The Sigma-Delta Modulator as a Chaotic Nonlinear Dynamical System

by

Donald O. Campbell

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Donald Campbell

Abstract

The Σ-Δ modulator is a popular signal amplitude quantization error (or noise) shaper used in oversampling analogue-to-digital and digital-to-analogue converter systems. This shaping of the noise frequency spectrum is performed by feeding back the quantization errors through a time delay element filter and feedback loop in the circuit, and by the addition of a possible stochastic dither signal at the quantizer. The aim in audio systems is to limit audible noise and distortions in the reconverted analogue signal. The formulation of the Σ-Δ modulator as a discrete dynamical system provides a useful framework for the mathematical analysis of such a complex nonlinear system, as well as a unifying basis from which to consider other systems, from pseudorandom number generators to stochastic resonance processes, that yield equivalent formulations.

The study of chaos and other complementary aspects of internal dynamical behaviour in previous research has left important issues unresolved. Advancement of this study is naturally facilitated by the dynamical systems approach. In this thesis, the general order feedback/feedforward Σ-Δ modulator with multi-bit quantizer (no overload) and general input, is modelled and studied mathematically as a dynamical system. This study employs pertinent topological methods and relationships, which follow centrally from the symmetry of the circle map interpretation of the error state space dynamics. The main approach taken is to reduce the nonlinear system into local or special case linear systems. Systems of sufficient structure are shown to often possess structured random, or random-like behaviour.

An adaptation of Devaney's definition of chaos is applied to the model, and an extensive investigation of the conditions under which the associated chaos conditions hold or do not hold is carried out. This seeks, in part, to address the unresolved research issues. Chaos is shown to hold if all zeros of the noise transfer function lie outside the circle of radius two, provided the input is either periodic or persistently random (mod Δ). When the filter

satisfies a certain continuity condition, the conditions for chaos are extended, and more clear cut classifications emerge. Other specific chaos classification cases are established. A study of the statistical properties of the error in dithered quantizers and Σ-Δ modulators is pursued using the same state space model. A general treatment of the steady state error probability distribution is introduced, and results for predicting uniform steady state errors under various conditions are found. The uniformity results are applied to RPDF dithered systems to give conditions for a steady state error variance of $\Delta^2/6$. Numerical simulations support predictions of the analysis for the first-order case with constant input. An analysis of conditions on the model to obtain bounded internal stability or instability is conducted. The overall investigation of this thesis provides a theoretical approach upon which to orient future work, and initial steps of specific inquiry that can be advanced more extensively in the future.

Acknowledgements

To begin, I thank my supervisor Dr. Stanley Lipshitz for his outstanding dedication and foresight in opening up a new field of applied mathematical research for me (signal processing for audio applications), and for his wise and incisive advice and guidance throughout the completion of this thesis, and my Ph.D. studies. I thank Dr. Sue Ann Campbell for her consistent service on my advisory and examining committees, and for her sound opinions and counsel on my thesis and research work from the perspective of our common field of dynamical systems. I thank the following other members of the examining committee for their helpful suggestions to improve this thesis: Dr. Pei Yu, who provided applied dynamical systems insight that integrated with my outlook; Dr. Andrew Heunis, who provided an independent engineering viewpoint, with particular experience in stochastic processes; and Dr. Bernhard Bodmann, who provided complementary ideas to those of my supervisor and I on signal processing and related theory.

In addition, I thank Dr. Edward Vrscay for his early service on my advisory committee, and for our valuable consultations on chaos in discrete mappings at an important stage of this research. I also thank Dr. John Vanderkooy for offering a unique, physicist's insight into some of the more applied aspects of my research.

I acknowledge the Ontario Graduate Scholarship, the University of Waterloo, the Department of Applied Mathematics, Dr. S. P. Lipshitz, and his research partner Dr. J. Vanderkooy, for their financial support during my Ph.D. studies. I acknowledge my office partners Mohamad Alwan, Gibin Powathil and Yanwei Wang for providing a collegial work environment, and Gabriel Esteves in computer science for helping with the thesis diagrams. I also want to acknowledge all of the associates of mine, faculty, staff, or other students at the University of Waterloo who contributed to my learning, or helped to make my long, arduous period of study here more satisfying.

Finally, I thank my parents for being a dependable and unqualified source of assistance and support, and an invaluable complementary presence for me to academic life in Waterloo Ont.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

The transmission, storage and reproduction of audio and visual information form the central components of any communications process in which human beings, responding through hearing and sight, are the end receivers. Technological, as well as physical and biological stages of this process are fundamental in many aspects of modern life. Simple examples of this include listening to music or speech from a radio or recording, and looking at printed photographs or visual images on a screen. In the most basic form, audio information is defined in one dimension over the time domain, while visual information is defined in up to four dimensions over a possible combination of the space and time domains for moving images. This information exists naturally in analogue form. By this we mean that it is quantified in amplitude over some continuous range of values, and that this value varies continuously over the respective domain defined as a continuum. The logic, operations and storage mode of a computer, however, are of a discrete nature. With the development and application of computer technology over the last half of the twentieth century, it has

thereby become more efficient and at least equally effective to represent this information in digital or discrete form for transmission and storage. A common audio system example of storage and reproduction is the compact disc and player. To summarize the process involved, analogue information is converted to a digital approximation, the information is transmitted or stored in this form, and it is then converted back to an analogue form in reproduction.

The conversion to digital form intrinsically involves two separate discretizations. First, the continuous domain (over time or space) must be discretized by a sampling of the information "signal" at specific points usually separated by equal intervals on a grid over the domain ("uniform sampling"). Second, the amplitude of the sampled information signal must be discretized or approximated by the point nearest to its value from a sequence of points or "levels" separated by equal intervals or "steps" over some range. This is called quantization. Quantization of the amplitude and sampling in this fashion are in fact the central discretization or "modulation" processes in all digital converters. The resulting digital information may then be stored on a computer disc type of assembly, where each signal sample with quantized value on one of 2^n levels is stored in a binary word of length n bits, $n \geq 1$. The two operations of sampling and quantization are usually performed together in a single device called an analogue-to-digital converter (ADC). For the analogue reconstruction of the transmitted or stored information signal that is required for reproduction, the following properties hold. If the frequency components of the signal are all smaller than a certain finite value, that is if the signal is band limited, then, in theory, the signal may be perfectly reconstructed from its samples, provided the sampling frequency is greater than twice that of the largest frequency component. This result comes from the Sampling Theorem, and justifies the modulation approach used for digital modulators. With sampling at a sufficiently high rate, it follows that the errors in the reconstructed

analogue signal will arise entirely from the amplitude quantization process. This latter process is in fact a lossy one. Sampling at a higher rate than required by the sampling theorem is called oversampling. The objective in analogue to digital modulator design is then to limit the distortion or errors introduced by amplitude quantization. The device which performs the conversion back to the analogue domain is called a digital-to-analogue converter (DAC).

The Σ-Δ modulator is a widely used and efficient type of oversampling quantizer that changes the frequency spectrum of quantization error or "noise", but leaves the input signal unchanged in its defined form within the modulator. This process of error or noise "shaping" is accomplished by feeding the quantization errors into a filter, and then feeding the "shaped" error from the filter output into a feedback loop within the modulator circuit. For audio systems, the shaped error is formed from the errors of time samples of the quantized error that are nearby in time, while for image systems, it is formed from the errors from neighbouring pixels of the quantized image in space. The use of oversampling allows the error noise to be shifted to higher frequencies, where it is less perceptible to human hearing or sight. In this thesis, we focus our attention on the Σ-Δ modulator where the application and corresponding model are concerned with audio systems. A different application of such a Σ-Δ modulator model is the pseudorandom number generator, as will be shown.

The Σ-Δ modulator is physically the signal amplitude quantizer and quantization error shaper component of the analogue-to-digital converter system. The Σ-Δ modulator input is the discrete time (for audio signals) samples of the signal amplitude. The output is the quantized or approximated forms of these sample amplitudes, to be transmitted or stored, and subsequently converted back to analogue form. The mathematical description and analysis of the Σ-Δ system may be carried out in either the real time state space

domain, or the frequency space domain. The frequency domain is best suited for the general model formulation, and for analysis of the system with the goal of improving design and performance. The use of frequency domain methods has thus been most prevalent in the development of Σ-Δ modulator related theory and work, and is the standard approach used by most electrical engineers who have contributed to this. The state space domain provides some advantages for the analysis of system dynamics and behaviour, and is best suited for the application of a dynamical systems approach.

A dynamical system is any physical system which can be described by a set of numbers or "state variables", which change with time according to deterministic law that may or may not be known to man [32]. The associated abstract definition normally involves a set of differential (continuous systems) or difference (discrete systems) equations described as mappings over a state space. The theory of dynamical systems is useful in understanding the behaviour of complicated nonlinear systems from a global or long term point of view [62]. For Σ-Δ modulators, such a description is particularly useful in giving an overview of equivalent systems which may have different implementations, and in providing the mathematical domain within which to specify a desired form prior to application [53]. Although the Σ-Δ modulator is mathematically a complicated nonlinear system, the dynamical systems approach is not widely used in electrical engineering, and has seen only limited use in the analysis of Σ-Δ modulator systems. As we shall see, analogue-to-digital converters, pseudorandom number generators, and even discrete modelling of some stochastic resonance processes, via their common Σ-Δ modulator topology, are all really discrete dynamical systems in disguise. The motivation and approach of the work of this thesis will be to use a dynamical systems formulation of the common Σ-Δ modulator model that underlies these examples, to pursue a study of the dynamical behaviour in a unifying fashion.

In this thesis, we study the Σ-Δ modulator where the amplitude quantizer Q is multi-

bit, and has an arbitrary number of levels or bits to accommodate its input. Such a general model retains applicability to practical, finite-bit systems. Bounded internal stability is always necessary practically, and will be examined briefly. This will form a prelude for the analysis approach of later topics.

One property of a dynamical system is that it may be formally characterized as chaotic or nonchaotic. A chaotic system may be thought of as a deterministic or nonrandom system that exhibits many of the important qualities of a stochastic or random system in its behaviour. Such characterizations, with suitable and precise mathematical definitions, thereby provide broad insights into the general dynamical behaviour of the system, and a delineation as to what type of specific behaviour to expect. In this thesis, we shall study the dynamical behaviour of the Σ-Δ modulator from the point of view of chaos, and seek to classify conditions under which chaos or nonchaos exists.

It is often desirable, in seeking to improve the performance of a Σ-Δ modulator, to add a deliberate, random noise or "dither signal" to some point in the modulator circuit. While the overall power of the quantizer errors or output noise will generally increase, other statistical properties of the noise, such as the degree of correlation with or dependence upon the input, or its own past, may be controlled by dither to improve the quality of the audio signal when the quantized output is converted back to analogue form. Partial dithering, as well as filter adjustment to bring about some properties of chaos, are ways of helping to prevent the occurrence of limit cycles in the system output. This is important because such limit cycles are synonymous with idle tones in the reconverted audio signal, which are particularly noticeable distortions to the human ear. This relationship exists because a limit cycle will contribute a prominent set of spurious frequency components to the audio signal. In this thesis, the chaos studies, and more broadly the dynamical systems approach, will then be extended to analyze the dithered Σ-Δ modulator system. Statistical

properties of the error dynamics will be investigated for the dithered case. In addition, a more theoretical analysis of stochastically described error dynamics will be carried out. This will serve as a conceptual bridge between the topics of deterministic chaos, and error statistics control.

The methods of analysis undertaken are naturally facilitated by the dynamical systems formulation. The circle map will be adopted as a convenient description of the quantizer error state space nature. The symmetry this provides will further support the use of topological relationships and techniques in formulating, proving and explaining results. The nonlinearity of the Σ-Δ modulator system, which is brought about by the quantization function, will be dealt with in the analysis by breaking the problem into linear subproblems, and by considering a certain class of systems to which linearity can easily be applied. In this approach, it will be seen that establishing continuity of the state space behaviour over initial conditions is sufficient to establish linearity. Stochastic or stochastic-like (i.e. chaotic) issues of the dynamical behaviour form a underlining theme that the theoretical approach in this thesis emphasizes. It will be seen, in general, that when there is some regularity, or structure, in the input or filter form of the Σ-Δ modulator, that theorems and results will arise to characterize a structured behaviour.

In the sections of this chapter to follow, the relevant research background and literature for this thesis will be presented in detail in the areas of signal processing, chaos in general for the Σ-Δ modulator, modelling and analysis of the Σ-Δ modulator for chaos, and the dithered quantizer and Σ-Δ modulator. The relevant mathematical background and methods for these areas will be presented in the respective sections as well. Further research motivations will be mentioned for chaos with the Σ-Δ modulator, and for the dithered quantizer and Σ-Δ modulator. The topology and basic mathematical description of the Σ-Δ modulator will be presented. The pseudorandom number generator, and

stochastic resonance will also be presented as examples of other practical or physical systems that function as Σ-Δ modulator-like dynamical systems. Finally, a brief overview of the structure of the body of the thesis will be given.

For the rest of the thesis, it will be assumed that the reader has sufficient mathematical background to either be aware of what a dynamical system is mathematically, or to be able to infer the essentials of its abstract definition and implications from the presentation of its application to the Σ-Δ modulator in this thesis. Therefore a rigorous definition of a dynamical system, along with its important attributes and properties will not be presented.

1.2 Signal Processing

In this section, we review some of the background and important developments in signal processing as this relates to the Σ-Δ modulator.

We begin with the Sampling Theorem, as introduced by Whittaker [66] and popularized by Nyquist and Shannon [60]. The definition of the Fourier transform is also given.

Fourier Transform *Let $x(t)$ be a complex valued, Lebesgue integrable function. Then the Fourier transform $X(f)$ of $x(t)$ is defined to be*

$$X(f) = \int_{-\infty}^{+\infty} x(t)e^{-2\pi ift}dt, \quad f \in \mathbb{R}.$$

The Sampling Theorem *Let $s(t)$ be an analogue signal in the time domain. Suppose the Fourier transform $S(f)$ of $s(t)$ exists and is square Lebesgue integrable. If $S(f) = 0$ for $|f| \geq f_s/2$, where f_s is the sampling frequency, then $s(t)$ is recoverable from its time samples $s(k/f_s)$, $k = \ldots, -2, -1, 0, 1, 2, \ldots$, according to*

$$s(t) = \sum_{k=-\infty}^{+\infty} s(\frac{k}{f_s}) \cdot \frac{\sin(\pi f_s(t - k/f_s))}{\pi f_s(t - k/f_s)}.$$

Note that the signal $s(t)$ is band limited, residing entirely in the base-band $-f_s/2 < f < f_s/2$. $f_s/2$ is called the Nyquist frequency. This theorem justifies the sampling process used for digital conversion and analogue reconstruction. The process of oversampling exceeds the conditions of this theorem and mathematically yields the spectrum space that permits noise shaping.

Oversampling techniques for converting signals between analogue and digital formats have become popular in recent years since they avoid many difficulties encountered in conventional methods for analogue/digital and digital/analogue conversion. Oversampling converters can use simple, relatively high tolerance analogue components to achieve high resolution, but they require fast and complex signal processing stages. They operate by converting (through modulation) the analogue signal into a simple code (typically 1-bit words) at a frequency many times the Nyquist rate (twice the signal bandwidth). Digital filters are then used to process the modulator output and reduce noise and high frequency components of the signal which could alias into the signal band when the code is resampled at the slower (Nyquist) rate. Oversampling is suitable for low-frequency signals and it takes advantage of VLSI (very large scale integration) technology which is best suited to provide faster digital but less precise analogue circuits. Oversampling is used in applications such as digital audio, digital video, digital telephony and instrumentation. Further applications in video and radar systems are imminent as faster technologies become available.

Work on oversampling originally turned to the Σ-Δ modulator from other types because its circuits are more robust. The Σ-Δ modulator shapes the noise spectrum by using the properties of oversampling to move the noise power to high frequencies, well outside the base-band signal, where it is removed by digital and/or analogue filtering. The important attraction of the Σ-Δ approach is the large amount of base-band noise reduction which may be obtained by this process with relatively few bits in the code. Other motivations for

Σ-Δ converter use are that it can be cheaper to use, and the potential for making it much better in the 3 to 5-bit case.

In 1960, Cutler [6] introduced the idea of a deterministically "dithered" system, where the filtered quantizer noise itself is used as a dither signal which is added to the input signal before quantization[1]. This intuitively may be thought of as approximating the adding of an independent identically distributed random process (white signal-independent process) to the signal before quantization. The system, in first order, may be described by the following nonlinear difference equation:

$$w_n = x_n - \varepsilon_{n-1} = w_{n-1} + x_n - Q(w_{n-1}), \quad n = 1,\ 2, \ldots,$$

where x_n is the original system input, w_n is the input to the quantizer Q, ε_n is the quantizer error, and n represents the nth element of the particular sequence. Candy and others (see [4] and references contained therein) developed much of the original analysis, the modern popularity of these systems and the name Σ-Δ reflecting the system viewed as the cascade of an integrator (Σ) and a differentiator (Δ) (although the name Δ-Σ was originally given by Inose and Yasuda [25], [24] in 1963 in their description of its properties, and remains most used). In general, quantization errors are called noise if the errors are uncorrelated from sample to sample and have statistical properties independent of the input signal. The case of input-independent additive white noise holds when the ε_n are independent of the

[1]Cutler patented the noise shaper topology in 1960. He "explained" its performance as being improved by the action of the recirculated error and acting as a "quasi-dither", although it is <u>not</u> an independent additive noise, but deterministically related to the signal.(see Figure 1.2)

Noise shaping (Cutler) can only produce the theoretical power spectral density (PSD) $|1 - H|^2$ independent of the signal, if the PSD of the error ε is made white and signal independent; and this can be done only if an independent additive dither ν is used in the circuit.

Oversampling will generally be necessary to restrict the number of levels needed in the quantizer. This is essential for 1-bit (i.e. 2-level) and other low-bit systems. [34, 35, 65]

w_k, ε_n is an independent identically distributed sequence, and ε_n is uniformly distributed (note that weaker versions of these conditions exist).

We make these statistical properties of a random process precise below:

Definition 1.1 (i.i.d.) *Let the set $S_X = \{X_n,\ n \geq 0\}$ define a random process. Then S_X is said to be independent and identically distributed (i.i.d.) if the elements of every finite subset of S_X are jointly independent, and each such element has the same marginal probability distribution.*

Definition 1.2 (Whiteness) *Suppose that $\lim_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T} E[X_n] = 0$, and $E[X_n^2]$ exists for all X_n in the set S_X defined above. Then S_X is said to be white if the elements of S_X are, on average, pairwise uncorrelated according to the following condition: $\lim_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T} E[X_n X_{n+\tau}] = 0$, for all $\tau \in \mathbb{Z} - \{0\}$. If S_X is i.i.d. with the above moment conditions holding, then S_X is white (converse is not always true).*

The whiteness property essentially extends to any set $S_Y = \{Y_n,\ n \geq 0\}$ with $\lim_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T} E[Y_n] = K$ for some $K \in \mathbb{R}$, if $Y_n = X_n + K$, $n \geq 0$, and S_X is white.

The definition of whiteness has the following relevance to signal processing: Suppose that the signal $s(t)$ is recoverable from its samples $s(k/f_s),\ k = 0, 1, 2, \ldots,$ according to the Sampling Theorem. Then the Fourier transform $E(f)$ of the quantizer error signal $e(t)$ associated with $s(t)$ will be constant over $-f_s/2 \leq f \leq +f_s/2$ (f_s the sampling frequency) if and only if the process $e(k/f_s),\ k = 0, 1, 2, \ldots,$ is white, when interpreted as a random process. This result may be shown mathematically. If this process is strictly deterministic, then the whiteness definition may still be applied with $E[e(k/f_s)] = e(k/f_s)$ and $E[e(k/f_s)e((k+\tau)/f_s)] = e(k/f_s)e((k+\tau)/f_s),\ k \geq 0$. Having the unshaped noise

signal $e(t)$ (associated with the reconstructed output) with all its frequencies equally represented in its power spectrum (i.e. in an physical sense "white") is then equivalent to the corresponding process $e(k/f_s),\ k = 0, 1, 2, \ldots,$ being white.

For simple single-loop systems with a 1-bit quantizer, Gray [17] shows that the Σ-Δ modulator does not yield a white quantization noise process as the "deterministically dithered" idea might suggest. In this analysis, the overall effect of the feedback loop appears as an affine operation (adding the previous inputs plus a bias) on the input followed by a memoryless nonlinearity (taking the fractional part), when one considers the expression for ε_n. Further results for the case of a constant irrational DC input (where this input is normalized with respect to the quantization step size), show that although the first-order statistical properties of the errors are consistent with those of uniform white noise, the second-order properties are not (i.e. the locations and amplitudes of the spikes in the quantizer error spectrum depend strongly on the input signal). Thus the quantizer noise is neither continuously distributed nor white and this approximation is incorrect.

No simple mathematical solution for the second-order or higher-order Σ-Δ modulator exists. Evaluation of the statistical properties of the errors to compare with white noise remains an open problem although some other results have been found by many researchers. The first-order case in fact is the only completely analyzable case.

In [11], [16] and [28], the case of a constant rational DC input to a single-loop Σ-Δ modulator is analyzed. Friedman [11] obtains noise spectra results for the second-order case which Iwersen [28] obtains more explicitly and quickly by applying other published results. Iwersen [28] says these results have limited applicability. Gray [16] obtains statistical results for noise for the first-order case with DC (i.e. constant) input analogous to the irrational DC input case above, and shows analogously that for rational DC input, the noise is not white. Rational inputs are normally of little interest as they occur statistically

with zero probability, but simulation of analogue/digital conversion on a digital computer necessarily creates rational input. In this case one must make the input "nearly" irrational (use a large denominator) to give limit cycles with long periods, which approximates the irrational generic case.

For the Σ-Δ modulator applications focused on in this thesis, the analogue input signal $s(t)$ defines an audio pattern as a function of time that may be physically converted directly back to sound in a speaker. Quantitatively, the numerical value of $s(t)$ is proportional to the pressure deviation, above or below atmospheric pressure, characterizing sound waves and measured in a microphone. Audio systems are generally electrical, and in such systems the pressure variations (of the deviation from atmospheric) are converted to voltage variations in a circuit, with the physical quantity $s(t)$ representing a voltage. If the voltage $s(t)$ is constant, for example, it is DC, and the input signal is referred to as being DC. Σ-Δ modulators in audio and other applications are generally electrical systems in this fashion, and hence their fundamental study in the discipline of electrical engineering.

1.3 Σ-Δ Modulators and Noise Shapers

In this section, the basic form of the Σ-Δ modulator will be introduced and the characteristics relevant to the work of this thesis described.

The most general Σ-Δ modulator, in frequency domain form, is defined by the topology given in Figure 1.1(a) below. Here, X is the transform of the input, Y is the transform of the quantized output, Q is the quantizer, N is the transform of the dither added to the quantizer input W, and F and G are filter transfer functions acting on the feedback Y. This topology may be rearranged to give the completely equivalent form shown in Figure 1.1(b) below. This equivalent system, with its transfer function blocks expressed in terms

of F and G, is called a noise shaper.

The purpose and advantage of using the noise shaper form of the topology of Σ-Δ modulators is that it explicitly expresses the error E of the dithered quantizer in the diagram. The error is the quantity of prime importance in the general Σ-Δ modulator in that it is the quantity to be studied and controlled when analyzing and designing the system in terms of performance. The term "noise shaper" comes from this concept of the Σ-Δ modulator as a system with a filter designed to shape or control the spectrum of the noise or errors arising from the quantization. The error is also generally the central state space quantity when considering the Σ-Δ modulator as a dynamical system, as shall be seen. In this section, we neglect the dither N, and equivalently set $N = 0$. The quantizer Q is shown as a multi-bit quantizer in Figures 1.1(a) and 1.1(b). The 1-bit Σ-Δ modulator or noise shaper is a special case. In this thesis we shall generally assume Q to be a multi-bit quantizer.

The transfer equation for the noise shaper is given in Figure 1.1(b). This can be obtained by applying simple algebra rules at each juncture in Figure 1.1(b), and then simplifying to get the resulting transfer equation. For this, we have

$$\begin{aligned} Y &= U + E = (X - R)\frac{G}{1+FG} + E \\ &= (\frac{G}{1+FG}X - \frac{FG}{1+FG}E) + E = \frac{G}{1+FG}X + \frac{1}{1+FG}E, \end{aligned}$$

where U and R are the internal signals at the indicated points. To simplify the noise shaper form in Figure 1.1(b) for analysis, the upper transfer function block may be pulled behind the left hand summation junction to then multiply with the X and the lower transfer function block F. We then relabel $\frac{G}{1+FG}X$ as X, and $\frac{FG}{1+FG}$ as H. This gives an equivalent form of noise shaper to that of Figure 1.1(b), with the upper block transfer function set to 1, and the lower block transfer function relabelled as H. In this form, the

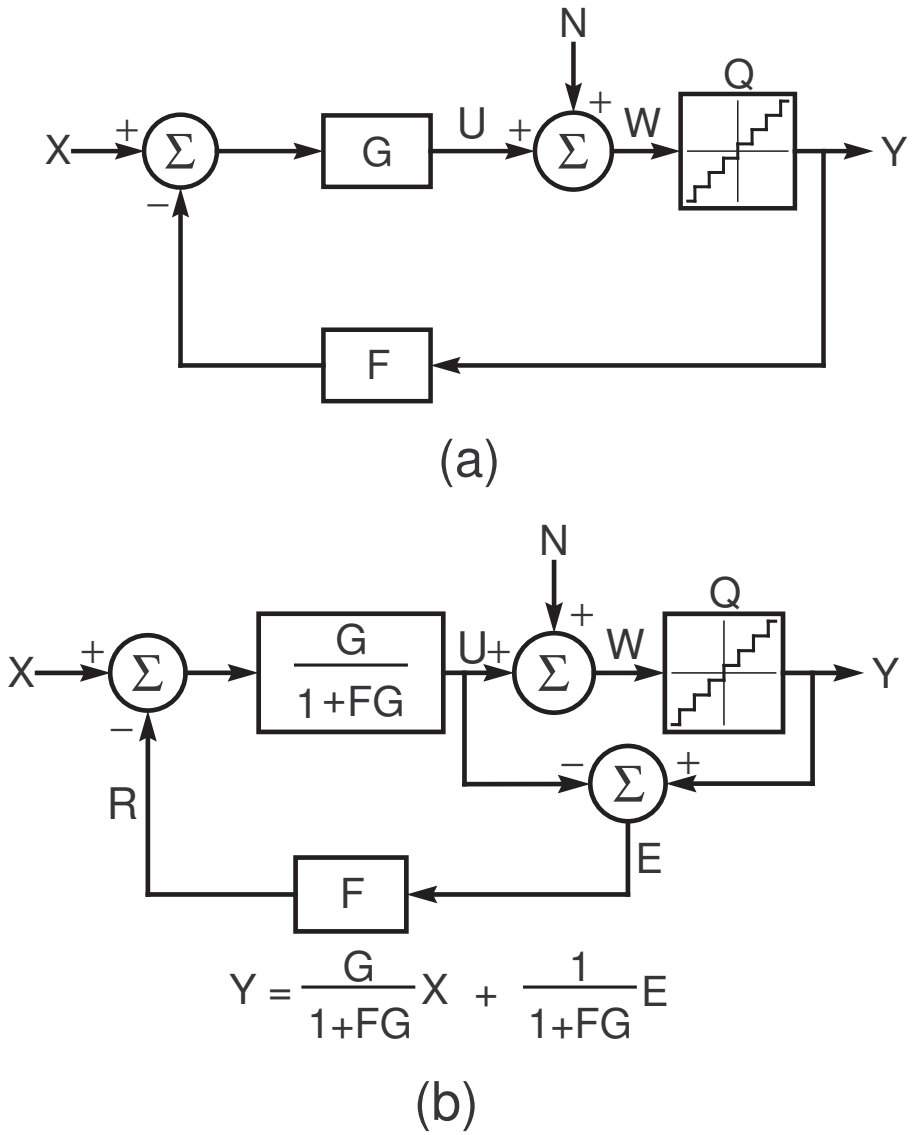


Figure 1.1: Σ-Δ modulator in (a) general form, and (b) noise shaper form

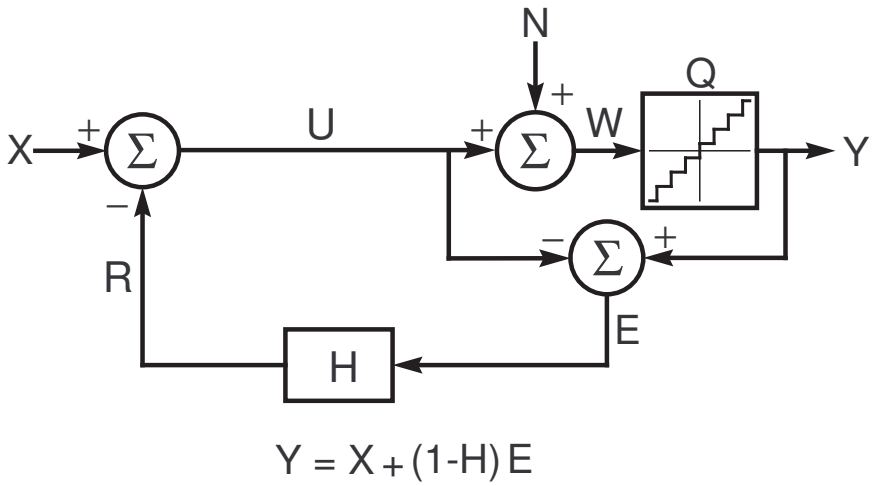


Figure 1.2: Simplified noise shaper

effective input X we are considering is equal to the actual input fed through the signal transfer function $\frac{G}{1+FG}$. For the types of noise shapers that are to be considered in the work of this thesis, there is no loss of generality if we set $G = 1 + FG$, so that the effective X is the original X, and the signal transfer function is 1. The transfer equation now has the form

$$Y = X + (1 - H)E.$$

The transfer function of the noise shaper, the noise transfer function, is thus $(1-H)$. That is, the error E appears in the output Y modified by the effective noise-shaping filter $(1-H)$ — hence the designation "noise shaper". This simplified circuit is shown in Figure 1.2.

