial state from which the limit cycle can be reached is said to belong to that limit cycle. In practice, this can be done concurrently with the first step.

Third, each limit cycle is examined in turn to determine which other limit cycles can be reached from it when dither is added. This can be done by examining each state on the limit cycle, finding which state can be reached from there when dither is present and noting the limit cycles to which these states belong.

Lastly, a directed graph called transition diagram⁽³⁹⁾ results, showing which limit cycles can be reached from which others. It can be discovered from this transition diagram whether or not it is possible to reach the origin from every one of the limit cycles.

In the following sections, we will describe the procedure with the example frequently used before, step by step. As will be seen, although we use the typical example, any specified pair of coefficient values A and B is suitable, i.e., the procedure mentioned is general.

1. Identify All the Limit Cycles in the Second-Order Filter Section With Coefficient Values $A=-1.74$, $B=0.95833$

The amplitude bound of the limit cycles in the filter section is found by the use of Jackson's bound formula. This bound is equal to 11. In the state plane, this

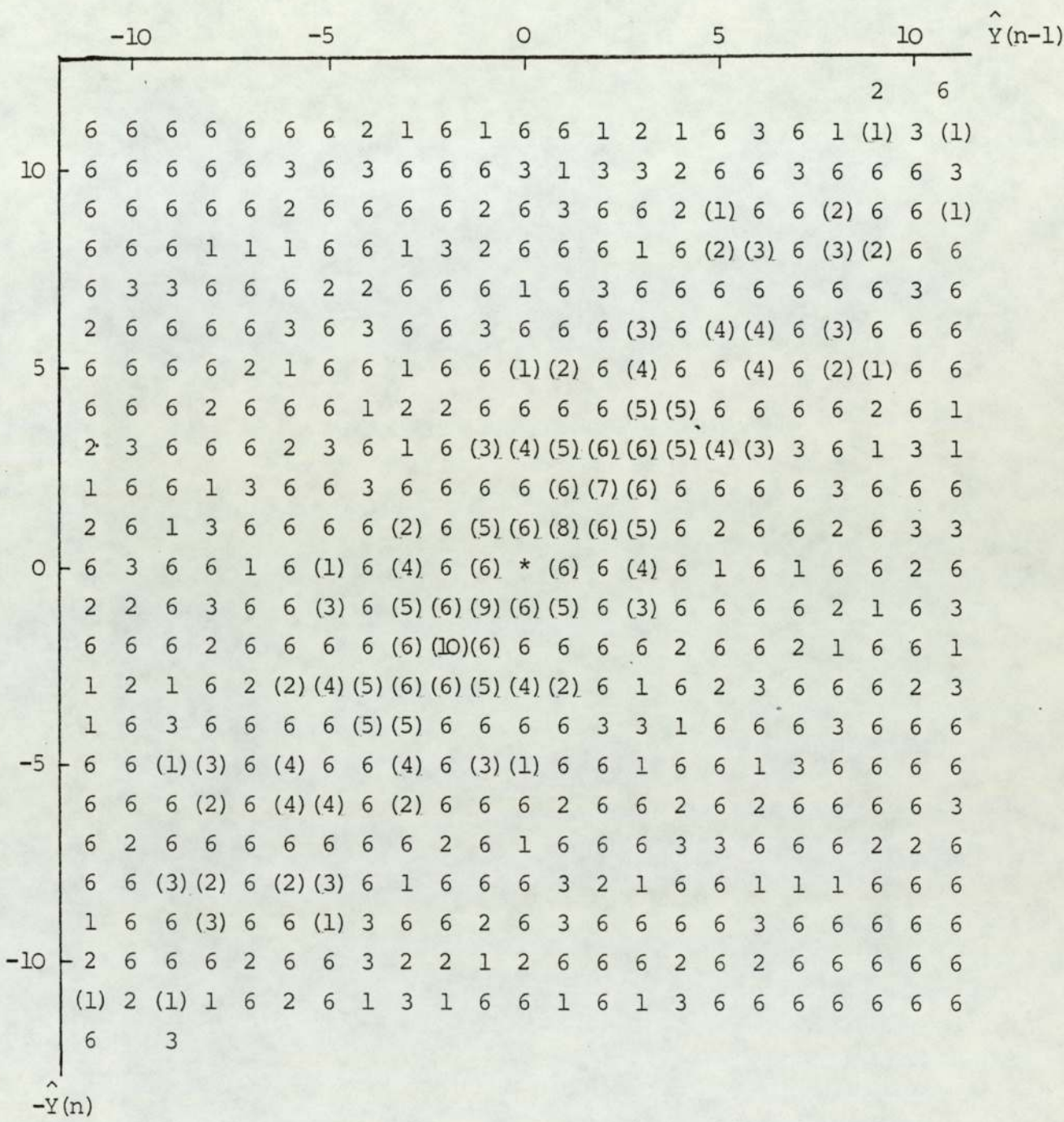
bound defines a zone which is bounded by $\hat{Y}(n) = \pm 11$ and $\hat{Y}(n-1) = \pm 11$. Apparently, there are $(11 \times 2 + 1)^2 = 529$ states altogether including two axes themselves. Having known the coefficients and initial condition it is easy to simulate the filter with the computer and find the limit cycles. All the limit cycles have been shown in Fig. 18 before.

2. Obtain the "Distribution Diagram" of Limit Cycles

First, each limit cycle is numbered successively from 1, 2, 3, 4, ... As mentioned earlier, for this example, there are 10 limit cycles, therefore each limit cycle from the largest periodic one to the smallest constant one is given by the number from 1 to 10 respectively. Each state in the zone defined by the amplitude bound of limit cycle is used in turn as an initial state. The limit cycle which the initial state belongs to can be found by simulation and the initial state is indicated by the same number with that of the limit cycle. After 529 steps the "distribution diagram" of the limit cycles can be obtained as shown in Fig. 21. Because the Jackson's bound may be exceeded, a check is necessary. No larger limit cycles have been found outside the zone for this example.

3. Determine Which Other Limit Cycles can be Reached from Each Limit Cycle when the Dither is Added

In this example, suppose a uniformly distributed dither



* origin state
 (*) limit cycle

Fig. 21 "Distribution diagram" of the limit cycles in the example shows to which limit cycle each state belongs.

is used. If this random dither is added at the nth instant then the difference of the output from the filter with and without dither can be written as

$$\hat{Y}'(n) - \hat{Y}(n) = [\delta(n) + d(n)]_R$$

Because the dither is uniform distributed in the open range $(-\frac{q}{2}, \frac{q}{2})$ and $|\delta(n)| \leq 0.5$ the following equation can be obtained.

$$\hat{Y}'(n) - \hat{Y}(n) = \begin{cases} +1 \\ 0 \\ -1 \end{cases} \quad (83)$$

It is reasonable to assume that the probability of occurrence of each value in the three possible numbers is equal to $\frac{1}{3}$. In other words, by the addition of the dither, there is a probability of $\frac{1}{3}$ that the filter will stay at the same limit cycle. Both probabilities of leave off the original limit cycle to two neighbouring states are also equal to $\frac{1}{3}$ respectively.

Suppose that there is a limit cycle which includes N states in the state plane. Then after adding dither, the number of states which can be reached is 3N. From the "distribution diagram" of limit cycles we know which state belongs to which limit cycle. Therefore, the probability of transition from a limit cycle i to limit cycle j can be calculated.

For example, limit cycle 1 includes 14 states. Hence by the injection of the dither, $14 \times 3 = 42$ states, can be

reached by the filter. From the "distribution diagram" of limit cycle we know that in the 42 states, 14 states belong to the limit cycle 1 itself, 6 states belong to limit cycle 2, 6 states belong to limit cycle 3, 16 states belong to the limit cycle 6. It is clear that by the use of dither, the filter either stays at the original limit cycle or moves from limit cycle 1 to limit cycle 2, 3 and 6. The probabilities of the transition are $\frac{14}{42}$, $\frac{6}{42}$, $\frac{6}{42}$, and $\frac{16}{42}$ respectively.

Along the same way, the probabilities of transition from any limit cycle to others can be calculated.

4. Drawing the Transition Diagram

As we have known that without dither a digital filter with coefficients A, B lying on some region in the parameter space will continue to oscillate in a particular limit cycle depending on the initial condition. However, when dither is injected it becomes possible for the filter to move between limit cycles. The state of the filter is, therefore, no longer trapped in a particular limit cycle but can move randomly from one to another. These transitions can be graphically represented by a transition diagram⁽³⁹⁾. In the transition diagram, each node is numbered to represent one limit cycle. A directed line segment or branch is drawn from each node i to each node j and labelled with the transition probability, $P_{i,j}$. Note that because

from the origin state the filter cannot move to any limit cycle except the origin state itself, the probability of transition from origin to origin is one. This state is called a trapping state.

Fig. 22 shows the direct graph corresponding to the filter section used as an example above. The nodes from 1 to 10 represent 10 limit cycles and node 11 represents the origin state. The probabilities of transition from each limit cycle to others have been calculated as above.

It can be seen from the transition diagram that each limit cycle can be reached from some other limit cycles. Although the origin state cannot be reached directly from all the limit cycles, there is no limit cycle from which the origin cannot be reached indirectly if necessary.

Once the filter reaches the origin the filter remains at this state as long as the input signal is zero. Thus, for this particular filter, it is proved that the dither signal stabilises it.

5.2 THE MAXIMUM TRANSITION TIME NEEDED TO MOVE FROM ANY LIMIT CYCLE TO THE ORIGIN STATE

As mentioned earlier, without dither for the second-order filter with the coefficient $|B| > 0.5$, the origin state is not a branch point. Therefore, each state except the origin in the state plane must belong to one limit cycle.

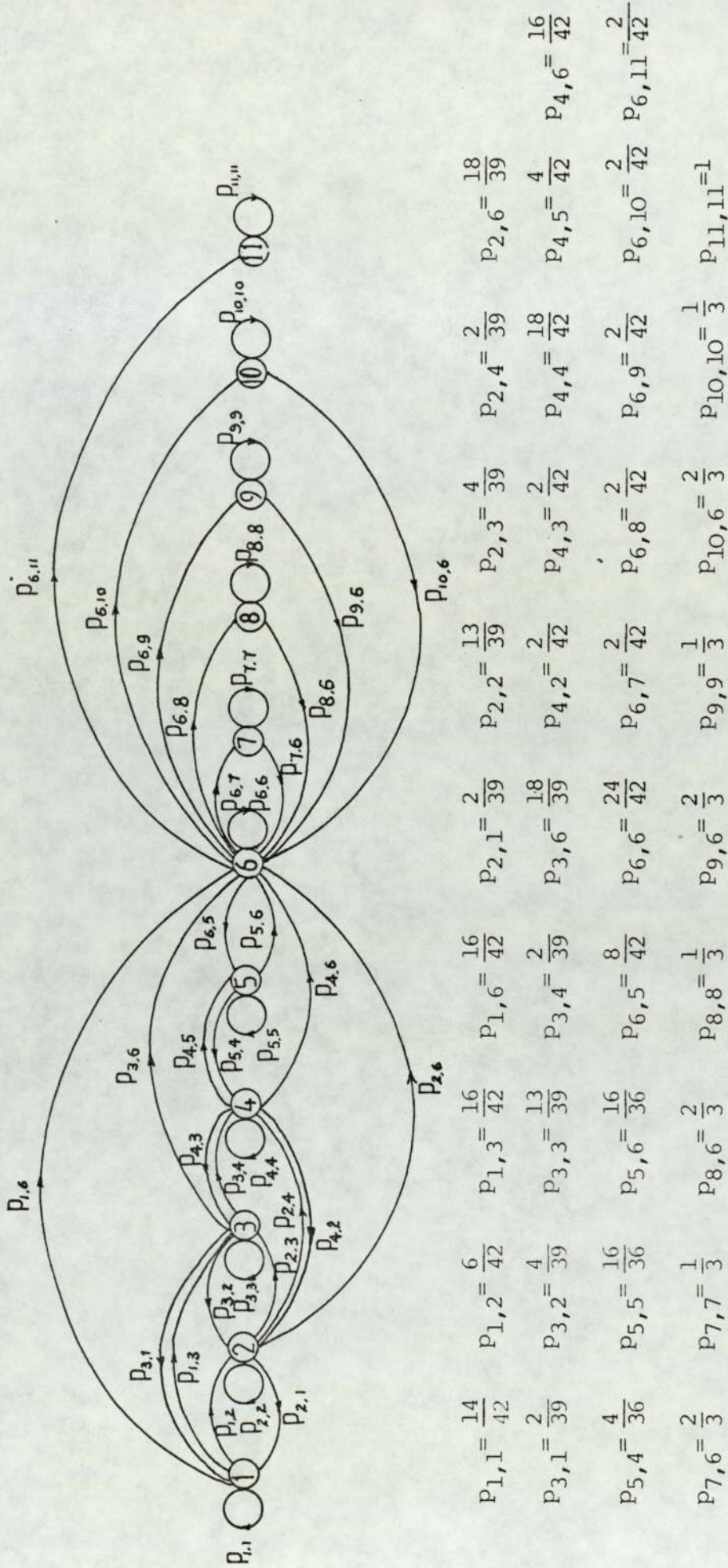


Fig. 22 Transition diagram for the second-order section ($A=-1.74$, $B=0.95833$) with uniformly distributed random dither.

The origin state only belongs to itself. Apparently, at any instant, a second-order filter must occupy a state. We say that a filter "occupies" limit cycle i when it occupies a state which belongs to the limit cycle i .

We have specified a set of conditional probabilities $p_{i,j}$ that a filter which now occupies limit cycle i will occupy limit cycle j after its next transition. As just mentioned, since the filter must occupy a limit cycle after its next transition, therefore

$$\sum_{j=1}^{N+1} p_{i,j} = 1 \quad (84)$$

where N is the total number of limit cycles which the filter may occupy. The upper limit of the summation is $(N+1)$ including the origin state $(0,0)$.

The probability that the filter will remain in i , $p_{i,i}$, has been included in the above equation. Apparently, since the $p_{i,j}$ are probabilities

$$0 \leq p_{i,j} \leq 1 \quad (85)$$

The transition probabilities $p_{i,j}$ may be ranged in matrix form called a transition probability matrix.

$$\bar{p}^{(1)} = \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots & p_{1,N+1} \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots & p_{2,N+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{N+1,1} & p_{N+1,2} & p_{N+1,3} & \cdots & p_{N+1,N+1} \end{bmatrix} \quad (86)$$

where $\bar{p}^{(1)}$ represents the first-step transition matrix and the sum of elements in each row is equal to unity.

For the example used above, the first-step transition matrix has been known and can be written as follows:

$$\bar{p}^{(1)} = \begin{bmatrix} \frac{14}{42} & \frac{6}{42} & \frac{6}{42} & 0 & 0 & \frac{16}{42} & 0 & 0 & 0 & 0 & 0 \\ \frac{2}{39} & \frac{13}{39} & \frac{4}{39} & \frac{2}{39} & 0 & \frac{18}{39} & 0 & 0 & 0 & 0 & 0 \\ \frac{2}{39} & \frac{4}{39} & \frac{13}{39} & \frac{2}{39} & 0 & \frac{18}{39} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2}{42} & \frac{2}{42} & \frac{18}{42} & \frac{4}{42} & \frac{16}{42} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{4}{36} & \frac{16}{36} & \frac{16}{36} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{8}{42} & \frac{24}{42} & \frac{2}{42} & \frac{2}{42} & \frac{2}{42} & \frac{2}{42} & \frac{2}{42} \\ 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (87)$$

As mentioned before, the injection of dither makes the filter section move from one limit cycle to another. The transition between limit cycles can be treated as a Markov process. According to Markov theory⁽³⁹⁾, the Mth step transition probability matrix can be expressed as

$$\bar{p}^{(M)} = [\bar{p}^{(1)}]^M \quad (88)$$

The matrix multiplication has been carried out by computer. The calculated results show that when the transition step M is greater than 310, all the elements in the last column in the Mth transition probability matrix tend to 1 and the others tend to zero with the error less than 1×10^{-4} , i.e.,

$$\bar{p}^{(M)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (88)$$

It means that in the sense of statistics, the filter should reach the origin state (0,0) from any limit cycle after about 310 step transitions. This has, of course, proved the stabilisation by the use of the dither for the specified filter. As will be seen later, in 1000 simulations the median value of the transition time from the largest limit cycle to the origin for the filter in this example is $330 T_s$ which nears the transition time, $310 T_s$, calculated above.

5.3 VERIFICATION OF THE STABILIZATION BY THE INJECTION OF DITHER FOR GENERAL CASES

1. Equivalent Quantizer

The only one nonlinear element (overflow has been excluded) is the rounding quantizer. Therefore, we will pay more attention on the rounding quantizer with dither.

Both dither adder (including the dither generator) and quantizer itself can be treated as an equivalent quantizer Q_e , as shown in Fig. 23. Its input and output signal are $-\hat{A}Y(n-1)-\hat{B}Y(n-2)$ and $[-\hat{A}Y(n-1)-\hat{B}Y(n-2)+d(n)]_R$ respectively. Fig. 24 shows the quantization characteristic of Q_e . In the figure, the solid line represents the quantization characteristic of the rounding quantizer, the dashed line shows the new possible values because of the use of dither, and the 45° line represents the ideal linear case, i.e., in case no quantization and dither are used.

The characteristic of Q_e consists of two parts as shown in Fig. 24; one part lies in the sector between the 45° line and X-axis which is just the magnitude-truncation characteristic, other part lies in the sector between the 45° line and Y-axis called a rounding up characteristic. The dither makes the characteristic of the quantizer jump randomly between these two parts.

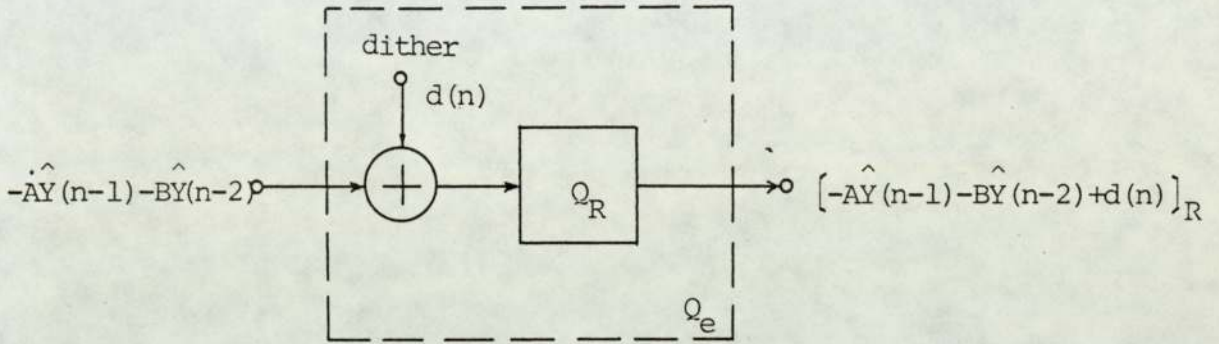


Fig. 23 Equivalent quantizer.

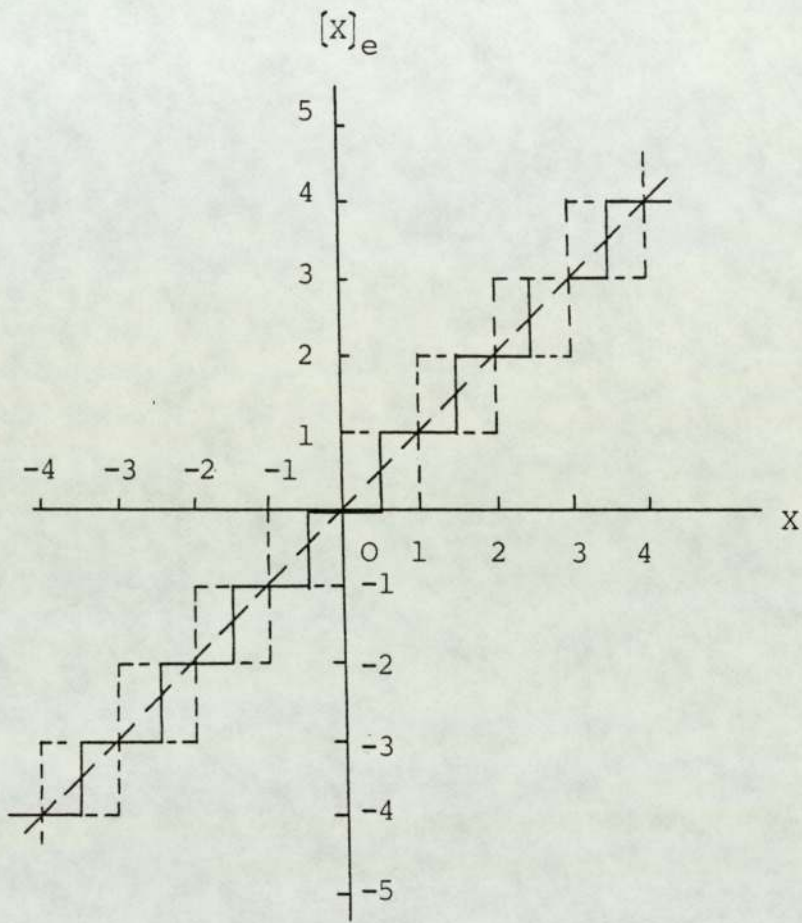


Fig. 24 Characteristic of equivalent quantizer Q_e .

2. The Magnitude-Truncation Part in the Characteristic of Q_e Tends to Stabilize the Filter

As mentioned in Chapter 2 and 3, the areas of asymptotic stability of the second-order digital filters with one and two magnitude-truncation quantizers in the parameter space have been derived by using the frequency domain criteria for absence of zero-input limit cycles. Comparing the filter with the rounding quantizer, the area of asymptotic stability of the filter with magnitude-truncation quantization is much bigger than that of the filter with roundoff quantization. Contrast with roundoff, whatever one or two magnitude-truncation quantizers are used the origin state is a branch point.

In particular, for one magnitude-truncation quantizer, limit cycles occur for only a very few values of the multiplier coefficients A and B. In these limit cycles, only about (25~40)% are accessible.

For the filter with two magnitude-truncation quantizers, only constant or alternating limit cycles have been observed by simulations and they are found in the linear stable triangle area for values $|A| > 1$. For high-Q poles ($B \leq 1$) only limit cycles with magnitudes equal to one quantization step are accessible. In this case, by the use of the dither, the filter may move to the origin state with a very big probability, because there are at least

two predecessors of the origin state defined by $(0, \pm 1)$.

The important fact which supports the argument on stabilization is that whatever one or two magnitude-truncation quantizers are used, even for those coefficients with which the limit cycles exist there are still a large number of initial states in the state plane from which the filter will move to the origin state $(0,0)$. Our simulations showed that averagely speaking, there are about 50% or more states in the state plane from which the origin can be reached by the filter with the coefficients which lie on the unstable area in the parameter space. In other words, for any second-order filter with magnitude-truncation quantization the origin state $(0,0)$ can be reached from all or at least a large number of initial states.

The conclusion is that if the filter is started with randomly chosen initial conditions, any second-order filter with magnitude - truncation quantization can reach the origin state $(0,0)$ with a big non-zero probability. As we have known, the dither makes the filter move state by state, randomly. Once the origin state has been reached, the filter sticks there as long as the input is zero.

3. The Rounding Up Part in the Characteristic of Q_e Makes the Filter be Unstable

This argument is readily proved, because in this case the origin state is not a branch point.

4. The Filter With the Characteristic of Q_e is, in a Broad Sense, Zero-Input Stable

As mentioned earlier, the dither makes the characteristic of the quantizer jump randomly between truncation and rounding up parts. We have concluded on the basis of experiment that when the magnitude-truncation characteristic part is used, the filter will tend to move with a big probability towards to the origin state and when the rounding up characteristic part is used, the filter will tend to move off the origin state. The dither signal makes the filter move state by state, randomly. These transitions continue until the origin state is reached. Since the transition is random after some finite time, the filter can move to the origin state. Once the origin state is reached, the filter will remain there as long as the input is zero.

Now the stabilization by the injection of dither has been proved.

By the use of the program listed in the Appendix 1, the suppressing procedure of limit cycles can be shown on

the screen of the computer PET. The trajectory shown on the screen represents the variation of the state occupied by the filter. As can be seen, by the use of the dither, the trajectory sometimes moves away from the origin, and sometimes moves towards the origin. This procedure continues until the origin state is reached. As soon as the filter reaches the (0,0) state, it remains there if the input keeps being zero.

This observation verifies what has previously been stated.

5.4 SUMMARY

In this chapter the necessity of the limit cycle suppression in the second-order digital filters by the use of dither has been proved though partly on the experimental basis. Because the quantization nonlinearities occurring in digital filters are highly discontinuous functions, it is difficult to prove the stabilization strictly. But for a specified pair of coefficient values A and B, it is possible to verify that the dither will stabilize the filter. By Markov theory, the maximum transition time needed for transition from any limit cycle to the origin state may be calculated. The result has been verified by simulation.

In principle, in the proposed method to suppress the limit cycles, a rounding quantizer with the random dither

is equivalent to a controlled quantizer where the quantization is controlled by the dither. Depending on the signal in the filter and the dither, the quantization is switched between magnitude-truncation and rounding up randomly. The magnitude-truncation part tends to stabilize the filter but the rounding up part makes the filter become unstable. Once the limit cycle has been suppressed the output from the filter keeps zero as long as the input keeps zero - it is not necessary for the filter to be asymptotically stable. The necessary condition of stabilization is that the origin state can be reached with nonzero probability. The magnitude-truncation part in the characteristic of the equivalent quantizer provides this possibility. This concept has been used to prove the stabilization of the filter by the injection of the random dither.

CHAPTER 6

DITHER SIGNALS

So far in this thesis only one kind of dither signal, uniformly distributed random dither, has been used. This chapter will proceed with further discussion on dither signal. Firstly, the previous work on the use of dither for limit cycle suppression will be reviewed and discussed. Then, the principal considerations of designing dither signal will be described. Finally, several dither signals which have been verified by simulation will be introduced.

6.1 REVIEW AND DISCUSSION OF THE PREVIOUS WORK ON THE USE OF DITHER FOR LIMIT CYCLE SUPPRESSION

In this section we will review and discuss the methods of limit cycle suppression proposed before to find out their advantages and disadvantages. This discussion is very helpful to propose the principal considerations of designing dither signals.

A number of methods for using added dither to suppress limit cycles have been suggested. In the following text the relevant methods will be discussed one by one.

1. Randomized Quantization Method⁽³⁾

In this method the quantization is switched randomly between truncating and rounding. This is equivalent to the addition of a random dither at the front of rounding

quantizer whose value is either 0 or $(-q/2 + \epsilon)$ where ϵ is a very small positive number compared with the quantization step q . The equivalent quantization characteristic of the whole quantization system is shown in Fig. 25.

It is readily proved that for truncation quantization the origin state is a branch point and its predecessors are $(0, \pm 1)$. But as we know for roundoff quantization the origin is not a branch point. Therefore, in order to reach the origin state $(0,0)$ truncation must have occurred while in the states $(0, \pm 1)$. Simulation has verified that this method can reduce or suppress limit cycles in digital filters.

In this method, when a roundoff quantization is used, the filter tends to form a limit cycle whose amplitude depends on the initial condition, but when the truncation quantization is used, the filter, in general, tends to move to the origin state, i.e., tends to decrease the amplitude of output signal from the filter. Therefore, when the rounding quantization is used again, the filter tends to form a smaller limit cycle because the filter starts at a new smaller initial state. In this way, whenever the filter is switched to truncation quantization, the amplitude of limit cycle oscillation will be decreased to some smaller value. This procedure will repeat randomly until the limit cycle is suppressed. Because of the monotonic decreasing characteristic of the oscillation during

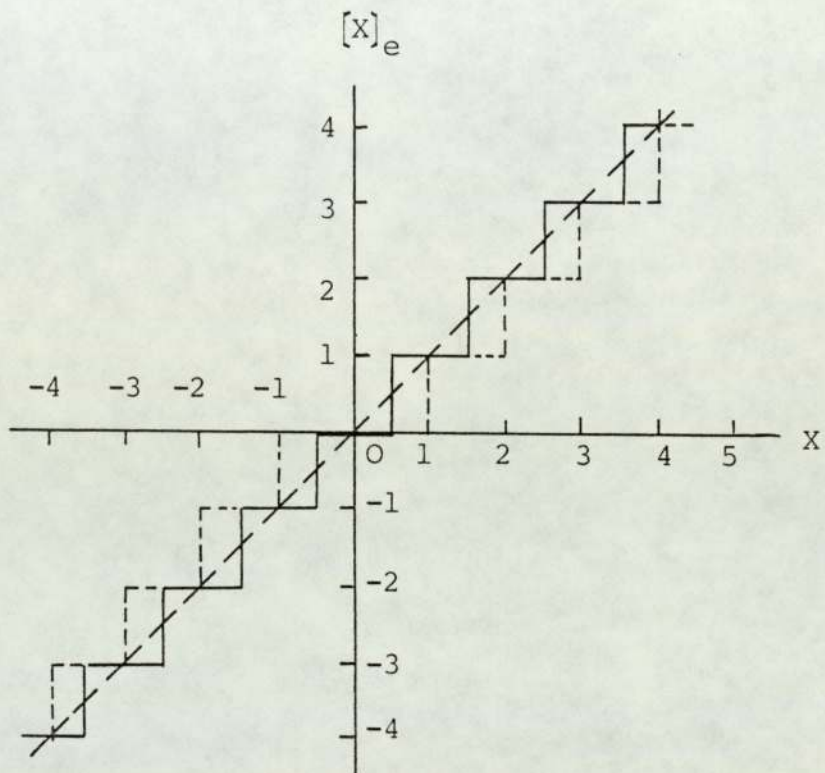


Fig. 25 The characteristic of equivalent quantizer in the randomized quantization method.

the switching period, this method can make the filter reach the origin state in a shorter time. Once a limit cycle has been suppressed, the output from the filter remains zero as long as the input signal is zero, i.e., there is no remaining noise at the output terminal. The disadvantage of this method is that some constant or alternating limit cycles cannot be suppressed. The reason is quite clear from the equivalent characteristic of quantization. As can be seen from Fig. 25 there is half a range of X , the input of the quantizer, where there is no difference between truncating and rounding. As mentioned earlier, when a constant or alternating limit cycle exists, the input of the quantizer is also a constant. Apparently, when this input signal of quantizer just lies in the range where there is no difference between truncating and rounding, then the quantization switching between truncating and rounding has no influence on the limit cycle. In this case, the limit cycle, of course, cannot be suppressed. When constant or alternating limit cycles exist, the roundoff quantization error $|\delta(n)|$ is a constant. As mentioned earlier, the difference of the quantizer output with and without dither is $[\delta(n) + d(n)]_Q$ where $d(n)$ is equal to either 0 or $-\frac{q}{2} + \epsilon$. It is clear that when $0 < \delta(n) < \frac{1}{2}q$ then whatever the rounding or truncating takes place $[\delta(n) + d(n)]_Q = 0$. In other words, the equivalent dither has no influence on the output value

of the quantizer.

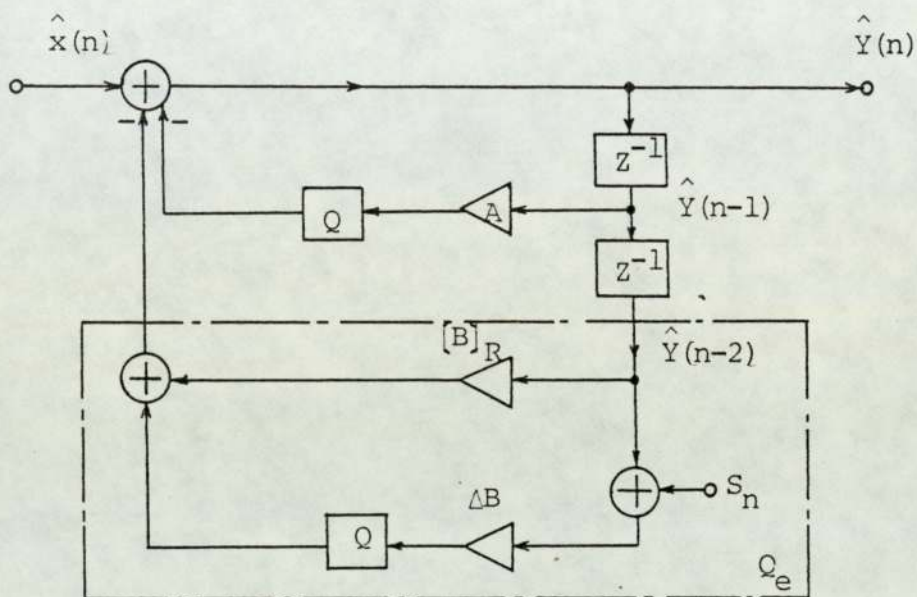
In order to suppress these kinds of limit cycle, we have to extend the range where the quantization characteristic is changeable. This purpose can be obtained by extending the range of the equivalent dither signal to $(-q/2, q/2)$. This is an important consideration in the dither signal designing.

2. The Method of Limit Cycle Suppression proposed by Rashidi and Bogner⁽⁵⁾

In this method, a stabilizing signal S_n , of amplitude A_s , is added to the coefficient B product branch as shown in Fig. 26. S_n is normally random.

If the whole coefficient B branch except delay block is treated as an equivalent quantizer, then its equivalent quantization characteristic is shown in Fig. 27. In this figure the shaded areas represent the regions where the transition of the quantization function can occur. The amplitude of the stabilizing signal needed for limit cycle suppression has been given by the formulae in the paper ⁽⁵⁾ which depends on the coefficients A and B.

The simulation shows that most limit cycles in the second-order digital filters can be suppressed by the use of this method. Once a limit cycle has been suppressed there is no remaining noise at the output terminal in zero-input condition. But as can be seen from the equivalent quantization characteristic shown



$$\Delta B = B - [B]_R$$

Fig. 26 Block diagram of the method for limit cycle suppression proposed by Rashidi and Bogner.

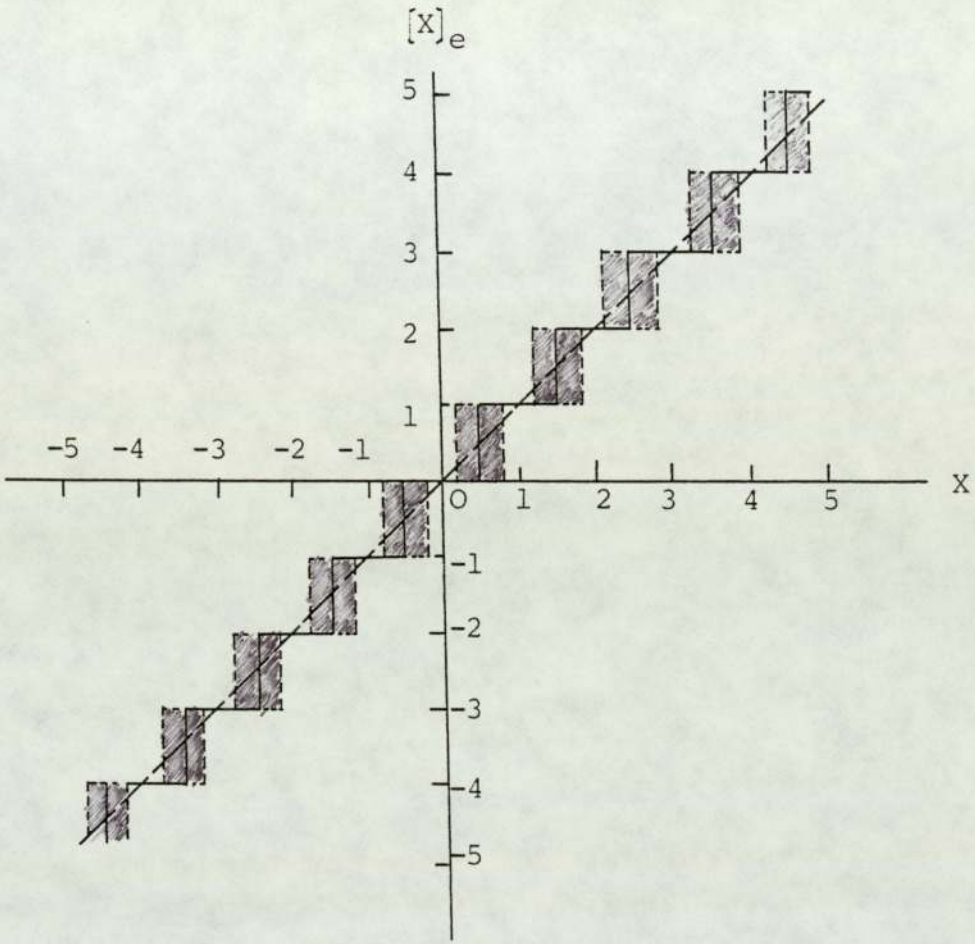


Fig. 27 The characteristic of equivalent quantizer in Fig. 26.

in Fig. 27, the region where the transition of the quantization function can occur is not wide enough. There are still some regions where the quantization characteristic has no difference with the roundoff quantization. It can be expected that there may be some constant or alternating limit cycles which cannot be suppressed by the use of this method. For example when this method is applied to the filter with the coefficient values $A = -\frac{109}{64}$ and $B = \frac{53}{64}$, i.e., $[A]_R = -2$, $[B]_R = 1$, and $\Delta B = B - [B]_R = -0.171875$. Reference ⁽⁵⁾ gives the formula to decide the amplitude of stabilizing signal. According to Eqn. (25a) in ⁽⁵⁾, $A_S = MT[0.5/|\Delta B|]q = 2q$. Suppose that the initial conditions are $Y(-1) = Y(-2) = 5$, it is readily verified that the output from the filter is also equal to 5. Therefore, the filter provides a constant limit cycle ($\dots, 5, 5, 5, \dots$) which cannot be suppressed by the use of the method. Other examples are the filters with coefficients $A = -1.58$, $B = 0.605$ and $A = -1.8125$, $B = 0.828125$. In these filters the constant limit cycles ($\dots, 5, 5, 5, \dots$) cannot be suppressed either.