Since we are considering the Σ-Δ or noise shaping stage to follow the sampling operation, the signals in Figures 1.1(a), 1.1(b) and 1.2 are discrete-time signals, and so the signal quantities and the function H in this equation correspond to the z-transform domain. They are obtained by taking the z-transform of the corresponding quantities in the time domain (sequences of input, output and error). The z-transform is the discrete version of

the Laplace transform, and is defined as follows:

z-Transform *Let $\{x_n\}$ be discrete sequence of real or complex numbers. Then the (bilateral) z-transform $X(z)$ of $\{x_n\}$ is defined to be*

$$X(z) = \sum_{n=-\infty}^{+\infty} x_n z^{-n}, \quad z \in \mathbb{C},$$

whenever the $\{x_n\}$ are summable.

Operationally, products in the frequency domain correspond to convolutions of sequences in the time domain. In a generalization of this, the product of the transfer function $(1-H)$ with E in the z-transform domain corresponds in the time domain to a convolution of the error sequence with a sequence of coefficients, representing the impulse response of the noise shaping filter $(1 - H(z))$. The impulse response is the sequence obtained from the transfer function by inverse z-transformation.

Considering the time domain, the filter H of the simplified Σ-Δ modulator of Figure 1.2 has the basic topology given in Figure 1.3. As shown in the diagram, the filter H can contain both feedforward and feedback elements which generate delay contributions from the filter input ε_n and output r_n respectively. With this specific filter architecture, the operation of multiplying the filter function H with the error E is then equivalent in the time domain to subtracting a convolution of the sequence of feedback quantities r_n with a sequence of feedback "gain" factors b_j, from a convolution of the feedforward error sequence ε_n with a sequence of feedforward gain factors a_i. These gain factors are associated with H. This relationship established by H, between a sum of a series of time delays (multiplied by a gain) of the error input and feedback output, is given as follows, where H is defined by the z-transform of the summable time domain sequences $\{h_{1,i}\}$, $\{h_{2,j}\}$:

$$h_{1,i} * \varepsilon_{n+i} - h_{2,j} * r_{n+j} = \sum_{i=-\infty}^{+\infty} a_i \varepsilon_{n-i} - \sum_{j=-\infty}^{+\infty} b_j r_{n-j} = 0, \tag{1.1}$$

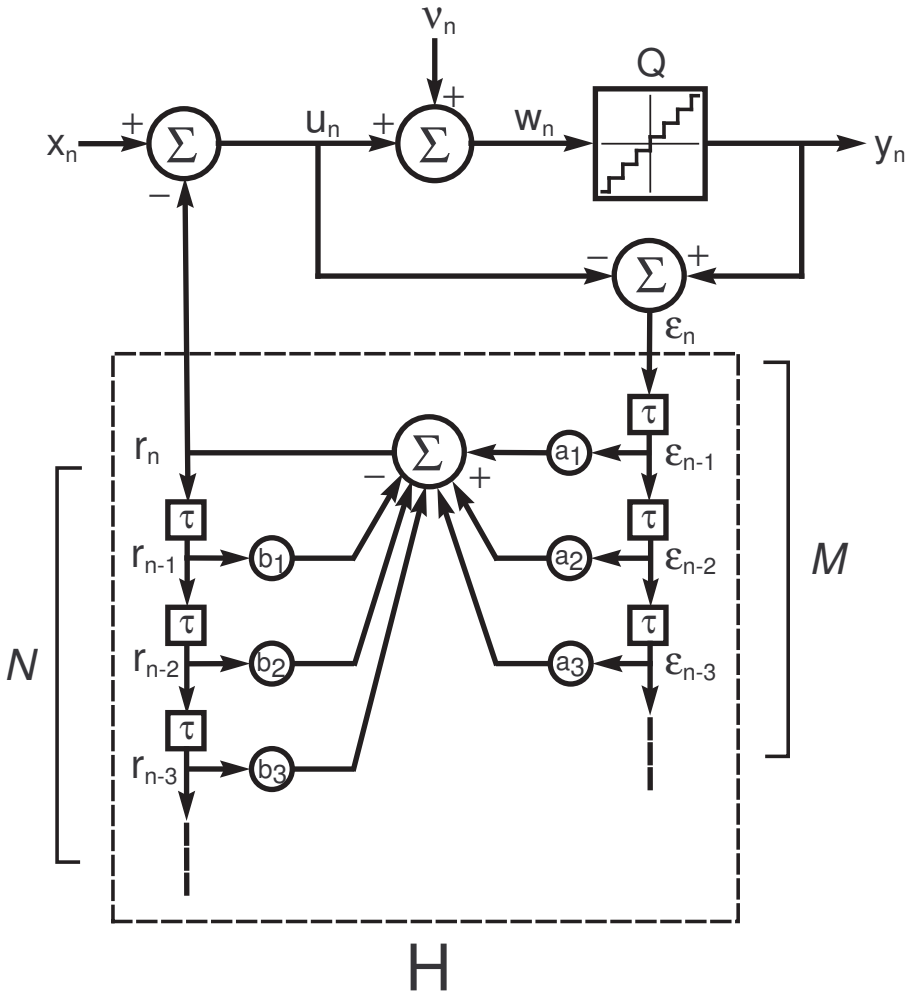


Figure 1.3: Topology of filter H and noise shaper in time domain

with $b_0 = 1$, $a_i = 0$, $b_j = 0$ for $i \leq 0$, $j < 0$, and $i > M$, $j > N$, for some $M > 0$ and $N \geq 0$. We also require that $a_i \neq 0$ for some i for the filter to be well defined. The latter condition ensures at least a 1-sample delay so that the circuit is recursively computable. The sequences $\{h_{1,i}\}$, $\{h_{2,j}\}$, $i, j \in \mathbb{Z}$, are defined by $h_{1,i} = a_i$, $h_{2,j} = b_j$. This description means H is an Mth order feedback and Nth order feedforward filter. The series of coefficients defined by the transfer function in the convolution is determined uniquely by the choice of the a_i and b_j in the filter H above. If all the b_j except b_0 are zero, then we have a finite impulse response (FIR) noise shaper. Otherwise, the filter will generally be infinite impulse response (IIR).

In the time domain the quantizer Q acts as a modulo operation with a fixed number of levels or steps that span the domain of its input, and with the modulo factor Δ being the distance between steps. Quantizers in use are typically of either mid-riser or mid-tread form. The functional form of the mid-riser and mid-tread quantizers are illustrated in Figure 1.4. The mid-riser form is generally used for Σ-Δ modulator noise shapers, and hence will be the form that we take for the model studied in this thesis, and present in this section. The mid-tread form tends to be used for Σ-Δ topologies with numerical applications, such as the pseudorandom number generator, as will be presented in Section 1.7.

Quantizers are generally designed to represent a certain number of bits of information. Specifically, an n bit quantizer will have 2^n levels and so quantizers will usually be designed with the number of levels a power of 2. Most typical Σ-Δ modulators in use have between 1 and 5 bits (i.e. 2 to 32 levels). Using the mid-riser form then ensures symmetry in the output domain, given this even number of levels. This is particularly important when the Σ-Δ system has a low number of bits, and is essential for the 1-bit case. The mid-tread form, with its integer roundoffs, is more natural numerically, and is standardly used in other applications.

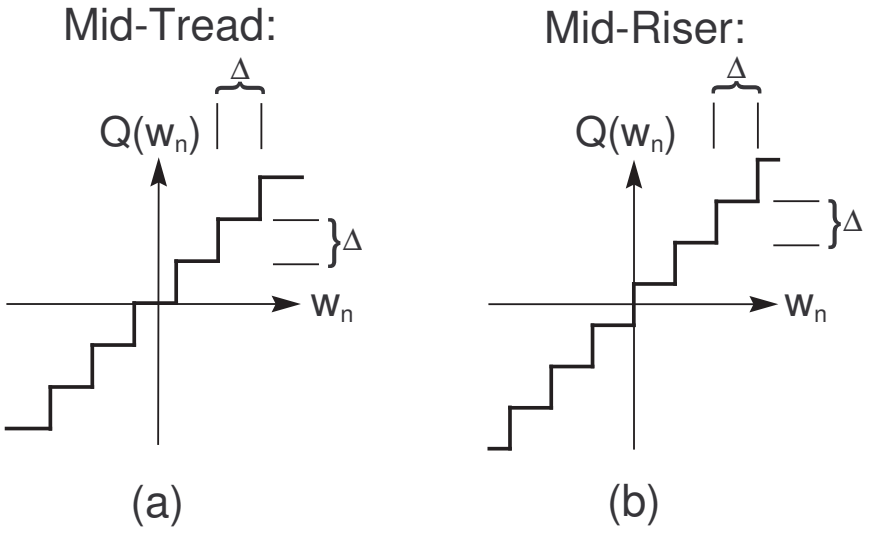


Figure 1.4: Functional form of mid-tread and mid-riser quantizer

In the system of equations (1.2) below we will assume that Q has an arbitrarily large number of levels (and hence bits) and can thus accommodate an input of arbitrarily large magnitude. This means that we will not be concerned with questions related to quantizer overload in this thesis. Such a multi-bit Q may still be used to model a finite bit case, however, if it is assumed a priori that quantizer overload does not occur. Specifically, suppose that the number of quantizer bits is $b \geq 1$ ($2 \leq b \leq 5$ is most common). The no overload condition means that the quantizer input must fall within the range of 2^b quantizer levels, that is within the interval $[-2^b\Delta,\ 2^b\Delta)$, where the 2^b levels reside at $-((2^b-1)/2)\Delta, \ldots, -(1/2)\Delta,\ (1/2)\Delta, \ldots, ((2^b-1)/2)\Delta$. With this condition, the b-bit case is just a special case of the general form as considered.

From the topology of Figure 1.3, the form of H and the quantizer, the following

$\max(N, M)$th-order system of difference equations may then be constructed:

$$\begin{aligned} r_n &= \sum_{i=1}^{M} a_i \varepsilon_{n-i} - \sum_{j=1}^{N} b_j r_{n-j} \\ \varepsilon_n &= Q(x_n - r_n) - (x_n - r_n), \end{aligned} \tag{1.2}$$

for $n \geq 0$, where the mid-riser quantizer Q is defined by

$$Q(w_n) = \Delta \lfloor w_n/\Delta \rfloor + \Delta/2, \quad w_n \in \mathbb{R}, \quad -\Delta/2 < \varepsilon_n \leq \Delta/2, \quad n \geq 0,$$

the system input is x_n, and the system output is given by

$$y_n = Q(x_n - r_n) \in (\mathbb{Z} \cdot \Delta + \Delta/2), \quad n \geq 0.$$

The $M + N$ initial conditions are: $\varepsilon_{-1}, \varepsilon_{-2}, \ldots, \varepsilon_{-M}$; $r_{-1}, r_{-2}, \ldots, r_{-N}$. For this system, we give $\varepsilon_i \in (-\frac{\Delta}{2}, \frac{\Delta}{2}]$; r_j, $x_n \in \mathbb{R}$; for all $i \geq -M$, $j \geq -N$, $n \geq 0$; with the a_i, $b_j \in \mathbb{R}$. We take the dither ν_n in Figure 1.3 to be zero in this formulation. The first line of system (1.2) follows directly from (1.1) as well. Such a system forms a basis for defining the Σ-Δ modulator in this form as a discrete dynamical system. This is called a $\max(N, M)$th order Σ-Δ modulator.

An n-sample time delay operation in the time domain corresponds to multiplication by z^{-n} in the z-transform domain. Applying the z-transform to the time domain relationship (1.1) given above for the function H (or equivalently the difference equation in (1.2)), and then rearranging the result to give a ratio of polynomials on one side, we have the following in the frequency domain:

$$R(z) = \frac{a_1 z^{-1} + \ldots + a_M z^{-M}}{1 + b_1 z^{-1} + \ldots + b_N z^{-N}} E(z) \equiv H(z) E(z).$$

The noise transfer function (NTF) may then be defined as follows:

$$(1 - H) = \frac{1 + (b_1 - a_1) z^{-1} + \ldots + (b_{\max(N,M)} - a_{\max(N,M)}) z^{-\max(N,M)}}{1 + b_1 z^{-1} + \ldots + b_N z^{-N}}. \tag{1.3}$$

In order to study Σ-Δ modulators as noise shapers, it is important to classify the transfer function $(1 - H)$ as minimum or nonminimum phase.

Definition 1.3 (Minimum/Nonminimum Phase) *A function $X(z)$ in the frequency domain, having all poles of magnitude less than one (inside the unit circle on the complex plane), is defined to be*

1. *strictly minimum phase if all the zeros of $X(z)$ have magnitude less than one;*

2. *marginally minimum phase if the zero(s) of largest magnitude of $X(z)$ have magnitude one (on the unit circle);*

3. *nonminimum phase if at least one zero of $X(z)$ has magnitude greater than one (outside the unit circle).*

If the noise transfer function $(1 - H)$ in (1.3) is minimum/marginally minimum/nonminimum phase, then we simply denote the Σ-Δ modulator system as minimum/marginally minimum/nonminimum phase respectively. By analogy to this definition, we will refer to a given zero of $X(z)$ as being minimum/marginally minimum/nonminimum phase if it has magnitude less than/equal to/greater than one respectively, and is counted once if it has magnitude one, multiplicity greater than one. A zero of magnitude one, multiplicity greater than one, will be loosely referred to as nonminimum phase when counted the extra (multiplicity $- 1$) times, while zeros with magnitude greater than one will be distinguished with the term "strictly nonminimum phase".

1.4 Chaos and the Σ-Δ Modulator

In this section, the approach to studying chaos in this thesis, and the relevant background and motivation will be discussed.

The study of the properties of chaos in the Σ-Δ modulator is important in seeking to gain a clearer picture of its overall dynamical behaviour, particularly since random-like or possible "chaotic" behaviour has been observed in simulations. Work in various electrical engineering papers, to be mentioned later, has investigated chaotic behaviour. A common suggestion is that the Σ-Δ modulator is chaotic if and only if it is nonminimum phase. It appears, however, that no adequately thorough explanation of this relationship has been established in this work. Furthermore, none of the research literature provides satisfyingly rigorous proof of the conditions for chaos based on thorough and precise definitions of chaos, for systems of such generality. There is a need for a comprehensive approach to investigate the conditions for chaos in the Σ-Δ modulator that seeks to remedy these deficiencies, and establish broad results concerning an array of chaotic properties of the dynamics. An important motivation and goal of this thesis is then to provide at least a fundamental first step in this endeavour. As a result of this, the approach of this thesis will be of a more theoretical and abstract nature, than is typical of previous work.

There are many possible definitions for deterministic chaos in a dynamical system, and no single generally accepted one. All approaches begin from the general requirements given by Li and Yorke (1975) [33] who first introduced the term "chaos" into the study of dynamical systems. The requirements on the practical behaviour of a dynamical system for it to be chaotic include [32]:

1. Solutions stay in a bounded region.
2. Solutions never settle down to periodic behaviour or an equilibrium point.
3. Any two solutions which start arbitrarily close to each other rapidly separate and

effectively behave independently.

4. It is impossible to predict any solution very far into the future, given finite initial information.

These conditions, taken together, imply an observed deterministic behaviour that may be thought of as mimicking that of a random process over time. In this sense, condition 4 is meant to imply, for any initial condition, an observed random-like behaviour after some finite period of time. Condition 2, in turn, implies that this type of behaviour persists for ever. Condition 1 would imply that the observed "process" is bounded in magnitude over time. Condition 3 implies that two different solutions would be observed to behave as independent random-like processes after a finite period of time. This third condition then implies a sensitivity to initial conditions. Formal mathematical definitions of chaos seek to incorporate these general, observational characteristics into precise conditions defined in a deterministic context. The relative importance of a particular definition is normally governed by the mathematical form of the particular dynamical system to which it is being applied. In this context, one could take into account theoretical questions, the ability to prove chaos results analytically or numerically, as well as the practical purpose at hand for studying chaos.

In seeking a way to define and identify chaos in the Σ-Δ modulator, we first consider Lyapunov exponents, defined below.

Lyapunov Exponents *Let $f : \mathbb{R}^n \to \mathbb{R}^n$, $f \in C^1$, and f invertible, so $\{\mathbb{R}^n, f\}$ is a dynamical system. Consider the orbit $\gamma(\vec{x}_0) = \{\vec{x}_k = f^k(\vec{x}_0) \mid k \in \mathbb{Z}\}$. Let the linearization be $J_k \equiv D[f^k(\vec{x}_0)]$, $k \in \mathbb{Z}$, where D is the operator that maps a C^1 vector valued function to its Jacobian matrix of partial derivatives. Let $r_1^k \geq r_2^k \geq \ldots \geq r_n^k \geq 0$ be the square roots of the eigenvalues of the matrix $J_k J_k^T$ (nonnegative real numbers called the singular values*

of J_k), for each k. Then the mth Lyapunov exponent of the orbit $\gamma(\vec{x}_0)$ is defined as

$$\lambda_m(\vec{x}_0) = \lim_{k\to\infty}(\frac{1}{k})\ln|r_m^k|, \quad 1 \leq m \leq n,$$

where existence of this limit follows from Oseledec's Theorem.

The existence of the Lyapunov exponent is ensured since the conditions of Oseledec's Theorem are met by f. Lyapunov exponents generalize to chaotic orbits the property that the eigenvalues at an equilibrium point measure rates of expansion (and contraction) in "eigendirections" at that point. From this and the definition of Lyapunov exponents above, we have the following result: if at least one Lyapunov exponent is bigger that zero, then there are solutions which diverge from the solution with the given initial condition. This is then a prerequisite for chaos, as it is required to uphold sensitivity to initial conditions.

A Lyapunov exponent approach naturally lends itself to the use of numerical methods for identifying chaos, as demonstrated by the following two papers. In [69], algorithms and Fortran programming code are presented that allow the estimation of non-negative Lyapunov exponents from an experimental time series and for systems of D.E.'s. For the time series, the technique of phase-space reconstruction with delay co-ordinates is used to obtain an attractor with the same Lyapunov spectrum. Using the methods from this paper, one could write a program to input a time series generated from a Σ-Δ system of known form, and output the largest one or two Lyapunov exponents. In [8], the QR based method (where "Q" and "R" refer to relevant matrices) for computing the first few Lyapunov exponents of continuous and discrete dynamical systems is considered. This also provides a possible way of finding Lyapunov exponents for some Σ-Δ systems, by applying the algorithms given to the analytic form of the Σ-Δ modulator as a dynamical system and then evaluating these numerically. In both papers, the existence of one Lyapunov exponent greater than zero is taken as a sufficient condition for the characterization of chaos, and

the algorithms of [69] are described as detecting and quantifying chaos. The definition of Lyapunov exponents, in analytical form, does not appear to be easily amenable to the analytical form of the discrete dynamical systems corresponding to Σ-Δ modulators, due to the complicated nonlinearity (modulo function) involved. Therefore, a specific definition of chaos that does not involve Lyapunov exponents was sought for use. The numerical approaches mentioned above were not pursued for this thesis.

Due to the difficulty of fitting realistic systems into rigorous mathematical definitions for chaos, working or operational definitions are usually used that relate to the four requirements given above and include some global characteristics. A positive Lyapunov exponent is one such characteristic. The positivity of topological entropy is another type that is used to define "topological chaos". Fractal attractor dimension is another type. In studying the modulo difference equations that define Σ-Δ modulators as discrete dynamical systems, it can be seen that these maps are essentially maps on the unit circle (after scaling). The definition of chaos used to define "rotational chaos" as given in Hao [20] was thus next considered.

Rotational chaos is defined to exist when the system possesses a rotation interval rather than a rotation number (global characteristic). The rotation number and rotation interval for a circle map are defined below.

Rotation Number and Rotation Interval *Let* $\theta_{n+1} = F(\theta_n) \mod 1 = f(\theta_n)$ *define a circle map, where* θ *is a normalized angle variable, and the function* $F(\theta)$ *takes into account the cumulative effect of rotation. The rotation number of the circle map* f *is then defined by*

$$\rho(\theta_0) = \lim_{n\to\infty} \frac{F^n(\theta_0) - \theta_0}{n}.$$

The rotation number may acquire different values for a different choice of the initial point

θ_0. The rotation interval of f is defined as the set of all rotation numbers of f generated by all the initial conditions $\theta_0 \in [0, 1)$. Note that in the definition given by Hao [20], the requirement $f(\theta + 1) = f(\theta)$ is made for f.

The rotation number essentially measures the average number of times that the mapping achieves a complete rotation of the circle per iteration. Ito [27] proves that the rotation interval is a closed interval. Newhouse et al. [46] study bifurcation and stability issues for circle mappings. These papers reveal the existence of a rotation interval, and its relevance to chaotic motion in circle mappings. Gambaudo et al. [12] demonstrate that the rotation interval can be used as a quantitative measure of chaos, and show that this interval can be determined by computing the rotation numbers of two monotonic maps related to F, which form the interval endpoints. MacKay et al. [40] describe all routes to positive topological entropy (chaos) from zero for circle maps and discuss the relevance to the transition to chaos. Topological entropy is compared with the rotation interval, which is asserted to be a more sensible criterion for chaos. Casdagli [5] investigates the chaotic attractors that arise using the rotation interval and investigates the relationships between the rotation interval and topological chaos. Numerical algorithms to calculate the rotation interval given an appropriate map, D.E. or time series are presented. The term "rotational chaos" is also first given. Although the notion of a circle map pertains to a one dimensional system, a means for defining rotational chaos for an n-dimensional system arising from an nth order Σ-Δ modulator (which would have an "n-dimensional" circle map or "torus" map) we presume would be possible.

In continuing to pursue a suitable definition of chaos to apply to Σ-Δ systems, the particular definition established by Devaney in [7] was arrived at as the most desirable. This definition, which possesses a balance between rigour and simplicity, was chosen because of its importance in applying to a large variety of dynamical system mappings and its relative

ease for analytical verification for circle maps of the form describing Σ-Δ systems. The definition is as follows:

Definition 1.4 (Devaney's Chaos) *Let $V \in \mathbb{R}$ be a set. $f : V \longrightarrow V$ is said to be chaotic on V if*

1. *f has sensitive dependence on initial conditions;*
2. *f is topologically transitive;*
3. *periodic points are dense in V.*

The appropriate definitions of sensitive dependence on initial conditions, topological transitivity, and density of periodic points used here are given below.

Definitions for Devaney's Chaos:

For the respective definitions below, let $J \in \mathbb{R}$ be a set.

Definition 1.5 (Sensitivity) *$f : J \to J$ has sensitive dependence on initial conditions if there exists $\delta > 0$ such that, for any $x \in J$ and any neighbourhood N of x, there exists $y \in N$ and $n \geq 0$ such that $|f^n(x) - f^n(y)| > \delta$.*

Definition 1.6 (Transitivity) *$f : J \to J$ is said to be topologically transitive if for any pair of open sets U, $V \subset J$ there exists $k > 0$ such that $f^k(U) \cap V \neq \emptyset$.*

Definition 1.7 (Density of P.P.s) *Suppose $f : J \to J$ has the set of periodic points $U \subseteq J$ where $U = \{x \in J \mid \exists\ n \geq 1$ with $f^n(x) = x\}$. The set U is dense in J if, for any $x \in J$ and any neighbourhood N of x in J, there exists $y \in N$ such that $y \in U$.*

In this conception of chaos, the three key ingredients identified by Devaney of unpredictability, indecomposability and an element of regularity are satisfied by conditions 1, 2, and 3 respectively, above. This definition of chaos is stronger and hence more restrictive

than the original definition set out by Li and Yorke, essentially by virtue of the third condition[2]. These three conditions capture the ideas of the observational requirements presented at the beginning of this section as follows: sensitivity captures requirement 3, transitivity requirements 1 and 2, while requirement 4 is captured by all three chaos conditions taken together. It is then the strength of Devaney's definition to get at the roots of chaos, without admitting "marginal" chaos, that is seen as an advantage. Devaney's definition is also well suited for characterizing chaos on subsets (of state space) or attractors. A disadvantage is that for continuous mappings, transitivity and density of periodic points taken together imply sensitivity, thus making the first condition redundant[3] [55]. For the more complex types of mappings that will arise from the Σ-Δ modulator model however, this redundancy will generally not be automatic, at least.

The definition of chaos used in this thesis is based on the basic Devaney form above. Devaney provides extensions of his definition to higher dimensions when considering the horseshoe map, hyperbolic toral automorphisms, and attractors in [7]. An extension of Devaney's chaos definition in our work to apply to $\mathbb{R}^n$, as needed to apply to an nth order Σ-Δ modulator system, along with other adaptations, are made as given in the formulation, and discussed, at the beginning of Chapter 5.

In the investigation and classification of Σ-Δ modulator systems for chaotic behaviour

[2]Strictly speaking, removing the third condition gives Wiggins definition of chaos. Chaos in the sense of Wiggins in turn implies chaos in the sense of Li and Yorke (which is a slightly weaker definition). The precise definition of Li and Yorke requires the existence of an uncountable scrambled set (a scrambled set may be thought of as a set where the orbits of any two elements are for ever becoming both arbitrarily close, and sufficiently separated from one another). Their original definition also required the existence of periodic points of all periods. See [55] and references contained therein.

[3]In fact for continuous interval maps transitivity implies both sensitivity and density of periodic points. On closed subsets, however, this is not true and Devaney's chaos is meaningful and gives an equivalency with positive topological entropy [55].

undertaken in this thesis, the scope of systems considered will be extended to include cases that are not purely deterministic. Specifically, we will consider some systems with a random or stochastic signal component (i.e. an input and/or dither component). While the concept of chaos is meant to apply to purely deterministically determined aspects of a system's dynamics, we see no inconsistency in applying a characterization of chaos or nonchaos to a system whose dynamics may be driven by both stochastic and deterministic processes. If the system is classified as nonchaotic, then clearly the legitimately nonchaotic deterministic aspects dominate the stochastic ones. If the system is classified as chaotic, then we do not presume that this necessarily results from its stochastic aspects. The chaos may arise from legitimately chaotic deterministic aspects, stochastic aspects alone, or by some combination of its stochastic aspects and some additional near chaotic deterministic aspects. No particular attempts will be made to determine how the chaos arises. We simply allow a classification of chaos or nonchaos to be made, so that a comparison of the system dynamics can be made with those of other systems, and a clearer understanding of the dynamics can be obtained.

1.5 Modelling and Analysis of the Σ-Δ Modulator for Chaos

In this section, we review the previous research involving chaos in the Σ-Δ modulator as relevant to this thesis. We also look at some of the other mathematical approaches to model a Σ-Δ modulator to analyze its dynamics, so as to put the approach of this thesis in some perspective.

We start with the paper by Keener [30] which analyzes solutions to the mapping $x_{n+1} = F(x_n)$, where F is a piecewise continuous, locally increasing mapping of the interval $[0, 1]$

to itself with one jump discontinuity. Using the definition of rotational chaos, it was shown that chaos generally exists when $F(0) < F(1)$ (overlapping mapping, not 1 to 1), and that chaos does not exist when $F(0) > F(1)$ (nonoverlapping, 1 to 1).

In the Feely and Chua paper [10], the one dimensional Σ-Δ case (single loop) with 1-bit quantizer and constant input is analyzed as a circle mapping. Properties of the rotation number/interval derived by Keener [30] are applied and rotational chaos (from Hao) [20] is shown to exist when the multiplier (from the transfer function) $a_1 > 1$. A study of the periodicity, stability and bifurcations in the dynamics (as the a_1 value changes) is also made.

In the thesis of Schreier [56], a proof that sensitivity to initial conditions exists if and only if the transfer function is nonminimum phase is given for the Σ-Δ modulator with 1-bit quantizer and general order transfer function and general input. This type of sensitivity, called predictability here, appears to be weaker than the definition used in condition 1 of Devaney's definition for chaos. In terms of the definition for condition 1, the proof only proves that sensitivity to initial conditions implies that the transfer function is minimum phase. Schreier uses this proof to justify the possible existence of chaos, but does not verify any other requirements of chaos (such as conditions 2 and 3). Limit cycle stability and Σ-Δ modulator stability (e.g. as a function of input) are also investigated from an engineering perspective along with other engineering issues. The paper [57] describes the treatment of the Σ-Δ modulator as a mathematical mapping in Schreier's thesis.

In the thesis of Wang [62], the 1-bit, first-order Σ-Δ modulator is analyzed in terms of a circle map that is similar to that of Keener [30], where the mapping of a trapping region is transformed to a mapping on the unit interval. Results for chaos (rotational definition), with constant input, are obtained that correspond to the cases in [30]. Wang replaces Keener's end point inequality conditions for $F(x)$ with conditions on the derivative

$DF(x)$. $1 < DF(x) < 2$ is associated with an expanding map of the circle and chaos, while $0 < DF(x) < 1$ with a contraction circle map and nonchaos. Stronger conditions on $F(x)$ seem to be used in the analysis. Evidence is provided that a small input signal is more resolvable from the noise spectrum background of the output, when the "chaotic" modulation of the expanding circle map case holds. For the rest of the thesis, a continuous embedding scheme is developed for the second and third-order Σ-Δ modulator to study other dynamical properties such as stability. In this scheme, solutions to the discrete system are obtained from their embeddings in solutions to a continuous system. The overall thesis stresses a dynamical systems approach.

In the thesis of Risbo [53], the 1-bit, general order Σ-Δ modulator with general input is analyzed. A formal dynamical systems formulation is defined and Devaney's three conditions for chaos are presented for the analysis. A noninvertible region for the mappings is defined and is needed to ensure that the quantizer output is bounded. Unlike the case of a multi-bit quantizer with an arbitrary number of levels, this boundedness does not automatically follow when the quantizer is constrained to 1 bit. Sensitivity to initial conditions is associated with state space stretching (divergence), and this is shown to hold when the system is nonminimum phase, by using an eigenvalue method with transition matrices. For such systems, a proof of the existence of a folding mechanism to map to a noninvertible region in state space is used to argue boundedness of orbits in state space (contracting effect). This second result is apparently used as a sufficient way of satisfying topological transitivity, although an explicit connection is not discussed. Periodic points and limit cycles are discussed, but chaos condition 3 is not addressed. An assumption of chaos in the Σ-Δ modulator is made when these two results hold, but this falls short of a rigorous proof that Devaney's three conditions hold. An analysis of the stability of "chaotic" Σ-Δ modulators is given, including a bifurcation between stability and instability. Risbo outlines

why he views the presence of what he terms chaos as interesting: it implies nonperiodic output, it can be used to suppress spurious tones in the output spectrum, and the question of stability is conceptually simpler. The rest of the thesis focuses on nonchaotic stability, modelling, design and optimization.

As seen in [62] and [53] above, the analysis of the 1-bit Σ-Δ modulator (where the quantizer Q has a signum function form) as a dynamical system requires consideration of a complex mapping structure involving a trapping region or noninvertible region, whereby the quantizer output remains bounded. Such an approach was initially considered for the study of chaos in this thesis, to apply Devaney's chaos definition to the 1-bit Σ-Δ modulator. This was found to be analytically difficult and overcomplicated when seeking to maintain the level of rigour that we demand for our chaos analysis. Therefore, partly owing to this circumstance, we focus our work in this thesis exclusively on the multi-bit Σ-Δ modulator, where the number of levels in the quantizer may be arbitrarily large and a bounded output is automatic. Such an approach is quite practical and is also commonly taken by others, e.g. Reiss [50]. One may extend such a system to the case of a finite bit modulator by simply assuming an appropriate no overload condition on the quantizer.

The paper [51] by Reiss and Sandler considers a multi-bit first-order Σ-Δ modulator with general input. Two ways of applying the filter gain are given, with the usual manner of applying this to the quantizer error shown to achieve greater quantizer accuracy. Chaos is associated with a gain in the range $1 < a_1 \leq 2$, and a bifurcation diagram demonstrating this is given. Use of a multi-bit quantizer is asserted to make the modulator stable over all inputs, and counter the potential for instability when operating in the "chaotic" regime. As with other papers, e.g. [58] (Schreier), [54] (Risbo), this chaos is argued to be beneficial for removing undesired idle tones in the output that may arise from periodic orbits or limit cycles. The thesis [50] of Reiss that presents this work is focused overall on an analysis of

chaotic time series. Time series from several applications were analyzed to construct and observe the dynamics, extract empirical quantities, and improve analysis techniques.

The work of Wang [62], Reiss [50] and others then is also inadequate in providing the rigorous analysis of chaos in the Σ-Δ modulator that we seek.

The following studies were conducted using simulations. In [58] by Schreier, it is shown that making the Σ-Δ modulator system nonminimum phase will destabilize limit cycles and consequently reduce the tonality of the quantizer noise, but that the signal to noise ratio will decrease (greater for higher-order systems). The paper [54] by Risbo applies the chaos ideas of his thesis to show that tone suppression in the output spectrum that could be achieved with the addition of dither, can also be achieved by making the Σ-Δ modulator "chaotic", i.e. nonminimum phase. The tone is weaker for the nonminimum-phase case than the dither case, although the signal to noise ratio is also less. For both cases, the SNR must decrease to maintain stability. The paper [36] by Lipshitz and Vanderkooy provides conclusions that contradict these results of Risbo. Here, the behaviour of a dithered minimum-phase and an undithered nonminimum-phase (i.e. "chaotic") Σ-Δ modulator is compared. It was found that the distortion performance (measured as the effective gain of the fundamental signal at output) of the dithered case was far better than that of the nonminimum-phase case. These results support the earlier work of Norsworthy [47] and are more complete.

The results of these studies differ because [36] and [47] kept the noise transfer function shaped the same and the input noise equal, for the cases being compared, while [54] did not. Lipshitz and Vanderkooy would argue that their comparisons are thus more meaningful. Researchers such as Schreier and Risbo are motivated to study chaos in the Σ-Δ modulator from a belief that tone or distortion reduction should be effected in practice by making the system nonminimum phase (i.e. "chaotic") if this can be shown to be successful. Such controversy serves to highlight the importance of the studies of Σ-Δ modulator chaos in

this thesis.

The paper [49] by Reefman et al. considers a 1-bit Σ-Δ modulator of general order feedforward form, with DC, (i.e. constant) input. A state space mathematical model for the description of limit cycles is presented, with the goal of providing tools other than simulations for obtaining insights into their behaviour. It is proved that periodicity in the bit output pattern implies a periodic orbit in state space variables. A recipe is given for finding limit cycles and relating them to state space initial conditions. The conditions for limit cycles and the minimum disturbance need to break them up are studied. Dithering is shown to be a suboptimal approach (i.e. compared with adding a small disturbance to the integrator state). Higher-order systems are shown to be less susceptible to limit cycles. Other issues, such as the relationships with DC zeros (i.e. at $z = 1$) of the transfer function (integrator), stability analysis and numerical results are covered. An odd aspect of this work is that the integrators used have zeros outside the unit circle, which tends to increase the noise gain.

The thesis [19] of Güntürk uses mathematical tools from analytic number theory, harmonic analysis and dynamical systems to provide a new framework and improved techniques for Σ-Δ modulator error analysis. The second part of the thesis is a function space approach to image compression.

The proceeding of Kalman [29] from 1956 provides an interesting perspective on discrete systems and the idea of a continuous imbedding mentioned by Wang [62]. Discrete deterministic systems that are both nonautonomous (i.e. sampled data systems) and nonlinear (i.e. quantized systems) such as the general Σ-Δ modulator, are considered, and the following assertions made. Having both properties creates more complexity and greater difficulty in analysis of the behaviour. Under certain circumstances, a nonlinear sampled data system can be paired with a continuous system so that the two behave similarly. When

these circumstances do not hold, the output can be treated as a random Markov process. In such a process, the probability of a state at a given point in time depends only on the state at the previous time iterate. The tools of analysis from probability theory may then be applied. In general, any continuous system can be paired to, and any Markov process can be synthesized, by means of a nonlinear sampled data system. From these results, we have that $\{\text{all nonlinear differential equations}\} \subset \{\text{all nonlinear difference equations}\}$ and $\{\text{all Markov processes}\} \subset \{\text{all nonlinear difference equations}\}$. These ideas will not be pursued in this thesis — the interested reader can check [29].

1.6 Dithered Quantizers and Σ-Δ Modulators

In this section, the approach to studying dithered quantizers and Σ-Δ modulators in this thesis, and the relevant background and motivation will be discussed.

The research background presented here forms some of the elements upon which the study of dithered Σ-Δ modulators will be pursued in this thesis. It is important to reinforce the validity and significance of basic known dither results from an abstract point of view that is consistent with the framework used to examine other dynamical issues such as chaos, but that can complement previous approaches for dither. The dynamical systems formulation that we will develop will provide such an abstract approach. It is also important to develop new approaches and arrive at new results for both specific and general statistical questions from such an abstract perspective. The general probability theory that we will develop and apply, and the analysis of a particular example, will serve to this end.

In a dithered process, the addition of a dither signal to the input signal prior to its entry into the quantizer is considered, as shown in the simplified system (with no feedback) in Figure 1.5. The total error $E = Y - U$ is formed by subtracting the predithered quantizer

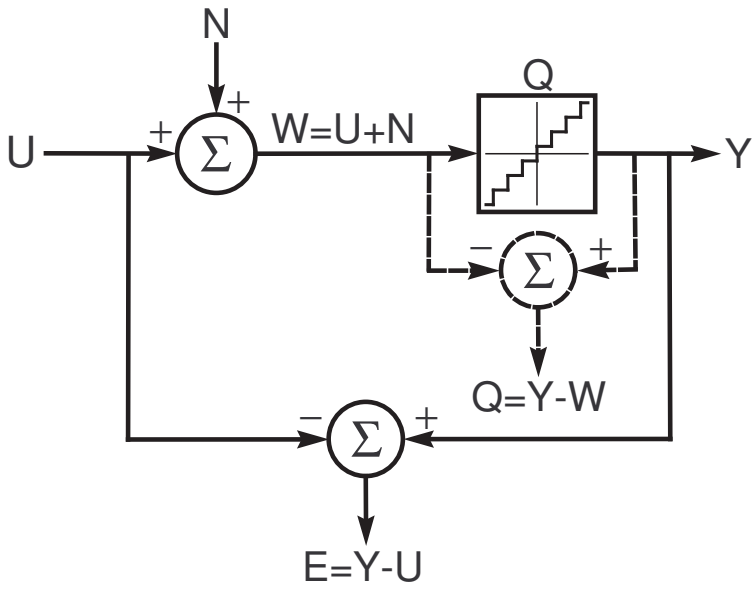


Figure 1.5: Dithered quantizer and total error

input from the quantizer output. For feedback systems (see Figure 1.2), it is this positioning of the added dither N, inside the total error loop, that gives it its unique mathematical properties for influencing the total error statistics of E, and distinguishes it from the input X, and the less consequential role one would have with a "dithered" input $X + N$, where $E = Y - (X + N)$. The new signal $Q = Y - W$ is identified as the internal quantizer error. This will be different from the overall (unshaped) system error E in a feedback system, when a nonzero dither N exists. The addition of dither is shown in the time domain in Figure 1.6.

The dither is typically defined by a prescribed stochastic probability distribution which we are free to choose, and is hence statistically independent of the input. In practice the dither ν_n is typically either independent and identically distributed (i.i.d.) for different values of n (white), or possesses some dependence on its previous values (coloured). In this

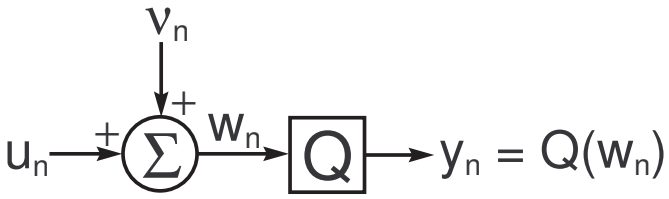


Figure 1.6: Dithered quantizer in time domain

thesis, we take the dither to be the former (white). This case is the most common and provides the simplest and sufficiently general approach. Strictly speaking, i.i.d. dither, or any other form of random dither, is normally approximated in practice by a pseudorandom number generator, so as to be sufficiently "near random", even though mathematically deterministic (see Section 1.7).

In general, we have $\nu_n \in \mathbb{R}$, for all $n \geq 0$. For a given dither value ν_n, the quantizer output is given by $Q(w_n) = Q(u_n + \nu_n)$, where the predithered quantizer input $u_n = x_n - r_n$ for the closed loop system. Then, from the statistical definition of expectation, we have that the expected value of the quantizer output y_n given the input u_n is given by

$$E[y_n|u_n] = \int_{-\infty}^{+\infty} Q(u_n + \nu_n) P_\nu(\nu_n) d\nu_n = \int_{-\infty}^{+\infty} Q(\eta) P_\nu(\eta - u_n) d\eta = Q * P_\nu^-(u_n),$$

where Q is the quantizer function, $P_\nu(\nu_n)$ is the probability density function (PDF) of the dither ν_n, $y_n|u_n = Q(u_n + \nu_n)$, and we define the function $P_\nu^-(\nu_n) \equiv P_\nu(-\nu_n)$, for all $\nu_n \in \mathbb{R}$. It is assumed here that $P_\nu(\nu_n)$ is piecewise continuous. If the dither distribution is discrete, then the following holds analogously:

$$E[y_n|u_n] = \sum_{\nu_n \in S_\nu} Q(u_n + \nu_n) P_\nu(\nu_n) = \sum_{\eta \in S_\eta} Q(\eta) P_\nu(\eta - u_n) = Q * P_\nu^-(u_n),$$

where S_ν is the set of all dither values of nonzero probability, $S_\eta = \{\eta \mid \eta = u_n + \nu_n, \forall\ \nu_n \in S_\nu\}$, and $P_\nu(\nu_n)$ is the probability mass function (PMF) of the dither ν_n.

From these expressions, we see that the expected output also takes the form of a convolution of the quantizer transfer characteristic with the distribution of the dither evaluated at the negative of its argument. The expected value of the error ε_n given the input u_n is given by $E[\varepsilon_n|u_n] = E[y_n|u_n] - u_n$. If the distribution of the dither is a rectangular probability density function (RPDF)[4] of width Δ , where Δ is the width of the steps in the quantizer, then the convolution equation above gives the result that $E[y_n|u_n] = u_n$, so that $E[\varepsilon_n|u_n] = 0$ for all u_n. Thus RPDF dither yields zero mean error for all inputs u_n within the domain of the quantizer, and any number of steps. It can be shown, however, that the mean squared error (error variance or noise power) will remain signal (input) dependent (i.e. there is no noise modulation), and is given by $E[\varepsilon_n^2|u_n] = \Delta^2/4 - \hat{u}_n^2$, where $\hat{u}_n = (u_n + \Delta/2)(\text{mod } \Delta) - \Delta/2$. If a triangular probability density function (TPDF) is used for the dither (the convolution of two RPDF densities, which corresponds to the addition of two RPDF random variables) with width 2Δ, then there will be zero mean error and a constant mean squared error given by $E[\varepsilon_n^2|u_n] = 3\Delta^2/12$. This implies, with the input and quantizer steps the same as for the RPDF case, that noise modulation no longer exists.

These established results come from a fully developed dither theory for dithered multi-bit quantizers taken either on their own, or incorporated in feedback loops of multi-bit noise shapers (Σ-Δ modulators), (see [35], [65] and the references therein). The paper "Dithered Noise Shapers and Recursive Digital Filters" [38] as well, gives a general analysis of the extension to dithered noise shapers. This theory includes the property that the 1st up to the pth moments of the error (i.e. $E[\varepsilon_n^i]$, $i = 1, \ldots, p$) can be made signal independent by using n convolutions of the RPDF for the density of the dither. This generalizes from the case of $n = 1$, 2 described above and corresponds to the use of dither obtained from

[4]The RPDF has the PDF $P_\nu(\nu_n) = \dfrac{1}{\Delta}$, $-\dfrac{\Delta}{2} \le \nu_n \le \dfrac{\Delta}{2}$.

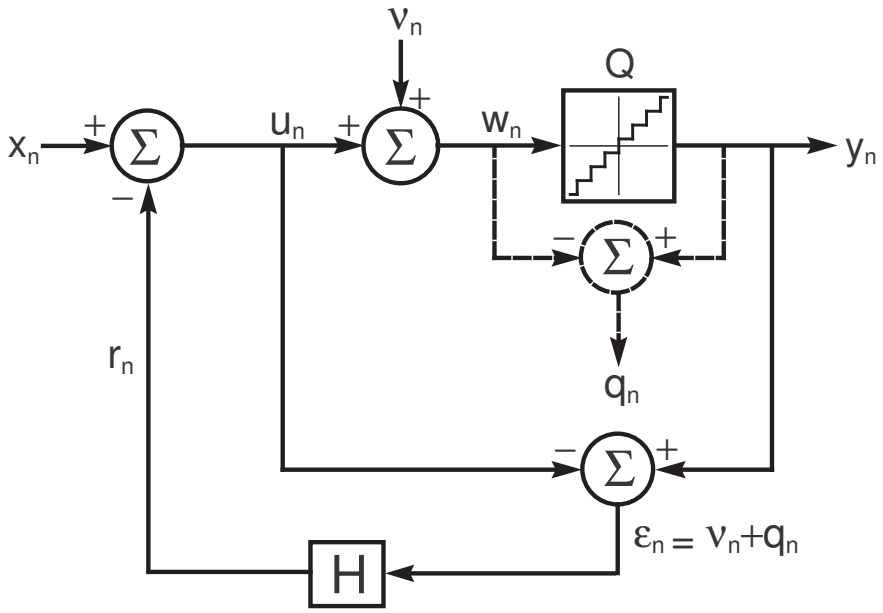


Figure 1.7: Topology of dithered noise shaper in time domain

the sum of p RPDF random variables. Normally, $p = 2$ is sufficient for digital audio practice. For digital images, it appears that $p = 2$ may be sufficient as well. For the one-bit dithered quantizer however, no theory will guarantee perfection in terms of making these error moments totally signal independent. This is because the quantizer will saturate or overflow, with input beyond its domain width of 2Δ.

The time domain topology for a Σ-Δ modulator noise shaper, with a dithered quantizer, is shown if Figure 1.7. For this, the internal quantizer error q_n is shown as well.

The following studies by Lipshitz et al. relate to the ideas introduced above for multi-bit quantizers and different types of dither. The paper [35] presents a theoretical survey of the overall case of dither. The paper [65] gives a mathematical investigation focusing on the nonsubtractive dither case, introducing new results. The paper [34] extends the study of dither to noise shapers and shows that the dithered quantizer theory is applicable. It is

shown that the presence of feedback is irrelevant to the theory, since the dithering linearizes the quantizer. Dithered noise shaper designs to achieve noise reduction are described. The paper [38] shows that when coloured dither (not i.i.d.) is used, more stringent conditions must be satisfied for error moment control, if a feedback circuit is used, than otherwise.

To focus more specifically on the background for the work of this thesis, we begin with the important early paper by Schuchman [59]. Here it is shown that the quantizer noise q_n will be independent of the input u_n, if the dither ν_n has a probability density function that is a convolution of RPDFs. Furthermore, q_n will have an RPDF (uniform probability density function); and it will be white (i.e. constant power spectrum) and i.i.d., if ν_n is i.i.d. In the case of subtractive dither, the dither ν_n is subtracted from the quantizer output y_n to give the final system output, which will then differ from the input u_n by just the quantizer error q_n. In this case, the whole PDF of the output error, along with all of its moments are independent of the input. This is a very desirable property to have since it means the lowest possible noise in the reconverted audio signals. Moreover, the error here is a purely signal-independent additive white noise process. The use of subtractive dither is not generally feasible, however, since the subtraction of dither after quantization cannot be accurately accomplished with the finite number of bits in physical systems. For nonsubtractive dither, we are left with the signal independent moment results for the output error, from dither theory given earlier.

The paper [52] of Reiss and Sandler looks at the dependence of the error moments on the input in dithered multi-bit Σ-Δ modulators. We note the paper for several errors, later admitted in correspondence to Lipshitz and Vanderkooy by the author. The paper incorrectly claims that the error ε_{n-1} and corresponding dither ν_{n-1} are independent. The paper Dither Myths and Facts [37] by Lipshitz and Vanderkooy addresses various common misunderstandings about the properties of dither that have persisted over the years, in-

cluding the claim above in [52]. A prominent misconception addressed is the related one which assumes that q_n and ν_n are independent. This cannot be true unless one assumes conditions exist for an application of Schuchman's result [59] above with the roles of u_n and ν_n reversed, that is with u_n assumed to be generated randomly with a PDF that is a convolution of RPDFs. There has been little practical role for such an assumption in a feedback system in previous work, although the theory presented in this thesis will allow for such an approach.

In [52] it is also incorrectly suggested that for a first-order Σ-Δ modulator with unity gain ($a_1 = 1$, $b_1 = 0$) and RPDF dither added, the long term time averaged error variance $E[\varepsilon_n^2] = \Delta^2/6$ with any constant input $x_n = c$. In fact, Lipshitz and Vanderkooy point out that numerical simulations suggest this result only for the case of irrational input, and that this definitely does not hold for the rational case. In this thesis, these results will be proven, and further analysis will be conducted on this Σ-Δ modulator form. The paper [22] of He et al. considers a higher-order, nondithered system of this form, with multiple transfer function zeros at DC (i.e. at $z = 1$), and with irrational constant input. The long term statistical behaviour of the errors is shown to be signal-independent, white, uniform noise. This result is consistent with the averaged error variance for the dithered first-order form above.

Finally, we mention the chapter on nonlinear maps in the Digital Signal Processing Handbook [26], which relates stochastic signals to chaotic ones. Here, eventually expanding maps are introduced, which are similar to those of Keener [30]. They are defined to be piecewise continuous, sufficiently bounded, and eventually expanding for every x, meaning essentially $|\frac{d}{dx}F^m(x)| > 1$ for some m. The map F is, in addition, a Markov map if it is affine on each continuous interval (i.e. of the form $f_i(x) = s_i x + d_i$ on each such interval i) and it maps interval partition points to partition points. Such maps are broadly appli-

cable to signal processing systems. The concept of an invariant density for random initial conditions is introduced, which will be considered as a "fixed point" functional in this thesis. Let f be an eventually expanding mapping that describes the evolution of a nonlinear system, such as the Σ-Δ modulator. When f has only one invariant density, then the time average $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} h(f^k(x_0))$, (where h is a sufficiently well behaved function) exists, and is generally independent of the initial condition x_0. From this property, transition probabilities can be defined. This result then yields the properties of a Markov map for f. Techniques then exist to analyze the statistics of Markov maps. Thus the statistics of any eventually expanding map can be modelled to arbitrary accuracy by those of some Markov map. This overall modelling approach may give more insight into behavioural aspects of chaotic signals or time series, and with less difficulties than with empirical averaging.

An important goal for future research would be to develop a counterpart mathematical theory for the one-bit quantizer and noise shaper to that of the multi-bit quantizer and noise shaper. It is expected that it would be impossible to derive such a theory that would give exact conditions for obtaining the exact results of making error moments signal-independent, as in the multi-bit case. The objectives would thus be to obtain results for conditions that put limits on the error quantities and give as good a prescription for error control and noise shaping as possible. The extent to which such partial dithering can prevent the occurrence of limit cycles will be an important aspect of this study. For this thesis, the Σ-Δ modulator with a multi-bit quantizer of arbitrary number of levels is studied, as explained in Section 1.5. These matters are therefore left to future studies, while statistical issues relevant to multi-bit case are explored here.

1.7 PRN Generators

In this section, the pseudorandom number (PRN) generator will be introduced as an important simple example of a Σ-Δ modulator system. The application is very different from that of signal processing in audio or image applications.

The most popular random number generators used today are the linear congruential type (generating "linear congruential sequences") introduced by D.H. Lemer in 1949. They have the following form:

$$\varepsilon_n = (a\varepsilon_{n-1} + c + \frac{\Delta}{2}) \pmod{\Delta} - \frac{\Delta}{2}, \quad -\frac{\Delta}{2} < \varepsilon_n \leq \frac{\Delta}{2}, \quad n = 1, 2, \ldots . \tag{1.4}$$

They may be written as

$$\varepsilon_n = (a\varepsilon_{n-1} + c) - Q(a\varepsilon_{n-1} + c)$$

using a "mid-tread" quantizer $Q(w_n) = \Delta\lfloor w_n/\Delta + 1/2 \rfloor,\ w_n \in \mathbb{R},\ n \geq 1$. This quantizer takes the functional form shown in Figure 1.4(a). The difference equation (1.4) is produced by a system with the topology shown in Figure 1.8. This system clearly is equivalent to a first-order noise shaper (i.e. a Σ-Δ modulator), where we have certain constraints on the parameters c, a, Δ and ε_0. System (1.4) may be considered in the following normalized form, which is simpler and more useful for analysis:

$$\frac{\varepsilon_{n+1}}{\Delta} = (a \cdot \frac{\varepsilon_n}{\Delta} + \frac{c}{\Delta} + \frac{1}{2}) \pmod{1} - \frac{1}{2}, \quad -\frac{1}{2} < \frac{\varepsilon_n}{\Delta} \leq \frac{1}{2}.$$

The sequences of numbers ε_n are not always "random" for all choices of c, a, Δ and the "seed" ε_0. Principles exist for choosing these parameters correctly and are given by Knuth in [31]. According to these we have the following: c may be arbitrary if a, Δ and ε_0 are suitably chosen, although generally we would want $c \neq 0$ (number processing is faster when $c = 0$ and may give sufficiently random sequences however). We wish to choose parameters to produce a period of maximum length (the maximum possible length being Δ), since this

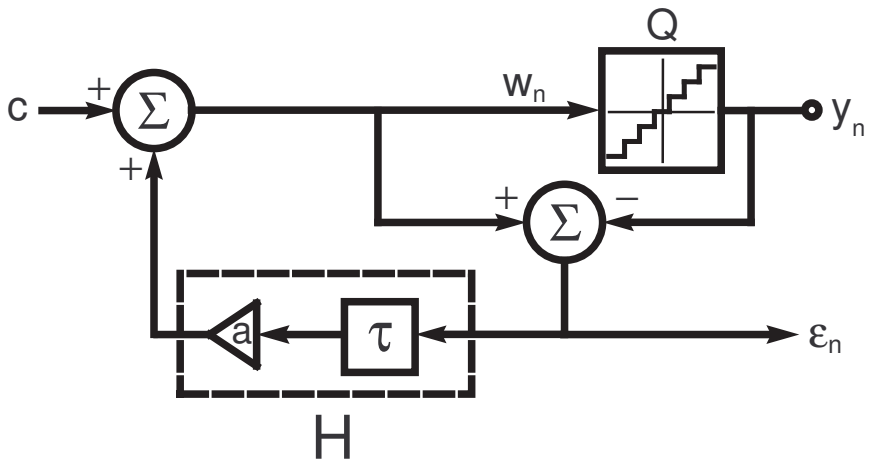


Figure 1.8: PRN generator of linear congruential type

is one necessary criterion for randomness. The following theorem [31] gives necessary and sufficient conditions for obtaining a period of maximum length.

Theorem A *The linear congruential sequence defined by* Δ, a, c *and* ε_0 *has period length* Δ *if and only if*

(i) c *is relatively prime to* Δ*;*

(ii) $b = a - 1$ *is a multiple of* p*, for every prime* p *dividing* Δ*;*

(iii) b *is a multiple of* 4*, if* Δ *is a multiple of* 4*.*

Another criterion for randomness is potency which is the dependency between consecutive numbers in a sequence (high potency would imply low dependence). This is defined as follows.

Potency *The potency of a linear congruential sequence with maximum period is defined to be the least integer s such that*

$$b^s \equiv 0 \mod \Delta.$$

From Theorem A, we would ideally choose Δ to be prime. We would also like to choose Δ to be of the form $\Delta = w \pm 1$, where $w = 2^e$, $(e \in \mathbb{Z}^+)$, on a binary computer with internal word-length e, since "$\Delta = w$" increases efficiency given this computer structure, and "$\Delta = w \pm 1$" creates sequences that are more random, given this structure. We would choose a to satisfy Theorem A and maximum potency.

A good choice for a set of parameters on a 32-bit computer is the Mersenne prime $\Delta = 2^{31} - 1 = 2,147,483,647$ (largest prime of the form $2^e \pm 1$, $e \leq 64$), $a = 62,089,911$. A bad choice would be $a = 1$, which does not give random sequences. The choice $\Delta = 2^{35}$, $a = 2^k + 1$, $k \geq 18$, leads to a potency of 2, which implies $\varepsilon_{n+1} - \varepsilon_n \equiv c + c(a-1)n$. This is not satisfactorily random and hence this parameter choice is not very good. Generally, small parameter numbers are also to be avoided.

A common mistake in designing PRN generators is the idea that by modifying a good generator slightly, one can get an "even more random sequence". Often one gets less random results, since the theory for a good generator breaks down, and the random choice of a generator likely doesn't give much randomness in the resulting sequences. Other types of PRN generators from the one considered above exist such as quadratic and "higher-order" recursive types.

1.8 Stochastic Resonance

In this section, stochastic resonance will be introduced as physical process whose models are continuous analogues to the dithered Σ-Δ modulator system.

Stochastic resonance refers to a phenomenon in which a nonlinear dynamical system subject to small input signals or "forcing", typically of a periodic nature, will show a greater response to this input in the presence of random perturbations or noise. This behaviour was first examined in [2] where the name "stochastic resonance" was coined. Research and applications of stochastic resonance (S.R.) subsequently occurred in a wide variety of areas. In [3], S.R. is considered in climate models and proposed as a possible explanation for the ice ages. Other areas where S.R. is used to explain system behaviour include electrical circuits (e.g. [9]), lasers, superconducting quantum interference devices and the neurological systems of animals such as crayfish. In many applications, S.R. reveals noise as beneficial. Although S.R. work began with Italian researchers, American researchers started work as well, such as with the paper [43].

Many S.R. papers consider first-order systems, however second-order systems will be considered here to more thoroughly convey the results. A typical second-order system is described by the following equation (stochastic D.E.):

$$m\ddot{x} = -b\dot{x} - k(x)x + f(t) + \text{“}noise\text{”}, \tag{1.5}$$

where the input f and "*noise*" are functions of time. Such systems have potential wells and associated stable equilibria, with the potential $V(x)$, where $F = -\dfrac{dV}{dx} = -k(x)x$ is the associated force. For example, in a "bistable" system with quartic potential, two stable equilibria exist, and the presence of noise may allow the small forcing signal to govern the transitions from one stable state to the other. These systems of (1.5) are generally similar to nonlinear spring systems (i.e. $k = k_0 + ax^2$, with $k_0 \geq 0$ or $k_0 < 0$, $a > 0$) but with some

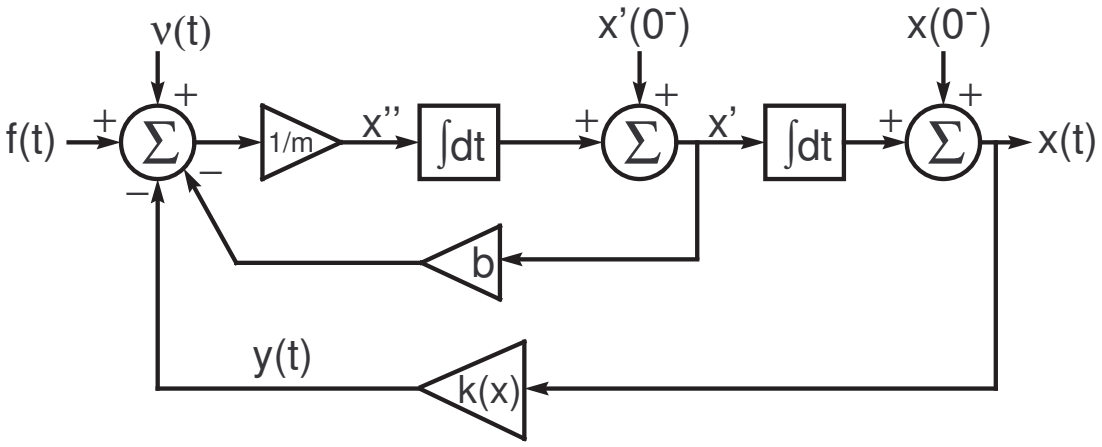


Figure 1.9: Analogue computer solves D.E.

differences.