Additionally, because the stabilizing signal S_n is added before the multiplication the amplitude A_S depends on the coefficient B . Therefore, when B changes, A_S should be adjusted also.

From this method one knows again that the region where the transition of the quantization function can

occur must be wide enough otherwise there must be some constant or alternating limit cycles which cannot be suppressed by the use of dither. The intuition tells us that it will be safe if this region is extended to the whole range of the input signal of the equivalent quantizer. How wide the region should be depends on the coefficient values. This problem will be discussed later.

3. The Method of Limit Cycle Suppression Proposed by Büttner⁽⁴⁾

In this method, two basic random roundings RR1 and RR2 are used.

In RR1, two added signals first add together precisely, then the bit having half of the weight of the LSB is replaced with the instantaneous value of a binary random sequence. Finally, the roundoff quantization is used.

Similarly, the replacement and the rounding quantizer can be treated as an equivalent quantizer and the equivalent quantization characteristic is the same with that shown in Fig. 24. It is clear that the transition region where the jump of the quantization function can occur is wide enough. As can be expected, all the limit cycles in the second-order filters can be suppressed by this method. But because the magnitude of the binary random sequence is equal to 0.5 with 50% probability, even the limit cycle has been suppressed, the output

from the filter does not vanish for zero-input. Instead, a noise-like signal is produced at the output. This extraneous noise is called the remaining noise. It is worth noting that because of the feedback in the recursive filter, this remaining noise is not necessarily very small.

Apparently, if the amplitude of added dither is near, but less than 0.5, then the remaining noise will vanish. This is another key point that should be considered in the dither signal design.

With method RR2, as mentioned in the original paper, some constant or alternating limit cycles still cannot be suppressed.

4. Folding-frequency Dither Method⁽¹⁾:

This method was proposed by Blackman in 1965. In this method, one folding-frequency dither, $(-1)^n D$, where $D = \frac{1}{2}q - \Delta$ and Δ is less than $(1-B)q$, is added before rounding. Blackman has asserted that by the use of the dither, the limit cycles (he called limit cycle phenomenon dead band effect of roundoff errors), will be suppressed. But he has only studied the first-order filters.

The simulations showed that this dither results in quicker limit cycle suppression than uniformly distributed random dither, due to the greater amplitude of the dither resulting in more frequent transitions between the various

limit cycles. But the simulations also showed that there were a lot of exceptions in the second-order sections in which some limit cycles could not be suppressed. For the typical example ($A = -1.74$, $B = 0.95833$, one quantizer) in the $(11 \times 2 + 1)^2 = 529$ states there are 102 initial states from which the origin could not be reached by the use of this alternative sign dither. For example, suppose the initial state is $Y(-1), Y(-2) = 3, 3$ then the outputs from the filter without dither are as follows,

..., 2, 1, 0, -1, -2, -3, -3, -2, -1, 0, 1, 2, 3, 3, ...

But the steady-state outputs from the filter with a dither $0.49(-1)^n$ are as follows,

..., 3, 2, 1, -1, -2, -3, -3, -3, -2, -1, 1, 2, 3, 3, ...

As can be seen from the above data, because of periodicity of this dither for some initial conditions the filter can never reach the origin state but another limit cycle. From this example another principal consideration in the dither signal design can be obtained i.e, the dither should not be periodic.

In this section, the main previous works on the use of dither for limit cycle suppression have been reviewed and discussed. As has been seen, although all these methods proposed before can be applied to suppress the limit cycles in the second-order filters but either with some methods some constant or alternating limit cycles cannot be suppressed, or with other methods even when the limit cycle has been suppressed there is remaining noise at the output terminal of the filter.

Apparently, one needs some methods with which all

the limit cycles in the second-order filters can be suppressed and once the limit cycle has been suppressed the output should keep being zero as long as the input is zero.

6.2 PRINCIPAL CONSIDERATIONS IN THE DITHER SIGNAL DESIGN

From the above review and discussion of the previous work on the use of dither for limit cycle suppression, some principal considerations in the dither signal design can be obtained.

(1) The amplitude of dither should be big enough to change the output of the quantizer.

As mentioned earlier, the difference of the output of the quantizer with and without dither is

$$\hat{Y}'(n) - \hat{Y}(n) = [\delta(n) + d(n)]_R$$

It is clear that in order to change the output of the quantizer, the dither must satisfy the following inequality

$$|[\delta(n) + d(n)]_R| > 1 \quad (89)$$

Apparently, the bigger the dither amplitude, the easier to change the output of the quantizer.

(2) The amplitude of dither should be small enough to satisfy

$$[d(n)]_R = 0$$

or

$$|d(n)| < 0.5 \quad (90)$$

Above equation insures that once the filter has reached the origin state (0.0) (in this case, the quantization error $\delta(n) = 0$) the dither does not change the state of the filter any more.

(3) The dither should have same probabilities of being positive and negative.

Because the input sequence to quantizer is sometimes rounded up where the quantization error $\delta(n)$ has the opposite sign with the output value of the quantizer $\hat{Y}(n)$ and sometimes rounded down where $\delta(n)$ has the same sign with $\hat{Y}(n)$. In a statistical sense, the probabilities of round up and round down are equal, only if the dither also has same probabilities of being positive and negative the dither can possibly change the sequence to be rounded down from round up or vice versa.

(4) The dither should be a non-periodic sequence. This requirement is from the experience of the use of folding-frequency dither method described in the last section. Combining the above requirements of the dither signal one knows that the dither signal should be a random signal distributed in the open range $(-q/2, q/2)$. With the dither, the transition region in the equivalent quantization characteristic will be extended to the whole range of the input sequence of quantizer. It could be expected that all the limit cycles in the digital filters would be suppressed and there would be no remaining noise at

output in the zero-input condition. As will be described later, the simulations have verified this expectation.

6.3 SOME USEFUL DITHER SIGNALS

According to above requirements of the dither signal, three types of dither signal have been proposed and used in our research.

1. Uniformly Distributed Dither Signals

This dither is a pseudo-random sequence whose values are distributed uniformly in the open range $(-q/2, q/2)$. Fig. 28 shows its probability distribution function. From the probability distribution function, it is clear that the uniformly distributed dither signal has mean value of zero. Its variance can be readily calculated as follows:

$$\begin{aligned}\text{Variance} &= E[(d-m_d)^2] \\ &= E[d^2] \\ &= \frac{1}{q} \int_{-q/2}^{q/2} d^2 dx \\ &= q^2/12\end{aligned}\tag{91}$$

There are several methods to generate the uniformly distributed pseudo-random sequence. In this research the program shown in Appendix 5 is used to generate the uniformly distributed pseudo-random sequence by computer.

By the use of this program the computer gives a pseudo-random real number taken from a uniform distribution between 0 and 1. The subroutine uses a multiplicative congruential

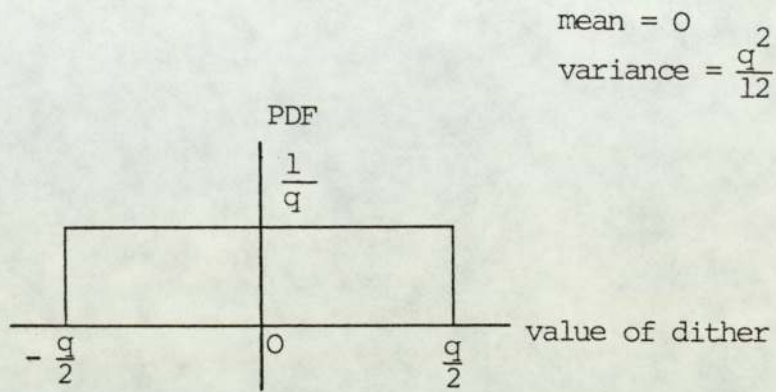


Fig. 28 The probability distribution function of the uniformly distributed random dither signal.

method.

$$P = 13^{13} \times P \bmod 2^{59} \quad (92)$$

The output sequence of the computer is equal to $P/2^{59}$ approximately. Here P is a variable whose value is preserved between calls of the subroutine. Its initial value is $123456789 \times (2^{32} + 1)$ but this may be altered. Because the output sequence P is distributed between 0 and 1, the sequence $(P - 0.5)$ is, of course, distributed between -0.5 and 0.5 . The sequence $(P - 0.5)$ is applied as the uniformly distributed dither. Fig. 29 shows its relative frequency histogram calculated from 1024 samples. As can be seen, principally, this sequence is uniformly distributed in the range $(-q/2, q/2)$.

Simulations have shown that the limit cycles in the second-order digital filter can be suppressed by the use of this dither. Once the limit cycle has been suppressed no remaining noise exists at the output terminal of the filter.

One disadvantage of this dither is that the time needed for the limit cycle suppression is long. For the example often used, this time is about $330 T_s$ where T_s is a sampling period. This is because the amplitude of the dither is distributed uniformly in the range $(-q/2, q/2)$. With relatively large probability the instantaneous value of dither is not big enough to change the values of the quantizer. Therefore, the transition of the filter state between limit cycles is not frequent. It can be

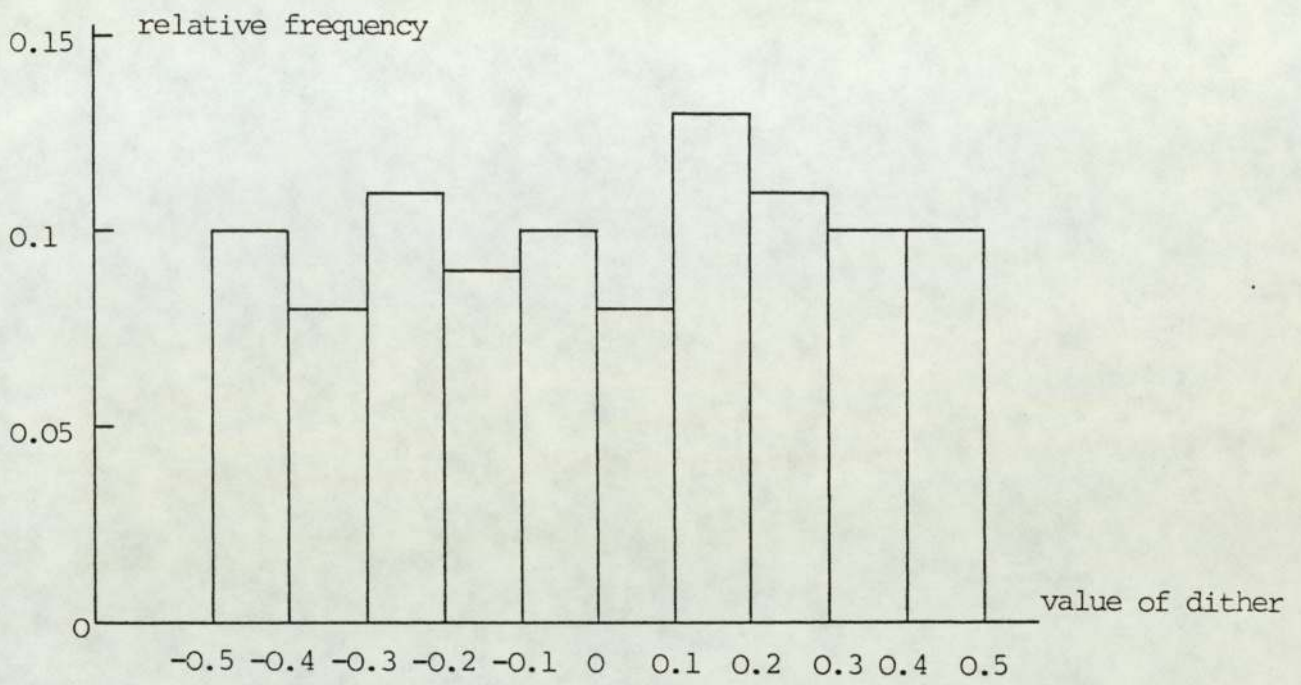


Fig. 29 The relative frequency histogram of the uniformly distributed random dither.

expected that if the probability when the absolute value of the dither tends to $0.5q$ is much larger than that when the dither is small, then the above disadvantage may be overcome. The extreme case leads to the second type of dither.

2. Binary Random Dither

As mentioned before, the advantage of the use of the folding-frequency dither is that the filter might reach the origin state in a shorter time due to the greater amplitude of the dither resulting in more frequent transitions between the various limit cycles. But because of the periodicity of this dither for some initial conditions the filter can never reach the origin but another limit cycle. Therefore, it would be hopeful to find a type of dither whose values keep at maximum, but are rid from the periodicity. Since the amplitude of the new dither should be fixed at the maximum value, the only way to get rid of the periodicity is changing the sign of the new dither randomly. This task can be reached by taking the sign of a random sequence (uniformly or Gaussian distributed) which distributes symmetrically between positive and negative values.

This new type of dither is called binary random dither which takes values $(-q/2 + \Delta)$ or $(q/2 - \Delta)$ with equal probability, where Δ is a positive quantity much smaller than q . We will discuss the value Δ later in this chapter.

The binary dither is readily generated from the uniformly distributed dither by the following equation.

$$D_b = \text{sign}(D_u) \cdot (q/2 - \Delta) \quad (93)$$

where D_b and D_u represent the binary and uniformly distributed dither respectively, and $\text{Sign}(x)$ represents the sign of x .

Fig. 30 shows its probability distribution. The mean value of the distribution is zero. The variance of this dither can be calculated as follows:

$$\begin{aligned} \text{Variance} &= E[(D_b - M_d)^2] \\ &= E[D_b^2] \\ &= \frac{1}{4} q^2 \end{aligned} \quad (94)$$

The variance of the binary random dither is three times bigger than that of the uniformly distributed dither. As could be expected, the simulation showed that when nonzero input signal was input, the noise output from the filter with binary random dither was also bigger than that with uniformly distributed dither. But as mentioned above, since the greater amplitudes of the dither result in more frequent transitions between the various limit cycles, the suppression of limit cycles with the binary random dither is faster than uniformly distributed dither. The increased noise at output is the penalty.

In summary, with the uniformly distributed dither the filter has a smaller noise output when a nonzero input

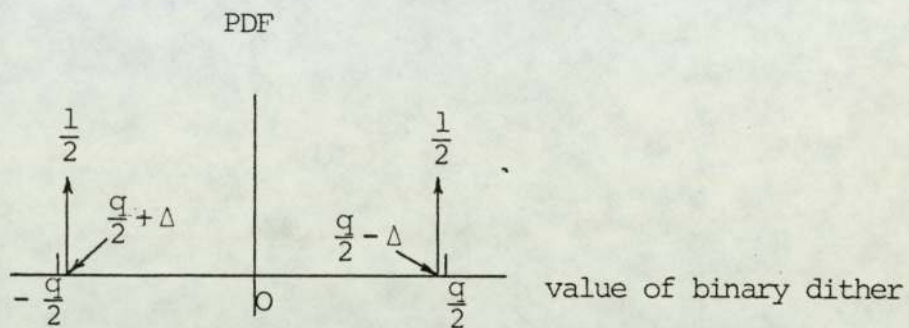


Fig. 30 Probability distribution function of the binary dither.

signal is added but it takes a longer time to suppress limit cycles. On the other hand, with the binary random dither, the limit cycles can be suppressed quicker, but the noise from the filter is bigger than that with the uniformly distributed dither, when the input signal is nonzero.

It is reasonable to ask the question that whether it is possible to generate a type of dither which has both the advantages, i.e., smaller output noise and quicker limit cycle suppression. This question leads to the third type of dither signal.

3. Band-stop Dither

When dither is used to stabilize a filter, two specifications are of particular interest. One of these is the increase in the output noise from the filter above the quantization noise without dither which is present when nonzero input signals are applied. The other specification of interest is the length of time taken for the filter, with zero input, to reach the state-plane origin from a limit cycle.

First, let us discuss the possibility of reducing the output noise caused by the dither.

As we have known that when the input signal is zero, after the transition time to the origin state $(0,0)$, the dither will not affect the states of the filter any more, i.e., once the limit cycles have been suppressed the output

from the filter stays at zero as long as the input is zero. There is no remaining noise at the output. On the other hand, when the input is nonzero, there is extra output noise caused by the dither. From Fig. 14, it is clear that the transfer function of the dither is the same with that of the input signal. Because of the filtering of the filter, only the frequency components of the dither which lie on the pass band can pass the filter, and all the components which lie on the stop band cannot pass the filter, or precisely speaking, for the later components the filter will give them a big attenuation. Hence, intuitively, we can expect that before the dither is added to the filter if we remove the frequency components of the dither which lie on the pass band of the filter, then it is possible to reduce the output noise caused by the dither. All the other components of the dither will be filtered out by the filter itself.

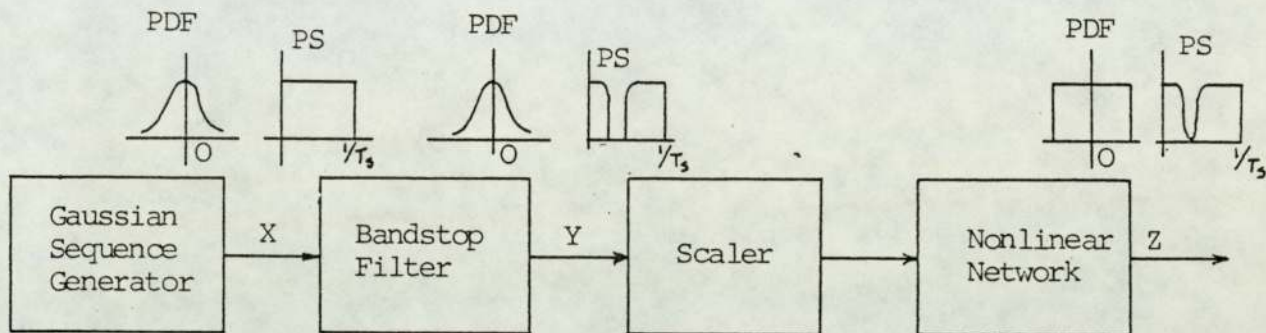
Second, let us discuss the possibility of reducing the transition time to the origin state in the state plane. For a second-order digital filter section, if the poles are close to the unit circle in the Z-plane, a periodic limit cycle is approximately sinusoidal with a frequency close to the resonant frequency of the corresponding section without the quantizer. When the poles are not close to the unit circle in the Z-plane, the periods of the limit cycles are divergent but, in general, they still ^{are} around the reciprocal of the resonant frequency of the section.

When the procedure of the limit cycle suppression on the screen of the computer PET was investigated, it was found that the periodic limit cycle sometimes tended to be suppressed, but sometimes tended to be enlarged at the frequency near the resonant frequency. This enlargement makes the transition time to the origin state be postponed. Therefore, it is helpful for shortening the transition time to remove the frequency components of the dither which lie around the resonant frequency. As discussed in Chapter 2, the second-order basic section has a bandpass characteristic and the central frequency of the pass band is the resonant frequency.

Fortunately, above two requirements, reducing the output noise and shortening the transition time to the origin, require the same thing, i.e. to remove the frequency components of the dither which lie around the resonant frequency of the section. At the same time, of course, the dither still has to satisfy the inequality $|d(n)| < \frac{1}{2}q$.

We have discussed two kinds of dither, uniformly distributed dither and binary random dither. It is difficult, if it is not impossible, to obtain a binary dither signal which has no components around the resonant frequency of the filter. Now, what we want to do is to find a type of dither signal which has uniform probability distribution in the range $(-q/2, q/2)$ and the frequency components lain around the resonant frequency have been removed.

Veltman, Van den Bos, de Bruine, de Ruiter and Verloren⁽⁴⁰⁾ have proposed a synthesis method of random signals with a prescribed amplitude probability density function and a prescribed power density spectrum. But for our purpose, a simpler method can be used. In Fig. 31, the input sequence X is assumed to be a wide band Gaussian noise. The sequence X is first added to a linear bandstop filter so as to remove the frequency components lain on the stop band whose centre is the resonant frequency of the filter to be stabilised. The output of the bandstop filter, Y, has Gaussian property too. Then the sequence Y passes a memoryless nonlinear network. This nonlinear characteristic transforms the Gaussian probability density function into a uniform distribution one. Because the nonlinear transformation will result in intermodulation products of the bandstop filtered dither falling within the stop band, this transformation also changes the power density spectrum of its input signal. Although Veltman et al. have described the method how to calculate the influence of the nonlinear network and compensate the linear filter so that the output of the nonlinearity produces the described frequency behaviours, but the simulation on a digital computer showed that in our case, the nonlinearity gave minor changes in the power density spectrum. The simulations have shown that for our purpose, the output of the nonlinear network has been



PDF : Probability Density Function

PS : Power Spectrum

Fig. 31 Block diagram of the bandstop dither generation.

a good dither signal already. The bandstop dither generating will be described in detail in Chapter 7.

As will be seen in the next chapter, the experiments have verified that the bandstop dither has the property of suppressing limit cycles quickly without causing much increase in output noise.

6.4 A NOTE ON THE DITHER SIGNALS

There are two questions about the dither signals that we have not yet discussed. The first one is how many bits are needed for storing the dither signal. For example the uniformly distributed random dither is distributed in the open range $(-q/2, q/2)$. The question is how many places of the dither value are needed for the limit cycle suppressing.

In general, the more places the dither has, the more bits it needs. The second question is how to decide two values, $(-q/2+\Delta)$ and $(q/2-\Delta)$, in the binary random dither signal design. These two questions will be discussed in this section. As mentioned earlier, one of the important considerations in the dither signal design is that the amplitude of the dither should be big enough to change the output of the quantizer. This requires the dither signal to satisfy the inequality

$$\left[|\delta(n) + d(n)| \right]_R \geq 1$$

or

$$|\delta(n) + d(n)| \geq 0.5$$

As pointed out earlier, for a periodic limit cycle, the quantization error $\delta(n)$ is also periodic and varies between $-q/2$ and $q/2$. In general, there is at least one of $\delta(n)$ in a period which is bigger than, for example say, $0.25q$. For the example often used, the maximum of $\delta(n)$ corresponding to the largest periodic limit cycle is equal to $0.48497q$. Therefore, generally speaking, the suppression of periodic limit cycles only requires small amplitude dither which needs 2 or 3 bits because there is at least one state of the periodic limit cycle which can be changed by the use of small amplitude dither. But it is clear that if one wants to shorten the transition time to the origin then big amplitude is still desirable so as to change as many states on the limit cycle as possible. In our research for insurance we choose that $|\bar{d}(n)|_{\max} = 0.499$, which needs 9 bits for storing. But it does not mean that this is necessary. In some cases, most critical situations occur when the constant or alternating limit cycles exist.

Suppose that without dither a constant limit cycle with value C exists, i.e.,

$$\begin{aligned} C &= [-AC - BC]_R \\ &= C(-A - B) - \delta \end{aligned} \quad (95)$$

where δ is the quantization error which is a constant when a constant limit cycle exists.

From the linear stable condition (the triangle in

the parameter space) the coefficient values A and B satisfy the inequality.

$$-A - B < 1$$

Let $-A - B = 1 - \beta$ (96)

where

$$\beta > 0$$

Substitute Eq. (96) into Eq. (95). We obtain

$$C = C(1 - \beta) - \delta$$

or $\delta = -C\beta = -C(1 + A + B)$ (97)

Equation (97) shows that the quantization error has a opposite sign as that of C because $\beta > 0$. And $|\delta|$ decreases with $|C|$ proportionally. Apparently, the minimum of $|C|$ is equal to unity. Therefore, the minimum of the quantization error, δ , satisfies

$$|\delta|_{\min} = 1 + A + B \quad (98)$$

when constant limit cycle exists.

Along the same way, when the alternating limit cycles exist the following inequality can be obtained

$$|\delta|_{\min} = 1 - A + B \quad (99)$$

Now it is clear from the $|\delta|_{\min}$ expression that in order to suppress all constant or alternating limit cycles the minimum amplitude of the dither should satisfy the inequality

$$|1 - |A| + B + d(n)| \geq 0.5$$

or

$$|d(n)| \geq 0.5 - (1 - |A| + B) \quad (100)$$

The inequality (100) whose variables are the coefficient values A and B determines the number of bits needed for the dither signal. As a very special example, suppose the value $(1 - |A| + B) < 0.01$, then it is clear that $|\hat{d}(n)|_{\max} = 0.499$ is still not big enough. We have to increase the places of the dither so as to suppress the constant or alternating limit cycles whose amplitude is lq . Apparently, in this case, the time needed to suppress this constant limit cycle or alternating limit cycle by the uniformly distributed dither will be long. For the binary random dither, $(-q/2 + \Delta)$ or $(q/2 - \Delta)$, Δ should satisfy the inequality

$$0 < \Delta < 1 - |A| + B \quad (101)$$

Similarly, only if the value $(1 - |A| + B) < 0.01$ $|\hat{d}(n)| = 0.499$ is not big enough and the increase of the bits needed for storing the dither is necessary.

For the example often used in the thesis, $A = -1.74$, $B = 0.95833$, Δ should satisfy that

$$\Delta < 1 - 1.74 + 0.95833 = 0.21833$$

In other words, the amplitude of the binary dither should be greater than 0.28167 which is much less than 0.499. In this case, the amplitude of 0.499 is big enough. In fact in this example, it is possible to reduce the places of the dither, i.e., reduce the bits needed for dither storing. Simulation has verified this binary random dither.

6.5 SUMMARY

In this chapter, the previous works on the use of dither for limit cycle suppression have first been reviewed and discussed. Then the principal considerations in the dither signal design have been described. The dither signal should be a random signal distributed in the open range $(-q/2, q/2)$. Three types of dither signal have been derived from the principal considerations of the dither signal design. They are uniformly distributed random dither, binary random dither and bandstop dither. All the limit cycles in the second-order digital filters can be suppressed by the use of these three dither signals and there is no remaining noise at the output terminal when the input signal is zero. Each dither has its own characteristics. The simplest one for generating is the uniformly distributed dither. But it takes longer time for the limit cycle suppression by the use of it. When the binary random dither is used, the transition time to the origin on the state plane is shortest but the penalty is the increase of the output noise with nonzero input signal. It seems that the best dither for limit cycle suppression is the bandstop dither which has the advantage of suppressing limit cycles quickly without causing much increase in output noise. The penalty for the advantage is the complexity in the dither generating. But in systems already using many identical digital filters, the extra complexity need not be significant.

It is possible to generate other dither signals from the principal consideration in the dither signal design. But it seems that these three types of dither may have satisfied different requirements already.

The process of generating bandstop dither needs further study. In particular, it could be possible to use bandstop dither signals which have been synthesised by computer and stored in ROM. Further studies are needed on the best form for the spectrum of the bandstop dither. In particular, it is not known how deep should be the notch in the spectrum. Any further improvements would probably be small.

It is very hopeful if one can find a method to generate the bandstopped binary dither signal. It is expected that the transition time to the origin state and the increase of the output noise with nonzero input signal would be reduced further by the injection of the bandstopped binary dither.

Finally, in Section 6.4, it is concluded that the number of bits needed for storing the dither signal and the values in the binary random dither depend on the coefficient values A and B.

In the next chapter, the experimental results will be presented.

CHAPTER 7

EXPERIMENTAL RESULTS

In order to verify the proposed methods for limit cycle suppression extensive simulations have been carried out with computers using the three types of dither signal mentioned above. In this research, three different computers have been used for different purposes; PET, PDP-11 and HARRIS 500. But most simulations were implemented with the HARRIS 500. Fortran language was mainly used, but basic was also applied occasionally.

In the research, three kinds of experiments have been done. In the first kind of experiments, some were for verifying the limit cycle theory such as the existing conditions, classifications, amplitude bounds and frequency expression etc. The other simulations were for verifying, roughly, the methods of limit cycle suppression proposed in this thesis. The purpose of the second kind of experiments was to obtain the outline of the time needed for the limit cycle suppression by the use of the dither. At the same time, the use of dither for stabilizing the digital filter was checked further. The emphasis of the third kind of experiments was to investigate the effect of dither on the output noise from the filter section.

In the following sections the three kinds of experiment

will be presented respectively.

7.1 GENERAL SIMULATIONS ON THE LIMIT CYCLES IN THE SECOND-ORDER DIGITAL FILTER SECTIONS

1. Second-Order Filter Section With Quantization

As mentioned in the second chapter, there are two different ways of implementing the quantizations in the second-order sections: one quantizer and two quantizer versions. In the zero input condition, these two versions are described respectively, by the difference equations

$$\hat{Y}(n) = [-A\hat{Y}(n-1) - B\hat{Y}(n-2)]_Q \quad (102)$$

and

$$\hat{Y}(n) = [-A\hat{Y}(n-1)]_Q + [-B\hat{Y}(n-2)]_Q \quad (103)$$

where $[\cdot]_Q$ represents the operation of quantization.

The roundoff quantization, $[X]_R$, can be simulated by the following arithmetic

$$[X]_R = \text{Sign}(X) \times \text{integer} [\text{absolute}(X) + 0.5] \quad (104)$$

where $\text{sign}(X)$, $\text{integer}(X)$ and $\text{absolute}(X)$ represent the sign of X , integer of X and the absolute value of X respectively.

Using fortran language, the expression can be written as

$$\text{SIGN}(1.0, X) * \text{INT}(\text{ABS}(X) + 0.5) \quad (105)$$

Using basic language, above expression can be written as

$$\text{SGN}(X) * \text{INT}(\text{ABS}(X) + 0.5) \quad (106)$$

The magnitude-truncating quantization, $[X]_{\text{MT}}$, can be simulated by the arithmetic

$$[X]_{\text{MT}} = \text{sign}(X) \cdot \text{integer}[\text{absolute}(X)] \quad (107)$$

Similarly, using fortran language the above expression can be written as

$$\text{SIGN}(1.0, X) * \text{INT}(\text{ABS}(X)) \quad (108)$$

Using basic language, the corresponding statement is

$$\text{SGN}(X) * \text{INT}(\text{ABS}(X)) \quad (109)$$

By the use of the above statements and the corresponding difference equations, the second-order filter sections can be readily simulated with the computer.

2. Limit Cycles in the Second-Order Filter Sections

In order to find out the limit cycles in various filters, first, we must determine the amplitude bound of the limit cycles. Usually, Jackson's bound expression is used because it is easy to calculate and it gives the minimum bound in the all bound expressions. As mentioned earlier, checking is necessary because this bound may be

exceeded in some special cases. Suppose this bound is K , then there are $(2K+1)^2$ states in the region defined by the bound in the state plane including the origin state itself.

Secondly, each of $(2K+1)^2$ states is assumed as an initial state of the second-order filter section in turn. The filter is operated until a limit cycle or the origin state is reached. The limit cycles can be readily recognised from the periodicity. It is clear from difference equation that for a filter if $\hat{Y}(n-1)$ and $\hat{Y}(n-2)$ change sign simultaneously then $\hat{Y}(n)$ also only change the sign. Therefore, if the initial state $(\hat{Y}(n-1), \hat{Y}(n-2))$ belongs to a Type A periodic limit cycle (see Section 3.2), then the initial state $(-\hat{Y}(n-1), -\hat{Y}(n-2))$ also belongs to the same limit cycle since Type A limit cycle has half-wave symmetry, i.e., the equation $\hat{Y}(n + \frac{N}{2}) = -\hat{Y}(n)$ exists.

Because the rounding and magnitude-truncation characteristics are described by odd functions, consequently, if there is a limit cycle of Type B in the filter, $\hat{Y}(1), \hat{Y}(2), \dots, \hat{Y}(N)$, then a limit cycle of Type B can also be maintained in the same filter described by $-\hat{Y}(1), -\hat{Y}(2), \dots, -\hat{Y}(N)$, which is different from the former. If the former is designated by Type B^+ and the later by B^- , then a similar conclusion can be obtained that if the initial state $(\hat{Y}(n-1), \hat{Y}(n-2))$ belongs to a Type B^+ limit cycle, then the initial state $(-\hat{Y}(n-1), -\hat{Y}(n-2))$ belongs to the Type B^- limit cycle.

These two conclusions can be used to save the time for searching limit cycles to half, because only half initial states of $(2K+1)^2$ need to be used in the simulations. The limit cycles in the example often used have been listed in Chapter 2. Fig. 21 and Fig. 32 show which limit cycle each initial state belongs to in one- and two-quantizer versions respectively.

3. Limit Cycle Displaying

By the use of the program in Appendix 6, the limit cycles can be displayed on the screen of computer PET. For a clear display, the initial states on the different limit cycles have to be chosen. In the typical example, there are 10 different limit cycles, therefore, we have to input successively 10 initial states which are respectively on the 10 limit cycles. After a certain time, all the 10 limit cycles can be displayed on the screen. By the use of the program in Appendix 7, the limit cycles can be printed out with computer and printer. Before running the program, the sequences which limit cycles consist of have to be input. The print of the limit cycles in the typical example is also shown in Appendix 7.

7.2 GENERATIONS OF THREE TYPES OF DITHER SIGNAL IN SIMULATION

In the simulation, three types of dither signal described in Chapter 6 have been used; uniformly distributed

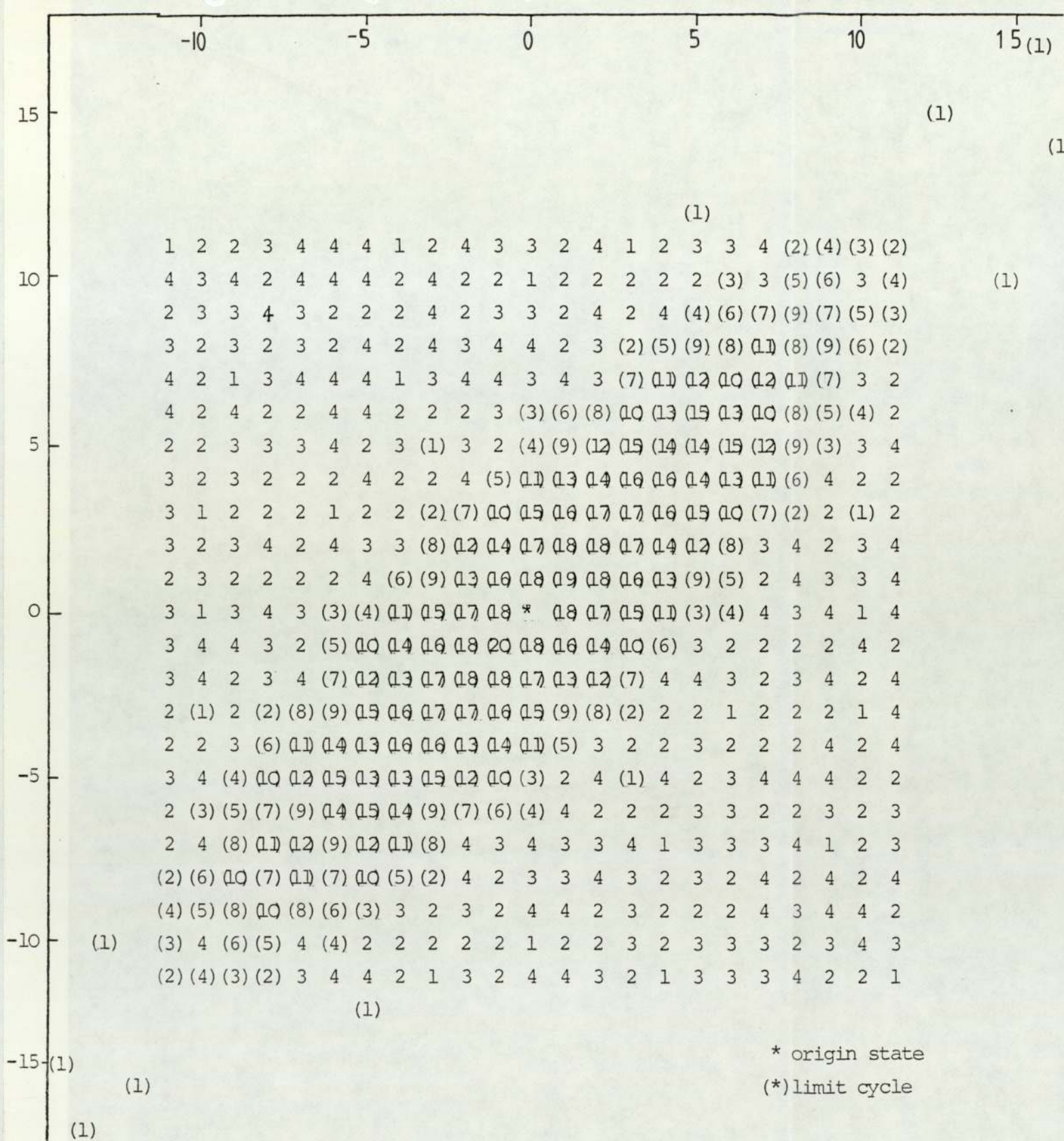


Fig. 32 The state plane plot shows which limit cycle each state belongs to for the filter section (A=-1.74, B=0.95833) with two quantizers.

random dither, binary random dither and bandstop dither. In this section, the methods how to generate these three types of dither signal will be first described and then the experimental results will be presented.

1. Uniformly Distributed Random Dither

When basic language is used a random sequence, RND, which is uniformly distributed between 0 and 1 can be delivered by the computer PET or HARRIS 500. In this case, the random sequence (RND-0.5) is used as uniformly distributed random dither.

For computer HARRIS 500, when fortran language is used a random sequence which is also uniformly distributed between 0 and 1 can be delivered by calling a subroutine, GO5CAF, in the NAG library. Appendix 5 shows the subroutine and how to call it by an example.