The differential equation (1.5) may be solved by an analogue computer in which the topology is such that it conforms with the form of the D.E. as shown in Figure 1.9. Here, "noise" has been relabelled as a dither $\nu(t)$. The Laplace transform of this circuit gives the result shown in Figure 1.10, where $H(s) = 1/(ms^2 + bs)$ is the "integrator", and $X(s)$, $Y(s)$, $F(s)$ and $N(s)$ are the Laplace transforms of $x(t)$, $y(t)$, $f(t)$ and $\nu(t)$ respectively. The Laplace transform is defined as follows:

Laplace Transform *Let $x(t)$ be a complex valued function defined on $\mathbb{R}$. Then the Laplace transform $X(s)$ of $x(t)$ is defined to be*

$$X(s) = \int_0^\infty x(t)e^{-st}dt, \quad s \in \mathbb{C},$$

whenever this integral exists.

The circuit in Figure 1.10 now can be seen to have a topology equivalent to that of a Σ-Δ system. A transformation will thus give the noise shaper form. The nonlinear spring

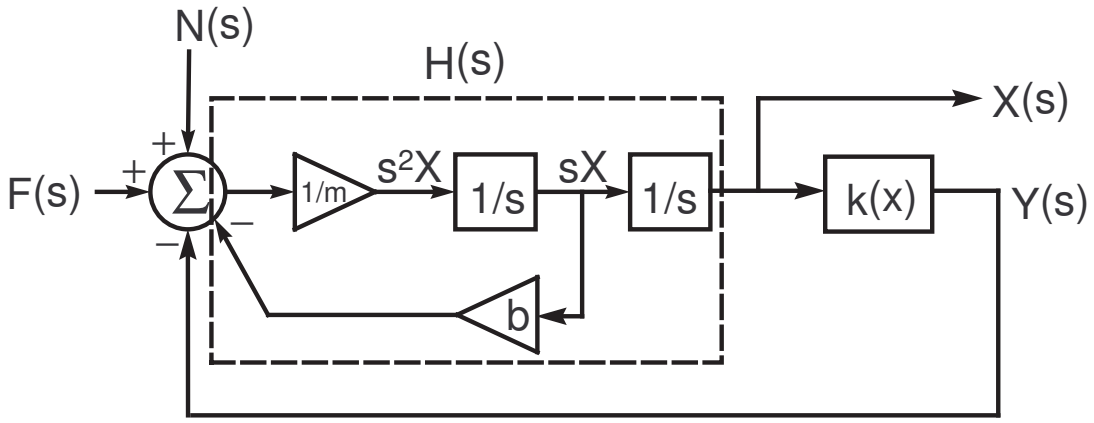


Figure 1.10: Laplace transform of analogue computer form of D.E.

constant $k(x)$ plays the role of the quantizer nonlinearity. This continuous system may also be approximated by the corresponding discrete system if the sampling rate for the "input" is fast enough. Such an approximation is given by the Σ-Δ system of the topology. Therefore the D.E. (1.5) may be rearranged to give a Σ-Δ system, implying a common method of analysis, although the nonlinearity $k(x)$ is different from the quantizer nonlinearity of Σ-Δ modulators. Considering the circuit of Figure 1.10, many aspects of S.R. behaviour may also be fully explained from dither theory. First-order systems may be analyzed in a similar manner (leaving out the "$m\ddot{x}$" term), to give a Σ-Δ system.

The paper [41] uses a second-order model, and presents the block diagram of the electronic circuit of an analogue computer to solve the D.E. of the problem by simulation. This block diagram is essentially equivalent to the form of the diagrams given above in that it accomplishes the same thing. In [13], the connection between the Σ-Δ nonlinearity as a feedback system with dither, and S.R. in threshold systems, is made, with both understood as noise activated processes. Here, it is mentioned that such S.R. systems are not real resonance phenomena as suggested by the name, but rather are a special case of the dither-

ing effect, consisting of a threshold crossing process aided by noise. The paper [64] shows that many S.R. systems are forms (via rearrangement) of dithered quantizer systems, and it shows how the results found from S.R. systems can be understood from dither theory. A hysteretic system is also considered (similar to that of [9]), to reinforce these ideas. A tutorial and update of S.R. research is given in [44]. A lengthy and in-depth review of S.R. research is given in the later paper [14]. An introductory overview of S.R. with a biological application, is given in [68]. The general research progress and importance of S.R. are also covered in the articles [15, 45] of Scientific American and [67] of Nature magazines.

As a result of the close connection between S.R. and Σ-Δ systems, the study of Σ-Δ modulators may allow for something to be said about S.R. (e.g. regarding dynamical behaviour). Studying Σ-Δ systems mathematically to obtain consequences that apply to and advance the research of S.R. systems is therefore a matter to be explored in future research. Indeed one might hope that such study would lead to a synthesis of understanding concerning S.R. in Physics and other areas, with the Σ-Δ systems of Engineering.

1.9 Thesis Overview

The body of this thesis will be organized as follows. In Chapter 2, the dynamical systems formulation of the Σ-Δ modulator, based on its mathematical description given in Section 1.3, is presented. In Chapter 3, the theorems and discussion regarding bounded internal stability are presented. In Chapter 4, the linear and continuous case model formulations are presented. In Chapter 5, the theorems and discussion concerning chaos in the nondithered case are presented. In Section 5.1, Devaney's definition of chaos is adapted to apply to the Σ-Δ modulator model of Chapter 2. Sections 5.2 deals with the treatment of the continuous case model and results. Section 5.3 does this for the general case model. Section

5.4 summarizes the chaos results and consequences. In Chapter 6, theorems and discussion concerning the dithered model and chaos are presented. Section 6.1 extends the formulation of Chapter 2 to include dither, and Section 6.2 studies chaos with dither . In Chapter 7, the theorems and discussion of the stochastically interpreted dynamics of Σ-Δ modulators are presented. Section 7.1 presents the background theory for long term error behaviour, and Section 7.2 applies this treatment to the Σ-Δ modulator model to formulate uniformity results. In Chapter 8, the theorems, discussion and analysis of the statistical error behaviour of dithered Σ-Δ modulators are presented. Section 8.1 deals with the fundamentals of dithered quantizers alone, while Section 8.2 extends this to the Σ-Δ modulator, applying work from Chapter 7 to give results. Section 8.3 applies this to an analysis and numerical simulations for a specific first-order case. The last sections of Chapters 7 and 8 expand the discussion of some pertinent issues. Finally, Chapter 9 provides main conclusions for the thesis work, and recommendations for further work that flow from this.

Chapter 2

Dynamical System Formulation

In this chapter, we formulate the Σ-Δ modulator system as a dynamical system. For the study of chaos in this thesis, the dynamical system formulation provides a mathematical framework through which a viable adaptation of a standard definition of chaos may be applied, and a thorough analysis to give general results on conditions for this chaos may be carried out. For the study of the statistical properties of the errors of dithered systems in this thesis, this formulation further provides a consistent and unifying approach with which to both seek new insights and results, and verify existing results from a new perspective.

State space descriptions of discrete-time processes are well established [49], [48], and their use in modelling the Σ-Δ modulator was seen in the literature reviewed in Section 1.5. For analytical simplicity, the state space used for such models typically involves different quantities from those of the observed output of the system. There is generally a close relationship between the behaviour of the state space and output variables, as shown for example with the phenomena of limit cycles in [49]. The formulation of the state space for the dynamical systems model in this thesis is given as follows.

From the frequency domain definition of the Σ-Δ modulator given in the first part of

Section 1.3, we have the transfer equation $Y = X + (1 - H)E$. In principle, X and H are always known. Therefore, for practical purposes, we have a direct relationship between E and Y. Specifically, if the nature and behaviour of E is known, we can automatically determine the nature and behaviour of Y from the transfer equation. This relationship carries through analogously to the time domain. Therefore, to describe the state and dynamics of the system (1.2) where the observable quantized output $y_n = Q(x_n - r_n)$ is the ultimate quantity of interest, it is sufficient to characterize system (1.2) in terms of the state and dynamics of the errors ε_n. This characterization is most useful, since ε_n is the quantity of interest for performance, and is also easily defined from the difference equations as the state space quantity. We now proceed to present the dynamical system formulation.

The formulation presented here follows from the developments in Section 1.3, where we take the dither signal N to be zero, and hence $\nu_n = 0$ for all $n \geq 0$. Extensions of the formulation to allow for dither will be treated in Section 6.1.

Dynamical System Model:

To begin, we let the difference equations of system (1.2), along with the given initial conditions, define a discrete dynamical system given by the form

$$\begin{aligned} \vec{x}_{n+1} &= \mathcal{F}(\vec{x}_n, x_n) \equiv f_n(\vec{x}_n) \\ y_n &= \mathcal{Q}(\vec{x}_n, x_n) \equiv Q_n(\vec{x}_n), \end{aligned} \tag{2.1}$$

for $n \geq 0$, where

$$\vec{x}_n = (r_{n-1}, \ldots, r_{n-N};\ \varepsilon_{n-1}, \ldots, \varepsilon_{n-M}) \in \mathbb{R}^N \times \mathcal{C}^M,$$

$x_n \in \mathbb{R}, \quad y_n \in (\mathbb{Z} \cdot \Delta + \Delta/2)$,

$$\begin{aligned} f_n : (\mathbb{R}^N \times \mathcal{C}^M) &\rightarrow (\mathbb{R}^N \times \mathcal{C}^M) \\ Q_n : (\mathbb{R}^N \times \mathcal{C}^M) &\rightarrow (\mathbb{Z} \cdot \Delta + \Delta/2), \qquad n \geq 0, \end{aligned}$$

and the initial condition is $\vec{x}_0$. We use the overbar to indicate a vector quantity. The state space of the system is $\mathbb{R}^N \times \mathcal{C}^M$, where $\mathcal{C}^M \equiv \underbrace{\mathcal{C} \times \mathcal{C} \times \ldots \times \mathcal{C}}_{M}$, and $\mathcal{C}$ is the "circle" defined by the interval $(-\Delta/2, \Delta/2]$, with $\lim_{x \to -\Delta/2} x \equiv +\Delta/2$ for $x \in \mathcal{C}$ holding. Note that with the circle $\mathcal{C}$ so defined, addition and scalar multiplication on $\mathcal{C}$ operate as follows: if $\alpha, \beta \in \mathcal{C}$, then $\alpha + \beta \equiv \alpha + \beta - Q(\alpha + \beta + \Delta/2) - \Delta/2 \in \mathcal{C}$, and $c\alpha \equiv c\alpha - Q(c\alpha + \Delta/2) - \Delta/2 \in \mathcal{C}$, where $c \in \mathbb{R}$. The functions f_n above define mappings from the state space to itself and while continuous, from the quantized nature of the Σ-Δ system, may lead to discontinuities (i.e. only piecewise continuity) on $\mathbb{R}^N \times \mathcal{C}^M$ over successive or composite mappings. The functions Q_n are the quantizing functions that define mappings from the state space to the observable output space $\mathbb{Z} \cdot \Delta + \Delta/2$. The functions f_n and Q_n are respectively formed by removing the input x_n as an independent variable in the associated functions $\mathcal{F}$ and $\mathcal{Q}$ of (2.1), and incorporating it into the functional form at each n. Specifically, for the definitions of f_n and Q_n we have

$$\begin{aligned}
f_{r(1),n}(\vec{x}_n) &= r_n = \sum_{i=1}^{M} a_i \varepsilon_{n-i} - \sum_{j=1}^{N} b_j r_{n-j} \\
f_{r(k),n}(\vec{x}_n) &= r_{n-k+1} \\
f_{\varepsilon(1),n}(\vec{x}_n) &= \varepsilon_n = Q(x_n - r_n) - (x_n - r_n) \\
f_{\varepsilon(p),n}(\vec{x}_n) &= \varepsilon_{n-p+1} \\
Q_n(\vec{x}_n) &= Q(x_n - r_n),
\end{aligned}$$

for $2 \leq k \leq N$, $2 \leq p \leq M$, and $n \geq 0$. The component functions of f_n are denoted above and as follows:

$$f_n(\vec{x}_n) \equiv (f_{r(1),n}(\vec{x}_n), \ldots, f_{r(N),n}(\vec{x}_n);\ f_{\varepsilon(1),n}(\vec{x}_n), \ldots, f_{\varepsilon(M),n}(\vec{x}_n)), \quad n \geq 0.$$

The component $f_{\varepsilon(p),n}(\vec{x}_n)$ contains the nonlinear quantizer element Q, which also leads to the discontinuities mentioned above.

We make the following clarification regarding the notational conventions used in this thesis. The symbols x_n and y_n, with no upper accents, denote the overall system input and output, respectively, as scalars, unless otherwise specified. When x_n or y_n have upper accents, they denote state space vectors in $\mathbb{R}^N \times \mathcal{C}^M$ ($\mathbb{R}^N \times \mathbb{R}^M$ if the system is dithered (see Chapter 6)), or some subspace thereof, unless otherwise noted. The vector symbol on top is used for the generic form in $\mathbb{R}^N \times \mathcal{C}^M$ (or $\mathbb{R}^N \times \mathbb{R}^M$).

For subsequent work, we provide the following notation for the set and composition of successive mappings of f_n over n:

Definition 2.1 (f) *Let f denote the mappings $\{f_n,\ n = 0, 1, 2, \ldots\}$ from (2.1). Let f^n denote $f_{n-1} \circ f_{n-2} \circ \ldots \circ f_1 \circ f_0$, $n \geq 1$. Also let f^0 denote the identity mapping so that $f^0(\vec{x}_0) = \vec{x}_0$ for all $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$.*

Given the importance of the error state space as already noted, we define the following projection mappings from the overall state space of $\vec{x}$ to the state space of errors, which will appear in subsequent work:

Definition 2.2 (g) *Let g be the projection of $\mathbb{R}^N \times \mathcal{C}^M$ onto $\mathcal{C}^M$, with $g : (\mathbb{R}^N \times \mathcal{C}^M) \rightarrow \mathcal{C}^M$ such that $g : (\vec{x}_n) \rightarrow (\varepsilon_{n-1}, \ldots, \varepsilon_{n-M})$, $n \geq 0$. Let g_k be the projection of $\mathbb{R}^N \times \mathcal{C}^M$ onto $\mathcal{C}$ with $g_k : (\mathbb{R}^N \times \mathcal{C}^M) \rightarrow \mathcal{C}$ such that $g_k : (\vec{x}_n) \rightarrow \varepsilon_{n-k}$, $k = 1, 2, \ldots, M$, $n \geq 0$.*

Metric:

To properly define the nature of the topology in the dynamical system for the Σ-Δ modulator, and to proceed with the later analysis for chaos in particular, the form of the metric on the state space $\mathbb{R}^N \times \mathcal{C}^M$ must be defined. We define the "distance" or metric between two points z_1, $z_2 \in \mathcal{C}$ by

$$\|z_1 - z_2\| = \min(|z_1 - z_2|, \Delta - |z_1 - z_2|).$$

This definition is geometrically sensible in that it simply represents the shortest distance between the two points z_1, z_2 on the circle. Now to generalize to the whole state space, the distance or metric between two points $\vec{x}_a = (z_{a1}, \ldots, z_{aN};\ \hat{z}_{a1}, \ldots, \hat{z}_{aM})$ and $\vec{x}_b = (z_{b1}, \ldots, z_{bN};\ \hat{z}_{b1}, \ldots, \hat{z}_{bM})$, where $\vec{x}_a,\ \vec{x}_b \in \mathbb{R}^N \times \mathcal{C}^M$, $z_{ai},\ z_{bi} \in \mathbb{R}$, and $\hat{z}_{ai},\ \hat{z}_{bi} \in \mathcal{C}$, is defined by

$$\|\vec{x}_a - \vec{x}_b\| = \sqrt{\sum_{i=1}^{N} \|z_{ai} - z_{bi}\|^2 + \sum_{i=1}^{M} \|\hat{z}_{ai} - \hat{z}_{bi}\|^2}.$$

The metric $\|z_{ai} - z_{bi}\|$, $i = 1, \ldots, N$, between two points on $\mathbb{R}$ is simply defined to be the usual Cartesian metric $|z_{ai} - z_{bi}|$. This metric has been extended to $\mathbb{R}^N$ above in the usual manner, and the metric on $\mathcal{C}$ has been extended to $\mathcal{C}^M$ above in the analogous manner. The overall metric in the above definition then just combines those of $\mathbb{R}^N$ and $\mathcal{C}^M$ to $\mathbb{R}^N \times \mathcal{C}^M$, in the same manner. Similarly, the distance or metric between two points x_a, $x_b \in \mathbb{R}$ or y_a, $y_b \in \mathbb{R}$ in observable input/output space is defined by $|x_a - x_b|$ or $|y_a - y_b|$ respectively, the usual Cartesian metric on $\mathbb{R}$. In all cases, the "magnitude" of a quantity, say s, is the metric between the quantity and zero on its particular state space, denoted $\|s\|$ or $|s|$. Thus we have a simply adapted metric to apply to the state space for the dynamical system of (2.1).

Circle Map Properties:

To start, we restate the circle map definition mentioned in Section 1.4:

Definition 2.3 (Circle Map) *A general normalized circle map is defined by* $C(\theta) = F(\theta) \bmod 1$, *where* $F : (\mathbb{R}) \to \mathbb{R}$, $C : (\mathbb{R}) \to [0, 1)$, *and* $C(\theta + 1) = C(\theta)$ *holds for all* $\theta \in \mathbb{R}$.

We have each error state ε_{n-i}, $i = 1, \ldots, M$, defined on the circle $\mathcal{C}$. This is consistent with the constraints and relative continuity in these errors as defined in system (1.2), and

demonstrates that in the structure of the dynamical system (2.1), we are using the concept of the circle map for the mappings of the state space to itself (the final property in the definition may or may not apply — see Section 4.1). Specifically, for a first order system, we would have $\varepsilon_n = C(\varepsilon_{n-1}, x_n)$ and $F(\varepsilon_{n-1}, x_n) = x_n - r_n = u_n$, from (1.2), without taking into account the normalization and translation simplifications in the definition above. The circle map approach we use is a generalization, to higher "circle" dimensions, of this and the one dimensional circle map approach of others, e.g. Wang [62], used to analyze first-order Σ-Δ modulator systems. This is related to results of Keener [30], and the application of the rotational chaos definition. We generalize here to deal with higher-order systems. Specifically, $\mathcal{C}^M$ represents the Cartesian product of M circles, or geometrically, an M dimensional "torus" in the state space. The part of the mappings that map onto the torus thus constitute generalized circle maps.

It is the cyclic symmetry of the quantizer element Q in the Σ-Δ modulator that enables the Σ-Δ system behaviour to be described using the circle map. The properties and symmetries of the circle map, in turn, provide for easier and more simplified analysis of the structure and dynamics of the system under description, as will be more apparent in the analysis for chaos in Chapter 5. The important consequence of this, and a strong justification for using the $\mathcal{C}^M$ error state space representation to capture the circle map property, is the continuity relationship. In particular, when ε_n is defined on $\mathcal{C}$, it is continuous as a function of the preceding coordinates (M previous ε iterates and N previous r iterates, each over $\mathbb{R}$) via r_n and x_n from (1.2). That the state space variables for this dynamical system correspond to the error ε_n, filter output r_n, and their delays up to Mth and Nth order respectively, follows directly from the structure of the difference equations in system (1.2) that this dynamical system formulation is based on. Since we are concerned with error or noise shaping as the function of the Σ-Δ modulator, the error ε_n is the natural

variable of the system to choose to study the system's dynamics. The equations (1.2) and the dynamical system 2.1 it yields above are hence formulated to incorporate this.

In light of the importance of dealing with the $\mathcal{C}^M$ and $\mathcal{C}$ domains, it shall be useful in later work of this thesis, to consider the "identity" circle map projection, where F is essentially the identity function, defined as follows:

Definition 2.4 ($\hat{P}_{\mathcal{C}}$) *Let $\hat{P}_{\mathcal{C}}$ be the projection of $\mathbb{R}^M$ onto $\mathcal{C}^M$, with $\hat{P}_{\mathcal{C}} : \mathbb{R}^M \rightarrow \mathcal{C}^M$, such that $\hat{P}_{\mathcal{C}}(\vec{x}) = (\vec{x} + (\Delta/2)^M)(\mathrm{mod}\ \Delta) - (\Delta/2)^M$, $\vec{x} \in \mathbb{R}^M$, where $(\Delta/2)^M$ is the M-d vector with $\Delta/2$ in all entries, and the* mod Δ *operation is applied independently to each vector entry in its argument. Let $\hat{P}_{\mathcal{C}^1}$ be the projection of $\mathbb{R}$ onto $\mathcal{C}$, with $\hat{P}_{\mathcal{C}^1} : \mathbb{R} \rightarrow \mathcal{C}$, such that $\hat{P}_{\mathcal{C}^1}(x) = (x + \Delta/2)(\mathrm{mod}\ \Delta) - \Delta/2$, $x \in \mathbb{R}$.*

If $\vec{x}$ is defined on $\mathbb{R}^q$, for $1 \leq q \leq M$, we shall refer to the projection of $\vec{x}$ on $\mathcal{C}^q$, as a generalization of the projection $\hat{P}_{\mathcal{C}}$ above, as "the value of $\vec{x}$ on $\mathcal{C}^q$", "$\vec{x}$ modulo $\mathcal{C}^q$", or simply "$\vec{x}$ on $\mathcal{C}^q$". Similarly, if X is a random variable defined on $\mathbb{R}^q$, whose projection on $\mathcal{C}^q$ is described by a uniform distribution over $\mathcal{C}^q$, we shall describe X or its PDF as being "uniformly distributed over $\mathcal{C}^q$" or "uniform over $\mathcal{C}^q$". ■

The system input x_n in (1.2) and (2.1) may have various characterizations in our treatment. We treat periodicity as follows:

Definition 2.5 (Periodic Input) *The input x_n is periodic if $x_{k_1 p+i} = x_{k_2 p+i}$, for all k_1, k_2, $i \in \mathbb{Z}^+ \cup \{0\}$, for some period $p \in \mathbb{Z}^+$.*

Hence x_n cycles with a period p. Our general characterization of a periodic point $\vec{x}_0$ will be looser, requiring only recurrence (see Definition 5.4 in Section 5.1 for a specific definition in the error state space). We say that an input x_0 is recurrent, or a point $\vec{x}_0$ is periodic (i.e. recurrent), if there exists a period $p \in \mathbb{Z}^+$ satisfying $x_{(k-1)p} = x_0$, or $\vec{x}_{(k-1)p} = \vec{x}_0$,

respectively, for all $k \in \mathbb{Z}^+$, where p need not be unique. A periodic point $\vec{x}_0$ lies on a limit cycle if $\vec{x}_0$ satisfies the stricter conditions of Definition 2.5 above, and hence has a cyclic orbit. This will hold for any periodic point, when the input is periodic. In addition, we may note that the composite mapping $f_{n+k+p-1} \circ \ldots \circ f_{n+k+1} \circ f_{n+k}$, with $k \in \mathbb{Z}^+ \cup \{0\}$, will be constant over all $n \geq 0$, and hence autonomous or input "independent" over n, if the input is periodic with period p.

Linear Difference Equation:

The dynamical system (2.1) constitutes a complex, nonlinear expansion upon a more basic linear dynamical system formed by its difference equations from (1.2), with the quantizer component neglected (e.g. set $Q(x) = 0$). The simplest form of this linear system may be given by a Pth order nonautonomous discrete difference equation with constant coefficients and P dimensional state space. A simplified difference equation representing a linear system of this type would be

$$z_n = \sum_{k=1}^{P} d_k z_{n-k} + c_n, \quad n \geq P, \tag{2.2}$$

with initial conditions $z_0,\ z_1, \ldots, z_{P-1} \in \mathbb{R}$, and with $d_k,\ c_n \in \mathbb{R}$, for $k = 1, \ldots, P$, and $n \geq P$. This simplified linear system is applicable to model local or subsystem behaviour of the system (2.1), and its form will arise recurrently in the analysis of stability, chaos, and the general dynamics of system (2.1) throughout this thesis. Therefore we outline here the nature of the solutions to system (2.2).

The general solution to the homogeneous difference equation corresponding to (2.2) is given by the following [42]:

$$z_{g,n} = \sum_{i=1}^{s}[\sum_{j=1}^{t_i} A_{ij} n^{j-1}]\mu_i^n + \sum_{i=1}^{\hat{s}}[\sum_{j=1}^{\hat{t}_i} n^{j-1}]|\mu_i^{\pm}|^n \cdot [B_{ij}\sin(\theta_i n) + C_{ij}\cos(\theta_i n)], \tag{2.3}$$

$n \geq P$, where

$$d(z) = z^P - \sum_{k=1}^{P} d_k z^{P-k}$$

is the characteristic polynomial of the difference equation, and has s real zeros μ_i with multiplicities t_i respectively, and $\hat{s}$ pairs of complex conjugate zeros $\mu_i^{\pm} = k_{1,i} \pm k_{2,i}\hat{i}$, $(k_{2,i} \neq 0)$ each with multiplicity $\hat{t}_i$ respectively. We let $\hat{i}$ denote $\sqrt{-1}$. The magnitude of these zeros is defined by $|\mu_i^{\pm}| = \sqrt{k_{1,i}^2 + k_{2,i}^2}$. The complex phase argument is defined by $\theta_i = \arctan(\frac{k_{2,i}}{k_{1,i}})$, $\theta_i \in (-\pi, \pi]$. If we express the formula above as $z_n = L_n(A_{ij}, B_{ij}, C_{ij})$, (where "$L_k$" denotes the appropriate linear combination of the A_{ij}, B_{ij}, C_{ij}), then the "P" constants A_{ij}, B_{ij}, C_{ij} are solutions of the linear system

$$z_k = L_k(A_{ij}, B_{ij}, C_{ij}), \quad k = 0,\ 1, \ldots,\ P-1,$$

which is a nonsingular system (i.e. solvable).

A particular solution is given by

$$z_{p,n} = c_n + \sum_{k=P}^{n-1} c_k L_{n-k-1}(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij}), \quad n > P, \quad z_{p,P} = c_P. \tag{2.4}$$

The "P" $\tilde{A}_{ij}$, $\tilde{B}_{ij}$, $\tilde{C}_{ij}$ are given as solutions to the following linear system:

$$L_k(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij}) = \sum_{i=1}^{k} d_i L_{k-i}(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij}) + d_{k+1},$$

with $k = 0,\ 1, \ldots,\ P-1$. This system is a nonsingular (solvable) system.

The full solution for z_n is the sum of a general solution $z_{g,n}$ and a particular solution $z_{p,n}$, given from (2.3) and (2.4) respectively, so that $z_n = z_{g,n} + z_{p,n}$.

Applying the Formulation:

The dynamical system formulation of the model in this chapter seeks to be a pragmatic, working application of the rigorous dynamical system definition (which we omit from exposition) — an approach we take in the formulations developed for chaos, dithered/random

error behaviour, stability and other such issues in the rest of this thesis, as well. In particular, the difference equation systems here are really being "driven" or "controlled" by an external input, making the systems generally nonautonomous. Thus the state space mappings are themselves functions of time n (generally different for different n), and hence orbits with different initial conditions may cross (at differing n), while remaining distinct afterward (unlike a strictly defined dynamical system).

In addition, we will allow our dynamical system model to be stochastic, by allowing any combination of a random input x_n, dither ν_n, or random initial condition $\vec{x}_0$ in some specified way. The formulation and analysis of the system, its dynamics and statistical properties, will be conducted in the same manner as for the deterministic dynamical system, or some natural adaptation or generalization thereof. In Chapters 2 to 6, $\vec{x}_0$ is assumed to be fixed (nonrandom), and the emphasis is on deterministic systems. For Chapters 7 and 8, the formulation in Chapter 2 will be extended in Section 7.1 to incorporate the concept of a long term steady state of $\vec{x}_n$, a stochastic $\vec{x}_0$, and functional mappings. Some general knowledge of stochastic processes in this thesis is drawn from [18].

It is important to clarify the mathematical modelling of Σ-Δ modulator systems with a stochastic signal component (i.e. an input and/or dither component) as treated in this thesis. We take a head on approach in our analysis, mathematically treating such signal components in pure stochastic form. Formal results for chaos and stability that follow from this may not be robust in extending to cases with "near random" (i.e. theoretically deterministic) signal approximations, as formed from PRN generators. Therefore, in Chapters 3 to 6 inclusive, the stochastic nature of such systems must be interpreted to exist in pure, theoretically random form. Systems that, in practice, approximate these aspects, must be interpreted as strictly deterministic (not random) when applying the chaos/stability results to be established. This is true, in general, for a stochastic steady state characterization as

introduced in Chapter 7 as well. The dither motivated error statistics analysis in Chapter 8 would be expected, however, to possess this robustness. Therefore, the error state space behaviour results in Chapters 7 and 8 for dithered or generally stochastic systems would be expected to sufficiently closely (for practical purposes) approximate what we would expect in practical PRN generator approximations of these systems.

With the development in this chapter, we may now proceed with the analysis of the subsequent chapters.

Chapter 3

Stability

In this chapter we present and discuss results concerning the stability of the Σ-Δ modulator. Conditions under which stability and chaos may coexist will be discussed in Section 5.4.

Internal and external stability are the two essential types of stability that exist for such a dynamical system. These relate to the state space coordinates $\vec{x}_n$, and system output y_n, respectively. The most basic concept of stability of concern (pertaining to both types) is standard bounded-input/bounded-output (BIBO) stability. We will simply call this bounded stability, which we define as follows:

Definition 3.1 (Bounded Stability) *The general Σ-Δ modulator is defined to be bounded internally stable if and only if the magnitude of the state space coordinate $\vec{x}_n$, given by $\|\vec{x}_n\|$, is bounded for all $n \geq 0$, whenever the magnitudes of the input x_n, and dither ν_n, given by $|x_n|$ and $|\nu_n|$ respectively, are bounded for all $n \geq 0$; that is there exists a constant $K > 0$ such that $\|\vec{x}_n\| < K$ for all $n \geq 0$, whenever there exists constants $\tilde{K}_x$, $\tilde{K}_\nu > 0$, such that $|x_n| < \tilde{K}_x$, $|\nu_n| < \tilde{K}_\nu$, for all $n \geq 0$. Bounded external stability has the same definition applied to the magnitude of the system output y_n given by $|y_n|$, instead of $\vec{x}_n$.*

From the definition of the metric, when the input x_n and dither ν_n are bounded for all

$n \geq 0$, these internal/external stabilities hold if and only if ε_n and r_n, or y_n are bounded, for all $n \geq 0$, respectively. Stricter forms of stability exist for each type, such as asymptotic stability about a fixed point or limit cycle, but these will not be considered here.

Bounded internal stability is an obvious physical and practical requirement of any Σ-Δ system. We cannot have the filter output r_n becoming unbounded, since this represents a physical overload of the feedback filter and loop, which will imply a physical breakdown of the system. Also, the development over time of unbounded errors ε_n means the accuracy of the amplitude quantization process is deteriorating to nil. Bounded external stability is also a natural requirement. If such stability could not be guaranteed, for example, then we would risk having the quantizer output y_n becoming arbitrarily large in magnitude over time; a situation which obviously renders the system useless as an analogue-to-digital converter under the standard operating assumption of a bounded input (at least when the available number of bits affords a magnitude of y_n well in excess of the upper bound on the input x_n). For practical systems with quantizers having a finite number of steps and hence the potential for overload, bounded external stability must always hold, at the price of having potentially large errors. With the general multi-step quantizer model having bounded dither ν_n, it is easy to see that bounded internal and external stability are equivalent. Identifying stability or nonstability in the standard fashion, with respect to a bounded input x_n, is reasonable since this is the expected condition of the system. An unbounded input clearly imposes a physical overload on the system essentially equivalent to that associated with a lack of bounded internal stability, for example, and thus could be regarded as a form of "induced" nonstability.

With the quantizer Q taken to have an arbitrarily large number of steps in our model (1.2) (i.e. general multi-step), there can be no overload of the quantizer. The errors ε_n will be bounded if and only the dither ν_n is bounded, and will always lie on $\mathcal{C}$ when no dither is

present. Hence bounded internal and external stability hold if and only if r_n and ν_n are bounded, for all $n \geq 0$. Thus we focus in this chapter on bounded internal stability via the boundedness of r_n. With such stability guaranteed, we are free to design the modulator to simply refine its capabilities further. The requirement we have is then that $|r_n| < K$, for all $n \geq 0$, for some $K > 0$. In the work of this chapter, we assume $\nu_n = 0$ for all $n \geq 0$, so that no dither is present. The theorems presented would directly extend in applicability to the case when an arbitrary bounded dither exists, however (see the discussion at the end of this chapter). In this case, the bound on the magnitude of the error ε_n would change from $\Delta/2$ to $\Delta/2 + K_D$, where the magnitude of the dither ν_n is bounded by $K_D > 0$. The term stability as used will be meant to imply bounded internal stability.

To proceed with the analysis of stability, we focus now on the nature of the filter output r_n. We see that difference equation (1.2) may be reorganized to give the following difference equation strictly in terms of r_n, where the errors ε_n have been incorporated into a relative input $\tilde{c}_n$:

$$\begin{aligned} r_n &= -\sum_{j=1}^{N} b_j r_{n-j} + \tilde{c}_n \\ \tilde{c}_n &= \sum_{i=1}^{M} a_i \varepsilon_{n-i}, \qquad n \geq 0, \end{aligned} \tag{3.1}$$

with initial conditions $r_{-1}, r_{-2}, \ldots, r_{-N}$ that we denote $\underline{r}_0 \in \mathbb{R}^N$. The solution is the sum of a general solution r_{gn} and a particular solution r_{pn}, so that $r_n = r_{gn} + r_{pn}$. The general and particular solutions may be obtained from applying (2.2) with (2.3) and (2.4); with $P = N$, $d_k = -b_k$, $k = 1, \ldots, N$, and with the subscripts of r increased by a factor of N. In the proofs of this chapter, we will implicitly assume that this subscript shift has been applied when using (2.2), (2.3) and (2.4), unless otherwise noted. We also have $d(z) = p_r(z)$, where

$$p_r(z) = z^N + \sum_{j=1}^{N} b_j z^{N-j}$$

is the characteristic polynomial of the difference equation for r_n in (3.1). If we multiply the noise transfer function (1.3) by $z^{\max(N,M)}$, we find that the denominator is given by the polynomial $p_r(z)z^{\max(N,M)-N}$. Thus the poles of the NTF (1.3) are given by the zeros of $p_r(z)$, together with $\max(N,M)-N$ poles at zero if $M > N$.

The equations in (3.1) have the following topological interpretation in the Σ-Δ modulator. Let H_r be a filter corresponding to the feedback elements of H only, that is the left side of the filter H in Figure 1.3. Then H_r has input $\tilde{c}_n$ and output r_n. The remaining part of the circuit, in turn, may be thought of as a "filter" that takes r_n, x_n, (and ν_n if a dither exists) as input, and gives $\tilde{c}_n$ as output.

With this development, we may now state and prove the theorems of this chapter regarding stability. For the proofs, we shall shift the subscripts of r and $\tilde{c}$ upward by N to be consistent with the formulas at the end of the last chapter that are used.

Theorem 3.2 *Suppose all the zeros of $p_r(z)$ have magnitude less than 1 (i.e. the poles of the NTF (1.3) are strictly inside the unit circle). Then the filter output r_n in (1.2) will remain bounded for all $n \geq 0$.*

Proof:

We have $|\mu_i| < 1$ and $|\mu_i^{\pm}| < 1$ for all zeros μ_i, $\mu_i^{\pm} = k_{1,i} \pm k_{2,i}\hat{i}$ of $p_r(z)$. Clearly $r_{g,n}$, $r_{p,n}$ and thus r_n are finite for any $n \geq 0$. Now $\limsup_{n\to\infty} |r_n| = \limsup_{n\to\infty} |r_{g,n} + r_{p,n}| \leq \limsup_{n\to\infty} |r_{g,n}| + \limsup_{n\to\infty} |r_{p,n}|$. From the zeros of $p_r(z)$ and (2.3), $\limsup_{n\to\infty} |r_{g,n}| = 0$. Now, from (2.4),

$$\begin{aligned}
\limsup_{n\to\infty} |r_{p,n}| &\leq \limsup_{n\to\infty} |\tilde{c}_n| + \limsup_{n\to\infty} \sum_{k=N}^{n-1} |\tilde{c}_k||L_{n-k-1}(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij})| \\
&\leq K_1 + \limsup_{n\to\infty} \sum_{k=N}^{n-1} K_1 |L_{n-k-1}(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij})|,
\end{aligned}$$

where $K_1 = \frac{\Delta}{2}\sum_{k=1}^{M}|a_k| \geq \sum_{k=1}^{M}|a_k||\varepsilon_{n-N-k}| \geq |\tilde{c}_n|, \ \ \forall\ n \geq N$. We also have

$$\limsup_{n\to\infty}\sum_{k=N}^{n-1}|L_{n-k-1}(\tilde{A}_{ij},\tilde{B}_{ij},\tilde{C}_{ij})| \ \leq\ \limsup_{n\to\infty}\sum_{k=N}^{n-1}\sum_{i=1}^{s}|[\sum_{j=1}^{t_i}\tilde{A}_{ij}(n-k-1)^{j-1}]\mu_i^{n-k-1}|$$
$$+\ \limsup_{n\to\infty}\sum_{k=N}^{n-1}\sum_{i=1}^{\hat{s}}|[\sum_{j=1}^{\hat{t}_i}(n-k-1)^{j-1}]|\mu_i^{\pm}|^{n-k-1}\cdot(|\tilde{B}_{ij}|+|\tilde{C}_{ij}|)|.$$

The R.H.S. of the above expression represents the sum of N convergent series. Let K_2 be the sum of the convergent values of these series, so that $\limsup_{n\to\infty}\sum_{n=N}^{n-1}|L_{n-k-1}(\tilde{A}_{ij},\tilde{B}_{ij},\tilde{C}_{ij})| \leq K_2$.

Putting the results together, we have $\limsup_{n\to\infty}|r_n| = \limsup_{n\to\infty}|r_{p,n}| \leq K_1 + K_1K_2 = K_3$. With r_n finite as well, we then have that r_n is bounded $\forall\ n \geq 0$. ■

This theorem gives the most basic and general result of this chapter. From this we have a simple natural condition on the zeros of $p_r(z)$, and hence on the b_j, for guaranteeing stability in the Σ-Δ modulator. A simple converse condition guaranteeing instability does not exist. In general, if $p_r(z)$ has a zero with magnitude greater than 1 (or magnitude 1, multiplicity greater than 1), instability will arise, although exceptions to this rule may exist.

We begin a further exploration by first considering the case where the largest magnitude zero of $p_r(z)$ equals 1 (multiplicity 1).

Theorem 3.3 *Suppose the largest magnitude zero(s) of $p_r(z)$ have magnitude 1 and multiplicity 1. Suppose also that the relative input $\tilde{c}_n$ is periodic with period $p \in \mathbb{Z}^+$. Suppose further that no zero μ_i of $p_r(z)$ is identically 1, is identically -1 when p is odd, or has $\theta_i = 2\pi m/p$ when $k_{2,i} \neq 0$, $m \in \mathbb{Z}$. Then the filter output r_n in (1.2) will remain bounded for all $n \geq 0$.*

Proof:

Let $r_n = r_{1,n} + r_{2,n} + \tilde{c}_n$, where $r_{1,n}$ and $r_{2,n}$ are the parts of the expression for r_n (involving L) corresponding to the zeros of $p_r(z)$ with magnitude less than 1 and equal to 1 respectively. Applying the proof of Theorem 3.2, we have that $r_{1,n} + \tilde{c}_n$ will be bounded. Let $L_{2,n}(A_{i1}, B_{i1}, C_{i1})$ represent the part of $r_{g,n}$ and $r_{2,n}$ corresponding to the zeros of $p_r(z)$ with magnitude 1. From (2.3) this will also be bounded. Let these bounds be K_4 and K_5 respectively. Now we have

$$|r_n| \leq K_4 + K_5 + |\sum_{k=N}^{n-1} \tilde{c}_k L_{2,(n-k-1)}(\tilde{A}_{i1}, \tilde{B}_{i1}, \tilde{C}_{i1})|, \quad \text{for } n > N. \tag{I}$$

Since $\tilde{c}_n$ is periodic with period p, we may represent it by $\tilde{c}_{qp+l+N} = c_l$, $l = 0, \ldots, p-1$, $q \geq 0$, for some such $c_l \in \mathbb{R}$, where $n = qp + l + N$. Assuming no zero of $p_r(z)$ is identically 1, then inequality (I) leads to

$$\begin{aligned} |r_n| \;\leq\; & K_4 + K_5 + \sum_{l=0}^{p-1} |c_l \sum_{k=N}^{\hat{q}} \tilde{A}_{21}(-1)^{n-(N+kp+l)-1}| \\ & + \sum_{l=0}^{\hat{q}} |c_l| \sum_{i=1}^{\hat{s}} (|\tilde{B}_{i1}||\Im(S(n-(N+\hat{q}p+l)-1, n-N-l-1, p))| \\ & \qquad + |\tilde{C}_{i1}||\Re(S(n-(N+\hat{q}p+l)-1, n-N-l-1, p))|), \quad n > N, \end{aligned}$$

where $S(\alpha, \beta, \gamma) = \sum_{k=\alpha}^{\beta} e^{\hat{i}\theta_i \gamma k}$, α, β, $\gamma \in \mathbb{Z}$; $\Re$, $\Im$ denote the real and imaginary parts respectively; and $\hat{q} = (n-l) - [(n-l) \bmod p\ ]$. Note that the order of summation is reversed in the trigonometric terms. The third term above will exist only if p is even. In this case, it is clearly bounded by $K_6 = \sum_{l=0}^{p-1} |c_l \tilde{A}_{21}|$, $\forall\ n > N$. The fourth and fifth terms above are bounded by $\sum_{l=0}^{p-1} |c_l| \sum_{i=1}^{\hat{s}} D_i \|(\frac{e^{\hat{i}(n-N)p\theta_i} - 1}{e^{\hat{i}p\theta_i} - 1})\| \cdot \|e^{-pl\theta_i}\|$, where $D_i = \max(|\tilde{B}_{i1}|, |\tilde{C}_{i1}|)$, which in turn is bounded by $K_7 = \sum_{i=1}^{\hat{s}} D_i |\frac{2}{\sqrt{2 - 2\cos(p\theta_i)}}|$, $\forall\ n > N$, if $\theta_i \neq 2\pi m/p$, m

$\in \mathbb{Z}$. Putting the results together, we have that $|r_n| \leq K_4 + K_5 + K_6 + 2K_7, \ \forall\, n > N$ so that r_n is bounded $\forall\, n \geq 0$. ■

Proposition 3.4 *Suppose the largest magnitude zero(s) of $p_r(z)$ have magnitude 1 and multiplicity 1. Suppose also that $c \in \mathbb{R} - \{0\}$ is a constant, and one of the following holds.*

(a) $\tilde{c}_n = c$ and no zero of $p_r(z)$ is identically 1;

(b) $\tilde{c}_n = c(-1)^n$ and no zero of $p_r(z)$ is identically -1;

(c) $\tilde{c}_{qp+l} = c(-1)^q$, $l = 0, \ldots, p-1$, $q \geq 0$, for some $p \in \mathbb{Z}^+$, $n = qp + l$; and no zero μ_i of $p_r(z)$ has $\theta_i = \pi(2m-1)/p$ when $k_{2,i} \neq 0$, $m \in \mathbb{Z}$, or is identically -1 when p is odd.

Then the filter output r_n in (1.2) will remain bounded for all $n \geq 0$.

Proof:

(a) This result follows directly from Theorem 3.3.

(b) Assuming $\tilde{c}_n = c(-1)^n$ and no zero of $p_r(z)$ is identically -1, then inequality (I) in the proof of Theorem 3.3 leads to

$$|r_n| \ \leq \ K_4 + K_5 + |\sum_{k=N}^{n-1} c(-1)^k \tilde{A}_{11}| \\ + \ |c| \sum_{i=1}^{\hat{s}} (|\tilde{B}_{i1}||\Im(S(0, n-N-1, 1))| + |\tilde{C}_{i1}||\Re(S(0, n-N-1, 1)|), \quad n > N.$$

The bounds of K_6 and K_7 from part (a) will hold for the third and fourth terms of the above respectively. Analogously to part (a), we then have that r_n is bounded $\forall\, n \geq 0$ here.

(c) Assuming $\tilde{c}_{qp+l+N} = c(-1)^q$, $l = 0, \ldots, p-1$, for some $p \in \mathbb{Z}^+$, $q \geq 0$, then inequality (I) leads to

$$
\begin{aligned}
|r_n| \;\le\; & K_4 + K_5 + |\sum_{k=N}^{n} \tilde{c}_k \tilde{A}_{11}| + |\sum_{k=N}^{n} \tilde{c}_k \tilde{A}_{21}(-1)^{n-k-1}| \\
& + |c| \sum_{i=1}^{\hat{s}} (|\tilde{B}_{i1}||\Im(S(n-N-p, n-N-1, 1)I_1 + S(N, n-N-1-2\tilde{q}p, 1))| \\
& + |\tilde{C}_{i1}||\Re(S(n-N-p, n-N-1, 1)I_1 + S(N, n-N-1-2\tilde{q}p, 1))|),
\end{aligned}
$$

where $n > N$, $n = qp + l + N$ from above, and $2\tilde{q}$ is the largest even number less than or equal to q ($\tilde{q} \in \mathbb{Z}$). We define $I_1 = S(0, \tilde{q}-1, -2p)(1 - S(1, 1, -p))$ if $\tilde{q} > 0$, and $I_1 = 0$ if $\tilde{q} = 0$.

The third term above is bounded by $K_6 = p|c\tilde{A}_{11}|$. The fourth term will be bounded by $K_7 = |c\tilde{A}_{21}|$ provided that p is even. If p is odd, we assume that no zero of $p_r(z)$ is identically -1 as stated in the theorem, so that the fourth term does not exist (if $\tilde{A}_{21} \neq 0$ and p is odd, this term would be unbounded). The subsequent terms above will be bounded by $|c| \sum_{i=1}^{\hat{s}} D_i(K_{8,i} I_{2,i} + K_{9,i})$, where $K_{8,i} = \|S(l-p, l-1, 1)\|$, $K_{9,i} = \|S(N, p(q-2\tilde{q})+l-1, 1)\|$, and $I_{2,i} = \|e^{\hat{i}qp\theta_i}\| \cdot \|(\frac{e^{-2\hat{i}\tilde{q}p\theta_i} - 1}{e^{-2\hat{i}p\theta_i} - 1})\| \cdot \|1 - e^{-\hat{i}p\theta_i}\|$ if $\theta_i \neq m\pi/p$, $m \in \mathbb{Z}$ and $\theta_i \neq 2m\pi/p$, $m \in \mathbb{Z}$; $I_{2,i} = 0$ if $\theta_i = 2m\pi/p$, $m \in \mathbb{Z}$. $I_{2,i}$ is then bounded by $K_{10,i} = |\frac{4}{\sqrt{2 - 2\cos(2p\theta_i)}}|$, if $\theta_i \neq m\pi/p$, $m \in \mathbb{Z}$ and $\theta_i \neq 2m\pi/p$, $m \in \mathbb{Z}$, and by $K_{10,i} = 0$ if $\theta_i = 2m\pi/p$, $m \in \mathbb{Z}$, $\forall$ $q \geq 0$.

Putting all the results together, we have that

$$
|r_n| \leq K_4 + K_5 + K_6 + K_7 + 2|c| \sum_{i=1}^{\hat{s}} D_i(K_{8,i} K_{10,i} + K_{9,i}), \quad \forall\, n > N,
$$

provided no zero of $p_r(z)$ has $\theta_i \neq (2m-1)\pi/p$, when $k_{2,i} \neq 0$, $m \in \mathbb{Z}$, and no zero of $p_r(z)$ is identically -1 when p is odd. r_n is then bounded $\forall$ $n \geq 0$ under these conditions. ■

With Theorem 3.3, we see how stability may be assured when the input $\tilde{c}_n$ is periodic. For a given such input, stability will fail to be assured only for special cases when the

natural "frequency" of the system, as determined by the position of the zeros of $p_r(z)$ on the unit circle in $\mathcal{C}$, is a rational multiple of the input frequency. These are "resonance" cases suggesting unboundedness. The conditions of Theorem 3.3 will generally be "if and only if" in nature, with exceptions to this possible only when certain combinations of input $\tilde{c}_n$ and/or the coefficients of the magnitude 1 zeros in the particular solution exist. Hence resonance generally means unboundedness. If all magnitude 1 zeros of $p_r(z)$ are not rational fraction multiples of 2π in phase on the unit circle, then stability would be assured for any periodic input $\tilde{c}_n$.

Proposition 3.4 parts (b) and (c) show, under a special form of constant magnitude periodic input $\tilde{c}_n$, how stability may be assured with less restrictive conditions. These essentially involve allowing a zero identically at 1, and generally half the added restrictions, since p here corresponds to a period of $2p$ in Theorem 3.3. Parts (a) and (b) are essentially special cases of part (c), with (a) corresponding to the limit as $p \to \infty$, and (b) corresponding to $p = 1$. Cases with more general periodic input $\tilde{c}_n$ with either constant or varying magnitude, could be analyzed by considering this input as a sum of inputs of the form in parts (a), (b) and (c). For such a decomposition, the union of all the conditions in the proposition to provide boundedness for each input case would provide boundedness for the case of the sum of the inputs as overall input. This result follows from the linearity of the difference equation for r_n in (3.1). With such added complexity, instability would be expected to be more prevalent as we see for general periodic input $\tilde{c}_n$ in Theorem 3.3.

A more direct relevance of Theorem 3.3 and Proposition 3.4 follows from the relationship between the input $\tilde{c}_n$ and the errors ε_n. Specifically, from (3.1), periodicity in the errors will clearly imply periodicity in the corresponding input. A more particular structure of periodic errors will be sufficient to yield an input of the form in Proposition 3.4 as well. From the consideration of dynamical behaviour in the Σ-Δ modulator that brings about

limit cycles (i.e. some error periodicity), one might then conclude that such behaviour would tend to be more stable in the marginally minimum-phase case.

We now have the following:

Proposition 3.5 *Suppose the largest magnitude zero(s) of* $p_r(z)$ *have magnitude* 1 *and multiplicity* 1. *Suppose also that either* $\sum_{k=-M}^{n} |\varepsilon_k| \leq K_A$, *for some* $K_A > 0$, *or* $\sum_{k=0}^{n} |\tilde{c}_k| \leq K_B$, *for some* $K_B > 0$, *hold for all* $n \geq 0$. *Then the filter output* r_n *in (1.2) will remain bounded for all* $n \geq 0$.

Proof:

To begin we note that if $\sum_{k=-M}^{n} |\varepsilon_k| < K_A, \forall\, n \geq -M$, then it follows, from the definition of $\tilde{c}_k$, that $\sum_{k=N}^{n} |\tilde{c}_k| \leq K_B, \forall\, n \geq N$, where $K_B = K_A \sum_{k=1}^{M} |a_k|$, $K_A, K_B > 0$, and where we increase the subscripts by a factor of N as in the previous notation, for $\tilde{c}_n$. Applying the initial approach of Theorem 3.3 here, we arrive at inequality (I). With the bound K_B above, this leads to $|r_n| \leq K_4 + K_5 + K_B[|\tilde{A}_{11}| + |\tilde{A}_{21}| + \sum_{i=1}^{\hat{s}}(|\tilde{B}_{i1}| + |\tilde{C}_{i1}|)], \ \forall\, n > N$, so that r_n is bounded $\forall\, n \geq 0$. ■

This result provides an extension of the guaranteed stability for minimum-phase systems to the marginally minimum-phase (multiplicity 1 for zeros on the unit circle) case, but under the restriction of having an error sequence ε_n whose partial sums are bounded in magnitude. Obviously for this requirement to hold it is necessary at least that the error $\varepsilon_n \to 0$ as $n \to \infty$.

Proposition 3.6 *The particular solution of the filter output r_n in (1.2) will be $r_{p,n} = \tilde{c}_n$, for all $n \geq K_N \geq 0$, if and only if $\sum_{j=1}^{N} b_j \tilde{c}_{n-j} = 0$ holds for all $n \geq K_N + N$, and $r_{p,n} = \tilde{c}_n$ for $n = K_N, \ldots, K_N + N - 1$. Moreover, if the largest magnitude zero(s) of $p_r(z)$ have magnitude 1 with multiplicity 1, and $\tilde{c}_n$ is such that the above holds, then the filter output r_n will remain bounded for all $n \geq 0$.*

Proof:

Suppose the particular solution is $r_{p,n} = \tilde{c}_n$, $\forall\ n \geq K_N \geq 0$. Then the result follows directly from the difference equation for r_n in (3.1). Suppose conversely that $\sum_{j=1}^{N} b_j \tilde{c}_{n-j} = 0$ holds $\forall\ n \geq K_N + N$, and $r_{p,n} = \tilde{c}_n$ for $n = K_N, \ldots, K_N + N - 1$. Then, by induction on the difference equation for r_n, we arrive at $r_{p,n} = \tilde{c}_n$, $\forall\ n \geq K_N \geq 0$. Now suppose that the zeros of $p_r(z)$ are as given. From the formation of inequality (I) at the beginning of the proof of Theorem 3.3, we have that the last term is zero for $n \geq \max(K_N, N)$. Thus $|r_n| \leq K_4 + K_5$ for $n \geq \max(K_N, N)$, so that r_n is bounded $\forall\ n \geq 0$. ■

This result provides an extension of guaranteed stability to these marginally minimum-phase cases in the simple and unique case when the input $\tilde{c}_n$ satisfies a special property with respect to the filter. Any such sequence $\tilde{c}_n$ would essentially constitute, for $n \geq K_N + N$, the orbit generated by a $(N-p)$th order difference equation with an initial condition formed from some $N - p$ iterates from the set $\{\tilde{c}_{K_N}, \ldots, \tilde{c}_{K_N+N-1}\}$, where p is the smallest integer such that $b_p \neq 0$. To have $r_{p,n} = \tilde{c}_n$ for $n = K_N, \ldots, K_N + N - 1$, requires appropriate control values for $\tilde{c}_{K_N-1}, \ldots, \tilde{c}_{K_N+N-2}$. The trivial case where $\tilde{c}_n = 0$ for all $n \geq 0$, holds if $K_N = 0$. It is under the trivial condition only, that the coefficients $\tilde{A}_{ij}$, $\tilde{B}_{ij}$, $\tilde{C}_{ij}$ in the particular solution of r_n all vanish.

Proposition 3.7 *Suppose that all the zero(s) of $p_r(z)$ have magnitude 1 or greater. Suppose also that the relative inputs $\tilde{c}_n$ are random, i.i.d., and described by a probability density/mass function that is either piecewise continuous on some domain, or discrete with at least 2 different values of nonzero probability. Then the filter output r_n in (1.2) will be unbounded with probability 1.*

Proof:

We consider r_n as a random variable where, from (2.4), we have $r_n = r_{g,n} + \tilde{c}_n + \sum_{k=N}^{n-1} \tilde{c}_k L_{n-k-1}(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij})$. First we have from (2.3), with some nonzero constants present, that $|L_{n-k-1}(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij})|$ will either be unbounded, or vary between zero and $|\tilde{A}_{11}| + |\tilde{A}_{21}| + \sum_{i=1}^{\hat{s}} (|\tilde{B}_{i1}| + |\tilde{C}_{i1}|)$ in magnitude, periodically or densely, as k goes from N to $n-1$ for arbitrarily large n. This implies the property $\lim_{m\to\infty} \sum_{l=1}^{m} L_l(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij})^2 = \infty$.

Now we suppose that $|E[r_n]| \leq K_m$, for all $n \geq 0$, and some $K_m > 0$. Taking the variance of r_n, we have that $\mathrm{Var}[r_n] = \mathrm{Var}[r_{g,n}] + \mathrm{Var}[\tilde{c}_N] \sum_{k=N}^{n-1} L_{n-k-1}(\tilde{A}_{ij}, \tilde{B}_{ij}, \tilde{C}_{ij})^2$, using the independence of the i.i.d. $\tilde{c}_k$ and the fixed $r_{g,n}$. Applying the property of L_{n-k-1} given above, it follows that $\lim_{n\to\infty} \mathrm{Var}[r_n] = \infty$. This then implies that for any $K > 0$, the probability that $\exists\ n_1 > N$ such that $|r_n| > K$ equals 1. If, conversely, $|E[r_n]|$ is unbounded as $n \to \infty$, then this final conclusion follows as well. Thus r_n is unbounded. ■

It is expected that the result of this proposition would generally hold under the relaxed condition that at least 1 zero of $p_r(z)$ have magnitude 1 or greater. Under such conditions, the proof of the proposition carries through as long as the nonzero constants of $r_{p,n}$ are not exclusive to the terms associated with the minimum-phase zeros of $p_r(z)$, as would be expected generically.

Proposition 3.7 illustrates how, for these marginally minimum-phase cases, we require some element either of regularity or dissipation in the magnitude of the input $\tilde{c}_n$, and by association the errors ε_n, in order to attain stability over at least broad sub-cases. A purely random i.i.d. input $\tilde{c}_n$ simply reinforces (i.e. assures) the tendency towards instability. Unsurprisingly, this property extends directly to the nonminimum-phase cases, where instability is generally expected.

Considering the overall system, it can be seen that a random i.i.d. system input x_n will bring about a random i.i.d. input $\tilde{c}_n$ for Σ-Δ modulators of general filter forms. Random i.i.d. errors ε_n are similarly related. It should be noted in this context that the definition for randomness of the $\tilde{c}_n$ in Proposition 3.7 is quite explicit. Thus while randomly interpreted long run error behaviour, as discussed in Chapter 7, and hence long run behaviour of the input $\tilde{c}_n$, may hint at some connections; Proposition 3.7 cannot be applied, in general, to simply any system with a defined random steady state error or $\tilde{c}_n$ (see Chapter 7).

We now proceed to the case where $p_r(z)$ has a zero with either magnitude greater than 1, or else magnitude 1, multiplicity greater than 1.

Theorem 3.8 *Suppose $p_r(z)$ has N_1 zeros of magnitude less than 1, and N_2 zeros of magnitude 1 and multiplicity 1, with $N_{MP} = N_1 + N_2$, where $0 \leq N_M \leq N$. Suppose also that $\tilde{c}_n$ is fixed with respect to the initial conditions r_i, $i = -1, \ldots, -N$, for all $n \geq 0$. Suppose further that the following hold:*

(a) the part of the solution r_n corresponding to the N_2 zeros of magnitude 1, multiplicity 1, is bounded for all $n \geq 0$;

(b) if zeros of magnitude 1 or greater, with multiplicity greater than 1, exist, and/or complex conjugate pairs of zeros of magnitude greater than 1 exist, then the part of the solution r_n corresponding to these zeros is bounded for all $n \geq 0$, for some initial condition $\underline{\hat{r}}_0 \in \mathbb{R}^N$.

Then the filter output r_n in (1.2) will remain bounded for all $n \geq 0$, over an N_M-dimensional subspace of the initial conditions r_i, $i = -1, \ldots, -N$.