2. Binary Random Dither

Because in the uniformly distributed dither signal both the sign and the magnitude are random, the sign information of the uniform dither can be used as the sign of the binary dither. The problem about how to determine the two values of the binary random dither has been discussed in Chapter 6.

Therefore, in order to generate the binary random dither,

first, the uniformly distributed random dither is generated, then let the sign of the binary random dither be the same with the sign of uniform dither. For the sake of simplicity, the amplitude of the binary dither is assumed equal to 0.499.

When basic language is used, the corresponding statements for the binary dither generating are as follows:

$$DU = RND - 0.5 \quad (110)$$

$$DB = SGN(DU) * 0.499 \quad (111)$$

When fortran language is used, the corresponding statements for the binary random dither generating can be written as

$$DU = GO5CAF(X) - 0.5 \quad (112)$$

$$DB = SIGN(1.0, DU) * 0.499 \quad (113)$$

3. Bandstop Dither

The block diagram for generating the bandstop dither has been shown in Fig. 31 before. Each block in this diagram will be described in detail as follows:

(A) The Generator of Input Gaussian Random Sequence

As mentioned in Chapter 6, the input sequence X in Fig. 31 should be a wide band Gaussian noise with a flat spectrum. In the simulation, this Gaussian random sequence

came from a subroutine in the NAG library of computer centre. The computer used was a HARRIS 500. The subroutine of generating the Gaussian noise X is shown in Appendix 8.

The standard deviation and mean value of the Gaussian random sequence used in the simulation were 1 and 0 respectively. As an example, some numbers of the Gaussian sequence is also listed in Appendix 8. Fig. 33 and Fig. 34 show the power spectrum and probability density function respectively. As can be seen, above requirements on the input sequence X are essentially satisfied.

(B) The Linear Bandstop Filter

Because the second-order basic filter section has a bandpass characteristic, the linear filter should be a bandstop filter. The centre of the stop band should be equal to the centre of the pass band of the filter to be stabilized.

In the experiments, two types of bandstop filter have been used: Butterworth and Elliptic. The bandstop filters were designed by the computer.

The Butterworth bandstop digital filter design program requires three input data: the order of lowpass prototype, lower cutoff frequency and upper cutoff frequency. The order of lowpass prototype is equal to half the order of bandstop filter. Two cutoff frequencies are in unit of

Power Density Spectrum

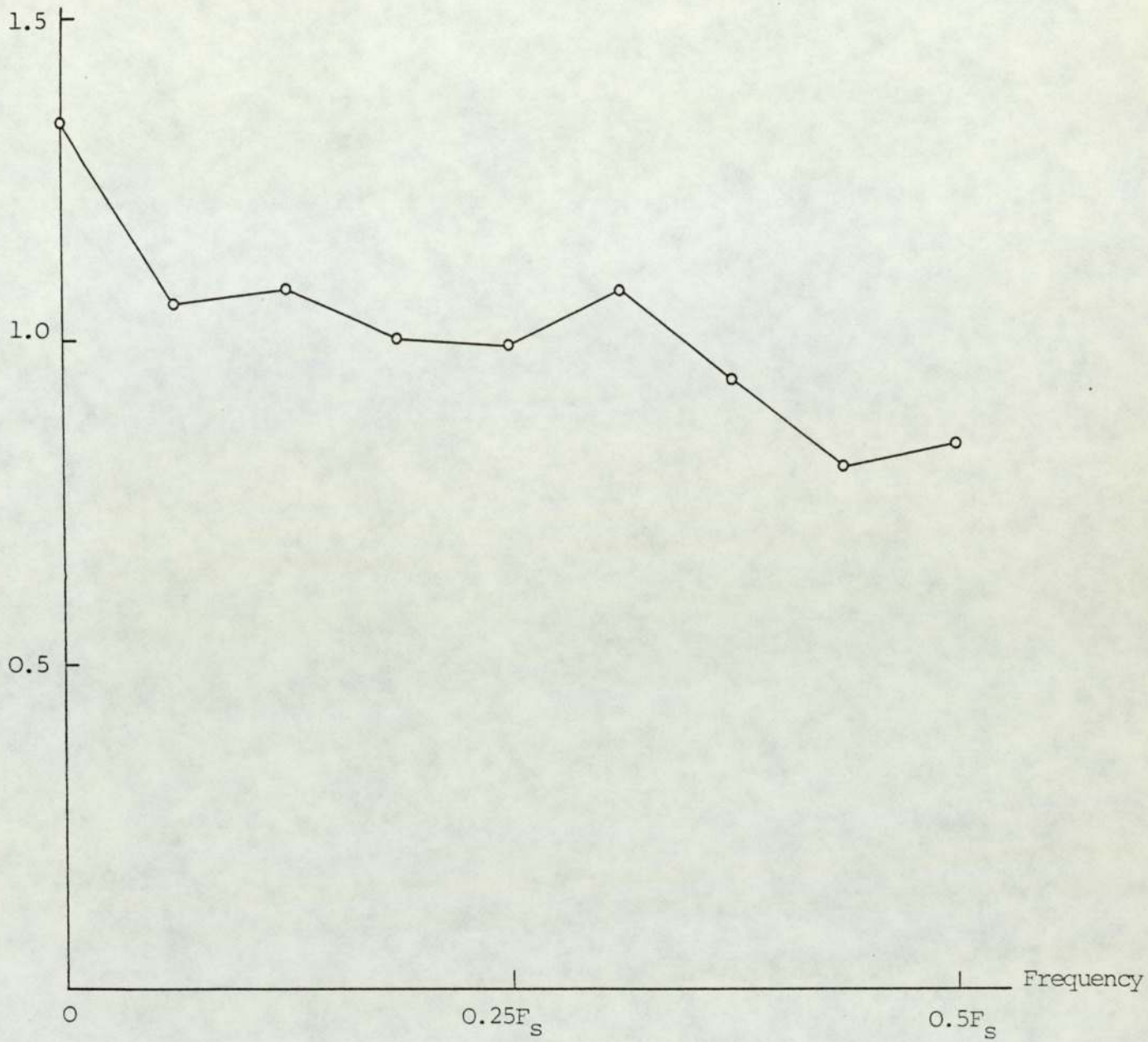


Fig. 33 The power density spectrum of Gaussian random sequence used in the research. The FFT with Hamming Window was repeated 127 times. The average values are shown here.

Probability density function

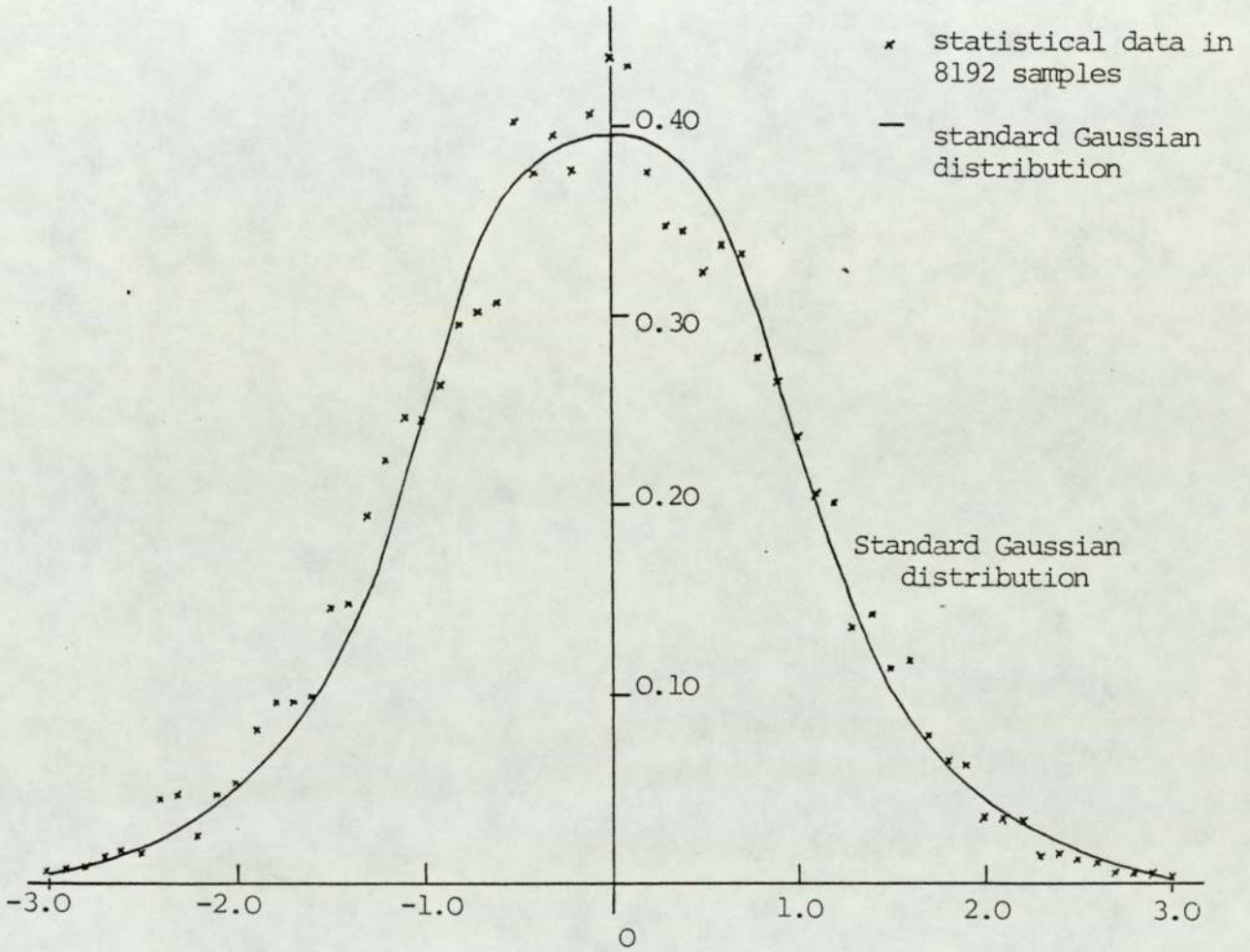


Fig. 34 The probability density function of the Gaussian random sequence generated by program.

Nyquist frequency. After the input data have been typed in, the computer will give the stopband width, the zeros and poles of lowpass prototype, the zeros and poles of bandstop filter and the coefficient values of the second-order sections. This program can give the frequency response of the designed filter if the operator requests.

The program for elliptic bandstop digital filter design has been shown in the reference (41). The parameters entered into the program are the filter order, the dB ripple in the passband, the sampling frequency, the stopband edge frequencies and a stopband attenuation in units of dB. The program requests an input by typing the line

N, DBR, FS, F1, F2, F3 OR DBDOWN

The variable N is the order of the filter in the s-plane. A zero or negative value entered for N terminates the program. For bandstop filters, the filter order in the Z-plane will equal to 2N. If the input data are entered with $0 < F1 < F2 < \frac{FS}{2}$, a bandpass filter is defined. If the input data are entered with $0 < F2 < F1 < \frac{FS}{2}$, a bandstop filter is defined. The final data entry is F3 OR DBDOWN. If it is positive and lies within the stopband, the entry defines one stopband edge frequency. If the entry is negative, it is used as the stopband attenuation. In the program output the denominator and numerator coefficients are listed for

the transfer function $T(Z)$ where

$$T(Z) = \frac{\sum_{i=1}^{M+1} p(i)Z^{1-i}}{\sum_{i=1}^{M+1} A(i)Z^{1-i}} \quad (114)$$

(C) The Nonlinear Memoryless Network

The purpose of the nonlinear network is to transform a Gaussian distributed sequence Y , into a uniformly distributed sequence, Z . The calculation of the nonlinear curve is straight forward.

$$dZ = \frac{p(Y)}{p(Z)} dY \quad (115)$$

with $p(Y)$ and $p(Z)$ respectively the Gaussian and the uniform probability density. Again, the quantization step, q , is assumed to equal to unity.

For Gaussian sequence

$$p(Y) = (\sigma\sqrt{2\pi})^{-1} \exp\left[-\frac{(Y-\mu)^2}{2\sigma^2}\right] \quad (-\infty < Y < \infty) \quad (116)$$

where the mean value μ is equal to the mean value of input and is equal to zero. σ is the standard deviation of Y which is not equal to the deviation of X , i.e., the σ of Y is not equal to unity any more. Therefore, in order to get a unity standard deviation, we have to scale the

sequence before it is input to the nonlinear network.

For uniformly distributed random output Z ,

$$p(Z) = 1 \quad (-0.5 < Z < 0.5) \quad (117)$$

Hence, the nonlinear gain can be written as

$$\begin{aligned} \frac{dZ}{dY} &= \frac{p(Y)}{p(Z)} = (\sigma\sqrt{2\pi})^{-1} \exp\left[-\frac{(Y-\mu)^2}{2\sigma^2}\right] \\ &= (\sigma\sqrt{2\pi})^{-1} \exp\left(-\frac{Y^2}{2\sigma^2}\right) \end{aligned} \quad (118)$$

or

$$dZ = (\sigma\sqrt{2\pi})^{-1} \exp\left(-\frac{Y^2}{2\sigma^2}\right) dY \quad (119)$$

Let $Y = \sigma Y_1$ where Y_1 has a unity standard deviation, or

$$dY = \sigma dY_1 \quad (120)$$

Therefore,

$$\begin{aligned} dZ &= (\sigma\sqrt{2\pi})^{-1} \exp\left(-\frac{Y_1^2}{2}\right) \sigma dY_1 \\ &= (\sqrt{2\pi})^{-1} \exp\left(-\frac{Y_1^2}{2}\right) dY_1 \end{aligned} \quad (121)$$

$$\begin{aligned} Z &= \int (\sqrt{2\pi})^{-1} \exp\left(-\frac{Y_1^2}{2}\right) dY_1 \\ &= \frac{1}{2} \operatorname{erf}\left(\frac{Y_1}{\sqrt{2}}\right) + C \end{aligned} \quad (122)$$

where C is constant.

When $-\infty < Y_1 < \infty$ the function $\frac{1}{2} \operatorname{erf}\left(\frac{Y_1}{\sqrt{2}}\right)$ distributes in the range $(-0.5, 0.5)$, it is just the function that we want. It is interesting that the function $\frac{1}{2} \operatorname{erf}(Y_1)$ itself is also in the range $(-0.5, 0.5)$. As will be discussed, for our purpose, it seems that the nonlinear function $\frac{1}{2} \operatorname{erf}(Y_1)$ is preferable to $\frac{1}{2} \operatorname{erf}\left(\frac{Y_1}{\sqrt{2}}\right)$. In the following sections, when the function $\frac{1}{2} \operatorname{erf}(Y_1)$ is used the dither is called bandstop dither 1 and when the function $\frac{1}{2} \operatorname{erf}\left(\frac{Y_1}{\sqrt{2}}\right)$ is used the dither is called bandstop dither 2.

The error function $\frac{1}{2} \operatorname{erf}(Y_1)$ can be expressed by series form⁽⁴²⁾

$$Z = \frac{1}{\sqrt{\pi}} \sum_{J=0}^{\infty} \frac{(-1)^J Y_1^{2J+1}}{J!(2J+1)} \quad (123)$$

Eqn. (123) is convenient to be simulated by computer. Fig. 35 shows the simulation characteristic of this nonlinear network. In the simulation, J is taken from 0 to 32. As can be seen from Fig. 35, the nonlinear network just like a suppressor which suppresses the Gaussian random signal whose values are distributed in ⁱⁿ⁻finite range into a uniform random sequence whose values are distributed in the range $(-\frac{q}{2}, \frac{q}{2})$.

(D) Scaler

In the simulation because J in Eqn. (123) is finite and taken from 0 to 32, when the value of Y_1 becomes bigger

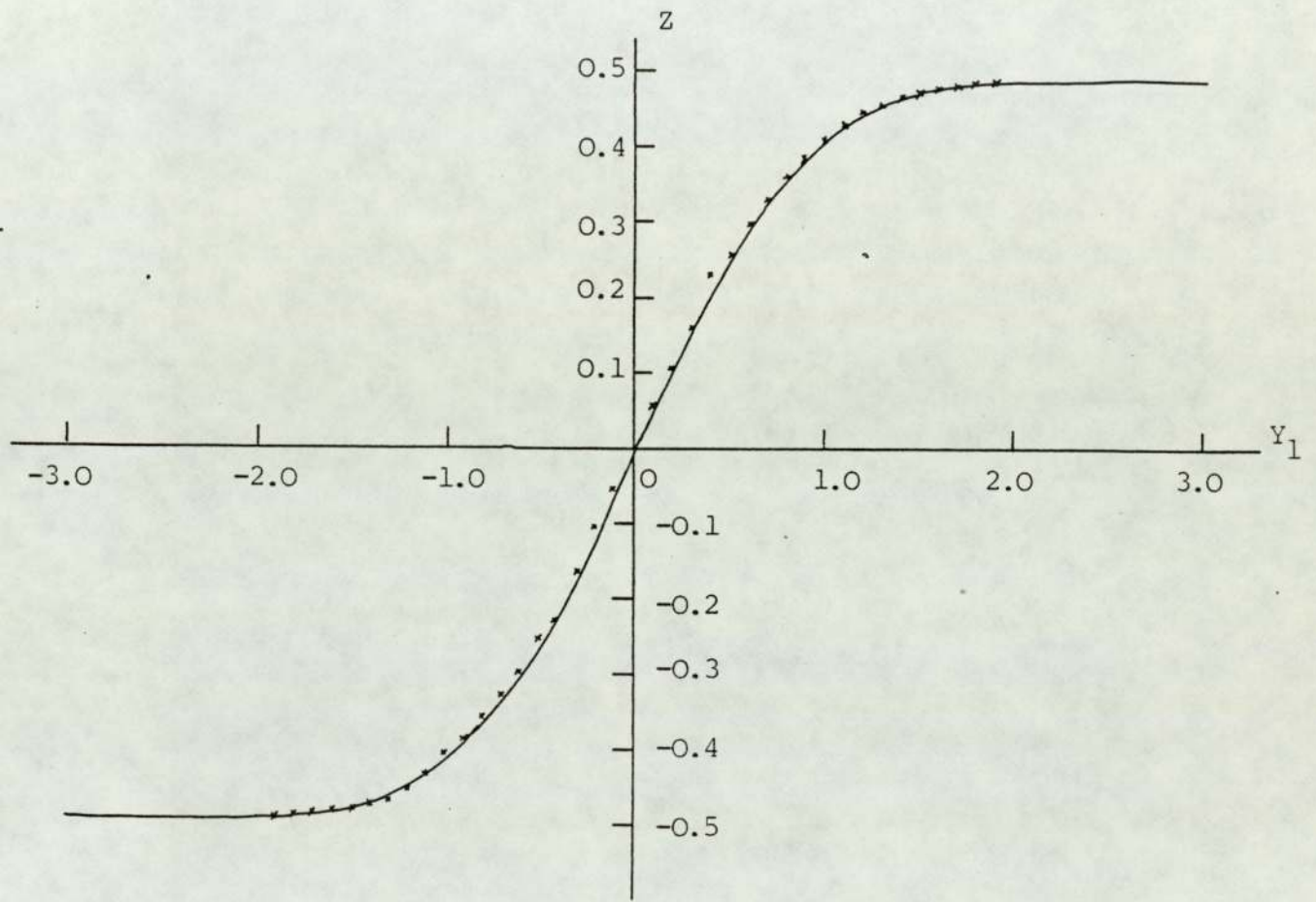


Fig. 35 The characteristic of nonlinear memoryless network in the bandstop dither 1 generator.

than 4, the error of the simulation becomes too big. Hence, it is necessary to limit the value of Y_1 . This can be done by scaling before the sequence input to the nonlinear network. In the simulation, the reciprocal of the standard deviation is used as a scaling factor. Therefore Y_1 has a unity standard deviation. In this case, the probability of $|Y_1|$ exceeding 3 is very small.

Now, referring to Fig. 31, it is clear that when the Gaussian random sequence with zero mean and unity standard deviation whose power spectrum is approximately flat is applied to the bandstop filter, the output from the bandstop filter is still Gaussian but the frequency components falling within the stop band have been attenuated. The stop band is greater than the pass band of the filter section to be stabilized and both have the same centre frequency, the components whose frequencies fall within the pass band of the basic section in the output sequence from the bandstop filter have been attenuated seriously. The purpose of the scaler is transforming the non-unity standard deviation input sequence into a unity standard deviation sequence so as to reduce the distribution range of the output sequence. This is requested by the limited dynamic range of the nonlinear network. The output from the nonlinear network is the bandstop dither which distributes approximately uniformly in the range $(-\frac{q}{2}, \frac{q}{2})$ and has very small frequency components falling within

the pass band of the filter section to be stabilized.

Fig. 36 and Fig. 37 show the examples of the probability distribution of the bandstop dither 1 and 2 respectively. As can be expected, the probability distribution of bandstop dither 2 is flat in the range $(-\frac{g}{2}, \frac{g}{2})$ because the function $\frac{1}{2} \operatorname{erf}(\frac{Y_1}{\sqrt{2}})$ satisfies Eqn. (122). It can be seen from Fig. 36, although the probability distribution of bandstop dither 1 is approximately flat but there are two peaks around the values -0.5 and 0.5. As mentioned before, these greater probabilities of big dither magnitude may shorten the time needed for the transition from any limit cycle to the origin state on the state plane due to the bigger amplitude of the dither resulting in more frequent transitions. Hence from reducing the transition time point of view, bandstop dither 1 is preferable to the bandstop dither 2.

On the other hand, because the function $\frac{1}{2} \operatorname{erf}(Y_1)$ causes stronger nonlinear than function $\frac{1}{2} \operatorname{erf}(\frac{Y_1}{\sqrt{2}})$ it must lead to more distortion in the power density spectrum of the bandstop dither. Fig. 38 and Fig. 39 show, respectively, the power density spectrums of the bandstop dither 1 and bandstop dither 2. It is clear from these two figures that the stop band attenuation of the bandstop dither 1 is indeed less than that of the bandstop dither 2. But this decrease of stop band attenuation is small, (about 2 dB in these examples).

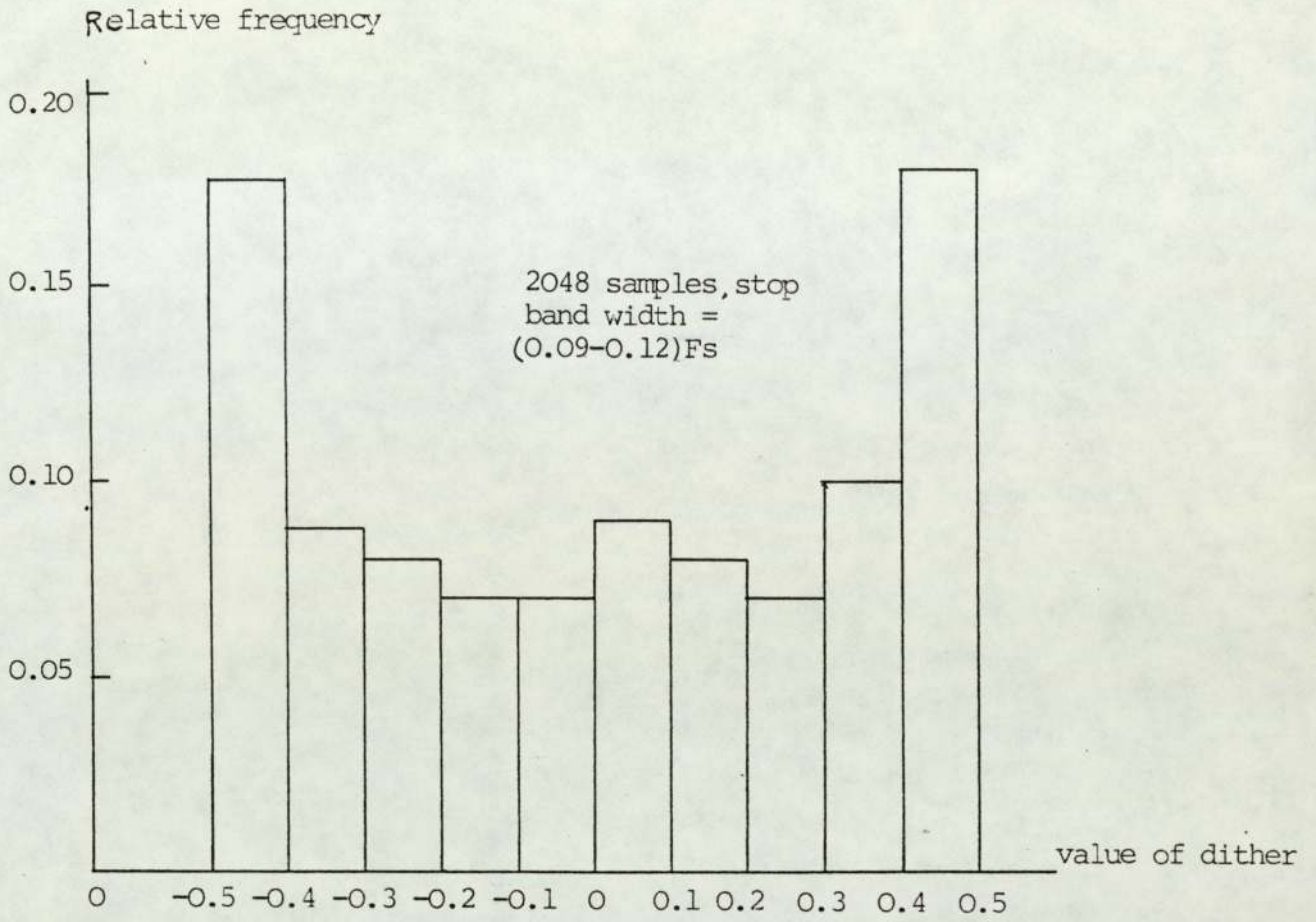


Fig. 36 The relative frequency histogram of the bandstop dither 1. The nonlinear network whose characteristic is $\frac{1}{2}\text{erf}(Y_1)$ has been used.

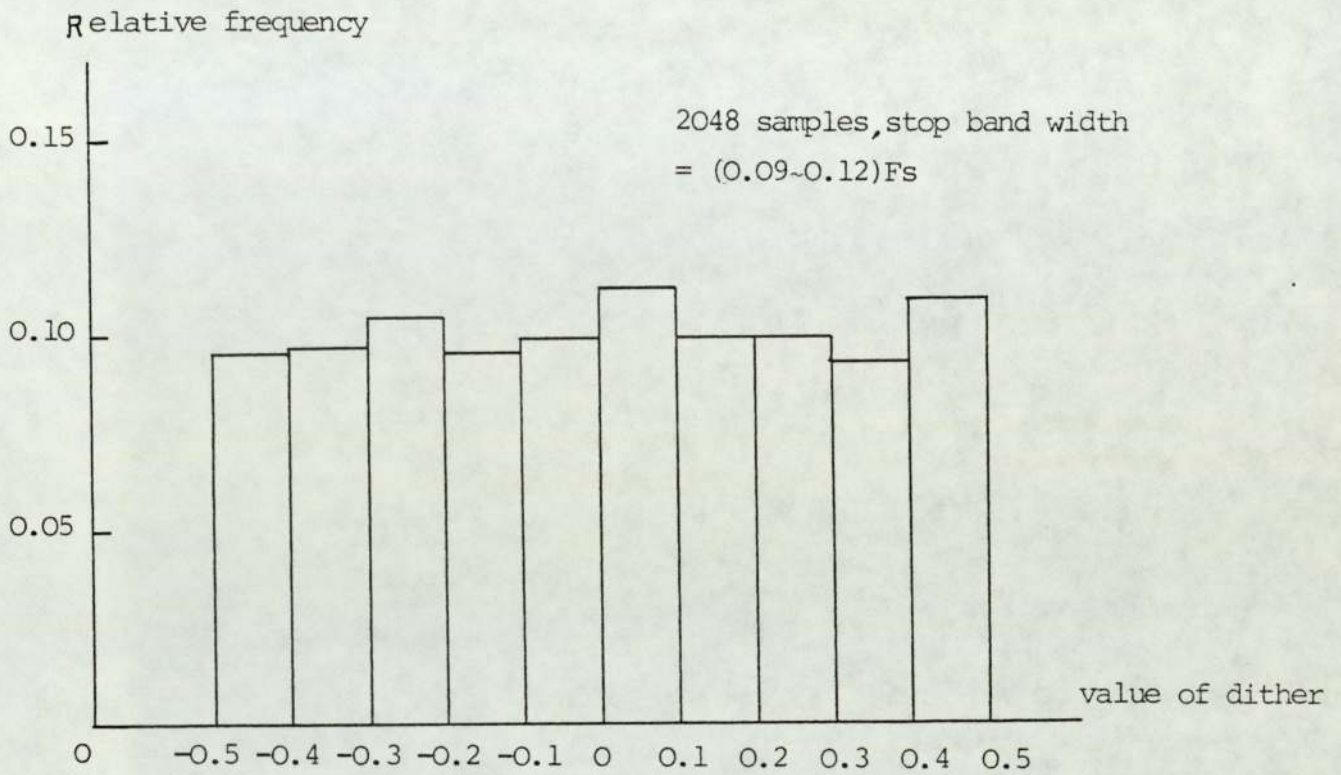


Fig. 37 The relative frequency histogram of the bandstop dither 2. The nonlinear network whose characteristic is $\frac{1}{2}\text{erf}\left(\frac{Y_1}{\sqrt{2}}\right)$ has been used.

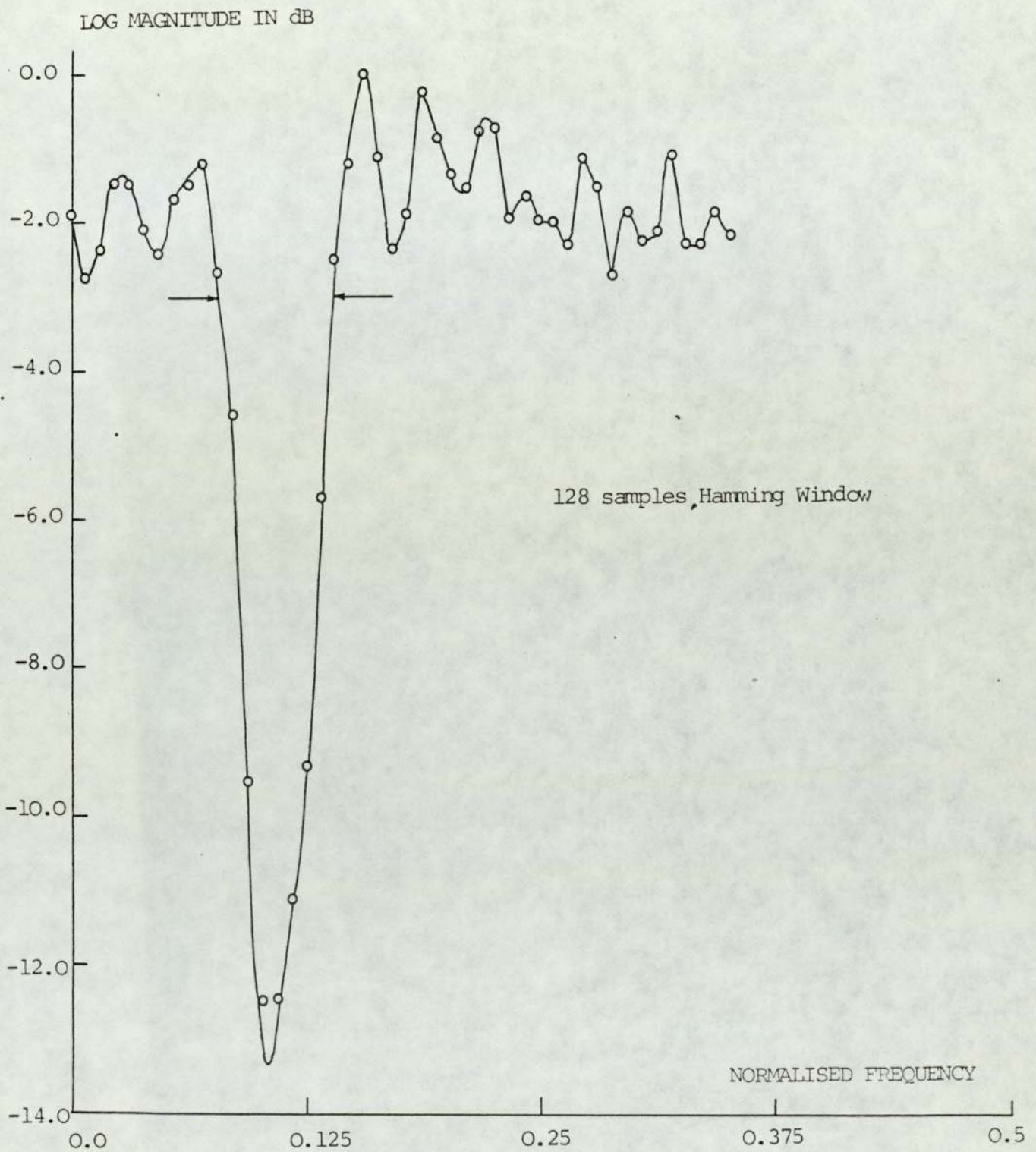


Fig. 38 The power density spectrum of the bandstop dither 1 whose stop band width was equal to $(0.08 \sim 0.14)F_s$. The nonlinear function $\frac{1}{2}\text{erf}(Y_1)$ was used.

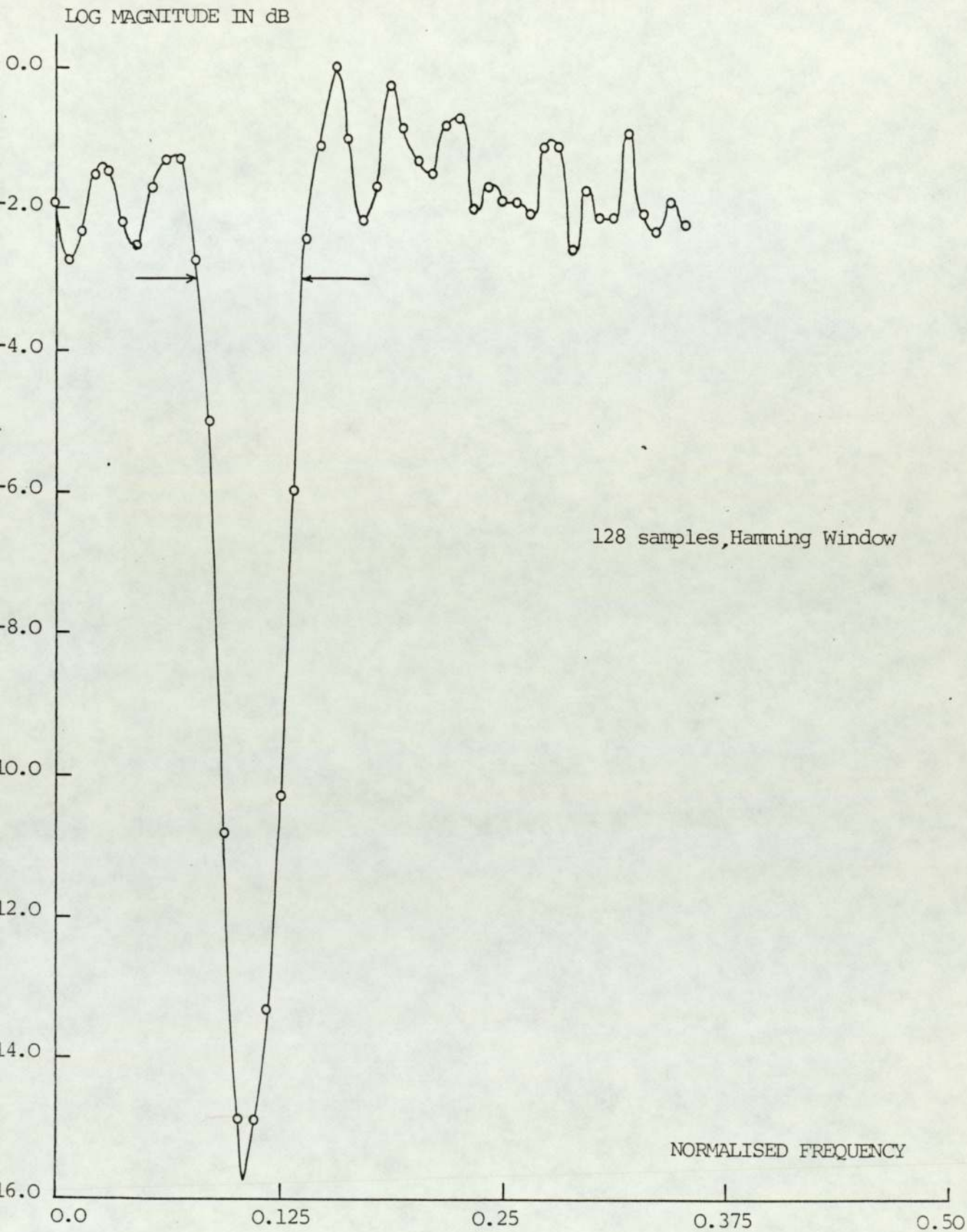
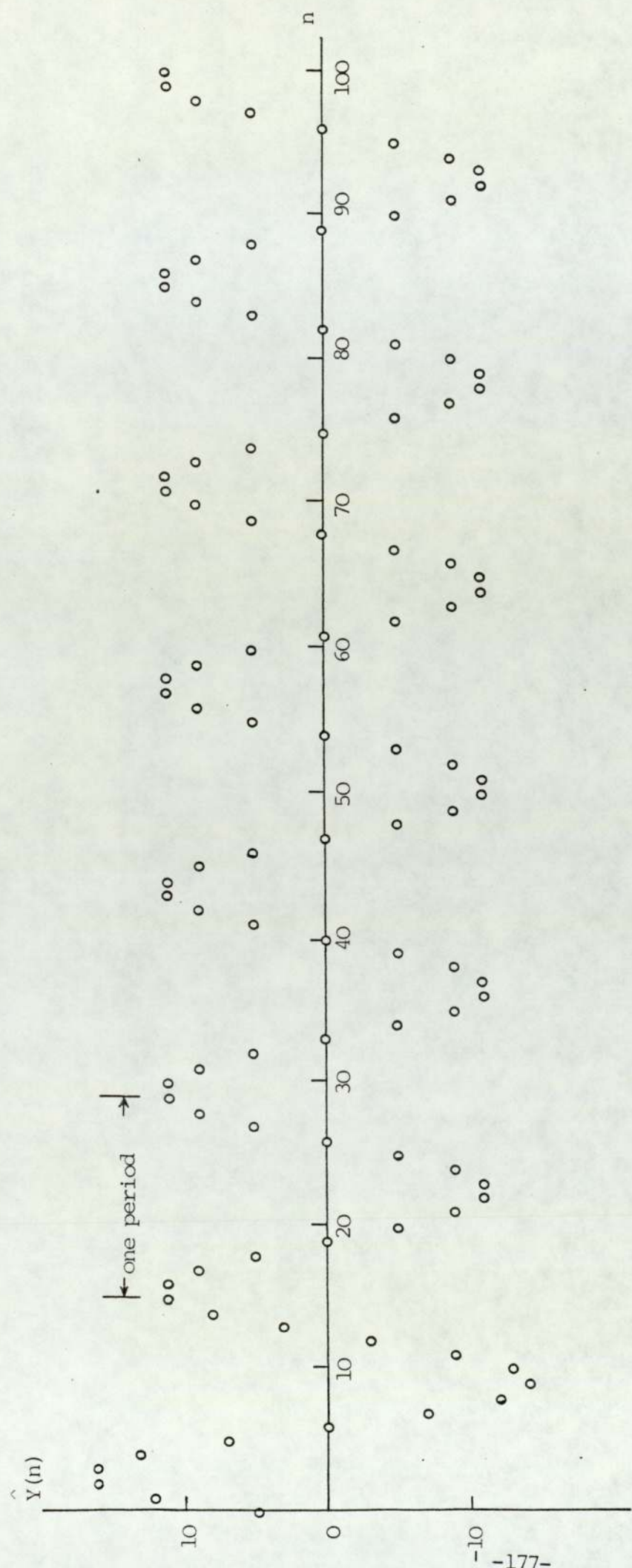


Fig. 39 The power density spectrum of the bandstop dither 2 whose stop band width was equal to $(0.08-0.14)F_s$. The nonlinear function $\frac{1}{2}\text{erf}\left(\frac{y_1}{\sqrt{2}}\right)$ was used.