Proof:

Let $r_n = r_{1,n} + r_{2,n} + r_{3,n} + r_{4,n} + \tilde{c}_n$, where $r_{1,n}$, $r_{2,n}$, $r_{3,n}$ and $r_{4,n}$ are the parts of the expression for r_n (involving L) corresponding to the zeros of $p_r(z)$ with magnitude less than 1, equal to 1 with multiplicity 1, equal to 1 or greater with multiplicity greater than 1 or complex with magnitude greater than 1, and real with magnitude greater than 1 with multiplicity 1, respectively. Applying Theorem 3.2 and conditions (a) and (b) of this theorem, we have that $r_{1,n} + r_{2,n} + r_{3,n} + \tilde{c}_n$ will be bounded. Let this bound be $\tilde{K}_1$ so that we have $|r_n| \leq \tilde{K}_1 + |r_{4,n}|$ for $n > N$. Denoting the part of L associated with $r_{4,n}$ by $L_{4,n}$, we have $r_{4,n} = L_{4,n}(A_{i1}) + \sum_{k=N}^{n-1} \tilde{c}_k L_{4,(n-k-1)}(\tilde{A}_{i1})$. Factoring out the μ_i^n from the terms of $r_{4,n}$ above, we obtain $r_{4,n} = \sum_{i=1}^{s} \mu_i^n (A_{i1} + \tilde{A}_{i1} \sum_{k=N}^{n-1} \tilde{c}_k (\mu_i^{-1})^{k+1})$. The $\tilde{c}_n$ above are bounded in magnitude by K_1 from the proof of Theorem 3.2. The series in the term of the sum over k above is the partial sum of a geometric series that thus converges as $n \to \infty$. Let S_{Ai} and $E_{Ai,n}$ be the convergent value of the series, and the error between this and the partial sum up to term n, respectively. Then we have $r_{4,n} = \sum_{i=1}^{s} \mu_i^n (A_{i1} + \tilde{A}_{i1}(S_{Ai} + E_{Ai,n}))$.

We may now choose $A_{i1} = -\tilde{A}_{i1} S_{Ai}$ to eliminate these terms. The error $E_{Ai,n}$ is $O(-n)$ in magnitude, which will cancel with the μ_i^n factor to make the remaining term $O(1)$ in magnitude $\forall$ n, such that we may conclude that $|r_{4,n}| \leq \tilde{K}_2$ for $n > N$ for some $\tilde{K}_2 > 0$. Thus, with the particular choice of A_{i1}, we have that $|r_n| < \tilde{K}_1 + \tilde{K}_2$, $\forall$ $n > N$, so that r_n is bounded $\forall$ $n \geq 0$. From the proof of Theorem 3.2 for the homogeneous part of r_n, we have that the homogenous parts of $r_{1,n}$ and $r_{2,n}$ will be bounded over all choices of initial conditions, as they determine all A_{ij}, B_{ij}, C_{ij}. Thus the boundedness of $r_{1,n}$ and

$r_{2,n}$ from Theorem 1 and condition (a) respectively, holds over all initial conditions. An initial condition $\underline{\hat{r}}_0$ at which condition (b) is satisfied imposes constraints on the constants of $r_{3,n}$ only. With the corresponding fixed constants A_{ij}, B_{ij}, C_{ij} to give the boundedness derived for $r_{4,n}$, and those we assume fixed corresponding to $r_{3,n}$ to satisfy condition (b), we are left with the N_M constants corresponding to $r_{1,n}$ and $r_{2,n}$ which may take on any values to keep r_n bounded $\forall\ n \geq 0$. The range of initial conditions $r_{-1}, \ldots, r_{-N}$ (subscripts increased by N for consistency with the formulas above) that allow any N_M constants A_{ij}, B_{ij}, C_{ij} while holding the other $N - N_M$ constants fixed constitutes an N_M dimensional subspace of initial conditions for the system.

If $\tilde{c}_n = 0, \forall\ n > N$, we may simply choose the $N - N_M$ A_{ij}, B_{ij}, C_{ij} all zero in $r_{3,n}$ and $r_{4,n}$ to satisfy conditions (b) and bound $r_{4,n}$. The particular solution part of $r_{2,n}$ is also bounded, and thus condition (a) is satisfied. ■

In Theorem 3.8 we have attempted to extend some sort of general guarantee for stability to the nonminimum-phase case (or marginally minimum-phase, with a zero on the unit circle with multiplicity greater than 1). We see greater possibility for this when some zeros of $p_r(z)$ are less than 1 in magnitude (or magnitude 1, multiplicity 1), and focus on a subspace of initial conditions of the r_i whose dimension reflects the relative degrees of freedom offered by these zeros. If no zeros of this type exist, then (with condition (b) satisfied) stability is guaranteed at only one unique initial condition of the r_i. Therefore we see that real zeros of magnitude greater than 1, multiplicity 1, impose no constraint on the nominal input $\tilde{c}_n$, but impose a constraint on the initial conditions of r_i to assure stability. Zeros of magnitude 1, multiplicity 1, still impose constraints on the input (e.g. the input $\tilde{c}_n$ of Proposition 3.5 would suffice), but not on the initial conditions; while zeros of magnitude 1 or greater, multiplicity greater than 1, or of magnitude greater than 1, and in complex

conjugate pairs, impose constraints on the initial conditions and even broader constraints on the input, to assure stability.

More trivially, both conditions of the theorem are satisfied when $\tilde{c}_n = 0$, for all $n \geq 0$. Note that conditions (a) and (b) in Theorem 3.8 simply reflect a general boundedness criterion relating to the part of the output r_n associated with the respective zeros in each case. From Theorem 3.8 it follows directly that if, for some system, r_n is bounded for all $n \geq 0$, for a particular initial condition $\underline{\hat{r}}_0 \in \mathbb{R}^N$, then r_n will remain bounded over an N_M-dimensional subspace of initial conditions containing $\underline{\hat{r}}_0$.

Since the stability result of Theorem 3.8 requires $\tilde{c}_n$ to be fixed with respect to changes in the initial conditions of the r_i, this generally means that the system input x_n, for each $n \geq 0$, must be restricted in some way that is at least as great as the subspace restriction of these initial conditions (in some cases only for finite iterations of n). This follows from the dependency of $\tilde{c}_n$ on the r_i initial conditions, as seen in (1.2).

Corollary 3.9 *Suppose the conditions of Theorem 3.8 hold for some initial condition* $\underline{\hat{r}}_0 \in \mathbb{R}^N$. *Then the filter output* r_n *in (1.2) will be unbounded over all initial conditions* r_i, $i = -1, \ldots, -N$, *not contained in the subspace of initial conditions for boundedness provided by Theorem 3.8.*

Proof:

From the proof of Theorem 3.8, we have that the subspace of initial conditions assuring boundedness provides for the constants A_{ij}, B_{ij}, C_{ij} corresponding to $r_{1,n}$ and $r_{2,n}$ to take on any values, and for the constants corresponding to $r_{3,n}$ and $r_{4,n}$ to be fixed. We denote these fixed constants by $\hat{A}_{ij}$, $\hat{B}_{ij}$, $\hat{C}_{ij}$. Suppose $\underline{r}_0^* \in \mathbb{R}^N$ is an initial condition not contained in the subspace of initial conditions assuring boundedness. We denote the constants corresponding to $r_{3,n}$ and $r_{4,n}$ arising from the initial condition $\underline{r}_0^*$ by A_{ij}^*, B_{ij}^*, C_{ij}^*.

Let $\hat{r}_n$ and r_n^* be the solutions corresponding to initial conditions $\underline{\hat{r}}_0$ and $\underline{r}_0^*$ respectively. We then have $|\hat{r}_n - r_n^*| \leq |\hat{r}_{g,5,n} - r_{g,5,n}^*| + |\hat{r}_{g,6,n} - r_{g,6,n}^*|$, where the subscripts refer to the general solution part of the solutions corresponding to parts $r_{5,n} = r_{1,n} + r_{2,n}$ and $r_{6,n} = r_{3,n} + r_{4,n}$ from the proof of Theorem 3.8. The particular solutions of $\hat{r}_n$ and r_n^* are equivalent and cancel in the above result. From the proof of Theorem 3.2, we have that $\limsup_{n\to\infty}(\hat{r}_{g,5,n} - r_{g,5,n}^*) = 0$.

Now we have $|\hat{r}_{g,6,n} - r_{g,6,n}^*| = |L_{6,n}(\hat{A}_{ij} - A_{ij}^*, \hat{B}_{ij} - B_{ij}^*, \hat{C}_{ij} - C_{ij}^*)|$, where $L_{6,n}$ denotes the part of L associated with $r_{g,6,n}$. From the definition of $\underline{r}_0^*$, it must hold that the differences $\hat{A}_{ij} - A_{ij}^*$, $\hat{B}_{ij} - B_{ij}^*$ and $\hat{C}_{ij} - C_{ij}^*$ are not all zero, because the corresponding constants of the two cases cannot all be equal. From the form of the zeros associated with $r_{6,n}$ and (2.3), it then follows that $\limsup_{n\to\infty} |L_{6,n}(\hat{A}_{ij} - A_{ij}^*, \hat{B}_{ij} - B_{ij}^*, \hat{C}_{ij} - C_{ij}^*)| = \infty$.

Thus, from above, we have that $\limsup_{n\to\infty} |\hat{r}_n - r_n^*| = \infty$. With $\hat{r}_n$ bounded, this implies that r_n^* is unbounded over n. Since $\underline{r}_0^*$ was chosen arbitrarily outside of the subspace of initial conditions for boundedness, we then conclude that r_n is unbounded over all initial conditions not in this subspace. ■

This result provides the strongest guarantee for instability that we have found. Basically, this corollary implies that any system with a zero of $p_r(z)$ with either magnitude greater than 1, or else magnitude 1, multiplicity greater than 1, cannot be stable over all initial conditions of the r_i, with $\tilde{c}_n$ fixed, or in general over all these initial conditions and all system input x_n. From our results we note that for any fixed input $\tilde{c}_n$, there exists an initial condition for the r_i which yields stability, provided all zeros of $p_r(z)$ have either magnitude less than 1, or are real with magnitude greater than 1, multiplicity 1. Furthermore, from the basic form of the difference equations in (3.1), we have that for any system (i.e. any $p_r(z)$) there exist initial conditions together with an input $\tilde{c}_n$ that yield stability. These

initial conditions would typically be small if the magnitude of the zeros is large, given the boundedness of the $\tilde{c}_n$. If no zeros of $p_r(z)$ have either magnitude greater than 1 or else magnitude 1, multiplicity greater than 1, then stability, when it exists, holds for all initial conditions.

We note that the stability results of Proposition 3.6, Theorem 3.8 and its corollary are generally not robust to small perturbations in the required input $\tilde{c}_n$ (or x_n) or initial conditions r_i, hence limiting direct practical use.

Discussion:

In this chapter, we have taken the natural point of view of assuming that the overall system has some input x_n, which then gives rise to errors ε_n, which in turn give rise to the "feedforward" H_r filter input $\tilde{c}_n$ in the difference equation for r_n in (3.1). We then presented theorems and results which imposed conditions on this input $\tilde{c}_n$ in order to draw conclusions about stability. We now start with an assumed feedforward input $\tilde{c}_n$ and move in reverse, first considering how the ε_n arise from a given input $\tilde{c}_n$. With given initial conditions for ε_n, $n = -1, -2, \ldots, -M$, and hence $\tilde{c}_0$ as well, ε_n satisfies the difference equation

$$\varepsilon_{n-p} = \tilde{c}_n/a_p - \Big(\sum_{i=p+1}^{M} (a_i/a_p)\varepsilon_{n-i}\Big), \quad n \geq 0, \tag{3.2}$$

where $p = \min(i \mid a_i \neq 0)$. The ε_n must be bounded by $\Delta/2$ in magnitude. Thus, for a given $\tilde{c}_n$, $n \geq 0$, to be feasible, it at least must also satisfy the theorems for boundedness when $p_r(z)$ is defined from the difference equation (3.2). If we have such an input $\tilde{c}_n$ which generates valid errors ε_n, we may then find a system input x_n which such errors would require. This is given by $x_n = r_n + (m + 1/2)\Delta - \varepsilon_n$, for $m \in \mathbb{Z}$. Clearly the input can be bounded whenever r_n is bounded and the $\tilde{c}_n$ are feasible. The point of viewing things in reverse like this is to show, both how natural limitations in attempts to satisfy the stability

theorems may be seen from the nature of the difference equation (3.2) for ε_n, and how a realizable feedforward input $\tilde{c}_n$ may be obtained with an appropriate corresponding overall system input x_n.

We now comment further on stability for dithered systems. A necessary condition for stability is that the random dither ν_n be bounded, i.e. $|\nu_n| < K_D$, for all $n \geq 0$, for some $K_D > 0$. Under these conditions Theorem 3.2 will always apply. If $K_D > \Delta$, then $\tilde{c}_n$ will essentially be a discrete random variable for all $n \geq M$. If $K_D \leq \Delta$, then $\tilde{c}_n$ may be a discrete random variable for all $n \geq M$, some $n \geq 0$, or no $n \geq 0$, depending upon the behaviour of $x_n - r_n$, $n \geq 0$. In either K_D case, if $\tilde{c}_n$ is random for all $n \geq n_K$ for some $n_K \geq M$, then at best, only Proposition 3.7 is potentially applicable (in addition to Theorem 3.2). If $\tilde{c}_n$ is purely nonrandom, then all results, excluding Proposition 3.7, are potentially applicable.

Although simple bounded stability requires only that r_n remain bounded (for a bounded input and dither); in practical systems, it is naturally desirable for this bound not to be too large, and necessary for it not to be arbitrarily large. We may address this issue in relation to the results of this chapter. For the stability result of Theorem 3.2, we have from the proof, that, for general $\tilde{c}_n$, the size of the bound on r_n will depend on the magnitude of the largest zero of $p_r(z)$, and the number of multiplicities. Specifically, the bound will increase as the number of multiplicities of at least the largest zero increases, and will tend to infinity as its magnitude tends to 1. For Theorem 3.3 and Proposition 3.4, it follows from the proofs, that the bound on r_n will be roughly on the order of magnitude of $\max(|\tilde{A}_{ij}|, \|\tilde{B}_{ij}|, |\tilde{C}_{ij}|)$, (a result which extends when zeros are allowed inside the unit circle); and for Theorem 3.2, it follows that the bound will be (at most) this order of magnitude divided by (1 − magnitude of the largest zero), (multiplicity of 1 assumed). It is generally expected, from the formulation of (2.4), that $\max(|\tilde{A}_{ij}|, \|\tilde{B}_{ij}|, |\tilde{C}_{ij}|)$ will be roughly on the

order of magnitude of the b_j coefficients in (3.1). Proposition 3.5 provides the bounds K_A, K_B, on the order of magnitude of the bound on r_n directly; while Proposition 3.6 assures control on the r_n bound via control of the r_i initial condition magnitudes, a condition that is implicit for all stability results. For the stability result of Theorem 3.8, the magnitude of the bound on r_n, $n \geq 0$, is assumed directly. The bound on the magnitude of the initial conditions for r_i is at least that given from $\underline{\hat{r}}_0$, and, from the proof, will possibly be larger, as driven by the magnitude of the A_{i1}. Specifically, the magnitude of the largest A_{i1}, and hence that of the initial conditions, will tend to infinity as the magnitude of the associated smallest zero with magnitude greater than 1 tends to 1. Thus our results show that we can control the stability bound on r_n, if the magnitude of the zeros are not too close to 1, under the most general stability conditions.

From a practical point of view, Theorem 3.2 gives the most fundamental and useful result in this chapter for assuring stability, and it is consistent with the general stability assumptions of previous Σ-Δ modulator work, and indeed of feedback or control systems generally. The subsequent theorems may be less useful in particular applications, but they help illustrate the nature of the issue of stability and how it is governed by the input $\tilde{c}_n$, the form of the noise transfer function, and the structure of the Σ-Δ modulator as a dynamical system. From this we have an expansion on the traditional requirement for stability, with a dynamical analysis applied to the Σ-Δ system.

Chapter 4

Continuity in the Model

In this chapter, we examine the state space of the error coordinate $g(\vec{x}_n)$ from the non-dithered Σ-Δ modulator dynamical system (2.1), and derive some linearity and continuity results for the mapping f that will form the foundation of the analysis methods and approach in later chapters. These arise from the symmetries inherent in the associated circle map topology, and the algebraic form of (1.2). Pertinent extensions to the dithered form of (2.1) are dealt with in Chapter 6.

We begin by presenting a proposition which provides the most basic linear structure of the error state space topology, and the broadest result upon which this foundation is based — in particular, the means by which the analysis of all the chaos conditions will be made for the proofs of the theorems.

Proposition 4.1 *Suppose* $\vec{x}_{1,0}$, $\vec{x}_{2,0} \in \mathbb{R}^N \times \mathcal{C}^M$, *and there exists* $n_1 \geq N$ *such that* $Q(x_k - r_{1,k}) = Q(x_k - r_{2,k})$ *for all* $0 \leq k \leq n_1$, *where* $r_{1,k}$, $r_{2,k}$ *correspond to the system with initial conditions* $\vec{x}_{1,0}$ *and* $\vec{x}_{2,0}$ *respectively. Then* $\Delta\varepsilon_n$ *satisfies*

$$\Delta\varepsilon_n = \sum_{k=1}^{\max(N,M)} (a_k - b_k)\Delta\varepsilon_{n-k}, \quad N \leq n \leq n_1, \tag{4.1}$$

where $\Delta\varepsilon_n = g_M(\vec{x}_{2,(n+M)}) - g_M(\vec{x}_{1,(n+M)}), \quad n \geq -M.$

Proof:

The systems with initial conditions $\vec{x}_{1,0}$, $\vec{x}_{2,0}$ are governed by the difference equations (1.2). Subtracting the respective equations of (1.2) corresponding to $\vec{x}_{1,0}$ from those corresponding to $\vec{x}_{2,0}$ gives

$$\begin{aligned} r_{2,n} - r_{1,n} &= \sum_{k=1}^{M} a_k(\varepsilon_{2,(n-k)} - \varepsilon_{1,(n-k)}) - \sum_{k=1}^{N} b_k(r_{2,(n-k)} - r_{1,(n-k)}) \\ \varepsilon_{2,n} - \varepsilon_{1,n} &= Q(x_n - r_{2,n}) - Q(x_n - r_{1,n}) + (r_{2,n} - r_{1,n}), \end{aligned}$$

where the variables with subscripts 1 and 2 correspond to the system with initial conditions $\vec{x}_{1,0}$ and $\vec{x}_{2,0}$ respectively. Using the fact that $Q(x_m - r_{2,m}) - Q(x_m - r_{1,m}) = 0, \quad \forall\ 0 \leq m \leq n_1$, we may substitute the second equation into the first for these values of m to eliminate the r_m terms and obtain

$$\Delta\varepsilon_n = \sum_{k=1}^{\max(N,M)^*} (a_k - b_k)\Delta\varepsilon_{n-k}, \qquad \text{for } N \leq n \leq n_1,$$

where $\Delta\varepsilon_n = \varepsilon_{2,n} - \varepsilon_{1,n} = g_M(\vec{x}_{2,(n+M)}) - g_M(\vec{x}_{1,(n+M)}), \quad n \geq -M$. From iterating through (1.2), the initial conditions applying to this difference equation will then be given by $\Delta\varepsilon_{N-k}$, for $k = 1, \ldots, \max(N, M)^*$. ■

This result shows that when we consider two initializations of the Σ-Δ modulator that are "nearby" in $\mathbb{R}^N \times \mathcal{C}^M$ ($Q(x_1 - r_{1,0}) = Q(x_1 - r_{2,0})$ holds if this is true) then the small difference in the error variables associated with each initialization, as n increases, will behave according to the simple difference equation above, for all n such that the quantized values of $x_k - r_k$ in the two cases remain equal for $k = 0, 1, \ldots, n$ and $n \geq N$. The $\max(N, M)^*$ "initial conditions" governing this are $\Delta\varepsilon_{N-1}$, $\Delta\varepsilon_{N-2}, \ldots,$ $\Delta\varepsilon_{N-\max(N,M)^*}$.

For this, we define

$$\begin{aligned} \max(N,M)^* &= \max\{k \geq 1 \mid a_k - b_k \neq 0\} \quad \text{if this exists, and} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Clearly $\max(N,M)^* = \max(N,M)$, unless $N = M$ and there is reduction of order due to cancellation. We also note in reference to the sum in (4.1), and all subsequent summations used involving the filter coefficients, that $a_i = 0$ and $b_j = 0$, for $i > M$, $j > N$, by definition from Section 1.3.

With this difference equation for $\Delta\varepsilon_n$ we have eliminated the dependence on the input x_n, which affects the position of $\varepsilon_{1,n}$ and $\varepsilon_{2,n}$ on $\mathcal{C}$ but not the distance between them. Thus we may apply the general solution to such difference equations to examine the behaviour of $\Delta\varepsilon_n$. This "autonomous" linear formulation of the error state space topology dynamics differs from the nonautonomous linear form of the internal variable r_n dynamics examined in Chapter 3 to study stability. Note that, under the conditions of the proposition, we always have $|\Delta\varepsilon_n| < \Delta$ for any n where the proposition holds. The behaviour of this error difference is central to the analysis of all the chaos conditions and constructing the theorem proofs. Later work in the thesis draws on this as well. The use of this approach breaks down of course when $n \geq n_2$ with $Q(x_{n_2} - r_{1,n_2}) \neq Q(x_{n_2} - r_{2,n_2})$ for some $n_2 \geq 0$, greatly complicating the analysis.

For subsequent analysis in this and later chapters, we need to focus on the behaviour of $\Delta\varepsilon_n$ given by (4.1), with n_1 taken to be arbitrarily large. The solution for $\Delta\varepsilon_n$ is given by the general solution obtained from applying (2.2) with (2.3); with $P = \max(N,M)^*$, $d_k = (a_k - b_k)$, $k = 1, \ldots, \max(N,M)^*$, and with the subscripts of $\Delta\varepsilon$ increased by a factor of $\max(N,M)^* - N$. We also have $d(z) = p(z)$, where

$$p(z) = z^{\max(N,M)^*} + \sum_{k=1}^{\max(N,M)^*} (b_k - a_k) z^{\max(N,M)^* - k}$$

is the characteristic polynomial of the difference equation for $\Delta\varepsilon_n$ in (4.1). If we multiply the noise transfer function (1.3) by $z^{\max(N,M)^*}$, we find that the numerator is given by the polynomial $p(z)$. Thus the zeros of the NTF are given by the zeros of $p(z)$. This fact, and the analogous result for the poles of the NTF and $p_r(z)$ considered in the stability analysis, shows the direct relationship between the noise transfer function and the dynamical analysis at hand.

To permit the greatest generality in the subsequent analysis in this thesis, we will essentially omit the pole condition from the definitions in our use the terms minimum/marginally minimum/nonminimum phase. Of course, it is generically required that all the poles of the noise transfer function (1.3) be inside the unit circle in order to satisfy these definitions, since this is the only broad condition that guarantees bounded internal stability, and this stability is required. Therefore, with the implicit assumption that this pole condition holds, our use of the terms minimum/marginally minimum/nonminimum phase will convey the desired meaning following the proper definitions as well.

4.1 Continuous Model Formulation

In this section, we construct a special simplified Σ-Δ modulator form, for which the circle map interpretation of the error state space dynamics, in one dimension, can essentially be applied for any error interval length, and for any mapping iteration $n \geq 0$. This topological nature will enable broad conclusions to be drawn about chaos as we define it in Chapter 5, as well as important results regarding long term or stochastic error behaviour, as studied later in the thesis.

The standard requirement that $C(\theta + k) = F(\theta + k) \mod 1 = F(\theta) \mod 1 = C(\theta)$ for $k \in \mathbb{Z}$ in the definition of a circle map f is essentially the property, in some more general

form, that we need to preserve in our dynamical system formulation if we are to extend a full fledged circle map interpretation to some general modulator form. For a simple first order system in (2.1) with $M = 1$, $N = 0$, the requirement implies immediately that $a_1 \in \mathbb{Z}$. For (2.1) in general, clearly the extension $a_i \in \mathbb{Z}$ and $b_j \in \mathbb{Z}$, for $i = 1, 2, \ldots, M$; $j = 1, 2, \ldots, N$; would then preserve this property in the generalized circle map on $\mathcal{C}^M$, where we specify this property as $g \circ f^n(r_{-1} + k_{b,1}, \ldots, r_{-N} + k_{b,N},\ \varepsilon_{-1} + k_{a,1}, \ldots, \varepsilon_{-M} + k_{a,M}) = g \circ f^n(r_{-1}, \ldots, r_{-N},\ \varepsilon_{-1}, \ldots, \varepsilon_{-M})$, for all $n \geq 0$, and for the $k_{a,i},\ k_{b,j} \in \mathbb{Z}$. For the topological analysis and applications to follow in this thesis, it will be sufficient to relax this generalized circle map property somewhat so as to be essentially equivalent to the property of continuity in the mapping f of (2.1). As will be subsequently established, a sufficient general requirement for this form of extended circle map property, is that the a_i and b_j be such that they satisfy the following:

Definition 4.2 ($\tilde{r}_k$)

$$\tilde{r}_1 = a_1, \quad \tilde{r}_k = a_k - \sum_{j=1}^{\min(N,k-1)} b_j \tilde{r}_{k-j}, \quad k \geq 2.$$

Definition 4.3 (Condition (R)) $\tilde{r}_k \in \mathbb{Z}$, *for all* $k \geq 1$. (R)

This then defines the simplified Σ-Δ modulator form. Intuitively, condition (R) provides a relaxation of the integer condition on the a_i and b_j, as afforded by the interrelationship between the respective ε_n and r_n recursive parts of the difference equation in (1.2). The complex recursive form of (R) arises from the manner in which the ε_n are associated with corresponding r_n that are delayed by one (or more) iterations, in the recursive structure of (1.2). It is this complexity that contributes to the relative restrictiveness in (R) as will be subsequently discussed. First, we present the connection between condition (R) and the notion of continuity.

In spite of the potential for using Σ-Δ modulators to approximate continuous processes such as stochastic resonance, and the abstract parallels to continuous systems therein, the dynamical system description (2.1) that we have is fundamentally (mathematically) a discrete time system in "time" n. Therefore, when we discuss having "continuity" in the model, we mean continuity (on $\mathcal{C}$) over the state space $\mathbb{R}^N \times \mathcal{C}^M$, or some subset thereof, for the discrete-time mappings $g_1 \circ f^n$, for some set of n in $\mathbb{Z}^+ \cup \{0\}$. It is such continuity that we see as the important consequence of the following theorem, which is fundamental to the proofs of the subsequent theorems pertaining to Σ-Δ modulators of condition (R) form.

Theorem 4.4 *Suppose the a_i and b_j satisfy condition (R), where $i = 1, 2, \ldots, M$; $j = 1, 2, \ldots, N$; holds. Then the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$, $n \geq 0$, is continuous on $\mathcal{C}$ at any $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ satisfying $\hat{\varepsilon}_{-k} = g_k(\hat{x}_0) \neq \Delta/2$, $k = 2, \ldots, M$. If $\hat{\varepsilon}_{-k_i} = g_{k_i}(\hat{x}_0) = \Delta/2$ for some $k_i \in \{2, \ldots, M\}$, then the function is continuous on $\mathcal{C}$ along paths in $\mathbb{R}^N \times \mathcal{C}^M$ that approach $\hat{x}_0$ from the left on the circles $\mathcal{C}$ associated with the coordinates ε_{k_i} of $\vec{x}_0$, and approach $\hat{x}_0$ from any direction on the remaining circles $\mathcal{C}$ and $\mathbb{R}^N$.*

Proof:

To begin, choose any $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$.

1. First, we show by induction how under special circumstances the continuity requirements of the theorem will hold.

(a) Suppose $\hat{\varepsilon}_{-1} = g_1(\hat{x}_0) \neq \Delta/2$. Then $\lim_{\vec{x}_0 \to \hat{x}_0} g_k(\vec{x}_0) = g_k(\hat{x}_0)$ for $k = 1, \ldots, M$, $R_1, \ldots, R_N$, where g_{R_k} is the projection of $\mathbb{R}^N \times \mathcal{C}^M$ onto $\mathbb{R}$ such that $g_{R_k}(\vec{x}_n) = r_{n-k}$, $k = 1, \ldots, N$. The limit is defined to be the limit from the left on the circles $\mathcal{C}$ associated with the coordinates ε_{-k_i} of $\vec{x}_0$, and the general limit on the remaining circles $\mathcal{C}$ and $\mathbb{R}^N$.

These definitions of the limit are consistent with the requirements for the form of continuity stated in the theorem, and thus we have that the projections $g_k(\vec{x}_0)$ for $k = 1, \ldots, M$, $R_1, \ldots, R_N$ meet the continuity requirements at $\hat{x}_0$.

(b) Suppose that $\hat{\varepsilon}_{m-1} = g_1 \circ f^m(\hat{x}_0) \neq \Delta/2$ and that $g_k \circ f^m(\vec{x}_0)$ for $k = 1, \ldots, M$, $R_1, \ldots, R_N$ meet the continuity requirements at $\hat{x}_0$, for some $m \geq 0$. From the form of equations (1.2) it then it follows that $\varepsilon_m = g_1 \circ f^{m+1}(\vec{x}_0)$ and $r_m = g_{R_1} \circ f^{m+1}(\vec{x}_0)$ meet the continuity requirements at $\hat{x}_0$. (Note from part 2 that this is true even if $g_1 \circ f^{m+1}(\vec{x}_0) = \Delta/2$.)

By induction, we then have that the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$, $n \geq 0$, meets the continuity requirements at $\hat{x}_0$, under the special circumstance that $\hat{\varepsilon}_{n-1} = g_1 \circ f^n(\hat{x}_0) \neq \Delta/2$ for $n \geq 0$.

2. Now we consider the situation where $\hat{\varepsilon}_{n_i-1} = g_1 \circ f^{n_i}(\hat{x}_0) = \Delta/2$ for any arbitrary set of $n_i \geq 0$, and show how the required continuity is preserved for any specific such case.

Suppose that $\hat{\varepsilon}_{n_i-1} = g_1 \circ f^{n_i}(\hat{x}_0) = \Delta/2$ for some $n_i \geq 0$. Let $n_1 = \min\{n_i\}$. If $n_1 = 0$ then this will come from the initial condition $\hat{\varepsilon}_{-1} = g_1(\hat{x}_0) = \Delta/2$, and if $n_1 > 0$ this will arise from the second equation in (1.2) when the quantizer input $(x_{n_1-1} - \hat{r}_{n_1-1}) = q\Delta$, where $q \in \mathbb{Z}$. This then implies that $\lim_{\vec{x}_0 \to \hat{x}_0} g_1 \circ f^{n_1}(\vec{x}_0) = \pm\Delta/2$. If $n_1 = 0$, the $+\Delta/2$ corresponds to the limit from the left on the circle $\mathcal{C}$ corresponding to the coordinate $\hat{\varepsilon}_{-1}$ (left sided continuity on this $\mathcal{C}$), and the $-\Delta/2$ corresponds to the limit from the right on this $\mathcal{C}$ (right sided discontinuity on this $\mathcal{C}$). If $n_1 > 0$, the $+\Delta/2$ corresponds to the limit along paths in $R^N \times \mathcal{C}^M$ that cause $(x_{n_1-1} - r_{n_1-1})$ to approach $(x_{n_1-1} - \hat{r}_{n_1-1})$ from the left, and the $-\Delta/2$ corresponds to the limit along such paths that cause this approach from the right. The definition $\lim_{x \to -\Delta/2} x \equiv +\Delta/2$, for $x \in \mathcal{C}$, from the beginning of Chapter 2, means that the points $\pm\Delta/2$ are equivalent on $\mathcal{C}$. From this, and an application of part

1 above for $0 \leq n \leq n_1$, we have that $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ at least meets the continuity requirements at $\hat{x}_0$ for $n = n_1 - 1$. When continuing with the algebra in applying (1.2), we must not regard the points $\pm\Delta/2$ as equivalent, however. We now show by induction how the continuity requirements of the theorem hold if $\{n_i\} = n_1$.

(a) Taking the limit of the first line of (1.2) for $n = n_1$, and applying the results of part 1 and (R), leads to

$$\lim_{\vec{x}_0 \to \hat{x}_0} r_{n_1} = \lim_{\vec{x}_0 \to \hat{x}_0} g_{R_1} \circ f^{n_1+1}(\vec{x}_0) = \pm\frac{\Delta}{2}a_1 + \hat{r}_{0,n_1} = \pm\frac{\Delta}{2}\tilde{r}_1 + \hat{r}_{0,n_1},$$

where $\hat{r}_{0,n_1}$ is the value of this limit when ε_{n_1-1} only is taken as 0. From the second line of (1.2), we then have $\lim_{\vec{x}_0 \to \hat{x}_0} \varepsilon_{n_1} = \lim_{\vec{x}_0 \to \hat{x}_0} g_1 \circ f^{n_1+1}(\vec{x}_0) = \hat{\varepsilon}_{n_1}$. This limit yields the fixed value $\hat{\varepsilon}_{n_1}$ since $\tilde{r}_1 \in \mathbb{Z}$. Thus $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ meets the continuity requirements at $\hat{x}_0$ for $n = n_1$.

(b) Suppose that $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ meets the continuity requirements at $\hat{x}_0$ $\forall$ n satisfying $n_1 \leq n \leq n_1 + m$, for some $m \geq 0$. Suppose further that the limit of the first line of (1.2), for $n = n_1 + k$, $k = 0, \ldots, m$, gives

$$\lim_{\vec{x}_0 \to \hat{x}_0} r_{n_1+k} = \lim_{\vec{x}_0 \to \hat{x}_0} g_{R_1} \circ f^{n_1+k+1}(\vec{x}_0) = \pm\frac{\Delta}{2}\tilde{r}_{k+1} + \hat{r}_{0,(n_1+k)},$$

where $\hat{r}_{0,(n_1+k)} \equiv -\sum_{j=1}^{\min(N,k)} b_j \hat{r}_{0,(n_1+k-j)} + \hat{r}_{1,(n_1+k)}$ for $k = 1, \ldots, m$, and $\hat{r}_{1,(n_1+k)}$ is the value of the limit, excluding the terms with the coefficients a_{k+1}, b_j, for $j = 1, \ldots, \min(N, k)$. Now using these results, the results of part 1, and (R), and taking the limit of the first line of (1.2), for $n = n_1 + m + 1$, gives

$$\begin{aligned}
\lim_{\vec{x}_0 \to \hat{x}_0} r_{n_1+m+1} &= \lim_{\vec{x}_0 \to \hat{x}_0} g_{R_1} \circ f^{n_1+m+2}(\vec{x}_0) \\
&= \pm\frac{\Delta}{2} a_{m+2} \mp \frac{\Delta}{2} \sum_{j=1}^{\min(N,m+1)} b_j \tilde{r}_{m+2-j} - \sum_{j=1}^{\min(N,m+1)} b_j \hat{r}_{0,(n_1+m+1-j)} \\
&\quad + \hat{r}_{1,(n_1+m+1)} \\
&= \pm\frac{\Delta}{2} \tilde{r}_{m+2} + \hat{r}_{0,(n_1+m+1)}.
\end{aligned}$$

From the second line of (1.2), we then have $\lim_{\vec{x}_0 \to \hat{x}_0} \varepsilon_{n_1+m+1} = \lim_{\vec{x}_0 \to \hat{x}_0} g_1 \circ f^{n_1+m+2}(\vec{x}_0) = \hat{\varepsilon}_{n_1+m+1}$. This limit yields the fixed value $\hat{\varepsilon}_{n_1+m+1}$ since $\tilde{r}_{m+2} \in \mathbb{Z}$. Thus $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ meets the continuity requirements at $\hat{x}_0$ for $n = n_1 + m + 1$.

By induction, we then have that the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$, $n \geq 0$, meets the continuity requirements at $\hat{x}_0$, if n_1 is the only such n_i.

3. Now suppose that we assume that the continuity requirements of the theorem hold when the set $\{n_i\}$ has m elements. Consider the case when the set $\{n_i\}$ has $m+1$ elements. Let $n_2 = \max\{n_i\}$. Then, from the conclusions of part 2, the continuity requirements are met for $0 \leq n \leq n_2 - 1$. We may then proceed with the induction proof of part 2, with n_1 replaced by n_2; and identifying the new $\hat{r}_{0,(n_2+k)}$, for $k \geq 0$, as that part (the term) of $\lim_{\vec{x}_0 \to \hat{x}_0} r_{n_2+k}$ that does not propagate the effects of $\hat{\varepsilon}_{n_2-1} = \Delta/2$ and hence, from the assumption that the continuity requirements hold when $\{n_i\}$ has m elements, does not contribute by itself to any discontinuity in ε_n, for $n \geq 0$. This induction will then show analogously that the continuity requirements hold when the set has $m + 1$ elements.

Thus, by the overall induction of parts 2 and 3, we have that the continuity requirements of the theorem hold when $\hat{\varepsilon}_{n_i-1} = g_1 \circ f^{n_i}(\hat{x}_0) = \Delta/2$, for any arbitrary set of $n_i \geq 0$. Therefore we conclude that the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$, $n \geq 0$, meets the continuity requirements of the theorem at $\hat{x}_0$. ■

This theorem guarantees continuity when mapping n times for all $n \geq 1$ from the domain $\mathbb{R}^N \times \mathcal{C}^M$ onto the first error variable in the mapping in $\mathcal{C}$, provided the "boundary points" $\Delta/2$ on the $M-1$ circles of $\mathcal{C}^{M-1}$ from the domain are excluded. In a system with a generally discontinuous quantizer element $Q(x)$ as in the Σ-Δ one here, being able to assert such continuity obviously provides major simplification in the system's structure and for its analysis. That we are able to obtain this follows directly from the existence of the integer $\tilde{r}_k$ in (R), which uses the quantizer's "many-to-one" nature to "cancel out" the effects of the discontinuities it creates.

A simple way to satisfy the requirement (R) is simply to require that $a_i,\ b_j \in \mathbb{Z}$, for $i = 1,\ 2, \ldots,\ M$; $j = 1,\ 2, \ldots,\ N$; that is, that the feedback and feedforward gains in the filter must all be integers. In this situation the continuity in Theorem 4.4 in fact holds over all $\mathbb{R}^N \times \mathcal{C}^M$. The problem with this is that if all the b_j are integers (not all zero) then, from the properties of polynomials with integer coefficients, we have that the polynomial $p_r(z) = z^N + \sum_{j=1}^{N} b_j z^{N-j}$ cannot have any zeros of magnitude less than 1 if it has no zeros of magnitude greater than 1. From the results of Chapter 3, this implies a situation where the system is generally unstable except under special circumstances (i.e. depending on the input x_n or initial conditions). To have meaningful results and hence be worthy of special consideration when $N \geq 1$, we need our simplified systems to at least be stable over general initial conditions and inputs. Therefore it is sensible to consider the general condition (R) as presented and the broader range of systems it allows, including in particular those that are stable.

One simple way to satisfy the requirement (R) would be to require $\tilde{r}_1 \in \mathbb{Z}$, $\tilde{r}_k = 0$, for $k = 2, \ldots, \max(N+1, M)$. This implies the rather constrained requirement $a_1 \in \mathbb{Z}$, $b_{k-1} = a_k/a_1$, for $k = 2, \ldots, M$, with $N = M - 1$ holding. Under this requirement, stable systems are clearly possible by choosing a_1 large in magnitude relative to a_k, $k = 2, \ldots, M$,

so as to make the b_j small in magnitude to yield zeros of the $p(z)$ above that are all inside the unit circle. A generalization of this is given by the following scheme:

Assume $M > N$. Choose the b_k so that $p(z)$ has zeros all inside the unit circle, so that the system is stable. Choose any a_k to make $\tilde{r}_k \in \mathbb{Z}$ for $k = 1, \ldots, M-N$. Then choose $a_k = \sum_{j=1}^{N} b_j \tilde{r}_{k-j}$ so that $\tilde{r}_k = 0$, for $k = M-N+1, \ldots, M$. Since the remaining $\tilde{r}_k$, $k > M$, arise from the Nth order difference equation in b_j, they will all be zero as well.

With this scheme, one has some flexibility to choose the a_i and b_j, $(i = 1, \ldots, M-N)$, so as to position the zeros of the noise transfer function (1.3) in particular regions on the complex plane, as will be of interest later in this thesis. More theoretically, a further generalization of this scheme exists when some of the zeros of $p_r(z)$ are integer, and hence we are not concerned with stability.

In summary, using condition (R) rather than an integer requirement on all the a_i and b_j coefficients provides potentially for a reduction in the number of conditions for continuity from $N + M$ to $\max(N, M)$. The fact that the number of $\tilde{r}_k$ iterates required in (R) to be integer may be arbitrarily large seriously erodes this advantage in general, however. Significant additional constraints arise when requiring stability as well, although the flexibility for choosing the poles and zeros of the noise transfer function $(1 - H)$ is improved considerably over the all integer-coefficient filter case.

When the Σ-Δ modulator exhibits continuity over $\mathbb{R}^N \times \mathcal{C}^M$ or any subset thereof, we arrive at extensions of Proposition 4.1 as follows:

Theorem 4.5 *Suppose the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ is continuous on $\mathcal{C}$ at any $\vec{x}_0 \in U$, where U is a set and $U \subset \mathbb{R}^N \times \mathcal{C}^M$, for any $n \geq 0$. Then for any $\vec{x}_{1,0} \in U$ and $\vec{x}_{2,0} \in \tilde{N}^*$, where $\tilde{N}^* = \tilde{N} \cap U$ and $\tilde{N}$ is a small neighbourhood about $\vec{x}_{10}$, system (4.1)*

$$\Delta\tilde{\varepsilon}_n = \sum_{k=1}^{\max(N,M)} (a_k - b_k)\Delta\tilde{\varepsilon}_{n-k} \quad \textit{with initial conditions}$$

$\Delta\tilde{\varepsilon}_n = g_M(\vec{x}_{2,(n+M)}) - g_M(\vec{x}_{1,(n+M)}) = \Delta\varepsilon_n$, $n = N-1, \ldots, N-\max(N,M)^*$, *has solutions that satisfy* $\Delta\tilde{\varepsilon}_n = \Delta\varepsilon_n + m_n\Delta \cong g_1(\vec{x}_{2,(n+1)}) - g_1(\vec{x}_{1,(n+1)}) = \Delta\varepsilon_n$, *for some* $m_n \in \mathbb{Z}$, *and* $n \geq N$.

Proof:

To start, we define the line segment joining $\vec{x}_{1,0}$ and $\vec{x}_{2,0}$ by $L = \{\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M \mid \vec{x}_0 = \vec{x}_{1,0} + \alpha(\vec{x}_{2,0} - \vec{x}_{1,0}) \text{ for } 0 \leq \alpha \leq 1\} \in \tilde{N}^* \subset U$. Define a point $\vec{z}_i \in L$ by $\vec{z}_i = \vec{x}_{1,0} + \alpha_i(\vec{x}_{2,0} - \vec{x}_{1,0})$, for $0 < \alpha_i < 1$. Define an open interval in L between $\vec{z}_i$ and $\vec{z}_j$ by $(\vec{z}_i, \vec{z}_j) = \{\vec{z}_k \mid \alpha_i < \alpha_k < \alpha_j\}$. Let $r_{*,n}$, $\varepsilon_{*,n}$ and $\varepsilon_{x,*,n}$ correspond to the system with initial condition of the form $\vec{z}_*$ and $\vec{x}_{*,0}$ respectively.

Choose any $n_1 \geq N$. Now consider an open interval $(\vec{z}_a, \vec{z}_b) \in L$. Suppose $\exists\ n_2$, with $0 \leq n_2 \leq n_1$, such that $Q(x_n - r_{a,n}) - Q(x_n - r_{b,n}) \neq 0$ for $n = n_2$, and $Q(x_n - r_{c,n}) - Q(x_n - r_{d,n}) = 0$ for $0 \leq n < n_2$ if $n_2 > 0$, $\forall\ \alpha_c$, α_d such that $\alpha_a < \alpha_c < \alpha_d < \alpha_b$. From the continuity of $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ on U and hence L for $n \geq 0$, we must have that r_n is continuous (mod Δ) and thus piecewise linear on L for $n \geq 0$. Thus there exists a finite number of points $\vec{z}_i$ in $(\vec{z}_a, \vec{z}_b)$ at which $x_{n_2} - r_{i,n_2} = q\Delta$, where $q \in \mathbb{Z}$. This implies that for these points $\vec{z}_i$, we have

$$\begin{aligned} Q(x_n - r_{c,n}) - Q(x_n - r_{1,a,n}) &= Q(x_n - r_{k,b,n}) - Q(x_n - r_{(k+1),a,n}) \\ &= Q(x_n - r_{K,b,n}) - Q(x_n - r_{d,n}) \\ &= 0, \qquad \text{for } 0 \leq n \leq n_2, \end{aligned}$$

$\forall\ \alpha_c,\ \alpha_d,\ \alpha_{i,a},\ \alpha_{i,b}$ such that $\alpha_a < \alpha_c < \alpha_{1a} < \alpha_1$, $\alpha_k < \alpha_{k,b} < \alpha_{(k+1),a} < \alpha_{k+1}$, $\alpha_K < \alpha_{K,b} < \alpha_d < \alpha_b$, where $k = 1, \ldots, K-1$, and K is the number of such points.

Now we apply this result iteratively (successively) with successive n_2 values from 0 through to n_1 for the open interval $(\vec{x}_{1,0}, \vec{x}_{2,0})$ and possible successive respective subintervals, to obtain the result that there exists a finite number of open intervals $(\vec{x}_{1,0}, \vec{z}_1)$, $(\vec{z}_k, \vec{z}_{k+1})$, $(\vec{z}_{K_1}, \vec{x}_{2,0})$ with points $\vec{z}_i$ (satisfying $x_{n_2} - r_{i,n_2} = q\Delta$, where $q \in \mathbb{Z}$), such that

$$\begin{aligned} Q(x_n - r_{c,n}) - Q(x_n - r_{1,a,n}) &= Q(x_n - r_{k,b,n}) - Q(x_n - r_{(k+1),a,n}) \\ &= Q(x_n - r_{K_1,b,n}) - Q(x_n - r_{d,n}) \\ &= 0, \qquad \text{for } 0 \le n \le n_1, \end{aligned}$$

$\forall\ \alpha_c,\ \alpha_d,\ \alpha_{i,a},\ \alpha_{i,b}$ such that $0 < \alpha_c < \alpha_{1,a} < \alpha_1,\ \alpha_k < \alpha_{k,b} < \alpha_{(k+1),a} < \alpha_{k+1},\ \alpha_{K_1} < \alpha_{K_1,b} < \alpha_d < 1,$ where $k = 1, \ldots, K_1 - 1$, and K_1 is the number of such points.

From the continuity of $\varepsilon_{x,n} = g_1 \circ f^{n+1}(\vec{x}_0)$ on U it follows that

$$\begin{aligned} \Delta\varepsilon_{(x,2),(x,1),n_1} &= \varepsilon_{x,2,n_1} - \varepsilon_{x,1,n_1} = g_1 \circ f^{n_1+1}(\vec{x}_{2,0}) - g_1 \circ f^{n_1+1}(\vec{x}_{1,0}) \\ &= g_1 \circ f^{n_1+1}(\vec{x}_{2,0}) - g_1 \circ f^{n_1+1}(\vec{x}_0)\ |_{\alpha=\alpha_{K_1}} \\ &\quad + \sum_{k=1}^{K_1-1} [g_1 \circ f^{n_1+1}(\vec{x}_0)|_{\alpha=\alpha_{k+1}} - g_1 \circ f^{n_1+1}(\vec{x}_0)|_{\alpha=\alpha_k}] \\ &\quad + g_1 \circ f^{n_1+1}(\vec{x}_0)|_{\alpha=\alpha_1} - g_1 \circ f^{n_1+1}(\vec{x}_{1,0}) \\ &= \lim_{\alpha\to 1}^{l} g_1 \circ f^{n_1+1}(\vec{x}_0) - \lim_{\alpha\to\alpha_{K_1}}^{u} g_1 \circ f^{n_1+1}(\vec{x}_0) - m_{K_1,u,n_1}\Delta \\ &\quad + \sum_{k=1}^{K_1-1} [\lim_{\alpha\to\alpha_{k+1}}^{l} g_1 \circ f^{n_1+1}(\vec{x}_0) + m_{k+1,l,n_1}\Delta - \lim_{\alpha\to\alpha_k}^{u} g_1 \circ f^{n_1+1}(\vec{x}_0) - m_{k,u,n_1}\Delta] \\ &\quad + \lim_{\alpha\to\alpha_1}^{l} g_1 \circ f^{n_1+1}(\vec{x}_0) + m_{1,l,n_1}\Delta - \lim_{\alpha\to 0}^{u} g_1 \circ f^{n_1+1}(\vec{x}_0), \end{aligned}$$

where the limits on the right of the final equality are taken along the line L, and they are from the right (u-upper) or the left (l-lower) for α as indicated. With the continuity on $\mathcal{C}$, the function limits as defined here will differ from their values at the point only if the value approaches $-\Delta/2$ in the limit, so that the difference (from the corresponding value $\Delta/2$ at the point) is Δ. Thus the values of m_{*,u,n_1}, m_{*,l,n_1} above are 1 if the corresponding respective limits are $-\Delta/2$, and zero otherwise. Here we denote $\varepsilon_{*_A,n} - \varepsilon_{*_B,n}$ by $\Delta\varepsilon_{(*_A),(*_B),n}$.

Proposition 4.1 holds on each interval $(\vec{x}_{1,0}, \vec{z}_1)$, $(\vec{z}_k, \vec{z}_{k+1})$, $(\vec{z}_{K_1}, \vec{x}_{2,0})$, for $k = 1, \ldots, K_1 - 1$, of L. We then apply Proposition 4.1 to each limit difference term above to get the following:

$$\begin{aligned}
\Delta\varepsilon_{(x,2),(x,1),n_1} &= \lim_{\alpha\to\alpha_1}{}^{l}\, \Delta\tilde{\varepsilon}_{1,(x,1),n_1} + \sum_{k=1}^{K_1-1}\Big[\lim_{\alpha\to(\alpha_{k+1}-\alpha_k)}{}^{l}\, \Delta\tilde{\varepsilon}_{(k+1),k,n_1}\Big] \\
&\quad + \lim_{\alpha\to 1-\alpha_K}{}^{l}\, \Delta\tilde{\varepsilon}_{(x,2),K_1,n_1} - \sum_{k=1}^{K_1}(m_{k,u,n_1} - m_{k,l,n_1})\Delta \\
&= \Delta\tilde{\varepsilon}_{1,(x,1),n_1} + \sum_{k=1}^{K_1-1}\Delta\tilde{\varepsilon}_{(k+1),k,n_1} + \Delta\tilde{\varepsilon}_{(x,2),K_1,n_1} - m_{n_1}\Delta \\
&= \Delta\tilde{\varepsilon}_{(x,2),(x,1),n_1} - m_{n_1}\Delta.
\end{aligned}$$

We define $m_{n_1} = \sum_{k=1}^{K_1}(m_{k,u,n_1} - m_{k,l,n_1}) \in \mathbb{Z}$ here. The $\Delta\tilde{\varepsilon}_{*,n_1}$ terms on the right in the final equality above are solutions to (4.1) with initial conditions in $\mathbb{R}^N \times \mathcal{C}^M$ given by $\alpha_1(\vec{x}_{2,0} - \vec{x}_{1,0})$, $(\alpha_{k+1} - \alpha_k)(\vec{x}_{2,0} - \vec{x}_{1,0})$ and $(1 - \alpha_{K_1})(\vec{x}_{2,0} - \vec{x}_{1,0})$, for $k = 1, \ldots, K_1 - 1$ respectively. Since (4.1) is a linear difference equation, it holds that the sum of any two solutions is itself a solution with initial conditions equal to the sum of the initial conditions of the two solutions. Applying this to the above, we have that $\Delta\tilde{\varepsilon}_{(x,2),(x,1),n_1}$ is a solution for $n = n_1$ to (4.1), with initial conditions equal to

$$\begin{aligned}
\Delta\tilde{\varepsilon}_{(x,2),(x,1),n} &= \alpha_1[g_M(\vec{x}_{2,(n+M)}) - g_M(\vec{x}_{1,(n+M)})] \\
&\quad + \sum_{k=1}^{K_1-1}[(\alpha_{k+1} - \alpha_k)(g_M(\vec{x}_{2,(n+M)}) - g_M(\vec{x}_{1,(n+M)})] \\
&\quad + (1 - \alpha_{K_1})[g_M(\vec{x}_{2,(n+M)}) - g_M(\vec{x}_{1,(n+M)})] \\
&= [g_M(\vec{x}_{2,(n+M)}) - g_M(\vec{x}_{1,(n+M)})],
\end{aligned}$$

for $n = N - 1, \ldots, N - \max(N, M)$. These are the form of the initial conditions specified for (4.1) in the theorem. Since the solution to (4.1) with these initial conditions must be unique, it follows that $\Delta\tilde{\varepsilon}_{(x,2),(x,1),n}$ satisfies (4.1) for $N \le n \le n_1$. Since $n_1 \ge N$ was chosen arbitrarily, it thus follows that system (4.1) has solutions that satisfy $\Delta\tilde{\varepsilon}_{(x,2),(x,1),n} = \Delta\varepsilon_{(x,2),(x,1),n} + m_n\Delta \cong g_1(\vec{x}_{1,(n+1)}) - g_1(\vec{x}_{2,(n+1)})$, $n \ge 0$, as required in the theorem. ■

Corollary 4.6 *Suppose the function* $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ *meets the continuity results of Theorem 4.4 over a set* U*, where* $U \subset \mathbb{R}^N \times \mathcal{C}^M$*, for any* $n \geq 0$*. Then for any* $\vec{x}_{1,0} \in U$ *with* $\varepsilon_{-k_i} = g_{k_i}(\vec{x}_{1,0}) = \Delta/2$ *for some* $k_i \in \{2, \ldots, M\}$*, there exists an* $\vec{x}_{2,0} \in \tilde{N}^*$*, where* $\tilde{N}^* = \tilde{N} \cap U$ *and* $\tilde{N}$ *is a small neighbourhood about* $\vec{x}_{1,0}$*, such that system (4.1) of Theorem 4.5 with its initial conditions as specified has solutions that satisfy* $\Delta\tilde{\varepsilon}_n = \Delta\varepsilon_n + m_n\Delta \cong g_1(\vec{x}_{2,(n+1)}) - g_1(\vec{x}_{1,(n+1)})$*, for some* $m_n \in \mathbb{Z}$*, and* $n \geq N$*.*

Proof:

Clearly $\exists$ an $\vec{x}_{2,0} \in \tilde{N}^*$ (in a particular sector of $\tilde{N}^*$) such that, with the line segment joining $\vec{x}_{1,0}$ and $\vec{x}_{2,0}$ defined by $L = \{\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M \mid \vec{x}_0 = \vec{x}_{1,0} + \alpha(\vec{x}_{2,0} - \vec{x}_{1,0})$ for $0 \leq \alpha \leq 1\} \subset \tilde{N}^* \subset U$, we have that $g_{k_i}(\vec{x}_{1,0} + \alpha(\vec{x}_{2,0} - \vec{x}_{1,0}))$ approaches $g_{k_i}(\vec{x}_{1,0})$ from the left as α goes to zero (from the right), for all the k_i. We now proceed with the steps of the proof of Theorem 4.5, where the existence of the continuity requirements of Theorem 4.4 allow all the terms involving limits to remain valid statements in the expressions. Thus the results of this corollary are satisfied. ■

Proposition 4.7 *The* $\Delta\tilde{\varepsilon}_n$*,* $n \geq N - \max(N, M)^*$*, from Theorem 4.5 and Corollary 4.6 satisfy*

1. $\operatorname{sgn}(\Delta\tilde{\varepsilon}_n)(|\Delta\tilde{\varepsilon}_n| \bmod \Delta) = g_1(\vec{x}_{2,(n+1)}) - g_1(\vec{x}_{1,(n+1)}) + \operatorname{sgn}(\Delta\tilde{\varepsilon}_n) I_n(g_1(\vec{x}_{1,(n+1)}),\, g_1(\vec{x}_{2,(n+1)}))\Delta,$

where I_n*,* $n \geq 0$*, is defined such that, for any* $a, b \in \mathcal{C}$*,* $I_n(a, b) = 0$ *if* $(b-a)[\Delta\tilde{\varepsilon}_n - (b-a)] \geq 0$*, and* $I_n(a, b) = 1$ *otherwise;*

2. $(\Delta\tilde{\varepsilon}_n) \bmod \Delta = \|g_1(\vec{x}_{2,(n+1)}) - g_1(\vec{x}_{1,(n+1)})\|$*, when* $(\Delta\tilde{\varepsilon}_n) \bmod \Delta \leq \Delta/2$*;*
3. $(\Delta\tilde{\varepsilon}_n) \bmod \Delta = \Delta - \|g_1(\vec{x}_{2,(n+1)}) - g_1(\vec{x}_{1,(n+1)})\|$*, when* $\Delta/2 < (\Delta\tilde{\varepsilon}_n) \bmod \Delta$*.*

Proof:

1. From Theorem 4.5, we have that $\Delta\tilde{\varepsilon}_n = \Delta\varepsilon_n + m_n\Delta$, and $\Delta\varepsilon_n = g_1(\vec{x}_{2,(n+1)}) - g_1(\vec{x}_{1,(n+1)})$, $n \geq N - \max(N, M)^*$, for some $m_n \in \mathbb{Z}$, with $m_n = 0$ for $n = N - 1, \ldots, N - \max(N, M)^*$. Thus $I_n(g_1(\vec{x}_{1,(n+1)}),\, g_1(\vec{x}_{2,(n+1)})) = 0$ if $(\Delta\varepsilon_n)[m_n\Delta] \geq 0$, and $I_n(g_1(\vec{x}_{1,(n+1)}),\, g_1(\vec{x}_{2,(n+1)})) = 1$ otherwise, for $n \geq N - \max(N, M)^*$.

First suppose, for a given n, that $I_n = 0$. This implies that

$$\begin{aligned} \text{LHS} &= \text{sgn}(\Delta\tilde{\varepsilon}_n)(|\Delta\tilde{\varepsilon}_n| \bmod \Delta) = \text{sgn}(\Delta\varepsilon_n)[(|\Delta\varepsilon_n| + |m_n|\Delta) \bmod \Delta] \\ &= \text{sgn}(\Delta\varepsilon_n)|\Delta\varepsilon_n| = \Delta\varepsilon_n = \text{RHS}. \end{aligned}$$

Now suppose that $I_n = 1$, with $\text{sgn}(\Delta\varepsilon_n) = -\text{sgn}(\Delta\tilde{\varepsilon}_n) = \pm$. Then we have

$$\begin{aligned} \text{LHS} &= \text{sgn}(\Delta\tilde{\varepsilon}_n)(|\Delta\tilde{\varepsilon}_n| \bmod \Delta) = -\text{sgn}(\Delta\varepsilon_n)[(\mp\Delta\varepsilon_n \mp m_n\Delta) \bmod \Delta] \\ &= (\mp)[\Delta \mp \Delta\varepsilon_n] = \Delta\varepsilon_n \mp \Delta = \Delta\varepsilon_n + \text{sgn}(\Delta\tilde{\varepsilon}_n)\Delta = \text{RHS}. \end{aligned}$$

2. Using the results from the beginning of the proof of part 1 above, we have the following: Suppose that $g_1(\vec{x}_{2,(n+1)}) - g_1(\vec{x}_{1,(n+1)}) = \Delta\varepsilon_n \geq 0$. Then LHS $= (\Delta\varepsilon_n + m_n) \bmod \Delta = \Delta\varepsilon_n$. If $\Delta\varepsilon_n \leq \Delta/2$, then this magnitude represents the shortest distance on the circle $\mathcal{C}$ between $\vec{x}_{1(n+1)}$ and $\vec{x}_{2(n+1)}$. Thus LHS = RHS. Now suppose that $\Delta\varepsilon_n < 0$. Then LHS $= \Delta + \Delta\varepsilon_n$. If $\Delta + \Delta\varepsilon_n \leq \Delta/2$, then this magnitude again represents the shortest distance on $\mathcal{C}$ between the two points, since $|\Delta\varepsilon_n|$ would be the longer distance. Thus LHS = RHS.