In words, bandstop dither 1 has such a probability density distribution that is helpful for reducing the time needed to stabilise the basic filter section but it may cause a bit more increase in the output noise due to more intermodulation products falling within the stop band of the dither. In contrast with bandstop dither 1, the bandstop dither 2 has bigger stop band attenuation than bandstop dither 1, therefore it causes smaller increase in the output noise but a longer time for filter section stabilization is needed. As will be seen later, the experimental results have verified this conclusion.

7.3 SIMULATIONS OF LIMIT CYCLE SUPPRESSION

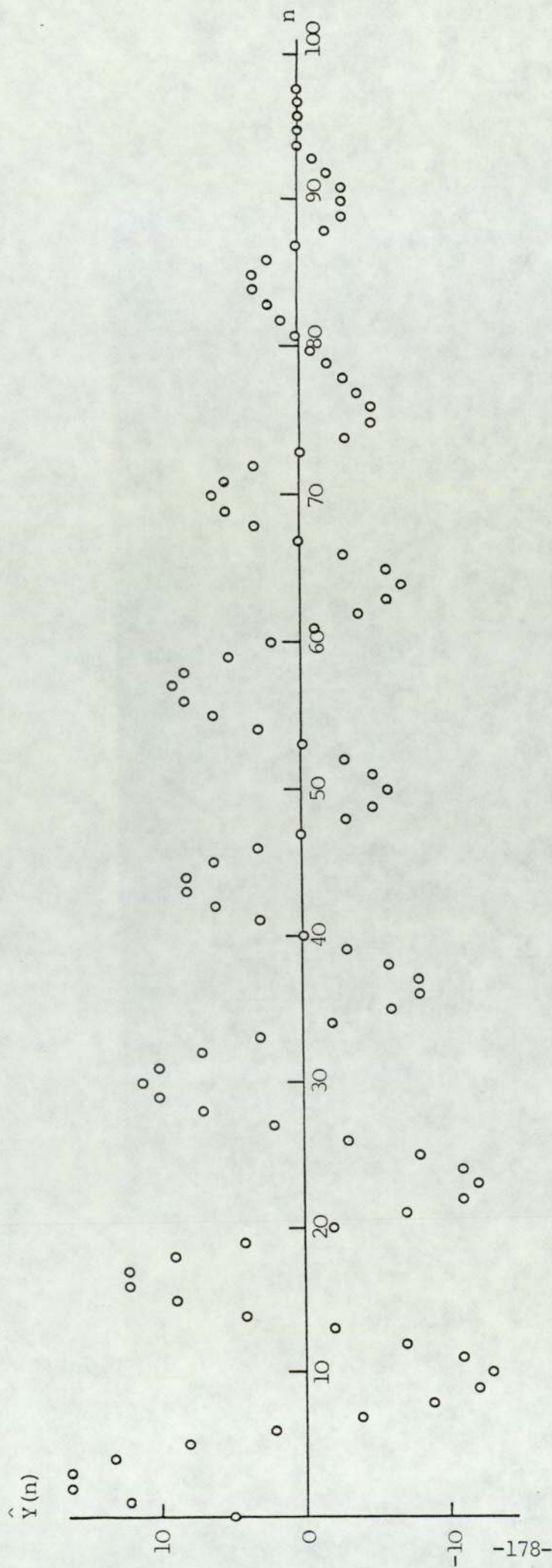
When the uniformly distributed or the binary random dither is used, the simulation of limit cycle suppression is simple. The only thing to do is to add the dither at the proper point as shown in Fig. 14. The simulations have verified that all the limit cycles in the second-order sections used can be suppressed without remaining noise in the zero input condition. Fig. 40 shows an example of limit cycle suppression. This is the example frequently used. The initial state is (5, 12). As can be seen from this figure, after 92 transitions, the filter reaches the origin state by the use of the uniformly distributed random dither. Once the limit cycle has been suppressed, the output keeps being zero in the zero input condition.



The zero-input response of the filter is as follows:

- (5,12) 16, 16, 13, 7, 0, -7, -12, -14, -13, -9, -3, 3, 8, 11, 11, 9, 5, 0, -5, -9,
- 11, -11, -9, -5, 0, 5, 9, 11, 11, 9, 5, 0, -5, -9, -11, -11, -9, -5, 0, 5, 9, 11, 11, ...

Fig. 40(a)



The output sequence from the filter with uniform dither is as follows:

(5,12) 16, 13, 8, 2, -4, -9, -12, -13, -11, -7, -2, 4, 9, 12, 12, 9, 4, -2, -7, -11,
 -12, -11, -8, -3, 2, 7, 10, 11, 10, 7, 3, -2, -6, -8, -8, -6, -3, 0, 3, 6, 8, 8, 6, 3, 0,
 -3, -5, -6, -5, -3, 0, 3, 6, 8, 9, 8, 5, 2, -1, -4, -6, -7, -6, -3, 0, 3, 5, 6, 5, 3, 0,
 -3, -5, -5, -4, -3, -2, -1, 0, 1, 2, 3, 3, 2, 0, -2, -3, -3, -3, -2, -1, 0, 0, 0, ...

(b)

Fig. 40 An example of limit cycle and its suppression (a) without dither (b) with uniform dither

For the sake of comparison, the zero-input response of the section without dither is also shown in the Fig. 40. ^{fold} Fig. 41 shows the output response in the same example but the binary random dither was used and the corresponding trajectory on the state plane from the initial state to the origin is shown in Fig. 42.

When the bandstop dither is used, the situation is a bit complicated. Fig. 43 shows the total experimental block diagram.

The transfer function of the linear bandstop filter can be written as

$$H_{ls}(Z) = \frac{F(1+ZAZ^{-1}+ZBZ^{-2}+ZCZ^{-3}+ZDZ^{-4})}{1+AZ^{-1}+BZ^{-2}+CZ^{-3}+DZ^{-4}} \quad (124)$$

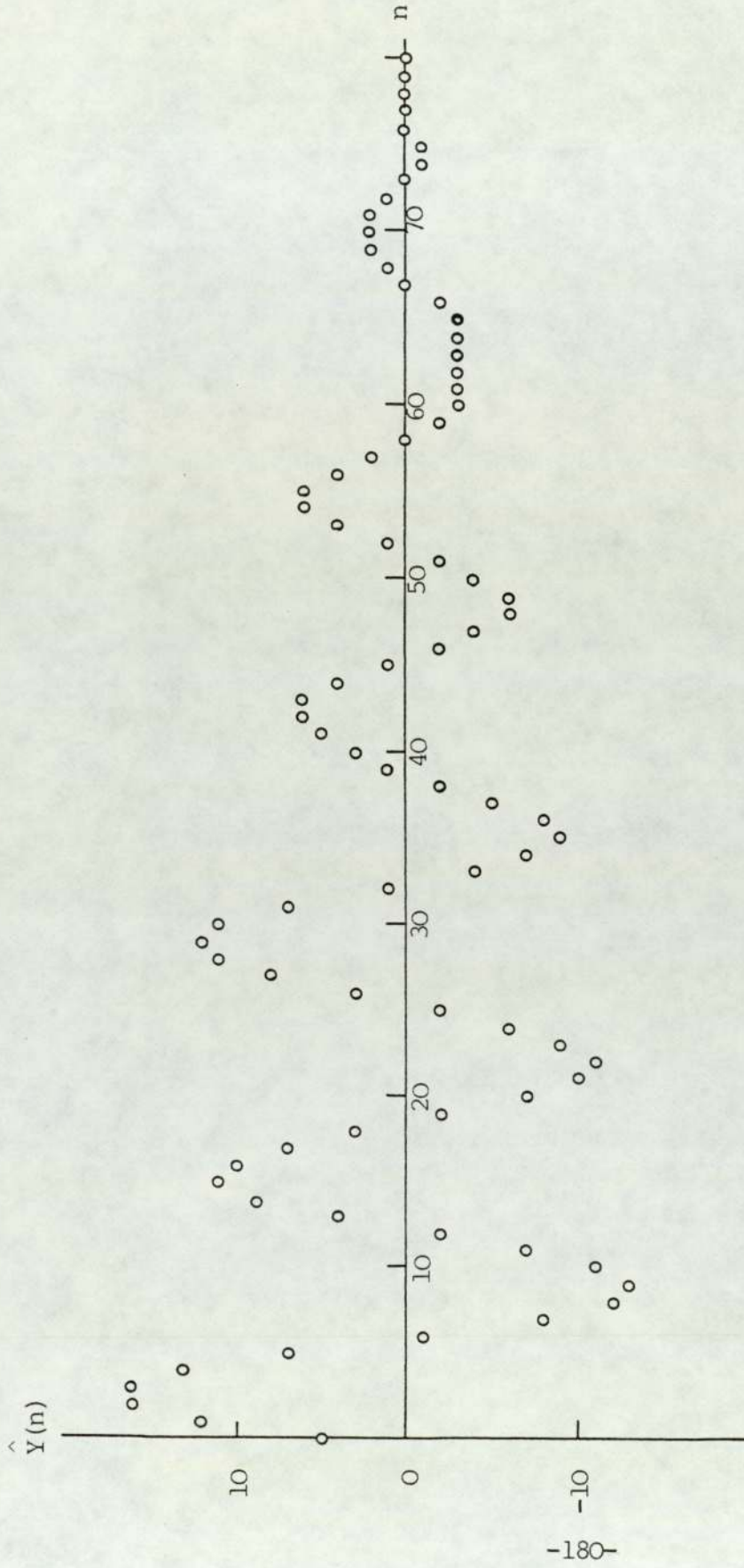
For a second-order bandstop filter

$$ZC = ZD = C = D = 0$$

As we have known that when a Gaussian random sequence with zero mean and unity standard deviation is applied into a linear filter, the standard deviation of the output from the filter, σ , is not equal to one. The standard deviation σ can be calculated as

$$\sigma = \sqrt{\sum_{n=0}^{\infty} h^2(n)} \quad (125)$$

where $h(n)$ is the impulse response of the linear bandstop filter.



The output sequence from the filter with binary dither is as follows:-

(5,12), 16, 16, 13, 7, -1, -8, -12, -13, -11, -7, -2, 4, 9, 11, 10, 7, 3, -2, -7, -10, -11, -9, -6, -2, 3, 8, 11, 12, 11, 7, 1, -4, -7, -9, -8, -5, -2, 1, 3, 5, 6, 6, 4, 1, -2, -4, -6, -6, -4, -2, 1, 4, 6, 6, 4, 2, 0, -2, -3, -3, -3, -3, -3, -2, 0, 1, 2, 2, 2, 1, 0, -1, -1, 0, 0, 0, ...

Fig. 41 An example of limit cycle suppression by the injection of binary dither.

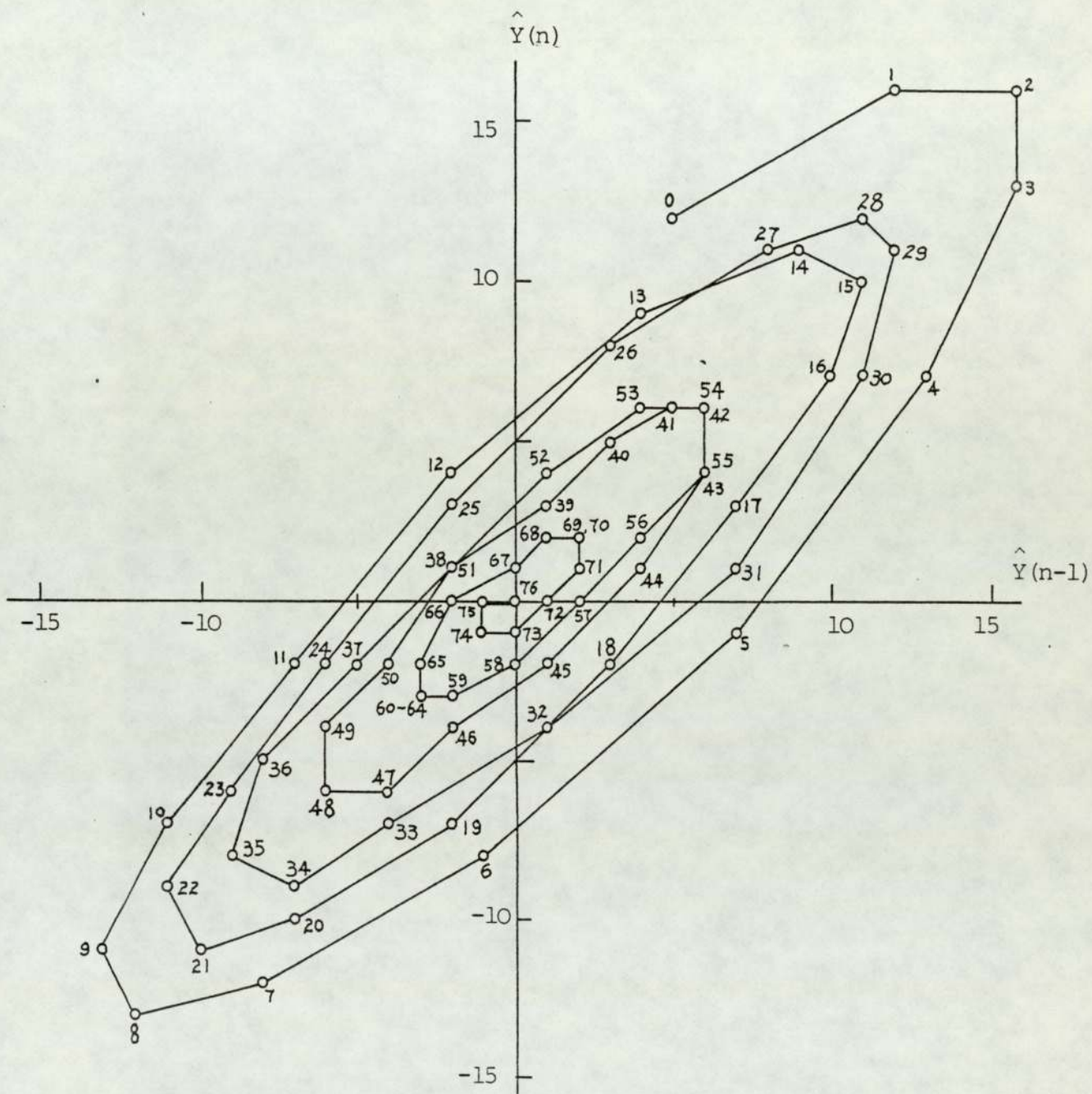


Fig. 42 The trajectory corresponding to Fig. 41 on the state plane.

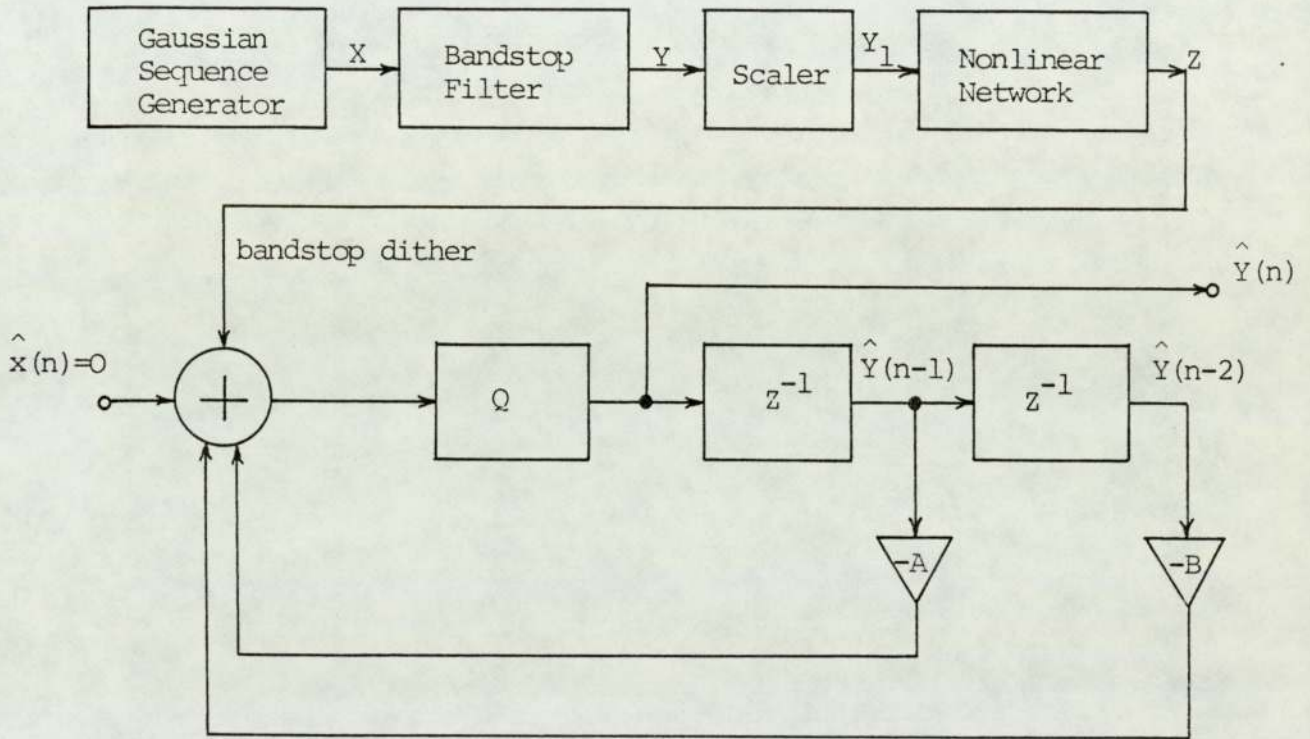


Fig. 43 The experimental block diagram of the limit cycle suppression in the second-order filter section with bandstop dither.

Only is the filter linear stable, the impulse response, $h(n)$, is convergent and when all the coefficients of the bandstop filter are known, σ can be calculated by computer easily. Appendix 9 shows the program for the calculation.

The procedures of doing a simulation are as follows:

1. Designing the Second-Order Basic Section to be Stabilized

The first step is choosing the pass band of the section, for example, $0.3 F_s - 0.31 F_s$ where F_s is the sampling frequency. Then by the use of the program mentioned before, a second order Butterworth bandpass digital filter can be designed.

Its transfer function is

$$\begin{aligned}
 H_p(z) &= \frac{F(1+OZA z^{-1}+OZB z^{-2})}{1+OA z^{-1}+OB z^{-2}} \\
 &= \frac{1-z^{-2}}{1+OA z^{-1}+OB z^{-2}} \qquad (126)
 \end{aligned}$$

when the order of the prototype filter which is now equal to one and the pass band edges of the filter, for example $0.3F_s$ and $0.31F_s$, have been typed in, the printer should print out the coefficient values $OZA=0.0$, $OZB=-1.0$, OA and OB . As mentioned earlier, because the zeros do not affect the number of limit cycles, they only affect the relative magnitude of the limit cycles, we can only

consider the poles of the second-order filter. Therefore, the transfer function of the second-order filter section can be expressed as

$$H_B(z) = \frac{1}{1+OAZ^{-1}+OBZ^{-2}} \quad (127)$$

2. Designing the Bandstop Filter

In the research, both Butterworth and elliptic bandstop filters were used. The orders were 2 or 4. According to different types of the filter, the different programs can be applied.

At this stage, one important question is how to choose the width of stop band.

There are at least three factors that should be considered. First, the stop band width should be greater than the pass band width of the filter section to be stabilized. The frequencies of the limit cycles are divergent although they are, in general, still near by the resonant frequency of the section and 3 dB attenuation at stop band edges is not enough for improving limit cycle suppression. Second, the stop band width should not be too wide, otherwise with the big probability the instantaneous values of the bandstop dither may not be big enough for limit cycle suppression because most

frequency components have been suppressed. Third, the stop band width should not be too wide, otherwise the distortion of the stop band characteristic will be more serious. More intermodulation products which are caused by the nonlinear network following the bandstop filter will fall within the wide stop band. Therefore, it is expected that there may be an optimum stop band width for the limit cycle suppression. Here "optimum" means the shortest time needed for limit cycle suppression.

In order to find the "optimum" stop band width, five different widths have been used. They are equal to 1, 2, 3, 4 and 5 times pass band of width of the second-order section to be stabilized respectively. The simulations have shown that generally when the stop band width is equal to three or four times pass band width of the section to be stabilized the performance of the limit cycle suppression is better.

3. Calculating the Standard Deviation of the Response of the Bandstop Filter by the Program shown in Appendix 9
4. Scaling the Output Sequence of the Bandstop Filter

The scaling factor is equal to $\frac{1}{\sigma}$.

5. Generating the Bandstop Dither Signal by the Program
Shown in Appendix 10

This program requests an input which consists of the parameters from step 1-4 above.

6. Simulating the Basic Second-order Filter Section,
Setting the Initial State

Then, the procedure of the limit cycle suppression by the use of the bandstop dither can be printed out by the computer.

7.4 TIME FOR STABILIZATION

The experiments to determine the time for dither to affect stabilization have been done by simulations.

Among the three types of dither signal mentioned before the uniformly distributed random dither is a basic type. The others come from it and are mainly for improving the performances of the limit cycle suppression. Therefore, first we pay attention to the use of uniformly distributed dither. Then, the results of the use of other dither signals will be presented.

1. Results of the Use of Uniformly Distributed Random
Dither

Several different methods to measure the time for

stabilization have been used.

In the first method, the average time from each initial state in the amplitude bound zone of the limit cycle to the origin of the state plane were measured.

In the second method, in each case the filter section started at the same state which was on or "outside" the largest limit cycle on the state plane. Here "outside" means that the distance from the initial state to the origin on the state plane is bigger than the maximum distance from the states on the largest limit cycle. Then the time for transition to the origin was measured, because the dither was random, this time was also random. Simulation was repeated 1000 times. Hence the cumulative distribution function (CDF) for the transition time could be obtained. The median value corresponding to 50% probability was interested.

The third method of measurement is similar with the second one but the mean value of the transition time was used.

It is expected that the requirements to the uniformly distributed random dither are not strict, i.e., no special strict requirements are needed on the correlation, probability distribution, etc. In order to obtain the impression about the effect of the use of different uniform random dither, the uniform random dither signals from

different computers have been used. As will be seen from the results, (See Figs.44 - 61), the performance of the limit cycle suppression does not much depend on the characteristic of the uniform random dither. Therefore, the uniformly distributed random dither can be generated by a simple method.

As mentioned in Chapter 2, there are two different ways of implementing the quantizations in the second-order digital filters: one quantizer and the two quantizer versions. In the two quantizer version, a dither can be added to the two coefficient products simultaneously or only to the B coefficient product. In other words, there may be three cases; one quantizer one dither, two quantizers two dithers and two quantizers B dither. In the simulations all the three cases have been included.

Table 3 shows the results when the uniformly distributed dither was used. In this table, six types of second-order filter section have been applied. The coefficient combinations were: $A=\pm 1.74$, $B=0.95833$, $A=\pm 1.875$, $B=0.91875$; $A=\pm 1.25$, $B=0.825$. Each pair of filters without dither respectively has periodic and constant (alternating); constant (alternating); and periodic limit cycles. In each case, the filter was initialised to the state (11,11). The simulation was repeated 1000 times, enabling an accurate estimate of the median time for stabilization to be made.

The CDF in each case is shown in Fig. 44 to Fig. 61.

TABLE 3

The transition times to the origin for various filter sections with uniform and binary dither signals

Filter Section Configuration	Type of Dither	Filter Section Coefficients A and B						Average transition time
		A=-1.74 B=0.95833	A=1.74 B=0.95833	A=-1.875 B=0.91875	A=1.875 B=0.91875	A=-1.25 B=0.825	A=1.25 B=0.825	
One Quantizer One Dither	Uniform dither	524	494	271	283	47	47	Average transition time
	Binary dither	175	172	50	53	34	33	Median transition time
	Uniform dither	330	425	185	250	40	50	Average transition time
	Binary dither	130	123	25	77	30	40	Median transition time
Two Quantizer Two Dither	Uniform dither	688	707	531	513	54	51	Average transition time
	Binary dither	176	174	246	260	48	48	Median transition time
	Uniform dither	480	515	370	430	48	55	Average transition time
	Binary dither	115	192	163	201	41	50	Median transition time
Two Quantizer B Dither	Uniform dither	329	312	576	633	41	41	Average transition time
	Binary dither	88	85	193	199	28	27	Median transition time
	Uniform dither	240	303	418	450	40	47	Average transition time
	Binary dither	62	134	115	179	28	36	Median transition time

Note: Times as multiples of the sampling period. The average transition time means the average transition time to the origin from each initial state which lies within the zone bounded by the amplitude bound of limit cycle on the state plane. The median transition time means the median transition time to the origin from the initial state (11,11) in 1000 simulations.

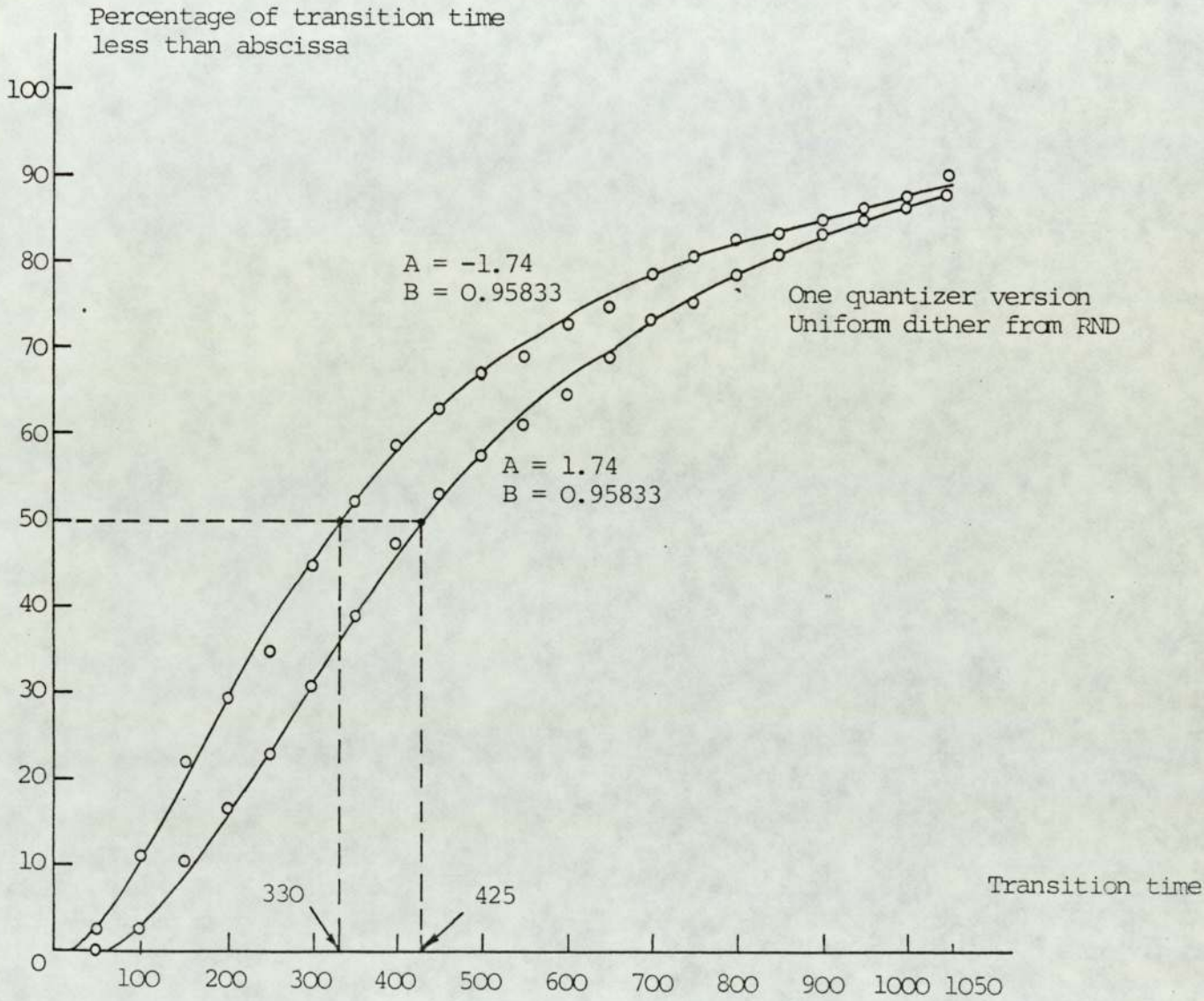


Fig. 44 Cumulative distribution function of the transition time to the origin state.

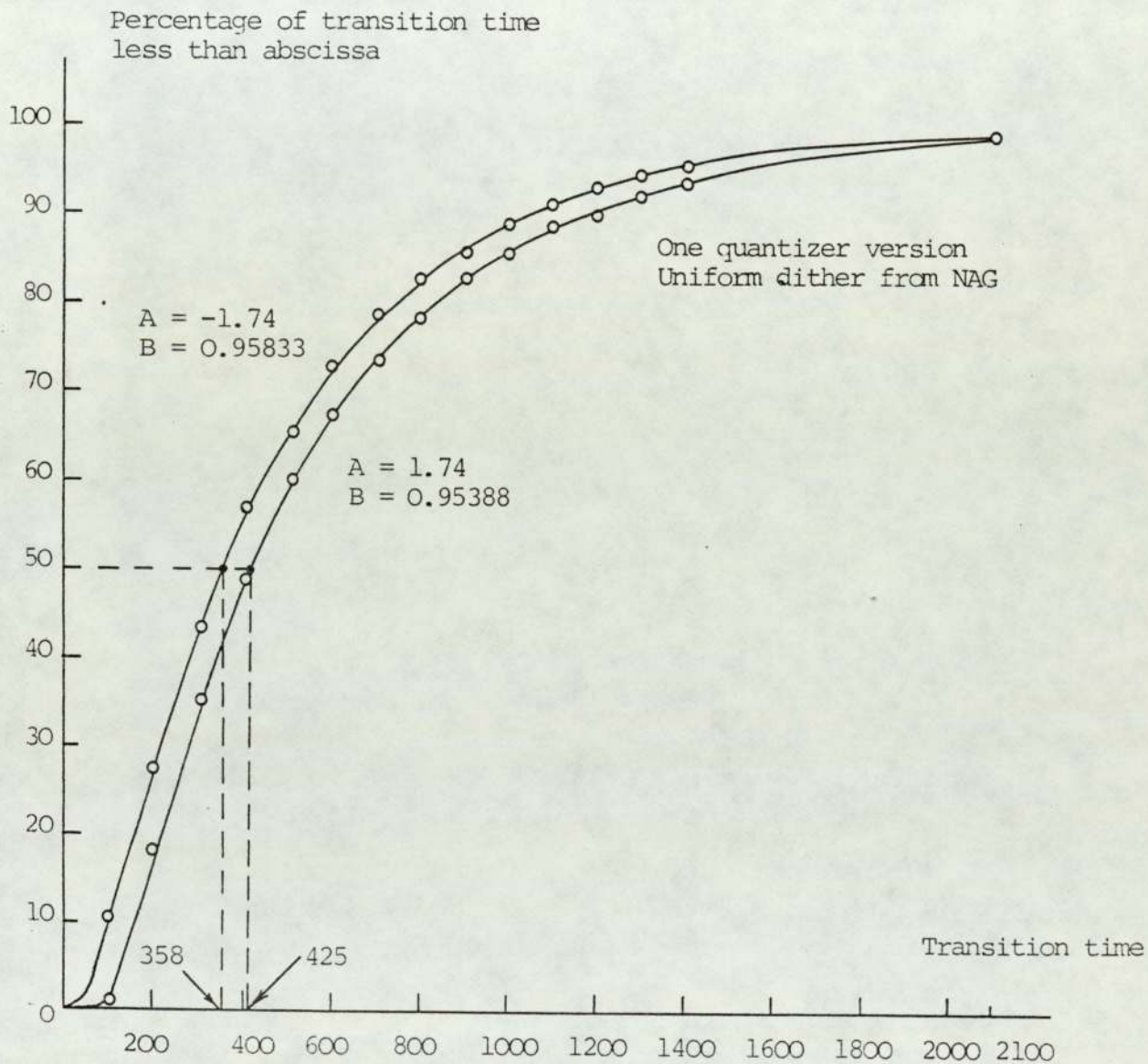


Fig. 45 The cumulative distribution function of the transition time to the origin state.

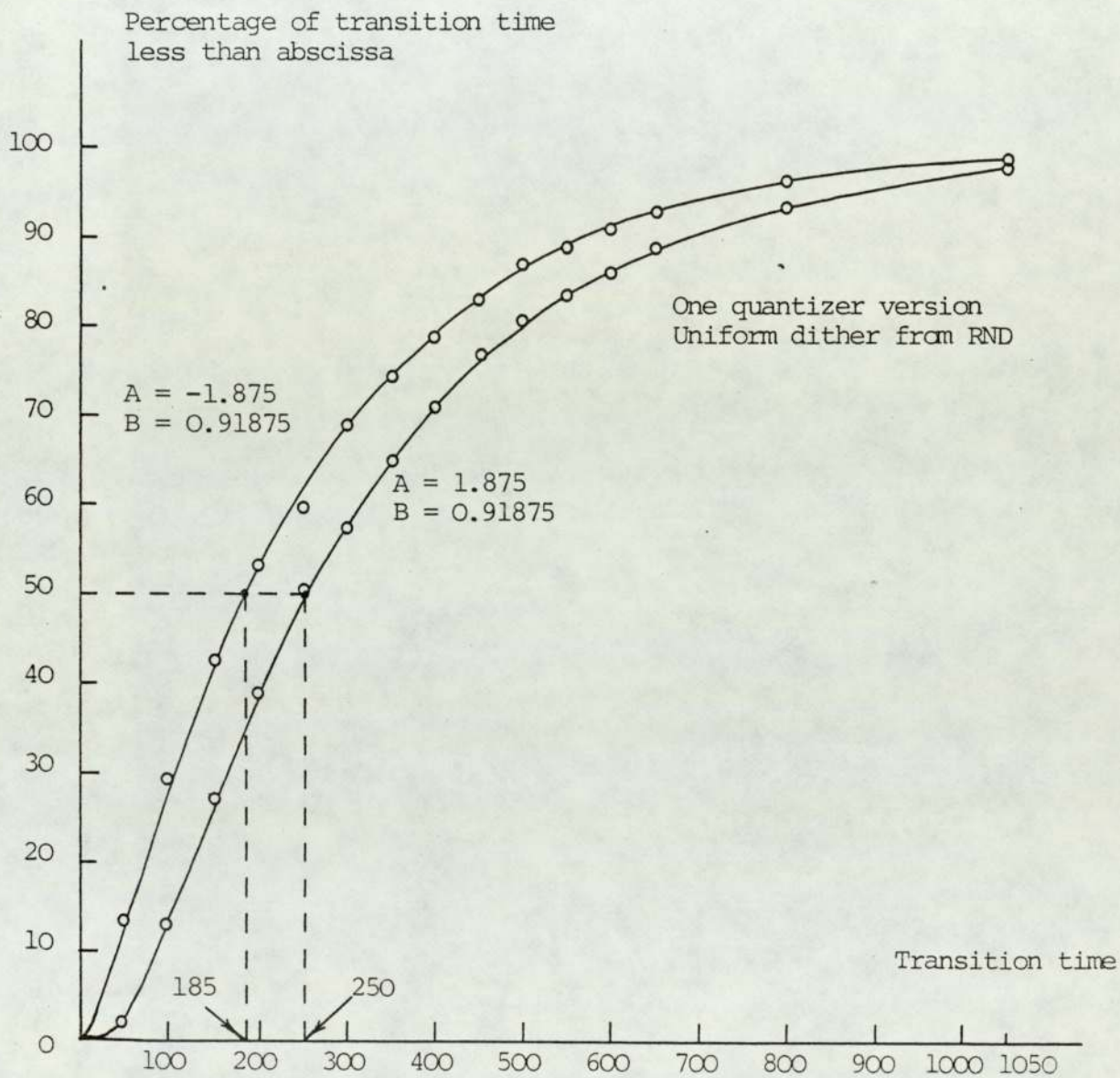


Fig. 46 Cumulative distribution function of the transition time to the origin state.

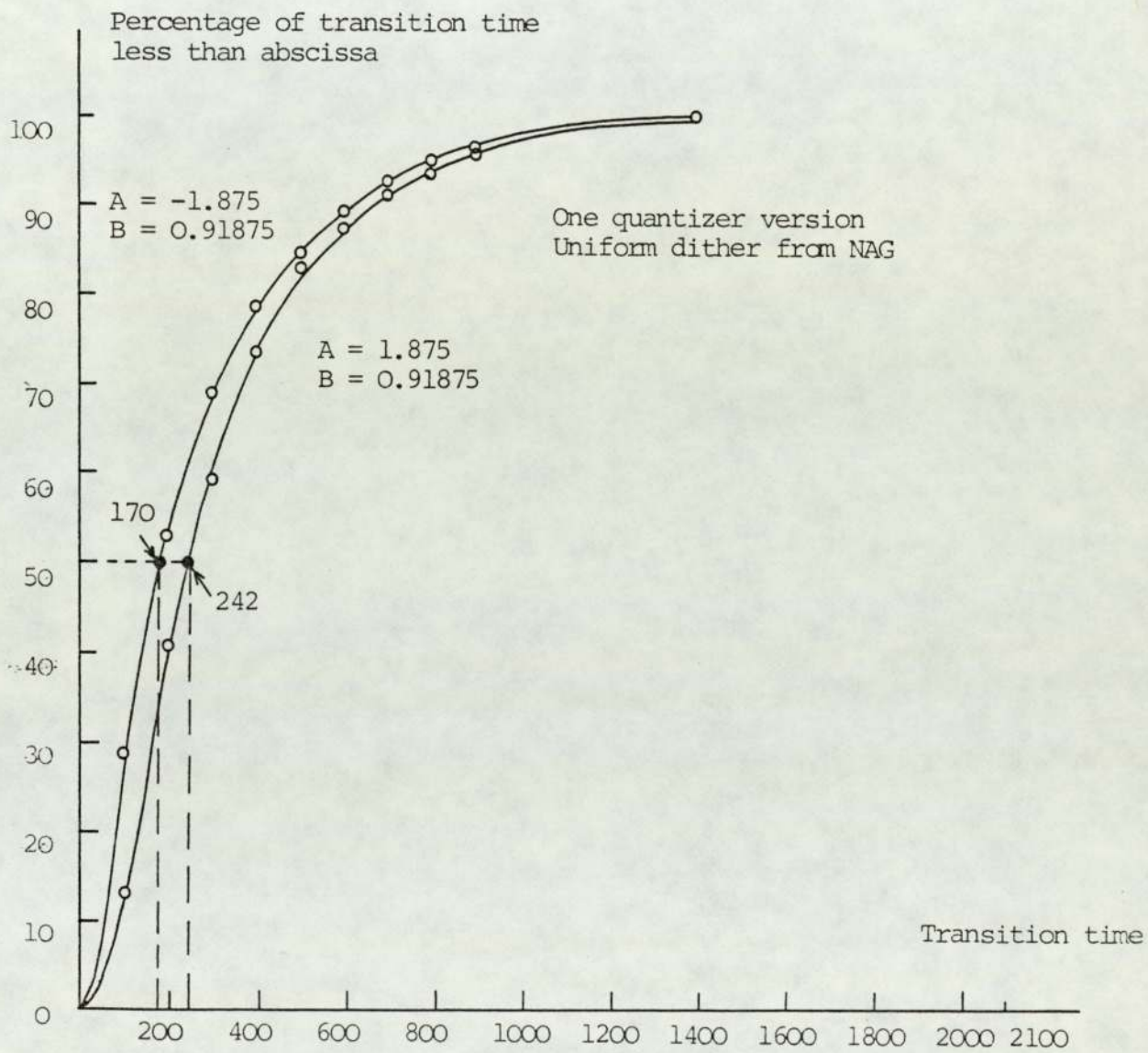


Fig. 47 Cumulative distribution function of the transition time to the origin state.

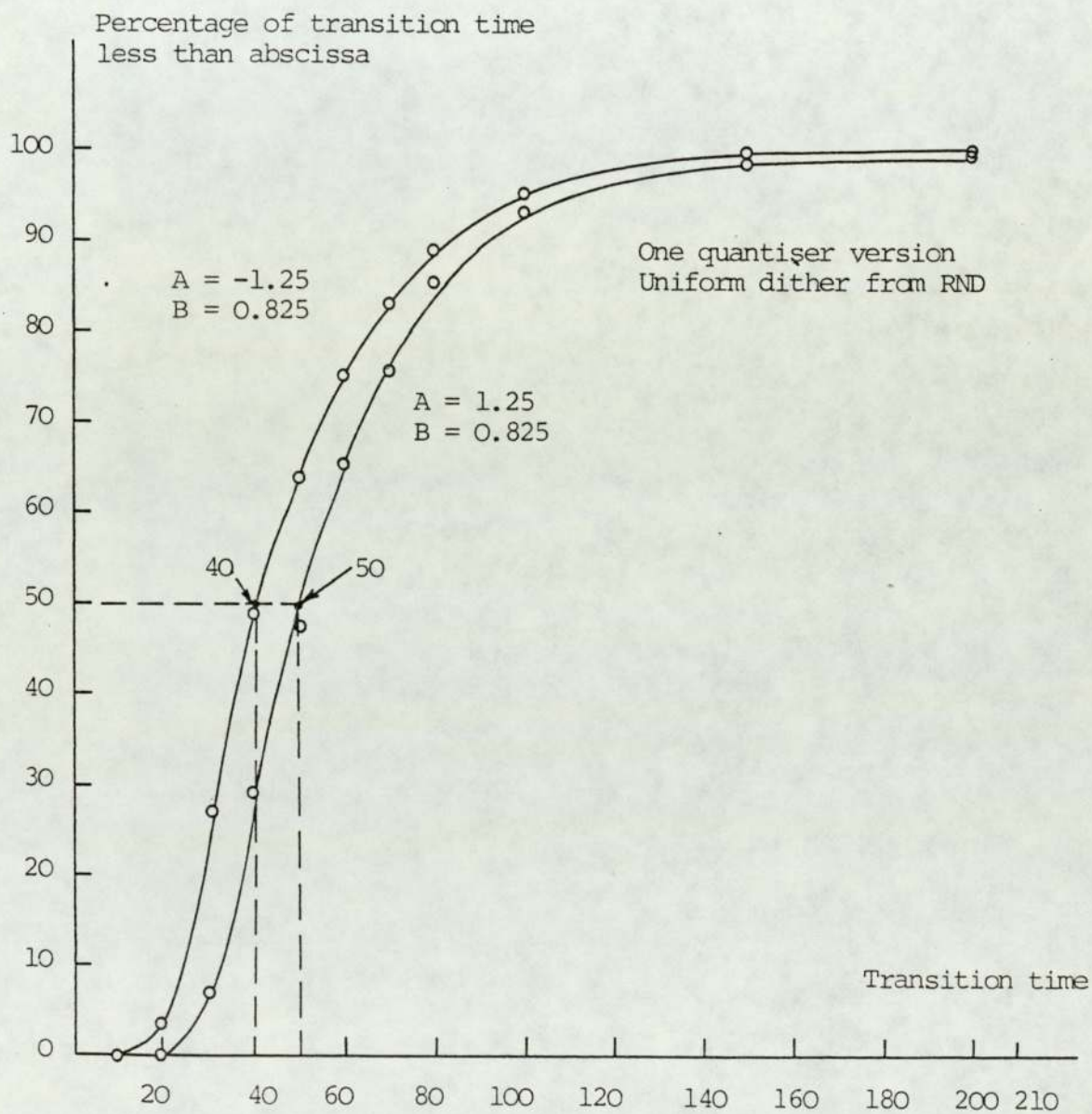


Fig. 48 Cumulative distribution function of the transition time to the origin state.

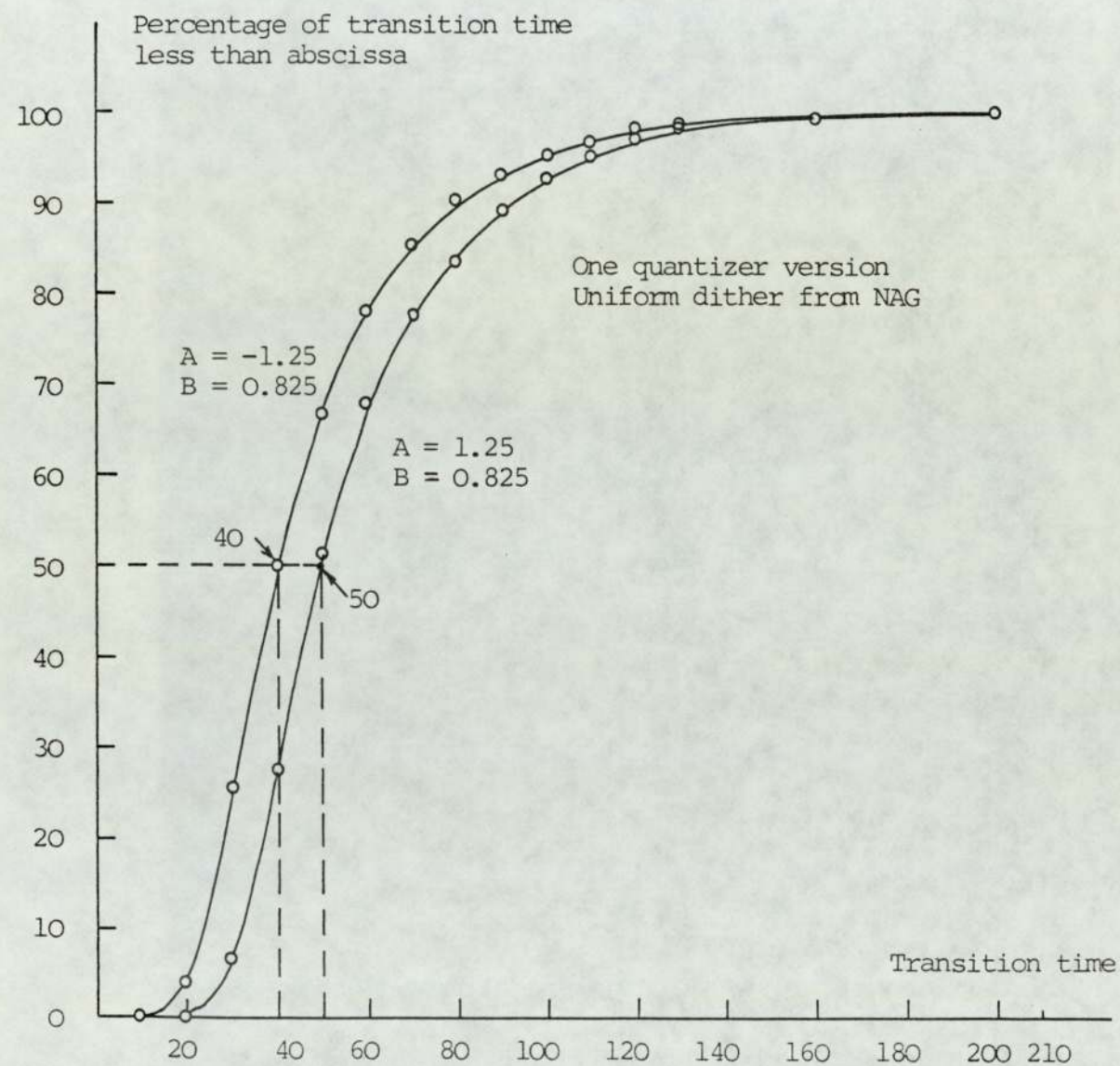


Fig. 49 Cumulative distribution function of the transition time to the origin state.

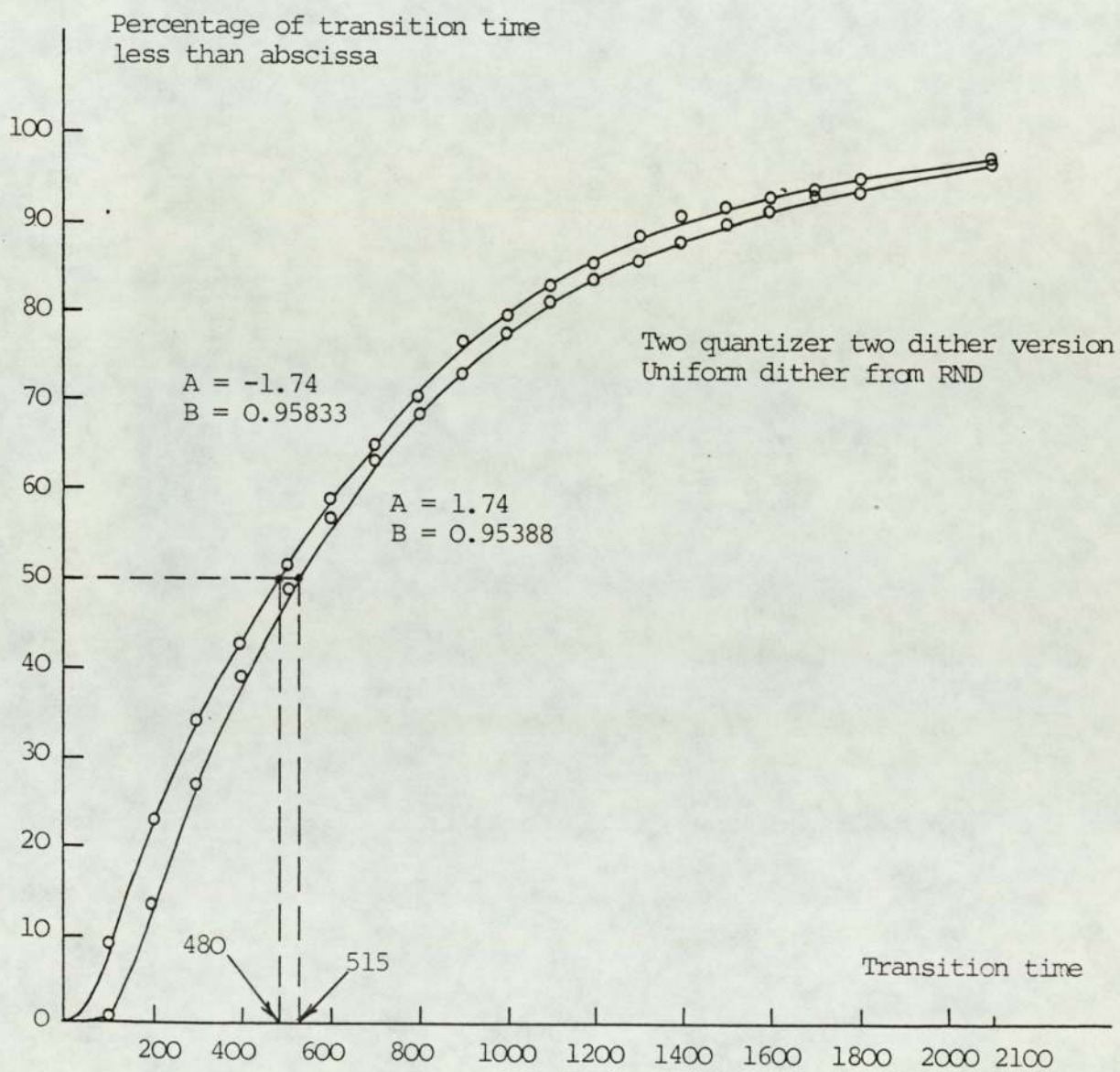


Fig. 50 Cumulative distribution function of the transition time to the origin state.

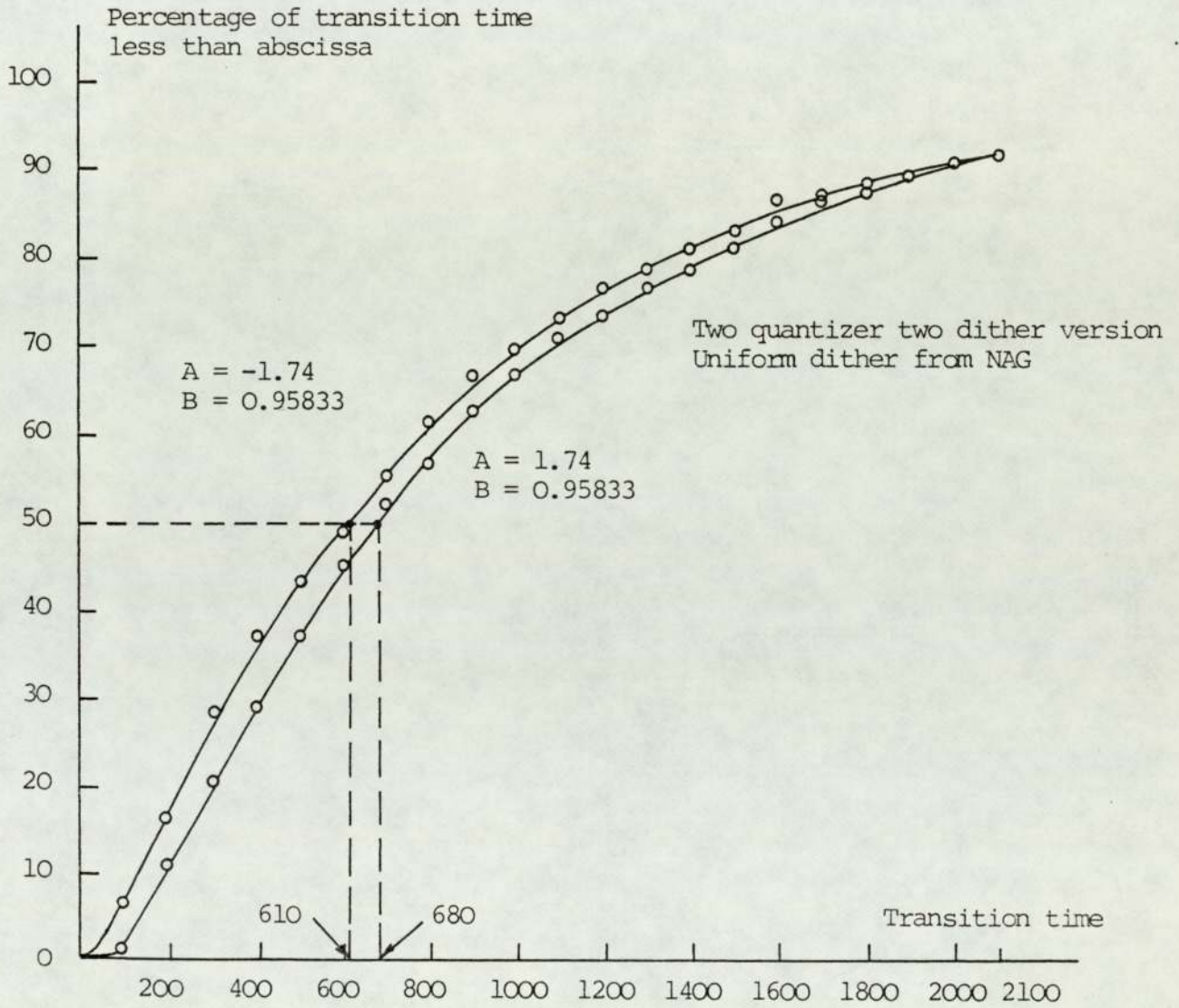


Fig. 51 Cumulative distribution function of the transition time to the origin state.

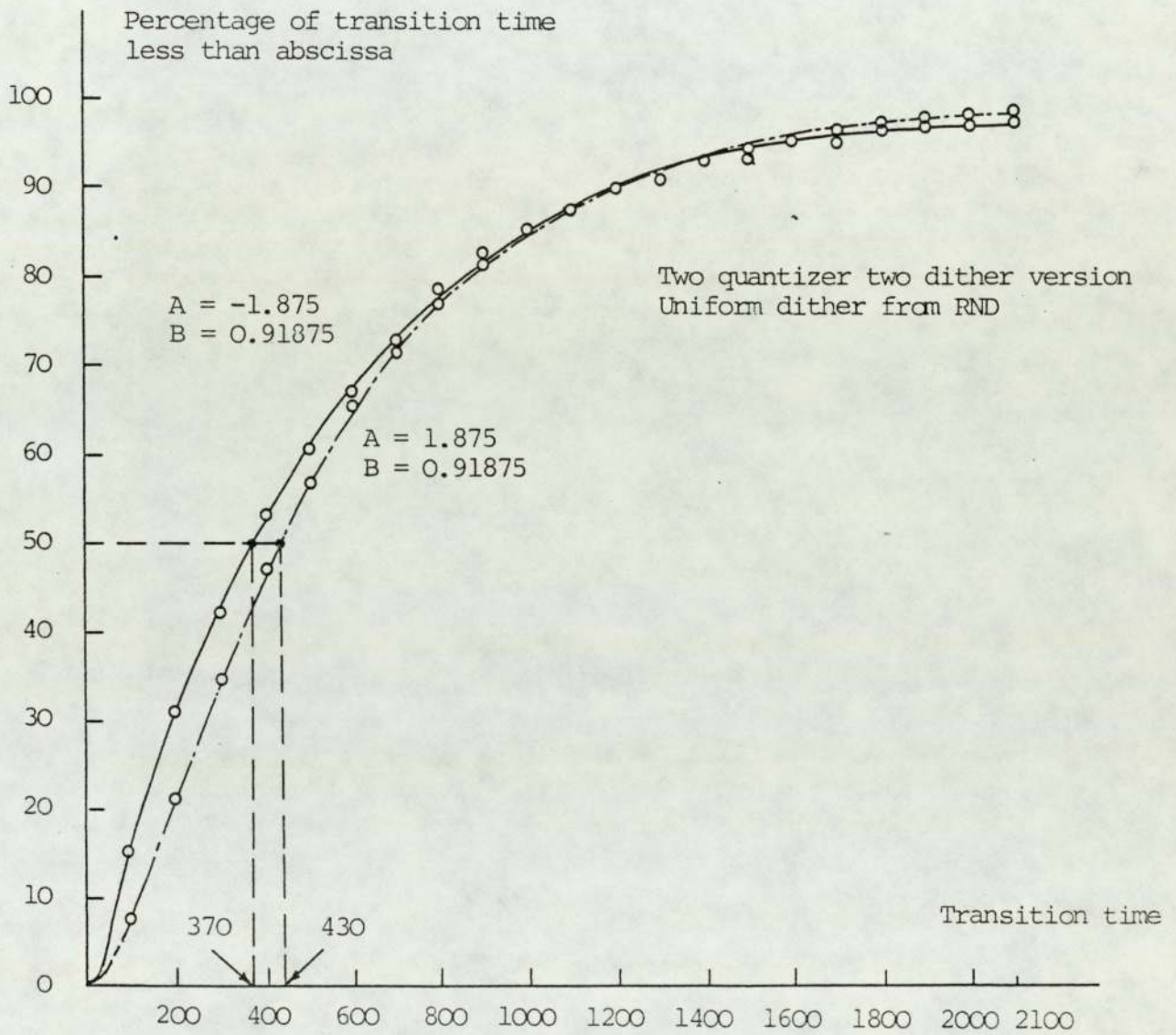


Fig. 52 Cumulative distribution function of the transition time to the origin state.

Percentage of transition time
less than abscissa

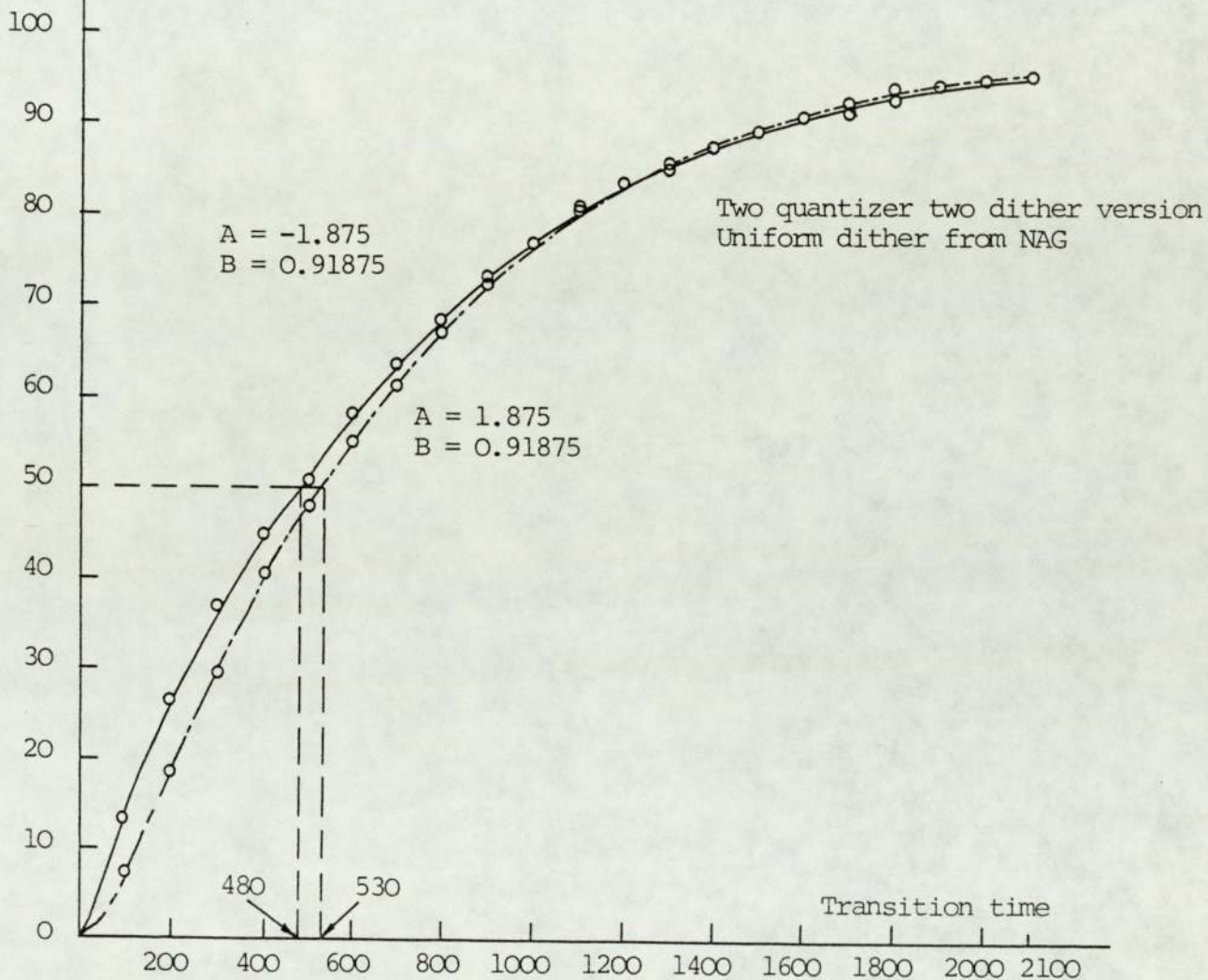


Fig. 53 Cumulative distribution function of the transition time to the origin state.

Percentage of transition time less than abscissa

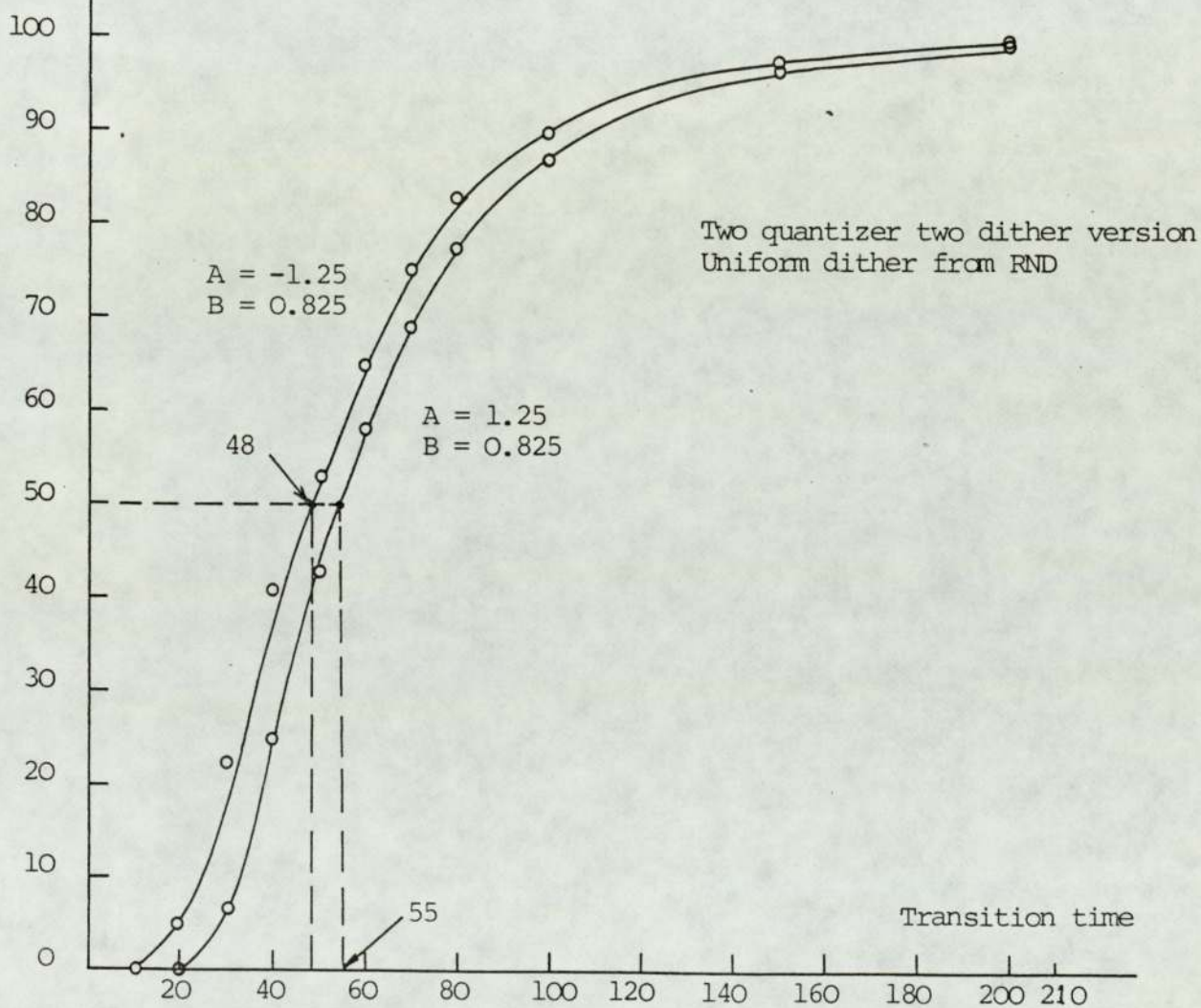


Fig. 54 Cumulative distribution function of the transition time to the origin state.

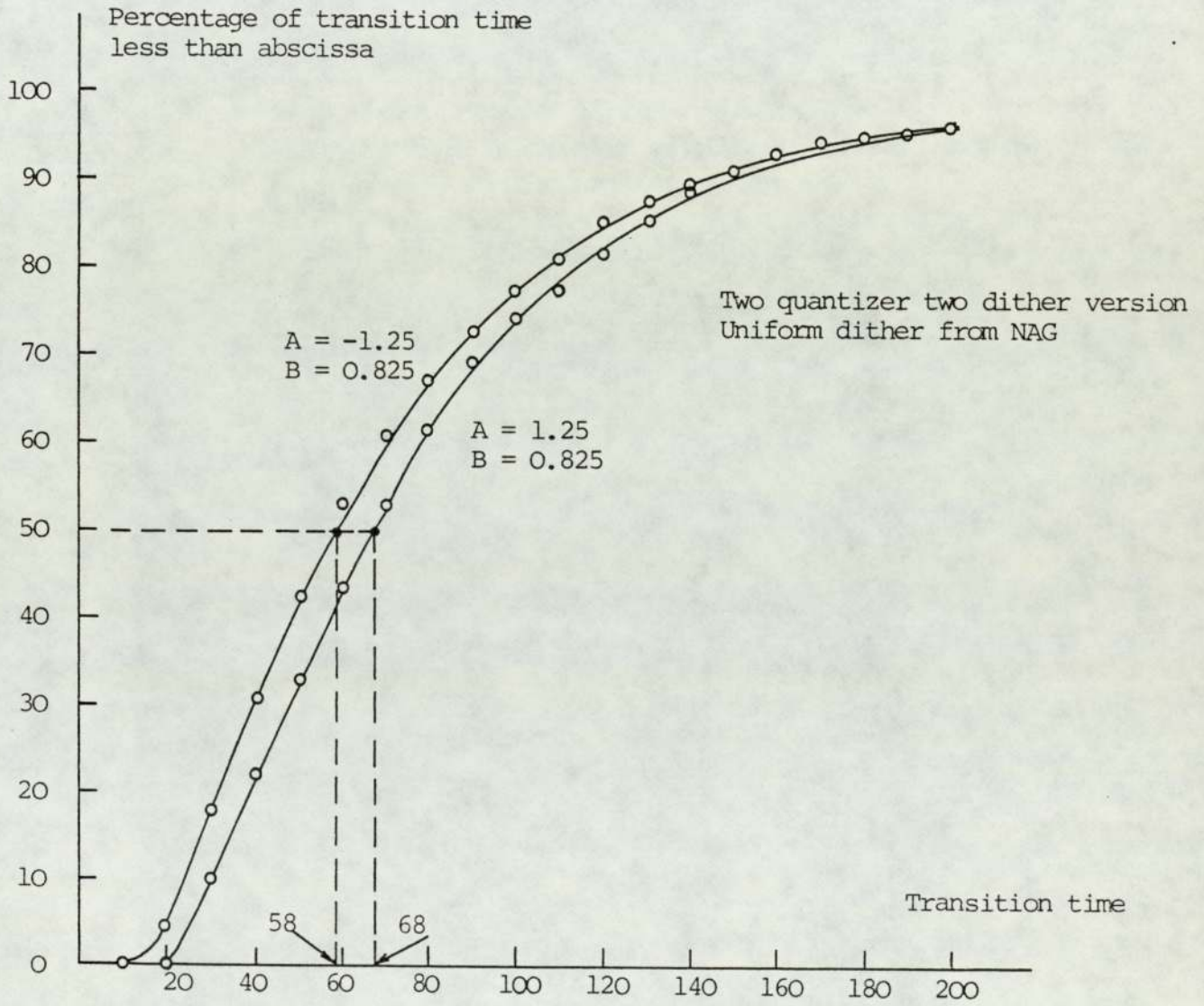


Fig. 55 Cumulative distribution function of the transition time to the origin state.

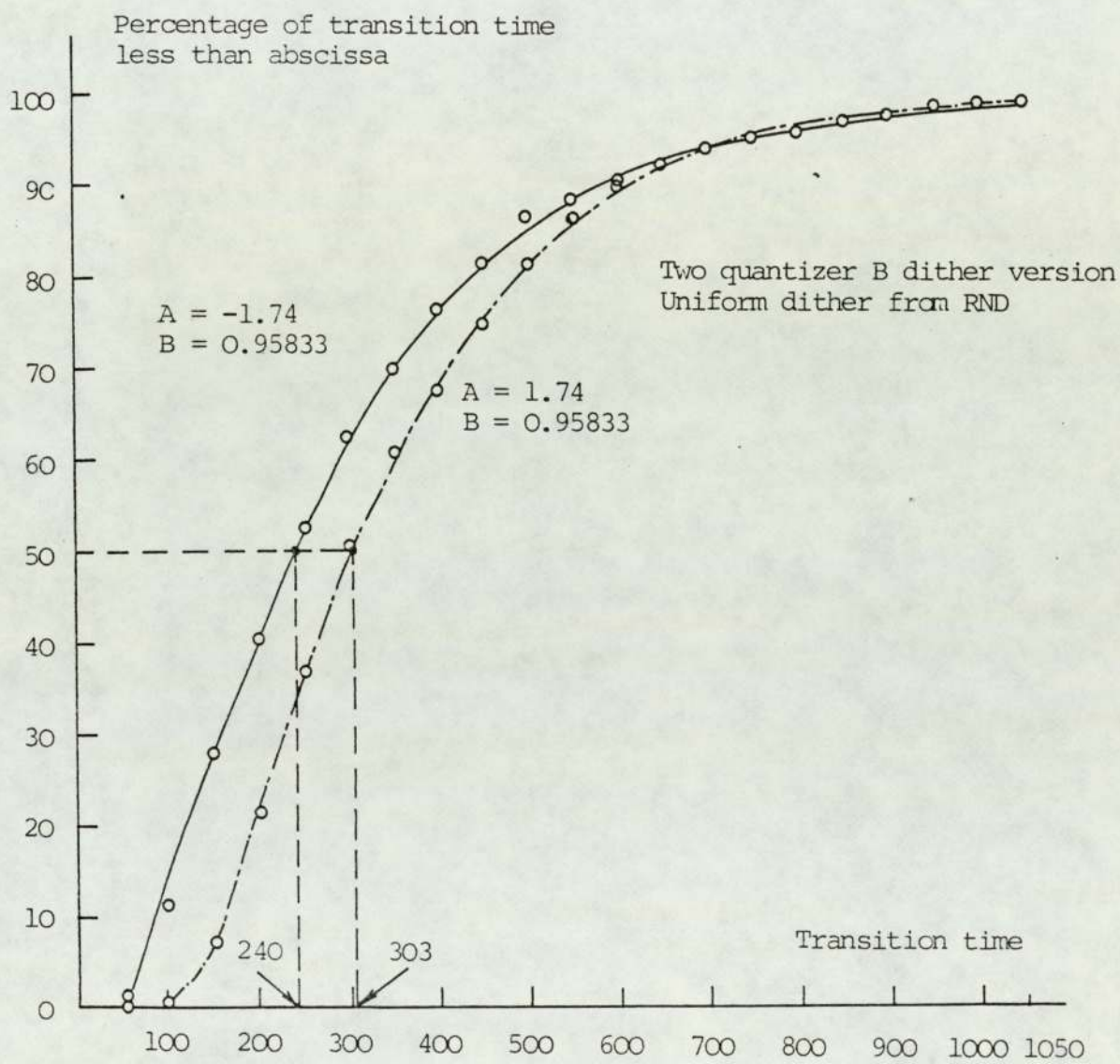


Fig. 56 Cumulative distribution function of transition time to the origin state.

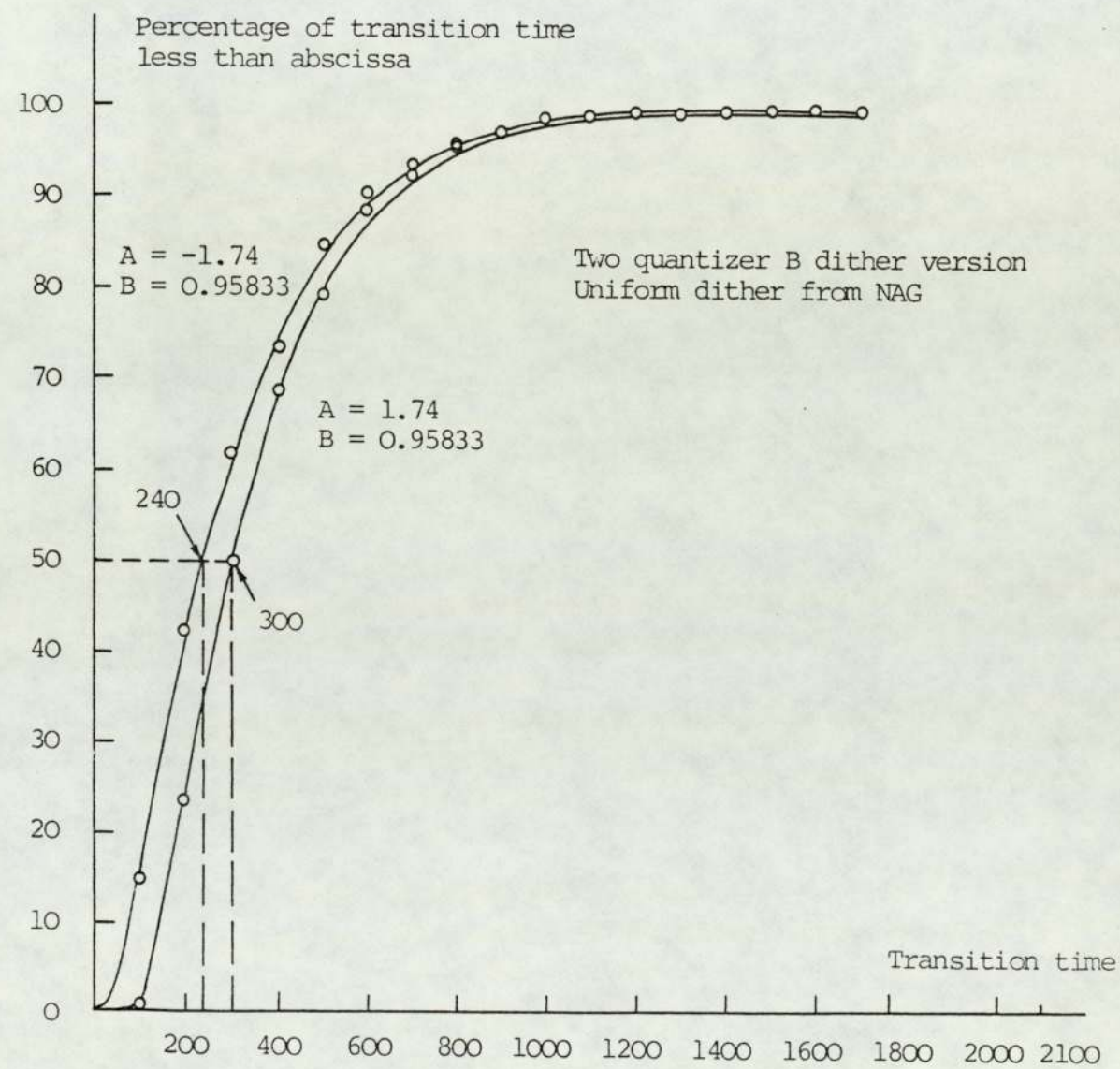


Fig. 57 Cumulative distribution function of the transition time to the origin state.

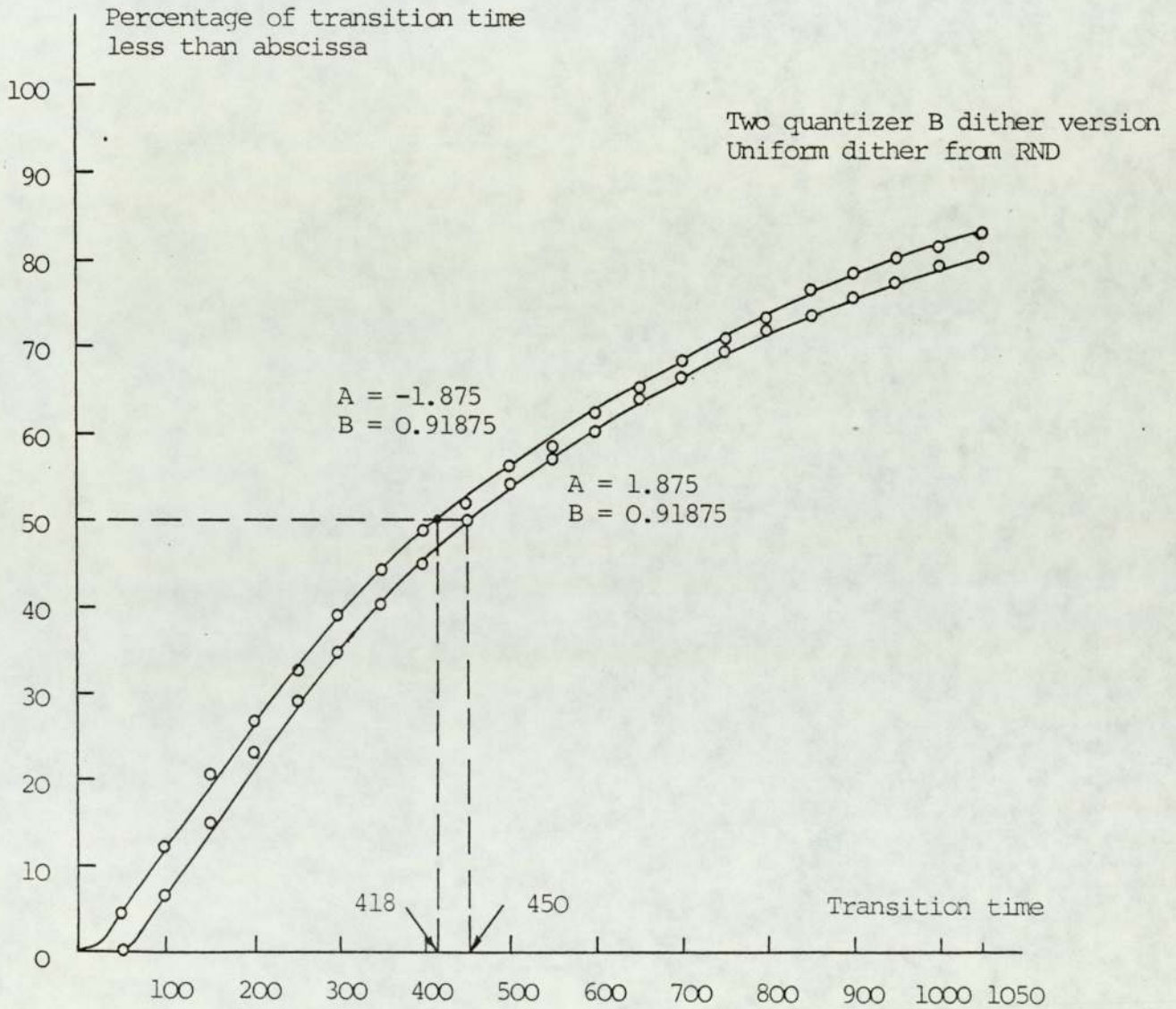


Fig. 58 Cumulative distribution function of the transition time to the origin state.

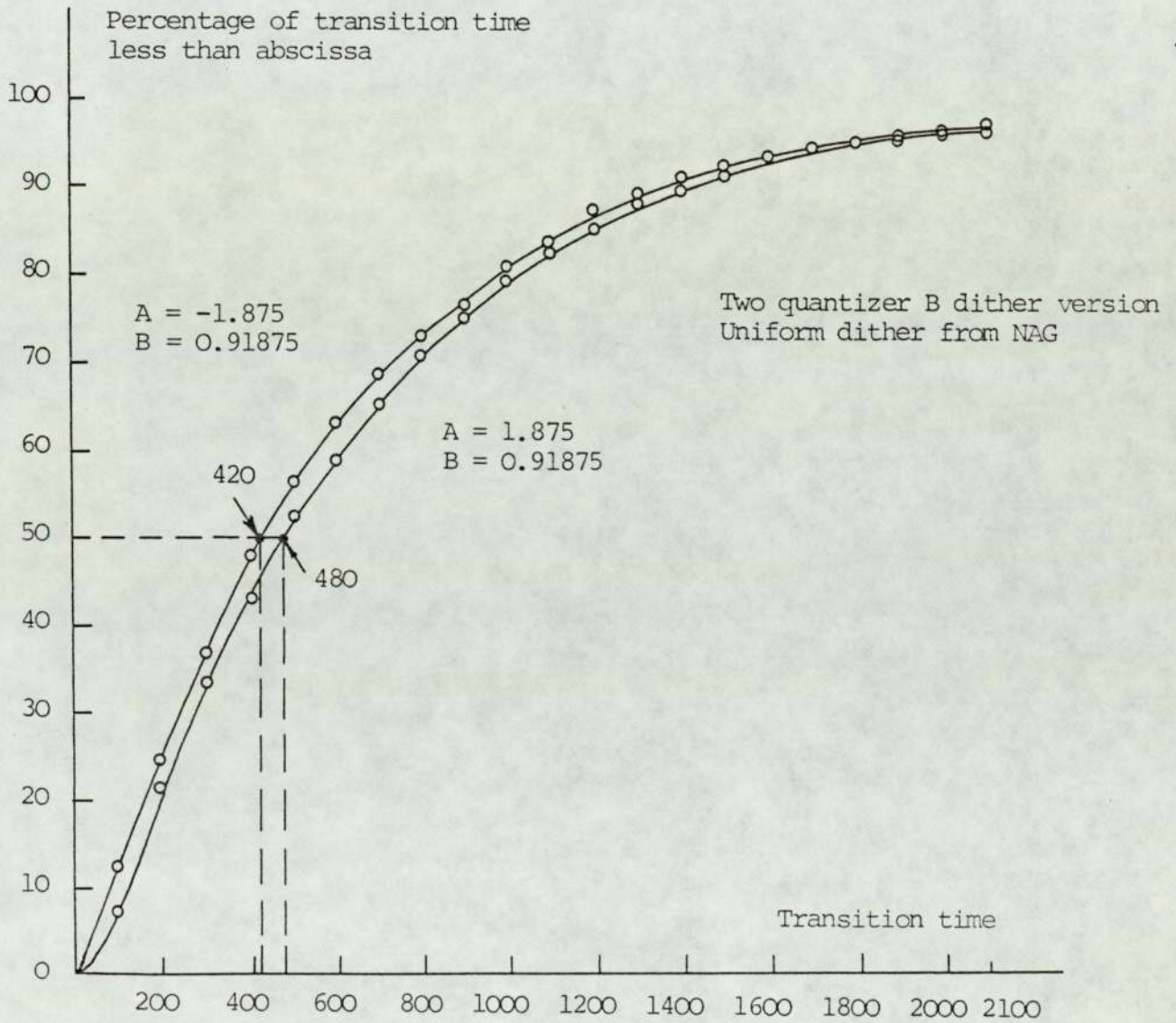


Fig. 59 Cumulative distribution function of the transition time to the origin state.

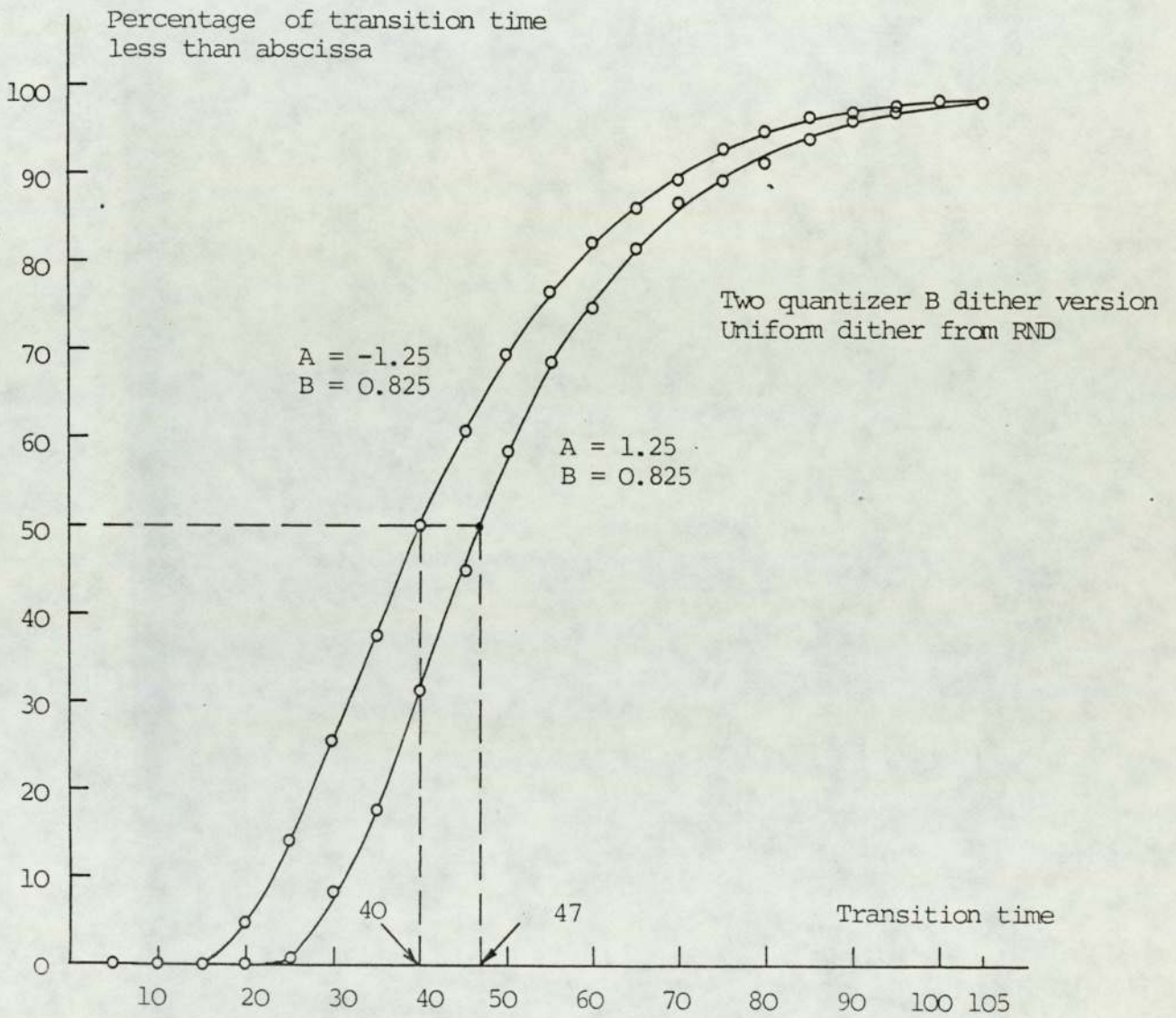


Fig. 60 Cumulative distribution function of the transition time to the origin state.

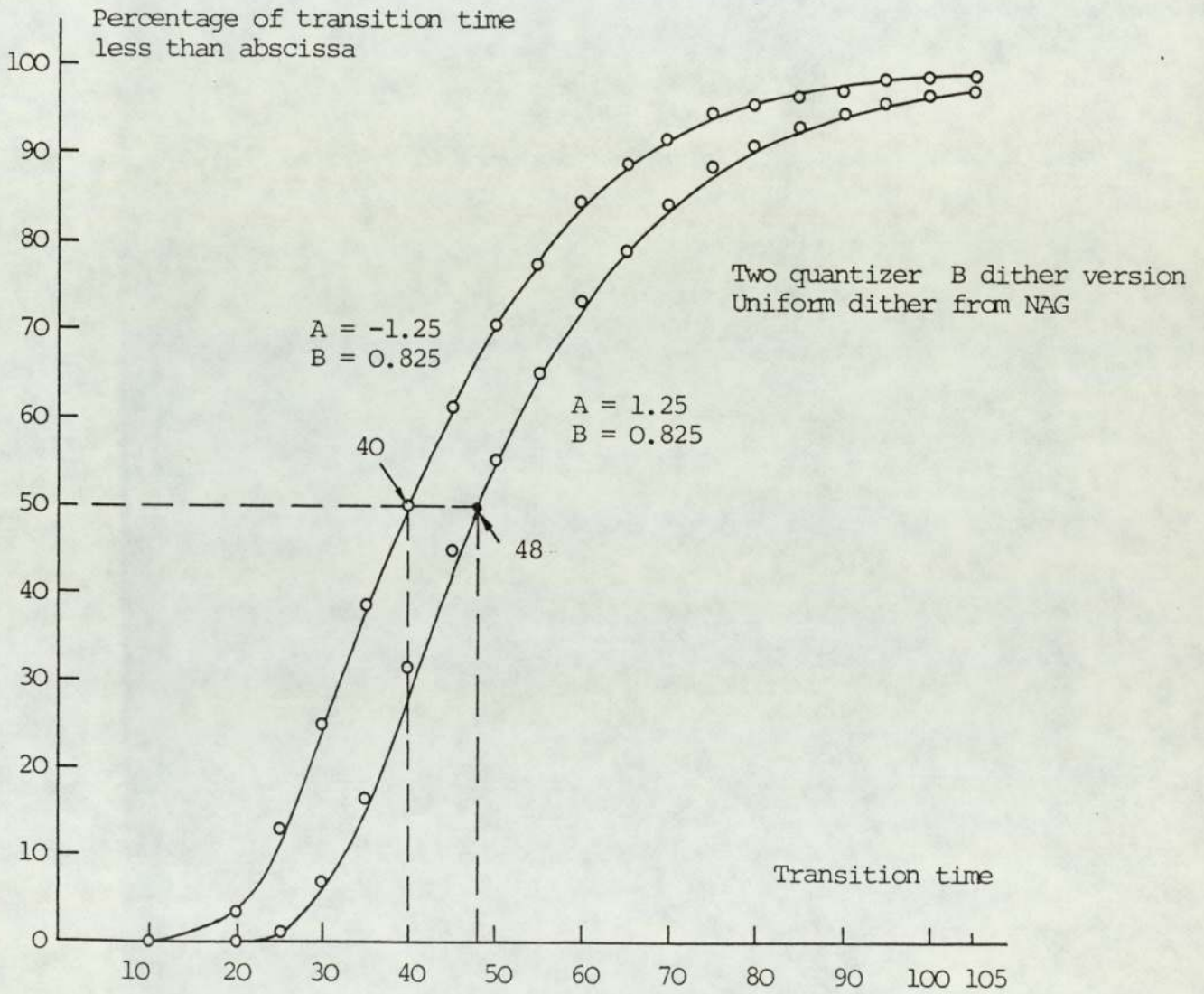


Fig. 61 Cumulative distribution function of the transition time to the origin state.

From these CDF it can be expected that the distribution of transition times to the origin are approximately normal.

2. Results of the Use of the Binary Random Dither

In this research, the amplitude of the dither was assumed to be 0.499. As mentioned in Chapter 6, this amplitude is, in general, big enough except in the very special case $|1-|A|+B| \leq 0.001$.

For convenience of comparing, the same filters and the same methods used in the simulation with uniformly distributed random dither have been applied. All the results are also shown in Table 3. As can be expected, the table shows that the transition time for the stabilization when the binary random dither is used is much less than that when the uniformly distributed dither is used. Specially, when the Q-value of the second-order section is high, this improvement is distinguished. For the example often used in this thesis, its improved transition time is only about one fourth to one third of that when the uniformly distributed random dither is used. This is particularly desirable, because only in high-Q cases, the transition times are very long, and the improvement is most remarkable.

3. Results of the Use of the Bandstop Dither

The program used in the simulation of limit cycle suppression by the use of the bandstop dither is shown in

Appendix 10. In this program, statement 55 calls a Gaussian random sequence with a unit standard deviation and zero mean. Statement 57 simulates the bandstop filter. Statement 66 scales the output from the bandstop filter so as to get a filtered Gaussian sequence with unity standard deviation. Statements 70-96 simulate the nonlinear network. The output from this nonlinear network is the required bandstop dither. Statements 100-117 simulate the basic second-order section. In the simulation, the initial condition of the basic filter section was (11,11) and the input signal was zero. The section was operated until the origin state was reached. The computer calculated the number of steps. Each simulation was repeated 100 times. The computer printed out the cumulative distribution function CDF. The median values of the transition time are shown in Table 4. For the sake of convenient comparing, the median values corresponding to the use of the uniformly distributed dither and the binary random dither are also listed in the same table.

An interesting simulation is the use of a bandpass filter instead of a bandstop filter to generate the dither. This type of dither has bigger frequency components falling within the pass band of the section to be stabilized. Let us call it bandpass dither. In contrast with the bandstop dither, it is expected that when bandpass dither is used the transition time to the origin should be increased very much. In the simulation, the fourth-order elliptic passband

TABLE 4

The median values of the transition time to the origin state corresponding to various filter sections and dithers

-3dB frequencies of filter section	Median value of the transition time to the origin state					
	Uniform Dither	Binary Dither	Second-order Butterworth Bandstop Dither 1	Fourth-order Butterworth Bandstop Dither 1	Fourth-order Elliptic Bandstop Dither 1	Fourth-order Elliptic Passband Dither 1
0.10, 0.11	164	86	140	98	110	
0.10, 0.12	68	45	54	50	52	1930
0.20, 0.21	145	82	115	107	101	
0.30, 0.31	154	98	132	110	115	8610
0.30, 0.32	67	53	67	53	58	
0.40, 0.41	223	110	180	149	152	7650
0.40, 0.42	93	65	87	72	74	
0.42, 0.43	240	125	178	155	158	

Note: Frequencies as multiples of the sampling frequency. Times as multiples of the sampling period.
The stop band width of the bandstop filter is equal to the pass band width of the second-order section to be stabilized. The initial state is (11,11).

filters were used. The experimental results have also been listed in the Table 4. As can be seen from these data, when bandpass dither signals were used, the transition times to the origin were increased by several ten times of that when the uniformly distributed dither signals were used. It is worth noting that the limit cycles in the second-order filters still can be suppressed by the use of the bandpass dither although the transition time to the origin is much longer.

It can be expected that when the input signal is not zero, the extra noise by the bandpass dither would be much greater. Therefore, we do not think that the bandpass dither is a useful dither signal. But these simulations have verified, from other points of view, the idea of the bandstop dither.

In order to find out the "optimum" stop band width of the bandstop filter from the shortest transition time point of view, various stop band widths have been used. Table 5 shows the median transition times with various bandstop dithers. As can be seen from the table, the "optimum" stop band width is equal to three (or possibly four) times pass band width of the filter section to be stabilized.

4. Summary About the Time for Stabilization

From the simulation, as far as the transition time to

TABLE 5

The median transition times to the origin with various stop band widths of the bandstop filter in bandstop dither 1 generation

-3dB frequencies of filter section	Median transition time to the origin			
	$W_s = W_p^*$	$W_s = 2W_p$	$W_s = 3W_p$	$W_s = 5W_p$
0.1, 0.11	110	98	95	97
0.1, 0.12	52	47	46	48
0.2, 0.21	101	94	98	93
0.3, 0.31	115	108	108	104
0.3, 0.32	58	58	58	55
0.4, 0.41	152	120	124	123
0.4, 0.42	74	67	68	64
0.42, 0.43	158	149	140	144
0.1, 0.101	1550	1375	1233	1250
0.2, 0.201	1060	1160	950	963
0.3, 0.301	1217	1163	1067	1150
0.4, 0.401	2300	1680	1450	1625

*Note: Frequencies as multiples of the sampling frequency.
 Time as multiples of the sampling period
 W_s represents the stop band width of the fourth-order elliptic bandstop filter in the bandstop dither 1 generation.
 W_p represents the pass band width of the filter section to be stabilized.
 The initial state is (11,11).

the origin state is concerned, the following conclusions can be obtained.

(A) The Shortest Time Needed for the Stabilization of Second-order Filter Section can be Obtained by the Use of the Binary Random Dither

As can be seen from the data in Table 4, for high-Q filter section, the improvement of the median transition times to the origin is outstanding. This is very desirable because in this case the transition time is long. For example, when the pass band of the filter section to be stabilized is from $0.4F_s$ to $0.41F_s$ where F_s is the sampling frequency, the transition time to the origin is reduced from $223 T_s$ with uniformly distributed random dither to $110 T_s$ with the binary random dither. But for low-Q filter section, the improvement of the median transition time is not apparent, in some cases even no improvement at all. For example, when the filter section with pass band $(0.3-0.38)F_s$ is used, no improvement has been found. (Both transition times are $17T_s$).

Practically, in low-Q case, the improvement of the transition time is not necessary because even with uniformly distributed random dither the transition time is short. For the example above, the transition time is only $17T_s$.

Hence, in high-Q filter section for a quick stabilization

the binary random dither is recommended.

- (B) For High-Q Filter Sections, the Transition Times with Bandstop Dither may be Near but in General, still Longer than that with the Binary Random Dither.

For example, when the filter section with pass band $(0.3-0.31)F_s$ is used, the median transition times with binary random dither and bandstop dither are respectively equal to $98T_s$ and $110T_s$. The difference is only 11%. As will be seen later, for mean transition time the difference is even smaller.

The disadvantage of the bandstop dither is the complication in generation. But as will be seen in the next section, its advantage is that it causes smaller extra output noise in the non-zero input condition.

As far as the transition time is concerned, for low-Q filter section, the uniformly distributed random dither is preferred because it is simple to generate and the transition time to the origin is short enough, for example less than $50 T_s$. For medium-Q basic section, the transition time with uniformly distributed dither is also medium, for example from $50 T_s$ to $100 T_s$. For this case, the second-order bandstop dither (in the generation a second-order bandstop filter is used) is recommended because higher-order cannot offer extra improvement. For high-Q basic section, the transition time with uniformly distributed

random dither is long, for example longer than $100 T_s$. In this case, the fourth-order bandstop dither signal is recommended because it can shorten the transition time further.

(C) From the View Point of the Shortest Transition Time to the Origin, the Recommended Stop Band Width is Equal to Three or Four Times the Pass Band Width of the Second-order Basic Filter Section in the Bandstop Dither Generation.

As can be seen from Table 5, when the bandstop dither signal is used, the transition time to the origin with a bandstop filter whose stop band width is equal to three times pass band width of the basic section to be stabilised is, in general, the shortest. For a medium-Q basic section with bandstop dither, when the stop band width is equal to or greater than two times pass band width of the basic section the difference of the transition time is not very apparent but the improvement does exist. Therefore, to put it briefly, a bandstop filter whose stop band width is equal to three times pass band width of the basic section is recommended.

In the next section, another important aspect of the experimental results in the dither application, i.e., the effect of dither on the output noise when the input signal is applied, will be presented.

7.5 THE EFFECT OF DITHER ON THE OUTPUT NOISE

Although when input is zero and the limit cycle has been suppressed the dither has no influence on the output from the filter, when the input is nonzero, the dither should be considered as an additive noise. If the amplitude of the input signal is much bigger than the quantization step, then the following assumptions are true.

- (1) Any two different samples from the same noise source are uncorrelated.
- (2) Any two different noise sources (i.e., associated with different multipliers) regarded as random process, are uncorrelated.
- (3) Each quantization noise and the dither are uncorrelated with the input sequence.

Thus each quantization noise source is modelled as a discrete stationary white random process with a uniform power density spectrum of $\frac{q^2}{12}$. If the uniformly distributed random dither is used, then the dither can also be treated like a quantization noise. For the other two dither signals, they can be treated as independent random noise sequence but have different power spectrums. The measurement shows that the power spectrum of the binary random dither still approximates uniform. Whatever dither signals are used, at output terminal each quantization noise power and dither

power can be added independently.

Suppose that there are P noise sources including the dither and quantization noise. Consider the Kth noise source $e_K(n)$. Let $h_K(n)$ be the impulse response from the noise source to the filter output. The output noise components $E_K(n)$, due solely to $e_K(n)$, may be obtained via convolution as

$$E_K(n) = \sum_{m=0}^n h_K(m) \bar{e}_K(n-m) \quad (128)$$

The variance of $E_K(n)$ may be obtained as

$$\begin{aligned} \sigma_{OK}^2(n) &= E \left[\sum_{m=0}^n h_K(m) e_K(n-m) \sum_{\ell=0}^n h_K(\ell) e_K(n-\ell) \right] \\ &= \sum_{m=0}^n \sum_{\ell=0}^n h_K(m) h_K(\ell) E [e_K(n-m) e_K(n-\ell)] \\ &= \sum_{m=0}^n \sum_{\ell=0}^n h_K(m) h_K(\ell) \delta(\ell-m) \sigma_e^2 \end{aligned}$$

or

$$\sigma_{OK}^2 = \sigma_e^2 \sum_{m=0}^n h_K^2(m) \quad (129)$$

where σ_e is the variance of input noise if the input noise is uniformly distributed in $(-\frac{q}{2}, \frac{q}{2})$ then $\sigma_e^2 = \frac{q^2}{12}$. In the limit, as n tends to infinity, the variance $\sigma_{OK}^2(n)$ tends to the steady-state limit.

$$\sigma_{OK}^2 = \sigma_e^2 \sum_{m=0}^{\infty} h_K^2(m) \quad (130)$$

The total steady-state noise variance σ_O^2 is then

$$\sigma_O^2 = \sum_{K=1}^P \sigma_{OK}^2 \quad (131)$$

Similarly, because each noise is independent with the input signal, the noise and signal power can be added independently at the output terminal.

To determine the effect of dither on the output noise, the filter was simulated with a sinusoidal input at a frequency close to its resonance and with an amplitude corresponding to full use of a ten-bit wordlength in the output.

The procedures in the simulation are as follows:

First, no dither was added. The amplitude of sinusoidal input signal was chosen such that ten-bit wordlength was fully used by the output sinusoidal signal. This sinusoid was applied to the second-order section with one quantizer. The power spectrum of the output from the filter was estimated by use of the FFT, with a Hamming Window. The Hamming Window is of the form

$$W_H(n) = \begin{cases} 0.54 + 0.46 \cos\left(\frac{2\pi n}{N}\right) & -\left(\frac{N-1}{2}\right) \leq n \leq \frac{N-1}{2} \\ 0.0 & \text{elsewhere} \end{cases} \quad (132)$$

Fig. 62 shows the frequency response of a Hamming Window. In the simulation, the total number of points, N , was equal to 2048. Although, the overall frequency response of the Hamming Window appears to have no ripples beyond $\omega = \frac{4\pi}{N}$, this is not the case. On the linear amplitude scale of Fig. 62, however, the ripples are not visible. The main lobe of the frequency response of the Hamming Window is twice the width of the main lobe of the frequency response with the rectangular window. Because the frequency of the input sinusoidal signal was just the integer multiples of the $\frac{F_S}{N}$, there were three points on the spectrum corresponding to the frequency of the sinusoid. These three points were used to measure the signal. There were, of course, some noise components included in these three points, but because the signal component was much bigger than the noise components and the pass band of the filter section included much more spectrum points, the error of measurement was small. The other points on the spectrum were used to provide an estimate of the mean square value of the noise. This mean square value of the noise was considered as the quantization noise and treated as a reference power level i.e., 0 dB.

Then, each type of dither signal was added respectively. The input signal was the same with above. In each estimation of power spectrum, the FFT were used thirty times successively and the average values were considered as the estimate of

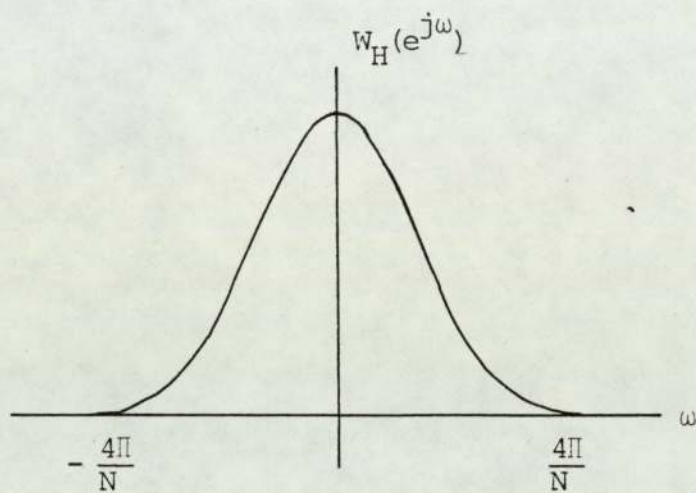


Fig. 62 Frequency response of a Hamming Window.

the power spectrum. The increase in the unit of dB was interesting. The program used is shown in Appendix 11.

The increase in output noise due to the uniform dither and the binary dither could easily be calculated if it could be assumed that the effect of the filter quantization is equivalent to the addition of a random error whose mean square value is $\frac{q^2}{12}$ and if nonlinear effects could be ignored, uniform random dither would then produce an increase in output noise of 3 dB and binary random dither would produce an increase of 6 dB. In reality, nonlinear effects cannot be ignored and the increase in noise will be somewhat different.

Table 6 shows the results for a number of second-order sections, corresponding to a variety of filter centre frequencies and bandwidth. In this table, two types of bandstop dither were used. The experimental results have verified the expectation mentioned before, i.e., bandstop dither 1 results in a quicker limit cycle suppression but it causes a bit bigger increase in the output noise. In contrast with this, bandstop dither 2 causes a bit smaller output noise increase but it needs longer time to suppress the zero-input limit cycles in the second-order filter sections. The difference between the two increases in output noise is less than 1 dB.

As can be seen from the table, the increase in output noise varies from case to case, but the binary random dither

TABLE 6

The mean time for stabilization and the increase in output noise with sinusoidal input for various filter sections with dither

-3dB frequencies of filter section	Mean time for stabilization (Standard error in parenthesis)				Decay time for linear filter	dB increase in output noise with sinusoidal input (Standard error in parenthesis)			
	Uniform random dither	Binary random dither	Bandstop dither 1	Bandstop dither 2		Uniform random dither	Binary random dither	Bandstop dither 1	Bandstop dither 2
0.07, 0.08	514 (15)	159 (4)	162 (3)	319 (9)	146	2.96 (0.19)	5.89 (0.15)	1.57 (0.13)	0.87 (0.13)
0.1, 0.11	216 (5)	93 (2)	96 (2)	151 (3)	87	1.65 (0.12)	4.56 (0.14)	0.42 (0.09)	-0.22 (0.11)
0.1, 0.12	62 (2)	32 (1)	31 (1)	50 (1)	28	7.09 (0.1)	9.90 (0.1)	5.74 (0.07)	5.49 (0.08)
0.2, 0.21	133 (4)	40 (1)	59 (1)	129 (2)	57	3.48 (0.11)	6.50 (0.12)	2.54 (0.08)	2.02 (0.12)
0.3, 0.31	144 (3)	67 (1)	72 (1)	142 (2)	70	5.57 (0.08)	8.16 (0.14)	4.19 (0.10)	3.74 (0.11)
0.4, 0.41	198 (6)	48 (2)	72 (2)	194 (4)	64	3.22 (0.13)	6.04 (0.14)	1.87 (0.11)	1.70 (0.09)
0.42, 0.43	247 (7)	71 (2)	97 (2)	230 (5)	75	2.93 (0.15)	5.83 (0.09)	1.63 (0.10)	1.32 (0.1)
0.2, 0.201	8032 (185)	1281 (16)	1731 (28)*	5225 (289)*	1669	2.41 (0.34)	5.15 (0.26)	1.12 (0.20)*	1.02 (0.25)*

Note: Frequencies as multiples of the sampling frequency.

Times as multiples of the sampling period.

* stop band width = (0.19-0.22).

consistently results in an increase in noise nearly 3 dB more than the increase with uniform random dither. The increase in output noise with bandstop dither is about 4 dB less than with the binary random dither. This is an encouraging result. As can be expected, the experimental results show that when the bandstop dither is used, the increase in the output noise is smallest. The increase in output noise with the bandstop dither is about (1.0~1.5) dB less than with uniformly distributed random dither. The biggest increase in the output noise appears when the binary random dither is used. But our simulations show that the practical increase in the output noise is less than that would be expected due to addition in a linear filter as a result of the nonlinear effects of quantization.

As a summary of the use of dither signals, the mean times for stabilization corresponding to different filters with different dither signals are also shown in Table 6. In each case, an initial state on the largest amplitude limit cycle was used. The simulation was repeated 1000 times, enabling an accurate estimation of the mean time for stabilization to be made. The standard errors of the mean times (the definition of this term will be given later) are also shown in the table together with the mean times.

The mean of a sample is a point estimate of the mean of the parent population. As mentioned earlier, the

distribution of transition time to the origin is approximately normal. But we know neither the mean value, μ , nor the standard deviation of the parent population, σ . From the theory of the sampling and estimation in statistics⁽⁴³⁾, we know that for large samples (e.g., $n \geq 30$) the standard deviation of the sample, s , may be used as an approximation to the standard deviation of the population σ . The sampling distribution of means of samples of size n from a population which is $N(\mu, \sigma^2)$ is $N(\mu, \sigma_{\bar{X}}^2)$ where $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ called the standard error of the mean. In our estimations, the mean of the time for stabilization \bar{X} and the standard deviation of the sample s could be obtained by calculating from the transition time distribution. The sample size, n , was equal to 1000. Hence it is accurate enough to use the standard deviation of the sample, s , as an approximation to the standard deviation of the population, σ , i.e., the standard error $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$ can be obtained. As well known for a normal distribution, the probabilities that a sample mean will lie between the limits $\mu - 1.96\sigma_{\bar{X}}$ and $\mu + 1.96\sigma_{\bar{X}}$ or $\mu - 2.57\sigma_{\bar{X}}$ and $\mu + 2.57\sigma_{\bar{X}}$ are 0.95 or 0.99 respectively. In other words, we can be 95 or 99 percent certain that a sample mean will not differ from the population mean by more than $1.96\sigma_{\bar{X}}$ or $2.57\sigma_{\bar{X}}$. If the mean value of a sample, \bar{X} , lies within the limits $\mu - 1.96\sigma_{\bar{X}}$ and $\mu + 1.96\sigma_{\bar{X}}$ (or $\mu - 2.57\sigma_{\bar{X}}$ and $\mu + 2.57\sigma_{\bar{X}}$) the mean value of the population, μ , must lie within the limits $\bar{X} - 1.96\sigma_{\bar{X}}$ and $\bar{X} + 1.96\sigma_{\bar{X}}$ (or $\bar{X} - 2.57\sigma_{\bar{X}}$ and $\bar{X} + 2.57\sigma_{\bar{X}}$), (see Fig. 63). Therefore, we can

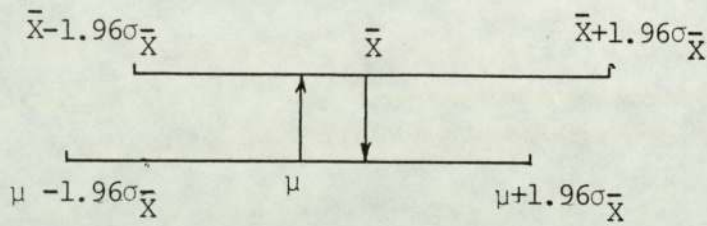


Fig. 63 If \bar{X} lies within the limits $\mu - 1.96\sigma_{\bar{X}}$ and $\mu + 1.96\sigma_{\bar{X}}$ then μ must lie within the limits $\bar{X} - 1.96\sigma_{\bar{X}}$ and $\bar{X} + 1.96\sigma_{\bar{X}}$.

be 95 (or 99) percent sure that the mean value of the population, μ , lies within the limits $\bar{X} - 1.96\sigma_{\bar{X}}$ and $\bar{X} + 1.96\sigma_{\bar{X}}$ (or $\bar{X} - 2.57\sigma_{\bar{X}}$ and $\bar{X} + 2.57\sigma_{\bar{X}}$). As can be seen from Table 6, the standard errors are much smaller than the mean \bar{X} , i.e., the estimations are accurate.

To provide a comparison for the transition time to the origin, the table also shows the time for the response of a filter without quantization to decay below the amplitude $\frac{q}{2}$ from the same initial state.

The results in this table show clearly that the binary random dither and the bandstop dither result in more rapid stabilization than uniform dither. The mean times for the binary random dither and the bandstop dither to effect stabilization are comparable with the decay time for the strictly linear filter.

7.6 SUMMARY

In order to verify the proposed methods for limit cycle suppression extensive simulations have been carried out. The main results of the simulations have been presented in this chapter.

Three types of dither signal: uniform random dither, binary random dither and bandstop dither have been obtained by the simulations. The use of these dither

signals for limit cycle suppression in a variety of filter sections have been studied experimentally.

When dither is used to stabilize a digital filter, two specifications are of particular interest. One of these is the length of time taken for the filter, with zero input, to reach the state plane origin from a limit cycle. The other specification of interest is the increase in the output noise from the filter above the quantization noise which is present when nonzero input signals are applied without dither.

Several different methods to measure the time for stabilization by the use of dither have been used: the average transition time from each initial state within the amplitude bound zone of the limit cycle in the state plane to the origin; the median and mean of the transition time from same initial state to the origin in 1000 times simulations. The time for the response of the linear version of the digital filter section (without quantization) to decay below the threshold of $\frac{q}{2}$ from the same initial state has been used as a reference time in the comparison of the transition time to the origin. As far as this transition time is concerned, in the three types of dither, the order of preference is the binary random dither, bandstop dither and uniform random dither signals. The mean times for the former two types of dither to affect stabilization are comparable with the decay time for the

strictly linear filter.

To determine the effect of dither on the output noise, the filter was simulated with a sinusoidal input at a frequency close to its resonance and with an output amplitude corresponding to full use of a ten-bit wordlength. The power spectrum of the output from the filter was estimated by use of the FFT, with a Hamming Window. The points on the spectrum corresponding to the frequency of the sinusoid were used to measure the signal power. The other points on the spectrum were used to provide an estimate of the mean square value of the noise. The quantization noise with the same signal but without dither is used as a reference of the noise output. The increase in output noise above the quantization noise is interesting.

As far as the increase in output noise is concerned in the three types of dither, the order of preference is the bandstop dither, the uniform random dither and the binary random dither.

The main conclusions about this research will be discussed in the next chapter.

CHAPTER 8

CONCLUSIONS AND SUGGESTIONS

FOR FUTHER RESEARCH

The purpose of this research has been to contribute methods of limit cycle oscillation suppression in the second-order filter sections by the use of somewhat different dither signals which have no disadvantages of the methods proposed before.

Despite the fact that these limit cycles can be made arbitrarily small by increasing the number of significant digits of the data, in practice, this increase is not desirable because it increases the cost and complexity of the filter.

When input signal is zero, only limit cycle oscillation exists. Because limit cycles are correlated noise it is more harmful than normal noise. Especially, when these integrated circuits need to be used for applications where limit cycles are not tolerable, an effective method to suppress limit cycle is very necessary.

The major contributions of this research are now summarized. In Chapter 3, two new existence conditions of constant and alternating limit cycles in the second-order filter section with one rounding quantizer have been derived on the experimental basis. In Chapter 4, the mechanism of

the limit cycle suppression by the injection of dither has been explained qualitatively. From the statistical average point of view, the injection of dither tends to linearize the nonlinear characteristic of rounding quantizer. In Chapter 5, the necessity of the limit cycle suppression in the second-order filter sections has been proved partly on the experimental basis. It has been shown that for any particular second-order digital filter section, i.e., one with specified coefficient values of A and B, it is straightforward to determine rigorously whether or not dither will suppress all limit cycles. By the transition matrix, the maximum transition time needed for transition from any limit cycle to the origin state on the state plane can be calculated. In Chapter 6, by checking the characteristic of the equivalent quantizer, the principal considerations of the dither signal design have been described. Three types of dither signal have been proposed. All the limit cycles in the second-order filter sections can be suppressed by the use of any one of the three dithers. Once the zero-input limit cycle has been suppressed, the output from the filter remains zero, i.e., no remaining noise left at all. Extensive simulations have been done. In none of the filter sections examined has dither, of the forms described here, failed to suppress limit cycles. The limit cycle suppression by the injection of the three dither signals were verified. The experimental results have been presented in Chapter 7. These results show

clearly that binary random dither results in more rapid stabilization than uniform and bandstop dither signals. The mean times for binary random dither and bandstop dither to affect stabilization are comparable with the decay time for the strictly linear filter. The increase in output noise varies from case to case, but binary random dither consistently results in an increase in noise, nearly 3 dB more than the increase with uniform random dither. This is expected by the fact that the variance of binary dither is three times bigger than that of uniform dither. The results with bandstop dither are encouraging. From the experimental results, it is apparent that bandstop dither has combined the advantages of binary dither and uniform dither. The stabilization times with bandstop dither are similar to those with binary dither which, as mentioned above, are themselves similar to the decay times of the filter without quantizer. The increase by the bandstop dither in output noise is, in each case, (1~1.5) dB smaller than the increase by uniform dither. Two types of bandstop dither have been used. By the injection of bandstop dither 1, the time needed for the stabilization of filter section can be reduced further but it will cause a bit more increase in output noise than that caused by bandstop dither 2. The increase in output noise due to bandstop dither 2 is smallest, but the time for the stabilization of filter section is longer than that when bandstop dither 1 is used. The difference between

two increases in output noise is less than 1 dB. The significance is that if the nonlinear function $\frac{1}{2}\text{erf}(CY_1)$ is used where C is a positive constant, C can be used as a control coefficient. By changing the value of C, the trade-off between transition time and output noise increase is possible. It can be expected when C is much greater than 1 the dither would have a similar performance with binary dither, and when C is less than $\frac{1}{\sqrt{2}}$ the transition time would be long though the output noise increase might be slightly smaller.

Each dither signal has its own advantages and disadvantages. We believe that each dither signal can find its own application. For uniformly distributed dither, it is easy to be generated and has smaller noise increase in the output but it needs a long time to stabilize the filter sections. Binary random dither can stabilize the filter section quickly, and it is still easy to be generated, but has a larger noise increase in the output. For bandstop dither, it can suppress the limit cycles quickly and has very small noise increase in the output but it is complicated to be obtained. However, bandstop dither can be applied to the system where many identical digital filters have been included already. In this case, the extra complexity need not be significant.

There remain several unanswered problems which have arisen as a result of this research and should be noted for

further investigation. Among them are the following:

(1) It is not clear whether there would be any advantage in achieving lower spectral levels within the stop band than those which have been used here. The results given, show that the forms of bandstop dither used in the experiments already result in only a small increase in output noise and the limit cycles disappear about as quickly as the normal resonant decay of a linear filter without quantization.

(2) Simpler methods for generating bandstop dither than described in this paper may exist. For example, Steiglitz's Markov process scheme for generating signals with a required spectrum⁽⁴⁴⁾ might be applicable. It could be possible to use bandstop dither signals which have been synthesised by computer and stored in ROM (read only memory).

(3) It would be interesting to find a method by which a binary dither signal with an approximate bandstop spectrum could be generated.

APPENDIX 1

Program used with Computer PET to display the
procedure of limit cycle suppression on
the state plane

APPENDIX 2

The existence conditions of limit cycles in the second-order basic filter section with one rounding quantizer

1. The existence condition of constant limit cycles

Suppose that a constant limit cycle with magnitude c exists. Then Eqn (58) becomes

$$\begin{aligned} C &= [-AC-BC]_R \\ &= C(-A-B) - \delta \end{aligned} \tag{A2.1}$$

where $|\delta| < 0.5$, δ is the quantization error.

According to the linear stable condition, Eqn (38), we know that

$$-A-B < 1$$

$$\text{Let } -A-B = 1-\beta \tag{A2.2}$$

where $\beta > 0$.

Hence, Eqn (A2.1) can be written as

$$C = C(1-\beta) - \delta$$

or

$$\delta = -C\beta = -C(1+A+B) \tag{A2.3}$$

Eqn (A2.3) shows that δ has the opposite sign as that of C because $\beta > 0$ and $|\delta|$ increases with $|C|$ proportionally. The minimum of $|C|$ satisfies

$$|C_{\min}| = 1$$

And the maximum of $|\delta|$ satisfies

$$|\delta|_{\max} = 0.5$$

Therefore, the ratio of δ_{\max}/C_{\min} is negative and satisfies

$$\frac{\delta_{\max}}{C_{\min}} = -0.5$$

From Eqn (A2.3) the existence condition of the constant limit cycles can be obtained as

$$1+A+B \leq \frac{-\delta_{\max}}{C_{\min}} = 0.5$$

or

$$A+B+0.5 \leq 0 \tag{A2.4}$$

The equation

$$A+B+0.5 = 0 \tag{A2.5}$$

describes the boundary line GN in Fig. 19.

In Eqn (A2.3), the constant C must be integer and the maximum of δ is 0.5, therefore, the amplitude bound of constant limit cycles can be derived as

$$B_C = C|_{\max} = \text{INT} \left(\frac{0.5}{1+A+B} \right) \quad (\text{A2.6})$$

Eqn (A2.6) is in unit of the quantization step, q .

It is clear from Eqn (A2.3) that because C_{\max} satisfies

$$C_{\max}(1+A+B) \leq 0.5 \quad (\text{A2.7})$$

all integers of C whose absolute values are less than C_{\max} also satisfy

$$C(1+A+B) \leq 0.5 \quad (\text{A2.8})$$

No integer which is greater than C_{\max} satisfies Eqn (A2.8).

In other words, all constant limit cycles in the second-order filter section are successive in magnitude, i.e., they must be $\pm 1, \pm 2, \pm 3, \dots, \pm C_{\max}$.

2. The existence condition of alternating limit cycles

Along the same way as above, it is readily verified that the existence condition of alternating limit cycles is

$$-A+B+0.5 \leq 0 \quad (\text{A2.9})$$

The equation

$$-A+B+0.5 = 0 \quad (\text{A2.10})$$

describes the boundary line HM in Fig. 19.

The amplitude bound of alternating limit cycles can

be written as

$$B_a = \text{INT} \left(\frac{0.5}{1-A+B} \right) \quad (\text{A2.11})$$

3. The existence condition of periodic limit cycles

Claasen et al⁽³⁷⁾ have proved that the second-order digital filters with coefficient B for which $|B| > 0.5$ will always exist limit cycles. Therefore, the existence condition of periodic limit cycles can be written as

$$|B| \geq 0.5 \quad (\text{A2.12})$$

The equations

$$B = \pm 0.5 \quad (\text{A2.13})$$

describe the boundary lines IL and RS in Fig. 19.

From the amplitude bound of periodic limit cycles proposed by Jackson⁽⁹⁾

$$B_p = \text{INT} \left(\frac{0.5}{1-B} \right)$$

the existence condition, Eqn (A2.12) can be also obtained because

$$B_p |_{\min} = 1.$$

4. Two extra boundary lines

It is possible that in some regions on the parameter

plane both constant (or alternating) and periodic limit cycles exist simultaneously. Let the amplitude bound of constant limit cycle be equal to that of periodic limit cycle we obtain

$$\text{INT} \left(\frac{0.5}{1-B} \right) = \text{INT} \left(\frac{0.5}{1+A+B} \right) \quad (\text{A2.14})$$

Let the amplitude bound of alternating limit cycle be equal to that of periodic limit cycle we obtain

$$\text{INT} \left(\frac{0.5}{1-B} \right) = \text{INT} \left(\frac{0.5}{1-A+B} \right) \quad (\text{A2.15})$$

From Eqn (A2.14), a new boundary line can be derived.

$$1-B = 1+A+B$$

or

$$A = -2B \quad (\text{A2.16})$$

Similarly, from Eqn (A2.15) we obtain

$$1-B = 1-A+B$$

or

$$A = 2B \quad (\text{A2.17})$$

Eqn (A2.16) and Eqn (A2.17) define the two boundary lines DJ and KE in Fig. 19, respectively.

In the regions DIJ and KLE on the parameter plane of Fig. 19, the amplitude bounds of periodic limit cycles are less than that of constant and alternating limit cycles respectively. The experiments found that for the filter

with only one rounding quantizer, the periodic limit cycle trajectories on the state plane surround all the constant or alternating limit cycles. In other words, the amplitude bound of periodic limit cycle is always greater than that of the constant or alternating limit cycle. This observation asserts that in the regions DIJ and KLE the periodic limit cycle do not exist.

APPENDIX 3

When the coefficient value B of the second-order filter section satisfies $1 > |B| > 0.5$, the origin state (0,0) on the state plane becomes a branch point by the use of dither

Refer to Fig. 14. Three cases can be defined.

In case 1, the second-order filter section has one rounding quantizer one dither as shown in Fig. 14(a). The following equation is satisfied

$$\hat{Y}(n) = [-A\hat{Y}(n-1) - B\hat{Y}(n-2) + D]_R \quad (\text{A3.1})$$

In case 2, the section has two rounding quantizers and two dither signals. The difference equation can be written as

$$\hat{Y}(n) = [-A\hat{Y}(n-1) + D]_R + [-B\hat{Y}(n-2) + D]_R \quad (\text{A3.2})$$

In case 3, the section has two rounding quantizers and one dither signal added at the front of coefficient B product quantizer. The difference equation is

$$\hat{Y}(n) = [-A\hat{Y}(n-1)]_R + [-B\hat{Y}(n-2) + D]_R \quad (\text{A3.3})$$

Apparently, for any cases above, if $\hat{Y}(n-1) = \hat{Y}(n-2) = 0$ then

$\hat{Y}(n)=0$. Hence, the origin state (0,0) is a stationary point, one of its predecessors is (0,0) itself. For the origin state (0,0) to be a branch point, there must be at least another predecessor (0,K), where the integer $K \neq 0$.

Suppose the filter section is at the state $(\hat{Y}(n-1), \hat{Y}(n-2)) = (0, K)$.

For case 1, the difference equation is

$$\begin{aligned}\hat{Y}(n) &= [-A\hat{Y}(n-1) - B\hat{Y}(n-2) + D]_R \\ &= [-BK + D]_R\end{aligned}\tag{A3.4}$$

For case 2,

$$\begin{aligned}\hat{Y}(n) &= [-A\hat{Y}(n-1) + D]_R + [-B\hat{Y}(n-2) + D]_R \\ &= [D]_R + [-BK + D]_R \\ &= [-BK + D]_R\end{aligned}\tag{A3.5}$$

For case 3,

$$\begin{aligned}\hat{Y}(n) &= [-A\hat{Y}(n-1)]_R + [-B\hat{Y}(n-2) + D]_R \\ &= [-BK + D]_R\end{aligned}\tag{A3.6}$$

As can be seen, above three cases lead to the same expression. Since the dither D is distributed in the open range $(-\frac{q}{2}, \frac{q}{2})$ when $1 > |B| > 0.5$, and $K = \pm 1$, the probability

of satisfying the equation

$$\begin{aligned}\hat{Y}(n) &= [-BK+D]_R \\ &= 0\end{aligned}\tag{A3.7}$$

is nonzero. In other words, when the random dither which distributes in the open range $(-\frac{q}{2}, \frac{q}{2})$ is used the states $(\hat{Y}(n-1), \hat{Y}(n-2)) = (0, \pm 1)$ may be the predecessors of the origin state $(0,0)$, i.e., the origin state becomes a branch point

In case 4, the filter section has two quantizers but the dither is added at the front of the coefficient A product quantizer. Its difference equation can be written as

$$\hat{Y}(n) = [-A\hat{Y}(n-1)+D]_R + [-B\hat{Y}(n-2)]_R\tag{A3.8}$$

At state $(0,K)$, the Eqn. (A3.8) becomes

$$\begin{aligned}\hat{Y}(n) &= [D]_R + [-BK]_R \\ &= [-BK]_R\end{aligned}\tag{A3.9}$$

Apparently, when $1 > |B| \geq 0.5$ and K is a nonzero integer

$$\begin{aligned}\hat{Y}(n) &= [-BK]_R \\ &\neq 0\end{aligned}\tag{A3.10}$$

In this case, the origin state is not a branch state. In

other words, the limit cycles cannot be suppressed by the use of the random dither when the dither is added at the front of A coefficient product quantizer. The simulation has verified this conclusion.

APPENDIX 4

The uniformly distributed random dither tends to linearize the roundoff quantization characteristic

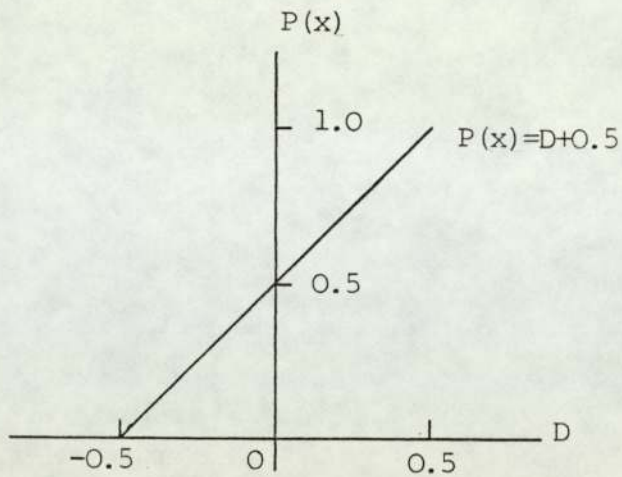
From the probability density function of uniform dither, its cumulative probability distribution can be readily obtained as shown in Fig. A4.1(a).

Suppose that the input signal of a rounding quantizer without dither is

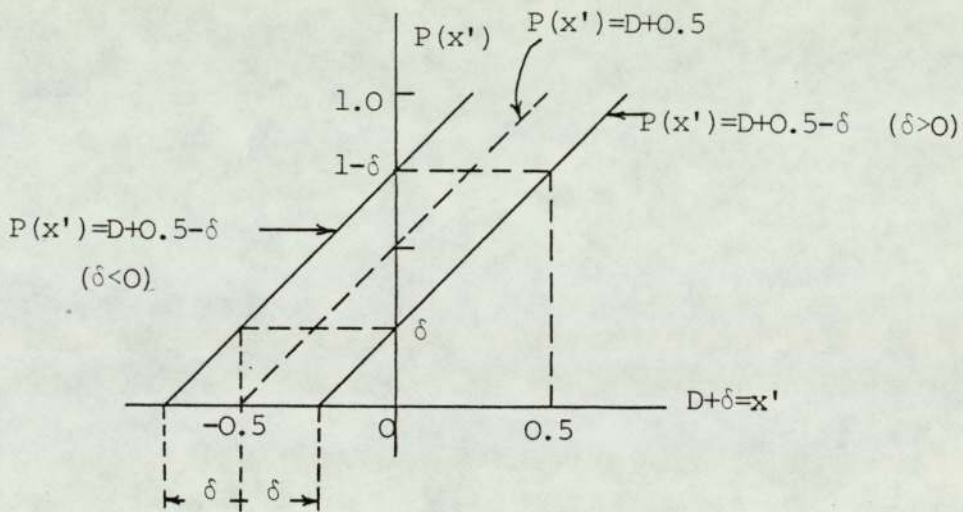
$$-A\hat{Y}(n-1) - B\hat{Y}(n-2) = \hat{Y}(n) + \delta \quad (\text{A4.1})$$

where the quantization error $|\delta| < 0.5$ and $\hat{Y}(n) = [-A\hat{Y}(n-1) - B\hat{Y}(n-2)]_R$. After adding dither, the composite signal $\hat{Y}(n) + D + \delta$ has a cumulative probability distribution as shown in Fig. A4.1(b). It is clear from this figure that when $\delta > 0$ the probability of the composite signal $\hat{Y}(n) + D + \delta$ being less than $\hat{Y}(n) + 0.5$ is $(1 - \delta)$ or the probability of the composite signal $\hat{Y}(n) + D + \delta$ being equal or greater than $\hat{Y}(n) + 0.5$ is δ , when $\delta < 0$ the probability of the composite signal $\hat{Y}(n) + D + \delta$ being less than $\hat{Y}(n) - 0.5$ is $|\delta|$ or the probability of $\hat{Y}(n) + D + \delta$ being equal or greater than $\hat{Y}(n) - 0.5$ is $1 - |\delta|$.

Now both the dither adder and the rounding quantizer itself can be treated as an equivalent quantizer. Its input signal is $\hat{Y}(n) + \delta$ and the output signal is



(a)



(b)

Fig. A4.1(a) Cumulative probability distribution of the uniform dither.
 (b) Cumulative probability distribution of the composite signal (dither and quantization error).

$[\hat{Y}(n)+\delta+D]_R$ as shown in Fig. A4.2 .

When $\delta > 0$, there are two possibilities.

- (1) If the composite signal $\hat{Y}(n)+\delta+D \geq \hat{Y}(n)+0.5$, i.e., $\delta+D \geq 0.5$ the equivalent quantizer Q_e outputs $\hat{Y}(n)+1$ and the corresponding probability is δ .
- (2) If the composite signal $\hat{Y}(n)+\delta+D < \hat{Y}(n)+0.5$, i.e., $\delta+D < 0.5$ the equivalent quantizer Q_e outputs $\hat{Y}(n)$ and the corresponding probability is $1-\delta$.

The mean value of the output is

$$\begin{aligned} M &= [\hat{Y}(n)+1] \delta + \hat{Y}(n) (1-\delta) \\ &= \hat{Y}(n) + \delta \end{aligned} \tag{A4.2}$$

Similarly, when $\delta < 0$, the mean value of the output is

$$\begin{aligned} M &= [\hat{Y}(n)-1] |\delta| + \hat{Y}(n) (1-|\delta|) \\ &= \hat{Y}(n) - |\delta| \\ &= \hat{Y}(n) + \delta \end{aligned} \tag{A4.3}$$

It is clear from Eqn (A4.2) and Eqn (A4.3) that when the uniform dither is used, the mean value of the equivalent quantizer output is varied linearly with the input of the equivalent quantizer, $\hat{Y}(n)+\delta$. Fig. A4.3(b) shows the statistical characteristic (mean value output versus input) of the equivalent quantizer. For comparing, Fig. A4.3(a) shows the characteristic with roundoff and possible output values after adding dither and rounding operation.

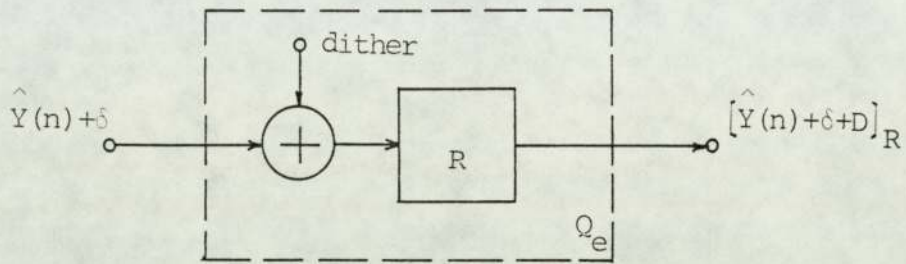


Fig. A4.2 Equivalent quantizer.

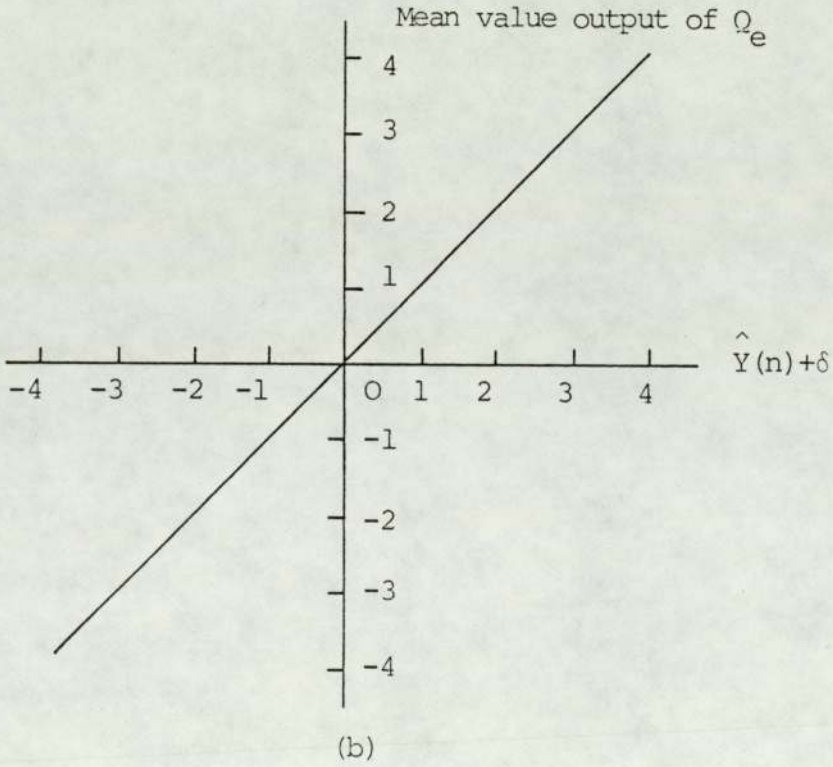
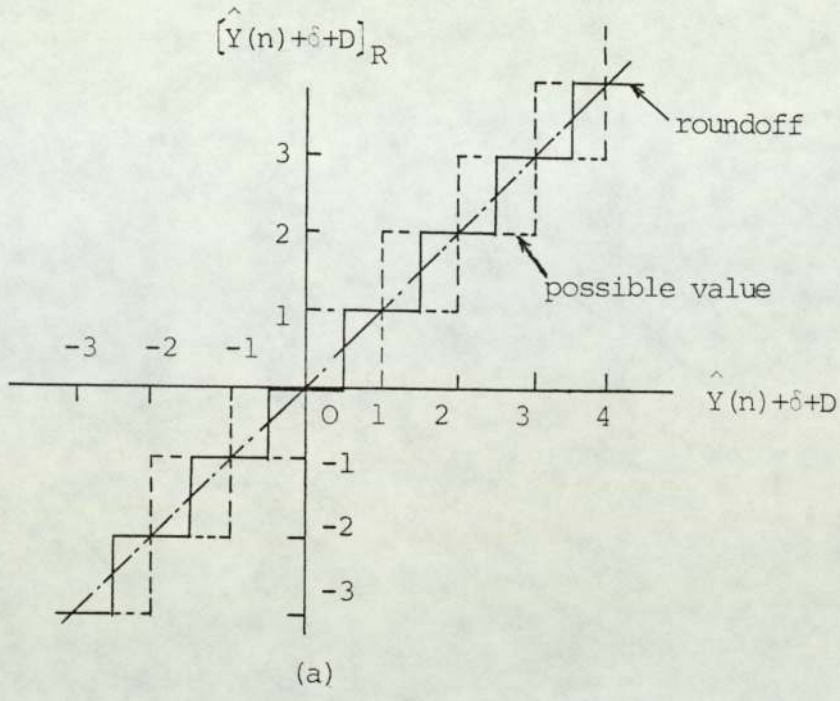


Fig. A4.3 (a) Roundoff characteristic with dither
 (b) Statistical characteristic of the equivalent quantizer.

APPENDIX 5

Program used for the generation of the
uniformly distributed random dither

13. ExampleProgram (contd)

```

C      G05CAF EXAMPLE PROGRAM TEXT
C      MARK 6 RELEASE NAG COPYRIGHT 1977
C      .. LOCAL SCALARS ..
C      REAL X
C      INTEGER I, NOUT
C      .. FUNCTION REFERENCES ..
C      REAL G05CAF
C      .. SUBROUTINE REFERENCES ..
C      G05CBF
C      ..
C      DATA NOUT /6/
C      WRITE (NOUT,99999)
C      CALL G05CBF(0)
C      DO 20 I=1,5
C          X = G05CAF(X)
C          WRITE (NOUT,99998) X
20 CONTINUE
C      STOP
99999 FORMAT (4(1X/), 31H G05CAF EXAMPLE PROGRAM RESULTS/1X)
99998 FORMAT (1X, F10.4)
C      END

```

Results

G05CAF EXAMPLE PROGRAM RESULTS

```

0.7951
0.2257
0.3713
0.2250
0.8787

```

APPENDIX 6

Program used with Computer PET to display the
limit cycles on the state plane

```

1 10 DIM Z(10),W(10)
2 20 C=0
3 30 A=-1.74
4 40 B=0.95833
5 50 PRINT""
6 60 FOR I=1 TO 10
7 70 PRINT"Z(";I;)"
8 80 INPUT Z(I)
9 90 PRINT"W(";I;)"
10 100 INPUT W(I)
11 110 NEXT I
12 120 REM
13 130 REM PLOT SPOTS.....
14 140 REM
15 150 PRINT""
16 160 FOR I=1 TO 24
17 170 PRINT"....."
18 180 NEXT I
19 190 REM
20 200 REM PLOT LINES
21 210 REM
22 220 FOR I=0 TO 25
23 230 POKE(32768+40*I+20),66
24 240 NEXT I
25 250 FOR I=0 TO 39
26 260 POKE(32768+480+I),45
27 270 NEXT I
28 280 FOR M=1 TO 10
29 290 X1=Z(M)
30 300 X2=W(M)
31 310 FOR K=1 TO 30:REM PLOT 30 POINTS
32 320 RXX=RND(2)-0.5
33 330 RX=SGN(RXX)*0.499
34 340 REM
35 350 REM COMPUTE NEXT STATE.....
36 360 Y1=-A*X1-B*X2+C*RX
37 370 Y=SGN(Y1)*INT(ABS(Y1)+0.5)
38 380 X2=X1
39 390 X1=Y
40 400 PRINT"
41 410 PRINT"X1: ";X1
42 420 PRINT"X2: ";X2
43 430 PRINT"K: ";K
44 440 REM
45 450 REM PLOT STATE PLANE POINT.....
46 460 REM
47 470 LOC=32768+(13-X1)*40+(X2-20)
48 480 FOR L=1 TO 2
49 490 FOR LA=1 TO 10:POKELOC,160:NEXT
50 500 FOR LA=1 TO 10:POKELOC,32:NEXT
51 510 NEXT L
52 520 POKELOC,M+47
53 530 NEXT K
54 540 NEXT M
55 600 END
EOF..
EOT..

```

APPENDIX 7

Program used to print out the limit cycles
on the state plane

THIS PROGRAM DRAWS THE PLOT OF THE LIMIT CYCLES IN DIGITAL FILTERS

```

C      INTEGER IL
C          THE TOTAL NUMBER OF THE LIMIT CYCLES BEING PLOTTED
C      INTEGER Y(100),X(100)
C      DIMENSION GRAPH(70,70)
C      DIMENSION AK(10)
C      Y(I) HOLDS THE DATA OF Y(N), I=THE MAXIMUM OF THE PERIOD OF THE LIMIT CYCLE
C      X(I) HOLDS THE DATA OF Y(N-1)
C      GRAPH(I,J) HOLDS THE WHOLE DATA IN Y(N)-Y(N-1) PLANE
C      I,J ARE EQUAL TO DOUBLE OF THE MAXIMUM OF Y(N)
C          INTEGER L(10)
C          INTEGER MAX
C          L(I) HOLDS THE PERIOD OF EACH LIMIT CYCLE
C          MAX IS THE MAXIMUM OF THE LIMIT CYCLES
C          DATA DOT/1H./, BLANK/1H /
C          DATA AK(1)/1H1/, AK(2)/1H2/, AK(3)/1H3/, AK(4)/1H4/, AK(5)/1H5/
C          DATA AK(6)/1H6/, AK(7)/1H7/, AK(8)/1H8/, AK(9)/1H9/, AK(10)/1H0/
C      READ TL FROM THE FILE 14
C          READ(14,50) TL,MAX
C          50 FORMAT(2I5)
C      SKIP TO THE TOP OF THE PAGE
C          WRITE(15,60)
C          60 FORMAT(1H1)
C          WRITE(15,70) IL,MAX
C          70 FORMAT(2X,' THE TOTAL NUMBER OF THE LIMIT CYCLES=',I5,' THE
C          1 MAXIMUM OF THE LIMIT CYCLES=',I5)
C      CLEAR THE ARRAY THAT HOLDS THE GRAPH
C          DO 100 I=1,70
C          DO 100 J=1,70
C          100 GRAPH(I,J)=BLANK
C      PUT DOTS INTO THE GRAPH SO AS TO FORM X,Y AXES
C          DO 200 J=1,70
C          200 GRAPH(36,J)=DOT
C          DO 300 I=1,70
C          300 GRAPH(I,36)=DOT
C      READ THE DATA OF L(I) FROM THE FILE 14
C          READ(14,400) (L(I),I=1,TL)
C          400 FORMAT(10I5)
C          500 CONTINUE
C      PUT THE DATA OF THE LIMIT CYCLES SUCCESSIVELY INTO Y(I) AND X(I), AND
C      'K' REPRESENTS THE KTH LIMIT CYCLE.
C      INPUT THE DATA OF THE FIRST LIMIT CYCLE
C          DO 900 N=1,TL
C          GO TO (505,510,515,520,525,530,535,535,535,535),N
C          505 READ(14,605) (X(I),I=1,L(1))
C          605 FORMAT(14I5)
C          GO TO 450
C          510 READ(14,610) (X(I),I=1,L(2))
C          610 FORMAT(13I5)
C          GO TO 650

```



```

515 READ(14,615) (X(I), I=1, L(3))
615 FORMAT(13I5)
GO TO 650
520 READ(14,620) (X(I), I=1, L(4))
620 FORMAT(14I5)
GO TO 650
525 READ(14,625) (X(I), I=1, L(5))
625 FORMAT(12I5)
GO TO 650
530 READ(14,630) (X(I), I=1, L(6))
630 FORMAT(14I5)
GO TO 650
535 READ(14,635) (X(I), I=1, L(N))
635 FORMAT(15)
650 Y(L(N))=X(1)
M1=L(N)-1
DO 700 I=1, M1
700 Y(I)=X(I+1)
C SET UP DO LOOP TO DEAL WITH THE KTH LIMIT CYCLE
DO 800 K=1, L(N)
C CONVERT THE DATA INTO SUBSCRIPT VALUES
I=3*(MAX+1+Y(K))
J=3*(MAX+1+X(K))
C CHECK THAT POINT LIES WITHIN GRAPH, AS EXPECTED, AND ENTER 'K' IF SO
800 IF (I.GE.1 .AND. I.LE.70 .AND. J.GE.1 .AND. J.LE.70)
1 GRAPH(I, J)=AK(N)
C CLEAR X(I), Y(I) SO AS TO PREPARE TO RECEIVE NEXT DATA OF LIMIT CYCLE
DO 850 I=1, L(N)
X(I)=0
Y(I)=0
850 CONTINUE
900 CONTINUE
C PRINT THE GRAPH
DO 930 I=1, 70
I1=71-I
910 WRITE(15, 920) (GRAPH(I1, J), J=1, 70)
920 FORMAT(1H, 70A1)
930 CONTINUE
STOP
END

```

SVX

dn 6 2499-HML TERMINAL: 68 15 FEB 82 22:16:40

APPENDIX 8

Program used for the generation of a
normal (Gaussian) distribution sequence

1. Purpose

G05DDF returns a pseudo-random real number taken from a normal (Gaussian) distribution with mean A and standard deviation B.

IMPORTANT: before using this routine, read the appropriate machine implementation document to check the interpretation of italicised terms and other implementation-dependent details.

2. Specification (FORTRAN IV)

```
      real FUNCTION G05DDF(A,B)
      C      real A,B
```

3. Description

The distribution has PDF (probability density function)

$$p(x) = \frac{1}{\sqrt{2\pi} B} \exp \left(- \frac{(x-A)^2}{2B^2} \right)$$

The routine uses the method of Brent [5].

4. References

- [1] KNUTH, D.E.
The Art of Computer Programming, Vol. 2.
Addison-Wesley, 1969.
- [2] HAMMERSLEY, J.M. and HANDSCOMB, D.C.
Monte-Carlo Methods.
Methuen, Published 1964, Reprinted 1967.
- [3] KENDALL, M.G. and STUART, A.
The Advanced Theory of Statistics, Vol. 1.
Griffin 3rd Edition, 1969.
- [4] NEAVE, H.
A Random Number Package.
Computer Applications in the Natural and Social Sciences, No. 14,
Department of Geography, University of Nottingham, 1972.
- [5] BRENT, R.P.
Algorithm 488.
C.A.C.M., p. 704, 1974.

G05DDF

5. ParametersA - *real*.

On entry, A specifies the parameter (mean) A of the distribution.

Unchanged on exit.

B - *real*.

On entry, B specifies the parameter (standard deviation) B of the distribution. If B is negative, the distribution of the generated numbers - though not the actual sequence - is the same as if the absolute value of B were used.

Unchanged on exit.

6. Error Indicators None.7. Auxiliary Routines

This routine calls the NAG Library routine G05CAF.

8. Timing

See appropriate implementation document.

9. Storage

Storage required by internally declared arrays, including those of auxiliary routines is 41 *real* elements.

10. Accuracy Not applicable.11. Further Comments

This routine uses a labelled COMMON block with the name BG05CA.

12. Keywords

Gaussian Distribution, Random Numbers.

Normal Distribution, Random Numbers.

Random Numbers, Gaussian Distribution.

Random Numbers, Normally Distributed.

13. Example

The example program prints the first five pseudo-random real numbers from a normal distribution with mean 1.0 and standard deviation 1.5, generated by G05DDF after initialization by G05CBF.

13. Example (contd)Program

This single precision example program may require amendment

- i) for use in a DOUBLE PRECISION implementation
 - ii) for use in either precision in certain implementations.
- The results produced may differ slightly.

```

C      G05DDF EXAMPLE PROGRAM TEXT
C      MARK 6 RELEASE NAG COPYRIGHT 1977
C      .. LOCAL SCALARS ..
C      REAL X
C      INTEGER I, NOUT
C      .. FUNCTION REFERENCES ..
C      REAL G05DDF
C      .. SUBROUTINE REFERENCES ..
C      G05CBF
C      ..
C      DATA NOUT /6/
C      WRITE (NOUT,99999)
C      CALL G05CBF(0)
C      DO 20 I=1,5
C          X = G05DDF(1.0,1.5)
C          WRITE (NOUT,99998) X
20    CONTINUE
C      STOP
99999  FORMAT (4(1X/), 31H G05DDF EXAMPLE PROGRAM RESULTS/1X)
99998  FORMAT (1X, F10.4)
C      END

```

Results

G05DDF EXAMPLE PROGRAM RESULTS

```

1.8045
2.9393
3.3701
0.9602
3.2751

```

APPENDIX 9

Program for the calculation of standard deviation
of output from bandstop filter when a Gaussian
random sequence $N(0,1^2)$ is input:

** Pdn 7 2499-H:ML TERMINAL: 70 20 JUL 82 14:43


```
1 C
2 C THIS PROGRAM GIVES THE SQUARE SUMMATION OF THE IMPULSE RESPONSE
3 C
4 DIMENSION X(1000),Y(1000)
5 NDOUT=15
6 A=0.932938
7 B=0.509526
8 C=1.236068
9 D=1.0
10 F=0.7547628
11 WRITE(NDOUT,50) A,B,C
12 50 FORMAT(2X,2HA=,F10.6,2X,2HB=,F10.6,2X,2HC=,F10.6)
13 WRITE(NDOUT,60) D,F
14 60 FORMAT(2X,2HD=,F10.6,2X,2HF=,F10.6)
15 DO 100 I=1,1000
16 X(I)=0.0
17 Y(I)=0.0
18 100 CONTINUE
19 X(3)=1.0
20 Y(1)=0.0
21 Y(2)=0.0
22 DO 200 N=3,1003
23 Y(N)=F*(X(N)+C*X(N-1)+D*X(N-2))-A*Y(N-1)-B*Y(N-2)
24 200 CONTINUE
25 DO 350 I=3,1003
26 Y(I)=Y(I)**2
27 350 CONTINUE
28 DO 400 N=3,1003
29 SIGMA2=SIGMA2+Y(N)
30 400 CONTINUE
31 SIGMA=SQRT(SIGMA2)
32 WRITE(NDOUT,500) SIGMA
33 500 FORMAT(2X,23HTHE STANDARD DEVIATION=,F15.8)
34 STOP
35 END
36 $VX
EOF..
```


** Pdn 7 2499-H:ML TERMINAL: 70 20 JUL 82 14:43

** Pdn 7 2499-HML TERMINAL: 70 20 JUL 82 14:43

1 A= 0.932938 B= 0.509526 C= 1.236068
2 D= 1.000000 F= 0.754763
3 THE STANDARD DEVIATION= 1.03888544

** Pdn 7 2499-HML TERMINAL: 70 20 JUL 82 14:40

APPENDIX 10

Program for the calculation of time needed for
stabilization by the use of dither

```

*****
*****
*****
** Pdn      7      2499-HML          TERMINAL: 54          26 JUL 82  9:4
*****
*****
*****

```

```

1 C THREE TYPES OF DITHER CAN BE USED.
2 C THEY ARE UNIFORMLY DISTRIBUTED RANDOM DITHER, BINARY RANDOM DITHER AND
3 C BANDSTOP DITHER.
4 C THIS PROGRAM GIVES THE MEAN VALUE AND STANDARD DEVIATION OF THE
5 C TRANSITION TIME FROM THE LARGEST LIMIT CYCLE TO THE ORIGIN
6 C STATE ON THE STATE PLANE.
7     DIMENSION Q(1000)
8     DIMENSION E(10)
9     REAL MS, MEAN
10    REAL G05CAF
11    REAL G05DDF
12    NOUT=15
13 C IF KRIT IS LESS THAN ZERO UNIFORM DITHER WILL BE USED.
14 C IF KRIT IS EQUAL TO ZERO BINARY DITHER WILL BE USED.
15 C IF KRIT IS GREATER THAN ZERO BANDSTOP DITHER WILL BE USED.
16    KRIT=2
17 C DA AND OB ARE THE COEFFICIENT VALUES OF THE BASIC FILTER SECTION
18    DA=-1.452679
19    OB=0.881619
20    IF (KRIT.LE.0) GO TO 111
21 C A, B, C, D, F, ZA, ZB, ZC AND ZD ARE THE COEFFICIENT VALUES OF
22 C THE BANDSTOP FILTER.
23    A=-2.78770841
24    B=3.55450670
25    C=-2.31321908
26    D=0.697549120
27    F=0.645716253
28    ZA=-2.02590537/F
29    ZB=2.88042372/F
30    ZC=ZA
31    ZD=1.0
32 C SIGMA SQUARE IS THE SQUARE SUMMATION OF THE IMPULSE RESPONSE
33 C OF THE BANDSTOP FILTER. IT IS USED AS A SCALE FACTOR SO AS TO
34 C NORMALIZE THE STANDARD DEVIATION OF THE SEQUENCE AT THE
35 C NONLINEAR NETWORK INPUT.
36    SIGMA=0.79647261
37    111 PI=4.0*DATAN(1.0)
38    CALL G05CBF(0)
39    DO 320 J=1,1000
40 C XX1 AND XX2 DEFINE A INITIAL STATE WHICH IS ON THE LARGEST
41 C LIMIT CYCLE.
42    XX1=4.0
43    XX2=4.0
44    Z1=0.0
45    Z2=0.0
46    Z3=0.0
47    Z4=0.0
48    X1=0.0
49    X2=0.0
50    X3=0.0
51    X4=0.0

```

```

52      K1=0
53      DO 50 K=1,100000
54      IF (KRIT.LE.0) GO TO 444
55      ZX=G05DDF (0.0,1.0)
56 C LET THE INPUT GAUSSIAN DATA PASS A SPECIAL DIGITAL FILTER
57      Z=-A*Z1-B*Z2-C*Z3-D*Z4+F*(ZX+ZA*X1+ZB*X2+ZC*X3+ZD*X4)
58      Z4=Z3
59      Z3=Z2
60      Z2=Z1
61      Z1=Z
62      X4=X3
63      X3=X2
64      X2=X1
65      X1=ZX
66      Z=Z/SIGMA
67      IF (ABS(Z).LT.3.0) GO TO 555
68      Z=SIGN(1.0,Z)*3.0
69      555 CONTINUE
70      Z=Z*SQRT(1.0/2.0)
71 C NOTE THAT THE DEVIATION OF THE NUMBERS AT THE OUTPUT HAS BEEN CHANGED
72 C THEREFORE THE NUMBERS HAVE TO BE DEVIDED BY THE SIGMA, STANDARD
73 C DEVIATION.
74 C PUT THE GAUSSIAN RANDOM NUMBERS (AFTER BEING DEVIDED BY SIGMA) INTO A
75 C NON-LINEAR NETWORK SO AS TO GENERATE A UNIFORM DISTRIBUTION RANDOM
76 C NUMBERS AND THE NON-LINEAR GIVES ONLY MINOR CHANGES IN THE POWER
77 C DENSITY SPECTRUM.
78      W1=1.0
79      E(2)=1.0
80      DO 26 JJ=1,32
81      FJ=FLOAT(JJ)
82      W2=W1*(-1.0)
83      ZZP1=W2*(Z**JJ)
84      ZZP2=Z**(JJ+1)
85      W1=W2
86      E(1)=FJ*E(2)
87      E(2)=E(1)
88      ZZ2=E(1)*(2.0*FJ+1.0)
89      ZZP3=ZZP2/ZZ2
90      ZZ3=ZZP3*ZZP1
91      ZZ4=ZZ4+ZZ3
92      26 CONTINUE
93      ZZ5=(ZZ4+Z)/SQRT(PI)
94      ZZ4=0.0
95 C XX IS THE BANDSTOP DITHER SEQUENCE.
96      XX=ZZ5
97      444 CONTINUE
98      IF (KRIT.LT.0) GO TO 222
99      IF (KRIT.EQ.0) GO TO 333
100     Y1=-0A*XX1-0B*XX2+XX
101     GO TO 666
102     222 XU=G05CAF(X)
103     Y1=-0A*XX1-0B*XX2+XU-0.5
104     GO TO 666
105     333 XU=G05CAF(X)-0.5
106     RX=SIGN(1.0,XU)*0.499
107     Y1=-0A*XX1-0B*XX2+RX
108     666 Y=SIGN(1.0,Y1)*AINT(ABS(Y1)+0.5)
109     K1=K1+1
110     XX2=XX1
111     XX1=Y

```

```
112      J1=INT(XX1)
113      J2=INT(XX2)
114      IF(J1.EQ.0) GO TO 45
115      GO TO 50
116      45 IF(J2.EQ.0) GO TO 60
117      50 CONTINUE
118      60 Q(J)=K1
119      AJ=FLOAT(J)
120      AJJ=AJ/100.0
121      IAJJ=INT(AJJ)
122      FIA=FLOAT(IAJJ)
123      IF (AJ.EQ.100.0*FIA) GO TO 394
124      GO TO 320
125      394 WRITE(3,396) AJJ
126      396 FORMAT(2X,F8.4)
127      320 CONTINUE
128      WRITE(NOUT,392) A,B,C,D,F,SIGMA
129      392 FORMAT(2X,6F12.8)
130      SUM=0.0
131      DO 400 I=1,1000
132      SUM=SUM+Q(I)
133      400 CONTINUE
134      MEAN=SUM/1000.0
135      MS=0.0
136      DO 450 I=1,1000
137      MS=MS+(Q(I)-MEAN)**2
138      450 CONTINUE
139      SD=SQRT(MS/999.0)
140      SDOM=SD/SQRT(1000.0)
141      WRITE(NOUT,500) MEAN,SDOM
142      500 FORMAT(2X,"THE MEAN=",F12.6,"THE STANDARD DEVIATION=",F12.6)
143      STOP
144      END
145 $VX
EOF.
```

```
*****
*****
*****
** Pdn      7      2499-H:ML      TERMINAL: 54      26 JUL 82  9:40
*****
*****
*****
```


* Pdn 7 2499-HML TERMINAL: 54 26 JUL 82 9:40:1

-2.78770841 3.55450670 -2.31321908 0.69754912 0.64571625 0.79647261
THE MEAN= 49.540000 THE STANDARD DEVIATION= 1.145246

* Pdn 7 2499-HML TERMINAL: 54 26 JUL 82 9:40:1

APPENDIX 11

Program for the calculation of the increase in output
noise from the basic filter section by the dither

```

*****
*****
*****
** Pdn      7      2499-HML                TERMINAL: 67                26 JUL 82 11:18
*****
*****

```

```

1 C THIS PROGRAM GIVES THE INCREASE IN THE OUTPUT NOISE (DB) FROM THE
2 C SECOND-ORDER FILTER SECTION DUE TO THE INJECTION OF DITHER.
3 C THIS PROGRAM ALSO GIVES THE STANDARD ERROR OF THE OUTPUT NOISE
4 C INCREASE. THE QUANTIZATION NOISE WITHOUT DITHER IS USED AS THE
5 C REFERENCE LEVEL. THE INITIAL STATE OF THE SECTION IS AT THE
6 C ORIGIN OF THE STATE PLANE. A SINUSOIDAL SIGNAL IS INPUT TO THE
7 C SECOND-ORDER FILTER SECTION. THE OUTPUT FROM THE SECTION IS
8 C ANALYZED BY THE FFT.
9 C HAMMING WINDOW IS USED AND N=2048
10      DIMENSION E(10)
11      DIMENSION X(2048),W(2048),XREAL(2048)
12      DIMENSION XIMAG(2048),P(2048)
13      DIMENSION TPN(50),TPS(50)
14      REAL G05CAF
15      REAL G05DDF
16 C IF KRIT IS LESS THAN ZERO UNIFORM DITHER WILL BE USED.
17 C IF KRIT IS EQUAL TO ZERO BINARY DITHER WILL BE USED.
18 C IF KRIT IS GREATER THAN ZERO BAND STOP DITHER WILL BE USED.
19      KRIT=2
20      NOUT=15
21 C QN IS THE QUANTIZATION NOISE WITHOUT DITHER WHEN A SINE WAVE IS INPUT
22      QN=694.2
23 C OA AND OB ARE THE COEFFICIENT VALUES OF THE BASIC FILTER SECTION.
24      OA=-1.452679
25      OB=0.881619
26      IF (KRIT.LE.0) GO TO 111
27 C A, B, C, D, F, ZA, ZB, ZC, AND ZD ARE THE COEFFICIENTS OF THE BANDSTOP FILTER.
28      A=-2.78770841
29      B=3.55450670
30      C=-2.31321908
31      D=0.697549120
32      F=0.645716253
33      ZA=-2.02590537/F
34      ZB=2.88042372/F
35      ZC=ZA
36      ZD=1.0
37 C SIGMA SQUARE IS THE SQUARE SUMMATION OF THE IMPULSE RESPONSE OF
38 C BANDSTOP FILTER. IT IS USED AS A SCALE FACTOR SO AS TO NORMALIZE
39 C THE STANDARD DEVIATION OF THE SEQUENCE AT THE NONLINEAR
40 C NETWORK INPUT.
41      SIGMA=0.79647261
42      111 PI=4.0*DATAN(1.0)
43 C HAMMING WINDOW FUNCTION, N=2048
44      DO 2 I=1,2048
45          H=FLOAT(I-1024)
46          W(I)=0.54+0.46*COS(2.0*PI*H/2048.0)
47      2 CONTINUE
48 C U=ENERGY IN HAMMING WINDOW
49      U=0.0
50      DO 4 I=1,2048
51          AE=W(I)**2

```



```

52      U=U+AE
53      4 CONTINUE
54      DO 5 I=1,2048
55      5 P(I)=0.0
56      CALL G05CBF(0)
57      DO 50 J=1,30
58      Z1=0.0
59      Z2=0.0
60      Z3=0.0
61      Z4=0.0
62      X1=0.0
63      X2=0.0
64      X3=0.0
65      X4=0.0
66      XX1=0.0
67      XX2=0.0
68 C IN ORDER TO OBTAIN A STEADY-STATE SINUSOIDAL OUTPUT THE FIRST 3584
69 C SAMPLES HAVE TO BE OMITTED.
70      DO 6 I=1,5632
71      FI=FLOAT(I)-1.0
72      IF (KRIT.LE.0) GO TO 444
73      ZX=G05DDF (0.0,1.0)
74 C LET THE INPUT GAUSSIAN DATA PASS A SPECIAL DIGITAL FILTER
75      Z=-A*Z1-B*Z2-C*Z3-D*Z4+F*(ZX+ZA*X1+ZB*X2+ZC*X3+ZD*X4)
76      Z4=Z3
77      Z3=Z2
78      Z2=Z1
79      Z1=Z
80      X4=X3
81      X3=X2
82      X2=X1
83      X1=ZX
84      Z=Z/SIGMA
85      IF (ABS(Z).LT.3.0) GO TO 555
86      Z=SIGN(1.0,Z)*3.0
87      555 CONTINUE
88      Z=Z*SQRT(1.0/2.0)
89 C NOTE THAT THE DEVIATION OF THE NUMBERS AT THE OUTPUT HAS BEEN CHANGED
90 C THEREFORE THE NUMBERS HAVE TO BE DIVIDED BY THE SIGMA, STANDARD
91 C DEVIATION.
92 C PUT THE GAUSSIAN RANDOM NUMBERS (AFTER BEING DIVIDED BY SIGMA) INTO A
93 C NON-LINEAR NETWORK SO AS TO GENERATE A UNIFORM DISTRIBUTION RANDOM C
94 C NUMBERS AND THE NON-LINEAR GIVES ONLY MINOR CHANGES IN THE POWER
95 C DENSITY SPECTRUM.
96      W1=1.0
97      E(2)=1.0
98      DO 26 JJ=1,32
99      FJ=FLOAT(JJ)
100     W2=W1*(-1.0)
101     ZP1=W2*(Z**JJ)
102     ZP2=Z**(JJ+1)
103     W1=W2
104     E(1)=FJ*E(2)
105     E(2)=E(1)
106     ZZ2=E(1)*(2.0*FJ+1.0)
107     ZP3=ZP2/ZZ2
108     ZZ3=ZP3*ZP1
109     ZZ4=ZZ4+ZZ3
110     26 CONTINUE
111     ZZ5=(ZZ4+Z)/SQRT(PI)

```

```

112      ZZ4=0.0
113      IF (ABS(ZZ5).LT.0.5) GO TO 30
114      Z5=SIGN(1.0,ZZ5)*0.4997
115 C XX IS THE BANDSTOP DITHER SEQUENCE. SZ IS THE SINUSOIDAL INPUT.
116      30 XX=ZZ5
117      444 SZ=38.49*SIN(2.0*PI*14.0/128.0*FI)
118      IF (KRIT.LT.0) GO TO 222
119      IF (KRIT.EQ.0) GO TO 333
120      Y1=-0A*XX1-0B*XX2+SZ+XX
121      GO TO 666
122      222 XU=G05CAF(X)
123      Y1=-0A*XX1-0B*XX2+SZ+XU-0.5
124      GO TO 666
125      333 XU=G05CAF(X)-0.5
126      RX=SIGN(1.0,XU)*0.4997
127      Y1=-0A*XX1-0B*XX2+SZ+RX
128      666 Y=SIGN(1.0,Y1)*AINT(ABS(Y1)+0.5)
129      XX2=XX1
130      XX1=Y
131      IF (I.LE.3584) GO TO 6
132      X(I-3584)=Y
133      6 CONTINUE
134      DO 40 I=1,2048
135      XREAL(I)=X(I)*W(I)
136      XIMAG(I)=0.0
137      40 CONTINUE
138      CALL FFT(XREAL,XIMAG,2048,11)
139      DO 42 I=1,2048
140      P(I)=P(I)+(XREAL(I)**2+XIMAG(I)**2)/U
141      42 CONTINUE
142      WRITE(3,44) J
143      44 FORMAT(2X,'J=',I2)
144      TOTP1=0.0
145      DO 70 I=1,223
146      70 TOTP1=TOTP1+P(I)
147      TOTP2=0.0
148      DO 80 I=224,226
149      80 TOTP2=TOTP2+P(I)
150      TOTP3=0.0
151      DO 90 I=227,1823
152      90 TOTP3=TOTP3+P(I)
153      TOTP4=0.0
154      DO 96 I=1824,1826
155      96 TOTP4=TOTP4+P(I)
156      TOTP5=0.0
157      DO 98 I=1827,2048
158      98 TOTP5=TOTP5+P(I)
159      TPN(J)=TOTP1+TOTP3+TOTP5
160      TPS(J)=TOTP2+TOTP4
161      DO 801 I=1,2048
162      801 P(I)=0.0
163      50 CONTINUE
164      DO 901 J=1,30
165      TPN(J)=10.0*ALOG10(TPN(J)/GN)
166      901 CONTINUE
167      AMEAN=0.0
168      DO 902 J=1,30
169      AMEAN=AMEAN+TPN(J)
170      902 CONTINUE
171      AMEAN=AMEAN/30.0

```

```

172     STAD=0.0
173     DO 903 J=1,30
174     STAD=STAD+(TPN(J)-AMEAN)**2
175 903 CONTINUE
176     STAD=SQRT(STAD/29.0)
177     STAR=STAD/SQRT(30.0)
178     WRITE(NOUT,52)
179     52 FORMAT(2X,' THE POWER SPECTRUM OF THE OUTPUT OF THE SECTION ')
180     WRITE(NOUT,54)
181     54 FORMAT(2X,' THE INPUT OF THE SECTION IS AS FOLLOWS ')
182     WRITE(NOUT,56)
183     56 FORMAT(2X,"38.49*SIN(2.0*PI*14.0/128.0*FI) ")
184     IF (KRIT.LT.0) GO TO 606
185     IF (KRIT.EQ.0) GO TO 707
186     WRITE(NOUT,57)
187     57 FORMAT(2X,' BANDSTOP DITHER ')
188     GO TO 148
189 606 WRITE(NOUT,667)
190 667 FORMAT(2X,' UNIFORM DITHER ')
191     GO TO 148
192 707 WRITE(NOUT,778)
193 778 FORMAT(2X,' BINARY DITHER ')
194 148 WRITE(NOUT,58) OA,OB
195     58 FORMAT(2X,' THE COEFFICIENTS OA=',F12.8,' OB=',F12.8)
196     WRITE(NOUT,123) QN
197 123 FORMAT(2X,' THE QUANTIZATION NOISE =',F14.4)
198     WRITE(NOUT,802)
199 802 FORMAT(2X,' THE INCREASE OF NOISE IN EACH SIMULATION ')
200     WRITE(NOUT,804) (TPN(J),J=1,30)
201 804 FORMAT(2X,F15.6)
202     WRITE(NOUT,905) AMEAN
203 905 FORMAT(2X,' THE MEAN OF NOISE INCREASE=',F10.6,' DB')
204     WRITE(NOUT,906) STAD
205 906 FORMAT(2X,' THE STANDARD DEVIATION=',F10.6,' DB')
206     WRITE(NOUT,907) STAR
207 907 FORMAT(2X,' THE STANDARD ERROR=',F10.6,' DB')
208     STOP
209     END
210
211
212
213     SUBROUTINE FFT(XREAL,XIMAG,N,NU)
214     DIMENSION XREAL(N),XIMAG(N)
215     N2=N/2
216     NU1=NU-1
217     K=0
218     DO 100 L=1,NU
219 102 DO 101 I=1,N2
220     P=IBITR(K/2**NU1,NU)
221     ARG=6.283185*P/FLOAT(N)
222     C=COS(ARG)
223     S=SIN(ARG)
224     K1=K+1
225     K1N2=K1+N2
226     TREAL=XREAL(K1N2)*C+XIMAG(K1N2)*S
227     TIMAG=XIMAG(K1N2)*C-XREAL(K1N2)*S
228     XREAL(K1N2)=XREAL(K1)-TREAL
229     XIMAG(K1N2)=XIMAG(K1)-TIMAG
230     XREAL(K1)=XREAL(K1)+TREAL
231     XIMAG(K1)=XIMAG(K1)+TIMAG

```

```

232 101 K=K+1
233     K=K+N2
234     IF (K.LT.N)GO TO 102
235     K=0
236     NU1=NU1-1
237 100 N2=N2/2
238     DO 103 K=1,N
239     I=IBITR(K-1,NU)+1
240     IF(I.LE.K) GO TO 103
241     TREAL=XREAL(K)
242     TIMAG=XIMAG(K)
243     XREAL(K)=XREAL(I)
244     XIMAG(K)=XIMAG(I)
245     XREAL(I)=TREAL
246     XIMAG(I)=TIMAG
247 103 CONTINUE
248     RETURN
249     END
250
251
252     FUNCTION IBITR (J,NU)
253     J1=J
254     IBITR=0
255     DO 200 I=1,NU
256     J2=J1/2
257     IBITR=IBITR*2+(J1-2*J2)
258 200 J1=J2
259     RETURN
260     END
261 $VX
EOF.

```

```

*****
*****
*****
** Pdn      7      2499-HML          TERMINAL: 67          26 JUL 82 11:19
*****
*****
*****

```


** Pdn 7 2499-HML TERMINAL: 67 26 JUL 82 11:18

1 THE POWER SPECTRUM OF THE OUTPUT OF THE SECTION
2 THE INPUT OF THE SECTION IS AS FOLLOWS
3 38.49*SIN(2.0*PI*14.0/128.0*FI)
4 BANDSTOP DITHER
5 THE COEFFICIENTS OA= -1.45267900DB= 0.88161900
6 THE QUANTIZATION NOISE = 694.2000
7 THE INCREASE OF NOISE IN EACH SIMULATION
8 6.208308
9 6.252003
10 5.791453
11 5.253524
12 4.961397
13 5.344134
14 5.477483
15 5.993696
16 4.760712
17 5.667707
18 4.928636
19 5.397100
20 5.793648
21 5.243185
22 5.715346
23 5.079068
24 5.736177
25 6.453139
26 5.519688
27 5.929192
28 5.039081
29 5.644546
30 5.606036
31 5.463432
32 5.224590
33 4.855541
34 5.251187
35 5.113903
36 5.434075
37 5.664541
38 THE MEAN OF NOISE INCREASE= 5.493418DB
39 THE STANDARD DEVIATION= 0.422882DB
40 THE STANDARD ERROR= 0.077207DB

** Pdn 7 2499-HML TERMINAL: 67 26 JUL 82 11:18

REFERENCES

1. Blackman, R. B., Linear Data-Smoothing and Prediction in Theory and Practice, Reading, MA: Addison-Wesley, 1965.
2. Parker, S. R. and Hess, S. F., "Limit-cycle oscillations in digital filters," IEEE Trans. Circuit Theory (special Issue on Active and Digital Networks), Vol. CT-18, pp.687-697, Nov. 1971.
3. Kieburtz, R. B., Lawrence, V. B. and Mina, K. V., "Control of limit cycles in recursive filters by randomized quantization," in Proc. 1976 IEEE Int. Symp. Circuits and Systems, Munich, Germany, pp.624-627, 1976.
4. Büttner, M., "A novel approach to eliminate limit cycles in digital filters with a minimum increase in the quantization noise," in Proc. 1976 IEEE Symp. Circuits and Systems, Munich, Germany, pp. 291-294, 1976.
5. Rashidi, P. and Bogner, R. E., "Suppression of limit cycle oscillations in second-order recursive digital filters," A.T.R., Vol. 12, No. 1, pp. 8-16, 1978.
6. 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, Signal Processing, Vol. ASSP-24, pp. 517-529, Dec. 1976.
7. Long, J. L. and Trick, T. N., "An absolute bound on limit cycles due to roundoff errors in digital filters," IEEE Trans. Audio Electroacoust., Vol. AU-21, pp.27-30, Feb. 1973.
8. Sandberg, I. W. and Kaiser, J. F., "A bound on limit cycles in fixed-point implementations of digital filters," IEEE Trans. Audio Electroacoust., Vol. AU-20, pp. 110-112, June 1972.
9. Jackson, L. B., "An analysis of limit cycles due to multiplication rounding in recursive digital filters," in Proc. 7th Annu. Allerton Conf. Circuit and System Theory, Monticello, IL, pp. 69-78, Oct. 1969.
10. 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," in Proc. IEE 1974 European Conf. Circuit Theory and Design, London, England, Conf. Publ. 116, pp. 354-358, July, 1974.

11. IEEE Trans. Audio Electroacoust. (Special Issue on Digital Filtering), Vol. AU-18, June 1970.
12. IEEE Trans. Circuits and Systems (Special Issue on Digital Filtering and Image Processing), Vol. CAS-22, March 1975.
13. Fettweis, A., "Digital filter structures, related to classical filter networks," Arch. Elek. Übertragung, Vol. 25, pp. 79-89, 1971.
14. Fettweis, A. and Meerkötter, K., "Suppression of parasitic oscillations in wave digital filters," IEEE Trans. Circuits and Systems (Special Issue on Digital Filtering and Image Processing) Vol. CAS-22, pp. 239-246, March 1975.
15. Fettweis, A. and Meerkötter, K., "Correction to suppression of parasitic oscillations in wave digital filters," IEEE Trans. Circuits and Systems (Lett.), Vol. CAS-22, pp. 575, June 1975.
16. Adams, P. F., Harbridge, J. R. and Macmillan, R. H., "An MOS integrated circuit for digital filtering and level detection," IEEE Journal of Solid-State Circuits, Vol. SSC-16, pp. 183-190, June 1981.
17. Butterweck, H. J., "Suppression of parasitic oscillations in second-order digital filters by means of controlled rounding arithmetic," Arch. Elek. Übertragung, Vol. 29, pp. 371-374, Sept. 1975.
18. Wong, K. W. and King, R. A., "Method to suppress limit-cycle oscillations in digital filter," Electronics Letters, Vol. 10, pp. 55-57, March 7, 1974.
19. Oldenburger, R. and Liu, C. C., "Signal stabilization of a control system," AIEE Trans. Vol. 78, Part II, pp. 96-100, May 1959.
20. Zames, G. and Shneydor, N. A., "Structural stabilization and quenching by dither in nonlinear systems," IEEE Trans. Automatic Control, Vol. AC-22, No. 3, pp. 352-361, June 1977.
21. Kaiser, J. F., "Some practical considerations in the realization of linear digital filters," Proceedings of the Third Annual Allerton Conference on Circuit and System Theory, Monticello, Illinois, October 1965.
22. Knowles, J. B. and Edwards, P., "Effect of a finite wordlength computer in a sampled-data feedback system," Proceedings IEE, Vol. 112, No. 6, June 1965.

23. Edwards, R., Bradley, T. and Knowles, J. B., "Comparison of noise performances of programming methods in the realization of digital filters," Proceedings of the Symposium on Computer Processing in Communications, Polytechnic Institute of Brooklyn, Brooklyn, New York, April 1969.
24. Hess, S, "A deterministic analysis of limit cycle oscillations in recursive digital filters due to quantization," Ph.D. dissertation, Naval Postgraduate School, Monterey, California, Dec. 1970.
25. Rabiner, L. R. and Gold, B., Theory and Application of Digital Signal Processing, Englewood Cliffs, N.J.: Prentice-Hall, 1975.
26. Antoniou, A., Digital Filters: Analysis and Design, McGraw-Hill Book Company, 1979.
27. Claasen, T. A. C. M., "Survey of Stability Concepts of Digital Filters," The Royal Inst. Tech., Stockholm, Sweden, Tech. Rep. 108 (Telecommunication Theory), 1976.
28. Kaneko, T., "Limit-cycle oscillations in floating-point digital filters," IEEE Trans. Audio Electroacoust., Vol. AU-21, pp. 100-106, April 1973.
29. Lacroix, A., "Underflow limit cycles in floating-point digital filters," in Proc. Florence Conf. Digital Signal Processing, Florence, Italy, pp. 75-84, Sept. 11-13, 1975.
30. Sandberg, I. W., "Floating-point-roundoff accumulation in digital filter realizations," Bell System Technical Journal, Vol. 46, No. 8, pp. 1775-1791, Oct. 1967.
31. Ebert, P. M., Mazo, J. E. and Taylor, M. G., "Overflow oscillations in digital filters," Bell Syst. Tech. J., Vol. 48, pp. 2999-3020, Nov. 1969.
32. Claasen, T. A. C. M., Mecklenbräuker, W. F. G., and Peek, J. B. H., "Quantization noise analysis for fixed point digital filters using magnitude truncation for quantization," IEEE Trans. Circuits and Systems, Vol. CAS-22, pp. 887-895, Nov. 1975.
33. Claasen, T. A. C. M., Mecklenbräuker, W. F. G., and Peek, J. B. H., "Second order digital filter with only one magnitude truncation quantizer and having practically no limit cycles," Electron. Letters, Vol. 9, pp. 531-532, Nov. 1, 1973.

34. 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 with one magnitude-truncation quantizer," *IEEE Trans. Acoust., Speech, Signal Processing (Corresp.)*, Vol. ASSP-23, pp. 240-242, April 1975.
35. Kao, C., "An analysis of limit cycles due to sign magnitude truncation in multiplication in recursive digital filters," in *Proc. 5th Asilomar Conf. Circuit and System Theory*, Pacific Grove, CA, pp. 349-353, 1971.
36. Claasen, T. A. C. M., Mecklenbräuker, W. F. G., and Peek, J. B. H., "Frequency domain criteria for absence of zero-input limit cycles in nonlinear discrete-time systems, with application to digital filters," *IEEE Trans. Circuits and Systems (Special Issue on Digital Filtering and Image Processing)*, Vol. CAS-22, pp. 232-239, March 1975.
37. 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, pp. 297-305, August 1973.
38. 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 Systems (Lett.)*, Vol. CAS-22, pp. 570-572, June 1975.
39. Howard, R. A., Dynamic probabilistic systems, Volume 1: Markov models, John Wiley & Sons, Inc., 1971.
40. Veltman, B., Van Den Bos, A, de Bruine, R., de Ruiter, R. and Verloren, P., "Some remarks on the use of auto-correlation functions with the analysis and design of signals," in Signal Processing, Griffiths, J. W. R., Stocklin, P. L. and Van Schooneveld, C., (Eds), Academic Press, 1973.
41. Gray, A. H., and Markel, J. D., "A computer program for designing digital elliptic filters," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-24, pp. 529-538, Dec. 1976.
42. Gradshteyn, I. S., and Ryzhik, I. M., Table of Integrals, Series and Products, Academic Press, 1980.
43. Hine, J. and Wetherill, G. B., A programmed text in statistics, Book Two: Basic Theory, Chapman and Hall, 1975.

44. Siegel, L. J., Steiglitz, K. and Zuckerman, M., "The design of Markov chains for waveform generation," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 6, pp. 558-562, Dec. 1976.
45. Ackroyd, M. H. and Liu, H. M., "Limit cycle suppression in digital filters," in Saraga Memorial Colloquium on Electronic Filters, pp. 5/1-5/6, 1982.
46. Liu, H. M. and Ackroyd, M. H., "Suppression of limit cycles in digital filters by random dither," to be published.