3. Suppose that $\Delta\varepsilon_n > 0$ so that LHS $= \Delta\varepsilon_n$. If $\Delta\varepsilon_n > \Delta/2$, then the magnitude $\Delta - \Delta\varepsilon_n$ represents the shortest distance on the circle $\mathcal{C}$ between $\vec{x}_{1(n+1)}$ and $\vec{x}_{2(n+1)}$. Thus LHS $= \Delta - (\Delta - \Delta\varepsilon_n) =$ RHS. Now suppose that $\Delta\varepsilon_n < 0$ so that LHS $= \Delta + \Delta\varepsilon_n$. If $\Delta + \Delta\varepsilon_n > \Delta/2$, then the magnitude representing the shortest distance on $\mathcal{C}$ between

the two points is given by $-\Delta\varepsilon_n > 0$, since $\Delta + \Delta\varepsilon_n$ represents the longer distance. Thus LHS $= \Delta - (-\Delta\varepsilon_n) =$ RHS. ■

These results show that with the continuity provided by condition (R) on the coefficients in the filter, the error differences arising from the two initializations may remain directly related (i.e. congruent) to the $\Delta\varepsilon_n$ described by the difference equation in Proposition 4.1, for all $n \geq 0$. For these results and related applications, we describe two quantities as being congruent with the symbol "$\cong$", if and only if they differ in value by an additive integer multiple of Δ. Essentially, the magnitude and sign of $\Delta\tilde{\varepsilon}_n$ represents the length and sign of the error difference interval on $\mathcal{C}$. If, for some n the magnitude of $\Delta\tilde{\varepsilon}_n$ exceeds Δ, lying say between $p\Delta$ and $(p+1)\Delta$ with $p \in \mathbb{Z}$ in a given $(+/-)$ direction, then the error difference interval is interpreted to have been wrapped p times around the circle $\mathcal{C}$ in the same direction, with $(\Delta\tilde{\varepsilon}_n \bmod \Delta)$ giving a net interval distance between error variables. Proposition 4.7 summarizes this relationship, and allows for a straightforward extension of the analysis of the chaos conditions, and constructing theorem proofs initiated by Proposition 4.1, without having to worry about the approach breaking down for any value of n. Some aspects of this may be more apparent when considering the mapping of intervals as will occur subsequently in the analysis.

It is easy to see that there always exists an initial condition difference $\Delta\vec{x}_0$ (where $\Delta\vec{x}_0 = \vec{x}_{2,0} - \vec{x}_{1,0}$ for some applicable $\vec{x}_{1,0}$ and $\vec{x}_{2,0}$) that will give rise to any particular set of initial conditions $\Delta\varepsilon_n$, $n = N-1, \ldots, N - \max(N, M)^*$, for (4.1). For example, we may simply set $\Delta r_i = \Delta\varepsilon_i$ for $i = -1, \ldots, -N$, set $\Delta\tilde{\varepsilon}_n$ equal to the required $\Delta\varepsilon_{N+n}$ for $n = -1, \ldots, -\max(N, M)^*$. When $\Delta\tilde{\varepsilon}_n = \Delta\varepsilon_n$ over some range of n, then we equivalently shift the subscripts of $\Delta\tilde{\varepsilon}_n$ down by a factor of N in (4.1) and its initial conditions. Such a system (4.1) takes the form of the difference equation (2.2) (with no input) and thus has

a solution given by (2.3). Without loss of generality, the initial $\Delta\varepsilon_i$, Δr_i that form the $\max(N,M)^*$ dimensional space of initial conditions $\Delta\vec{\varepsilon}_0 = (\Delta\varepsilon_{-1}, \ldots, \varepsilon_{-\max(N,M)^*})$ may be specified in this manner. In future treatment, we will generally drop the tildes from $\Delta\tilde{\varepsilon}_n$, and let $\Delta\varepsilon_n$ denote the solutions of (4.1), $n \geq 0$, as distinct from the value of $\varepsilon_{2,n} - \varepsilon_{1,n}$, unless otherwise specified.

To emphasize how condition (R) brings the simplifying circle map, and the properties mentioned after Proposition 4.7 into play, we describe the solution for ε_n with the following:

Proposition 4.8 *Suppose the a_i and b_j satisfy condition (R). Then we have that*

$$\varepsilon_n = Q(\tilde{w}_n) - \tilde{w}_n, \quad n \geq 0, \tag{4.2}$$

where $\tilde{w}_n$ is the solution to the difference equation

$$\tilde{w}_n = \sum_{k=1}^{\max(N,M)^*} (a_k - b_k)\tilde{w}_{n-k} + \sum_{j=0}^{N} b_j x_{n-j}, \quad n \geq \max(N,M)^*,$$

with initial conditions given by $\tilde{w}_n = x_n - r_n + \sum_{k=0}^{n-1} \tilde{r}_{n-k}\tilde{Q}_k$, *where* $\tilde{Q}_n = Q(x_n - r_n) + \sum_{k=0}^{n-1} \tilde{r}_{n-k}\tilde{Q}_k$, *for* $1 \leq n \leq \max(N,M)^* - 1$, *and* $\tilde{w}_0 = x_0 - r_0$, $\tilde{Q}_0 = Q(\tilde{w}_0)$, $b_0 \equiv 1$.

Proof:

We extend the definitions of $\tilde{w}_n$ and $\tilde{Q}_n$ in the proposition to all $n \geq 0$. Now we use these transformed variables, along with $\varepsilon_n = Q(x_n - r_n) - (x_n - r_n)$, to eliminate r_n and ε_n in the difference equation of (1.2), when $n \geq 0$. We find that LHS equals RHS and, in particular, the difference equation of (1.2) reduces to that of (4.2) in the proposition when $n \geq \max(N,M)^*$. The quantizer equation of (1.2) then becomes $\varepsilon_n = Q(\tilde{w}_n - \sum_{k=0}^{n-1} \tilde{r}_{n-k}\tilde{Q}_k) - (\tilde{w}_n - \sum_{k=0}^{n-1} \tilde{r}_{n-k}\tilde{Q}_k)$, $n \geq 0$. Condition (R) holds, and hence the $\tilde{r}_k$

are all integers. The $\tilde{Q}_k$ are thus all integer multiples of $\Delta/2$. The summation terms thus drop out of the quantizer equation, which hence reduces to that of (4.2). ■

When (R) holds, the behaviour of the error $\varepsilon_n = g \circ f^{n+1}(\vec{x}_0)$ is explicitly governed by the solution of the $\max(N, M)^*$th order difference equation of (4.2) with an input driven by the system input x_n. ε_n is simply the value of the solution wrapped around the circle $\mathcal{C}$. When (R) does not hold, the extra convolution term $\sum_{i=0}^{n-1} \tilde{r}_{n-i}\tilde{Q}_i$ appears in the input of the quantizer equation (4.2), and the solution lacks a simple interpretation. An alternative formulation of the most general form, with dither added, is given by (6.2) in Chapter 6.

We close out this chapter with some items that will be utilized primarily in the approach and analysis of Chapter 5. The following more technical results will be of use in proving subsequent theorems:

Definition 4.9 *For a point $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$, define the set $\tilde{N}_x$ by $\tilde{N}_x = \tilde{N} \cap \tilde{M}$, where $\tilde{N}$ is some small neighbourhood about $\vec{x}_0$ and $\tilde{M} = \{\vec{y}_0 \in \mathbb{R}^N \times \mathcal{C}^M \mid g_{k_i}(\vec{x}_0 + \alpha(\vec{y}_0 - \vec{x}_0))$ approaches $g_{k_i}(\vec{x}_0)$ from the left as α goes to zero (from the right), for all the $k_i\}$, where $\varepsilon_{-k_i} = g_{k_i}(\vec{x}_0) = \Delta/2$ for some $k_i \in \{2, \ldots, M\}$.*

Lemma 4.10 *Suppose the a_i and b_j satisfy condition (R). Then we have the following:*

1. *For any $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and any $\vec{y}_0 \in \tilde{N}_x$, the error differences $\Delta\varepsilon_n$, $n \geq N$, arising from the initial conditions $\Delta\vec{x}_0 = \vec{y}_0 - \vec{x}_0$ will be solutions to (4.1) given by (2.3).*

2. *Furthermore, if $p(z)$ has zeros with magnitude greater than 1, or magnitude 1 with multiplicity greater than 1, then there exists $\vec{y}_0 \in \tilde{N}_x$ such that the constants A_{ij}, B_{ij}, C_{ij}, associated with the terms on the right side of (2.3) corresponding to these zeros, are not all zero.*

Proof:

1. We apply Theorem 4.4 followed by Theorem 4.5 and Corollary 4.6. In the case of 4.6, from the proof of this corollary, it holds that the $\vec{x}_{20}$ is satisfied by any $\vec{y}_0 \in \tilde{N}_x$.

2. The range of initial conditions $\Delta\vec{\varepsilon}_0 = \{\Delta\varepsilon_n \text{ for } n = N-1, \ldots, N - \max(N,M)^*\}$ of (4.1) that lead to the constants A_{ij}, B_{ij}, C_{ij} (corresponding to the designated zeros in the lemma) all equalling zero constitutes a $\tilde{q}$ dimensional subspace of initial conditions $C^{\tilde{q}}$, where $\tilde{q}$ is the number of zeros of $p(z)$ with magnitude less than 1 or equal to 1 with multiplicity 1, and $0 \leq \tilde{q} < \max(N,M)^*$. Thus there exists a $\max(N,M)^*$ dimensional space of initial conditions, excluding the set contained in the $\tilde{q}$ dimensional subspace, $\mathcal{C}^{\max(N,M)^*} - C^{\tilde{q}}$, over which not all of these constants are zero. Clearly, from its definition, $\tilde{N}_x$ is a $\mathbb{R}^N \times \mathcal{C}^M$ dimensional sector of $\tilde{N}$. There exists a $\max(N,M)^*$ dimensional subsector yielding initial conditions $\Delta\vec{\varepsilon}_0$ of (4.1) in a $\max(N,M)^*$ sector of $\mathcal{C}^{\max(N,M)^*}$. Thus there must exist $\vec{y}_0 \in \tilde{N}_x$ such that the initial condition difference $\Delta\vec{x}_0 = \vec{x}_0 - \vec{y}_0$ will give rise to such an initial condition $\Delta\vec{\varepsilon}_0$ of (4.1) in a $\max(N,M)$ dimensional sector of $\mathcal{C}^{\max(N,M)^*} - C^{\tilde{q}}$, (i.e. not in the $\tilde{q}$ dimensional subspace $C^{\tilde{q}}$). ■

Matrix Notation in Linear Analysis:

When analyzing system (2.1) from a multidimensional state space point of view, and applying (4.1) or Theorem 4.5 or its corollary in a linear analysis approach, the following matrix notation is useful for the analytic description and work.

We consider (2.3) as the solution to (4.1). Now we define $[R_k]$, $k \geq 0$, to be the $\max(N,M)^* \times \max(N,M)^*$ matrix, such that $[R_k]$ has entry (i,j) corresponding to the term of the RHS of (2.3) associated with zero μ_j of the $\max(N,M)^*$ ordered zeros of $p(z)$, with $n = k - i$, and with the arbitrary constant set to 1. We define $\vec{\alpha} = (A_{ij}, B_{ij}, C_{ij})^T$ to

be the $\max(N,M)^* \times 1$ vector of arbitrary constants, with the same ordering as that of the corresponding terms in each row of $[R_k]$ (i.e. the ordering of the columns corresponding to each zero μ_j). We define $(\Delta\varepsilon_i)^T$ to be the $\max(N,M)^* \times 1$ vector of initial conditions $\Delta\vec{\varepsilon}_0 = (\Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-\max(N,M)})$ of (4.1). Note that the subscripts on the $\Delta\varepsilon_i$ decrease from top to bottom of the vector (i.e. with increasing i). Thus we have the relations $(\Delta\varepsilon_i)^T = (\Delta\vec{\varepsilon}_0)^T = [R_0]\vec{\alpha}$ and $(\Delta\vec{\varepsilon}_k)^T = (\Delta\varepsilon_{k-1}, \ldots, \Delta\varepsilon_{k-\max(N,M)}) = [R_k]\vec{\alpha}$ for $k \geq 0$.

Now suppose we partition the $\max(N,M)^*$ zeros of $p(z)$ into a group of q zeros (e.g. by magnitude), and a remaining group of $\tilde{q} = \max(N,M)^* - q$ zeros, with $0 \leq q \leq \max(N,M)^*$. We define a decomposition of $[R_k]$ into blocks, where $[R_{kq}]$ and $[R_{k\tilde{q}}]$ denote the first q and last $\tilde{q}$ rows respectively, of the q columns of $[R_k]$ corresponding to the group of q zeros identified above (upper and lower left blocks). Similarly, we define $\vec{\alpha}_q$, $(\Delta\varepsilon_i)_q^T$ and $\vec{\alpha}_{\tilde{q}}$, $(\Delta\varepsilon_i)_{\tilde{q}}^T$ to be the corresponding vectors of the first q and last $\tilde{q}$ entries of $\vec{\alpha}$ and $(\Delta\varepsilon_i)^T$ respectively. Thus we have $(\Delta\varepsilon_i)_q^T = [R_{0q}]\vec{\alpha}_q$ and $(\Delta\varepsilon_i)_{\tilde{q}}^T = [R_{0\tilde{q}}]\vec{\alpha}_q$ when $\vec{\alpha}_{\tilde{q}} = 0$.

Now we briefly introduce the idea of a lower state space dimensional form of (4.1) as may be obtained by modifying the filter coefficients a_i, b_j.

Proposition 4.11 *A given system of the form (4.1) with coefficients a_i, b_j, and $p(z)$ having $\max(N,M)^* > 0$ zeros may be modified to form a new "reduced" system of the form (4.1) with respective coefficients $\tilde{a}_i$, $\tilde{b}_i$, and characteristic polynomial $\tilde{p}(z)$ having q zeros that form a subset of the zeros of $p(z)$ as defined above. This will hold if $[\tilde{a}_i]_q - [\tilde{b}_j]_q = [a_i]_q - [b_j]_q + ([a_i]_{\tilde{q}} - [b_j]_{\tilde{q}})[R_{0\tilde{q}}][R_{0q}]^{-1}$ and $\tilde{a}_l - \tilde{b}_l = 0$, $l > q$, where $[a_i]_q$, $[a_i]_{\tilde{q}}$ denote the vectors $(a_1, \ldots, a_q)$, $(a_{q+1}, \ldots, \max(N,M)^*)$, respectively, and similarly for the b_j vectors.*

Proof:

Such $\tilde{a}_i$, $\tilde{b}_j$ and $\tilde{p}(z)$ clearly exist. Such a system must have solutions that satisfy $(\Delta\vec{\varepsilon}_k)^T = [R_{kq}]\tilde{\alpha}$, $k \geq 0$, for some $\tilde{\alpha}$ of dimension q. This is satisfied uniquely by $(\Delta\vec{\varepsilon}_k)^T =$

$[R_k]\vec{\alpha}$ in the original system when $\vec{\alpha}_{\tilde{q}} = 0$. Using the results above, this implies that $(\Delta\varepsilon_i)^T_{\tilde{q}} = [R_{0\tilde{q}}][R_{0q}]^{-1}(\Delta\varepsilon_i)^T_q$ for the initial conditions, where clearly $[R_{0q}]$ is invertible. Substituting these initial conditions into the original system in (4.1), and requiring it to also satisfy the reduced system gives

$$\Delta\varepsilon_0 = ([a_i]_q - [b_j]_q + [([a_i]_{\tilde{q}} - [b_j]_{\tilde{q}})[R_{0\tilde{q}}][R_{0q}]^{-1})(\Delta\varepsilon_i)^T_q = ([\tilde{a}_i]_q - [\tilde{b}_j]_q)(\Delta\varepsilon_i)^T_q.$$

Equating terms gives the relationship in the proposition. Since this is the unique characterization of $[\tilde{a}_i]_q - [\tilde{b}_j]_q$ for this equation that satisfies the necessary condition that $(\Delta\varepsilon_i)^T_{\tilde{q}} = [R_{0\tilde{q}}][R_{0q}]^{-1}(\Delta\varepsilon_i)^T_q$, and since a unique definition for $[\tilde{a}_i]_q - [\tilde{b}_j]_q$ must exist, we conclude that the definition is given by this relationship. ■

Chapter 5

Chaos

5.1 Chaos Definition and Preliminaries

In this chapter we present and discuss results concerning chaos in the undithered Σ-Δ modulator. Thus we take $\nu_n = 0$, for all $n \geq 0$. In this form, the main treatment and body of chaos conclusions will be arrived at. Extensions to examine chaos with dither present are made in Section 6.2.

The basic Devaney definition for chaos, as presented in Section 1.4, is given for a simple 1-dimensional mapping dynamical system. We wish to somehow extend and apply this definition to a multidimensional dynamical system with more complex dimensional structure as represented by the Σ-Δ modulator system of (2.1). To do this we need to modify both the form of the sensitivity, topological transitivity and density of periodic points definitions in Devaney's chaos, and how the corresponding chaos conditions are defined. Such a modified version of Devaney's chaos definition, to apply to the Σ-Δ modulator dynamical system is now given.

Definition 5.1 (Adapted Devaney Chaos) *$g \circ f : (\mathbb{R}^N \times \mathcal{C}^M) \to \mathcal{C}^M$ is said to be chaotic on $\mathcal{C}^M$ if*

1. *$g \circ f$ has sensitivity to initial conditions;*
2. *$g \circ f$ is topologically transitive;*
3. *periodic points are dense[1] in $\mathcal{C}^M$.*

These three conditions are correspondingly defined as follows:

Definitions for Adapted Devaney Chaos:

Definition 5.2 (Sensitivity) *$g \circ f : (\mathbb{R}^N \times \mathcal{C}^M) \to \mathcal{C}^M$ has sensitive dependence on initial conditions if there exists $\delta > 0$ such that for any $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and any neighbourhood $\tilde{N}$ of $\vec{x}_0$, there exists $\vec{y}_0 \in \tilde{N}$ and $n \geq 0$ such that $\|g \circ f^n(\vec{x}_0) - g \circ f^n(\vec{y}_0)\| > \delta$.*

Definition 5.3 (Transitivity) *$g \circ f : (\mathbb{R}^N \times \mathcal{C}^M) \to \mathcal{C}^M$ is topologically transitive if for any pair of open sets $U \subset \mathbb{R}^N \times \mathcal{C}^M, V \subset \mathcal{C}^M$, there exists $k > 0$ such that $g \circ f^k(U) \cap V \neq \emptyset$.*

Definition 5.4 (Periodic Points) *The points considered "periodic" are the set $U \subseteq \mathcal{C}^M$ where $U = \{\vec{x}_0^* \in \mathcal{C}^M \mid \exists\ n \geq 1$ with $g \circ f^{kn}(\vec{x}_0) = \vec{x}_0^*$, where $k \in \mathbb{Z}^+$,*
$\vec{x}_0 \equiv (r_{-1}, r_{-2}, \ldots, r_{-N}, \vec{x}_0^*) \in \mathbb{R}^N \times \mathcal{C}^M$ for some $(r_{-1}, r_{-2}, \ldots, r_{-N}) \in \mathbb{R}^N\}$.

Definition 5.5 (Adapted Density) *The set U is dense in $\mathcal{C}^M$ if, for any $\vec{x}_0^* \in U$ and any neighbourhood $\tilde{N}$ of $\vec{x}_0^*$ in $\mathcal{C}^M$, there exists $\vec{y}_0^* \in N$ such that $\vec{y}_0^* \in U$.*

To start with, we have defined the chaos conditions above in terms of the "projection" of the $N+M$ dimensional mappings f onto the M dimensional subspace $\mathcal{C}^M$. The dynamical behaviour and qualities of the system of interest are expressed by the external state variables

[1]The "adapted" definition of density that we use is weaker than the standard definition which follows as a generalization of that given in Definition 1.7 of Section 1.4.

of the system, that is the errors $(\varepsilon_{n-1}, \ldots, \varepsilon_{n-M}) = g(\vec{x}_n) \in \mathcal{C}^M$. It is thus both sufficient and sensible for our analysis to interpret chaos as existing and being defined on these state variables, i.e. on $\mathcal{C}^M$. The remaining internal variables $(r_{n-1}, \ldots, r_{n-N})$ are of little dynamical interest and figure in importance only for stability considerations in Chapter 3. Hence, we screen out the internal variables from consideration here with the projection used for the chaos conditions above. We choose to keep $R^N \times \mathcal{C}^M$ as the underlying state space in order to maintain a direct connection with the practical initial conditions of the system, despite the fact that the projected mappings $g \circ f^n$, $n \geq 1$, $N \geq 1$, will generally be many-to-one. The corresponding definitions for sensitivity, topological transitivity and density of periodic points above are then constructed in the natural way using $g \circ f$ as the mappings under consideration. Thus we have an adapted version of Devaney's definition for chaos to conform to the Σ-Δ system under analysis.

For our version of chaos, we have gone from a constant map from $\mathbb{R}$ to $\mathbb{R}$, to a map $g \circ f_n$ depending on n from $\mathbb{R}^N \times \mathcal{C}^M$ to $\mathcal{C}^M$. Note that from the definition of periodic point above, these are not strictly periodic points of the entire dynamical system of the Σ-Δ modulator in (2.1), but simply points with the corresponding periodic property on $\mathcal{C}^M$ that we choose to regard sufficiently as "periodic points" for the purpose of what is meaningful in our definitions for chaos. With this adapted version, we have a definition for chaos that captures essentially all of the fundamental aspects of Devaney's definition expressed for a simple 1 dimensional mapping in spite of the added complexity introduced.

The underlying extension we make in Devaney's definition from one to higher dimensions is straightforward and consistent with that made by Devaney to apply to hyperbolic toral automorphisms of higher dimension in [7]. Our definition of the "torus" state space $\mathcal{C}^M$, and its metric (including its Cartesian product structure) in Chapter 2, is also equivalent to that made by Devaney for the n-dimensional torus state space of hyperbolic toral automorphisms

in [7]. More uniquely, the notion and definition of "projected" chaos, as incorporated in our chaos definition and discussed above, was specially devised for the studies of this thesis. We know of no other instances where this has been introduced or applied, although such a formulation would clearly be applicable to a broad variety of practical systems that are described as dynamical systems.

Density of Periodic Points:

The one area in which a more significant departure from the strict Devaney definitions has been made concerns defining density of periodic points. The concept of what we mean by density of periodic points thus requires further clarification. In Devaney's definition, it is implicit that periodic points must exist as a prerequisite for the density condition 3 of chaos to hold. The assumption of this is clear from applications of Devaney's chaos by others, and is consistent with the original Li and Yorke definition requiring periodic points as well. Devaney's definitions were primarily intended for autonomous systems, where the mapping f_n in the dynamical system is the same at every iteration. In such systems, the desired "regularity" property of chaos is more easily associated with a necessarily nonempty dense set of periodic points over the state space. The systems we have in (2.1) are generally nonautonomous however, with the mappings f_n depending upon the input x_n at a given n. For systems with a relatively arbitrary, or random external input (i.e. not periodic or possessing some structure), periodic points will generally not exist. In this context, Devaney's definition seems overly restrictive — we don't want to completely exclude chaos from systems with simply a more arbitrary general input.

A simple and sensible property of "regularity" to allow for chaos would then be simply the nonexistence of periodic points as well. Therefore we interpret the definition inclusively, so that a system with no periodic points satisfies, in effect, this density and thus chaos condition 3 by default. An effect of this interpretation is that some autonomous systems

will be classified as satisfying density of periodic points that would not otherwise have been — e.g. the quasiperiodic first-order system (nonchaotic) analyzed in Section 8.3. It is less clear though not discountable that such a change in classification would apply to full chaos, for some cases as well. In any event, we will have a consistent basis upon which to compare chaos properties across a broad variety of Σ-Δ modulator systems.

Of potentially more controversy, is the manner in which we have defined density of periodic points when they exist. Using the standard, strict definition of a dense set, given in the Introduction, would require that periodic points be densely distributed everywhere on $\mathcal{C}^M$ to satisfy chaos condition 3. In developing analytical methods and conceptual approaches for establishing theorems concerning chaos condition 3, we were driven to exercise extra flexibility in how we ultimately defined condition 3, so as to obtain more clear cut and conceptually consistent results. In particular, it is difficult to extend general conditions for strict density, and hence chaos, to systems with more than one feedforward element ($N > 1$), especially if they are nonautonomous with an arbitrary external input. We can presume that Devaney's definition was less intended for systems with these sorts of complexities, and it seems overly restrictive in this context. We might expect, again, that broad conditions would arise under which chaos condition 2 would be satisfied, but not condition 3, which would make the Devaney conditions for chaos much stronger than that of Li and Yorke here. The work of others, e.g. Wang [62], seems to suggest an allowance of non-density of periodic points, by at least the strict definition, in classifying chaos.

In the course of developing more conceptually unified chaos inclusive results, a natural relaxation of Devaney's definition emerged through a further loosening of the notion of a dense set. Following this, for our definition of a dense set in chaos condition 3 above, we then require only that any existing periodic point have another periodic point that is arbitrarily close. This gives strict density only on a subset of $\mathcal{C}^M$, such as a submanifold

or a Cantor-like set, for example. In other words, for chaos condition 3 to be violated, we will require the necessary lack of "regularity" in the system to be that given when there exists a periodic point with no other periodic point arbitrarily close to it. The relaxation of the density condition to one requiring only an internally dense set clearly allows chaos condition 3, and perhaps full chaos, to apply to an array of relatively simple higher order systems (e.g. autonomous or with $N = 0$) to which they would not otherwise apply. This bears particularly on some of the theorems and results to follow. We are nevertheless in a position to build more consistent and comprehensive theoretical results as a contribution of this thesis. In the formal results to follow that assert conditions under which chaos condition 3 holds, footnotes will be given to note the effects of using the standard definition of density, and potentially requiring the existence of periodic points as well, on the results. Condition 3 will always fail under these stricter definitions, whenever it fails under our definition.
■

We now turn our attention towards finding cases under which various conditions in our established definition of chaos will or will not hold, with the ultimate aim of trying to classify conditions for chaotic or nonchaotic behaviour for the Σ-Δ modulator. The theorems presented in this chapter will prove results towards this end. To begin with, the following lemmas are given which will be useful in constructing the proofs of the subsequent theorems:

Lemma 5.6 *Chaos condition* 1 *(sensitivity) as defined with the projection g will hold if and only if this condition* 1 *also holds when g is replaced with g_1 in the definition.*

Proof:

1. Suppose that chaos condition 1 holds when g is replaced with g_1 in the definition. Then choose any $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and any neighbourhood $\tilde{N}$ of $\vec{x}_0$. Now $\exists\ \vec{y}_0 \in \tilde{N}$ and $n_1 > 0$ such that $\|g_1 \circ f^{n_1}(\vec{x}_0) - g_1 \circ f^{n_1}(\vec{y}_0)\| > \delta_1$, where $n = n_1$ and $\delta = \delta_1$ for the version of the definition. Now we have

$$\begin{aligned} \|g \circ f^{n_1}(\vec{x}_0) - g \circ f^{n_1}(\vec{y}_0)\| &= (\sum_{k=1}^{M} \|g_k \circ f^{n_1}(\vec{x}_0) - g_k \circ f^{n_1}(\vec{y}_0)\|^2)^{\frac{1}{2}} \\ &\geq \|g_1 \circ f^{n_1}(\vec{x}_0) - g_1 \circ f^{n_1}(\vec{y}_0)\| \qquad > \delta_1. \end{aligned}$$

Thus chaos condition 1 defined with g holds with $\delta = \delta_1$.

2. Suppose that the normal chaos condition 1 holds as defined with g. Choose any $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and any neighbourhood $\tilde{N}$ of $\vec{x}_0$. Now $\exists\ \vec{y}_0 \in \tilde{N}$ and $n_1 > 0$ such that $\|g \circ f^{n_1}(\vec{x}_0) - g \circ f^{n_1}(\vec{y}_0)\| > \delta_1$, where $\delta = \delta_1$ and $n = n_1$. Thus

$$(\sum_{k=1}^{M} \|g_k \circ f^{n_1}(\vec{x}_0) - g_k \circ f^{n_1}(\vec{y}_0)\|^2)^{\frac{1}{2}} > \delta_1.$$

Now we must have $\|g_{\hat{k}} \circ f^{n_1}(\vec{x}_0) - g_{\hat{k}} \circ f^{n_1}(\vec{y}_0)\| > \dfrac{\delta_1}{\sqrt{M}}$ for some $\hat{k} \in \{1, 2, \ldots, M\}$. This gives that $\|g_1 \circ f^{n_1-\hat{k}+1}(\vec{x}_0) - g_1 \circ f^{n_1-\hat{k}+1}(\vec{y}_0)\| > \dfrac{\delta_1}{\sqrt{M}}$. Without loss of generality, we assume that $\tilde{N}$ is small enough so that $\|g_k(\vec{x}_0) - g_k(\vec{y}_0)\| < \dfrac{\delta_1}{\sqrt{M}},\ \forall\ k = 1, 2, \ldots, M$, which then implies that $n_1 - \hat{k} + 1 > 0$. Thus chaos condition 1 holds when g replaced with g_1 in the definition, and with $\delta = \dfrac{\delta_1}{\sqrt{M}}$. ■

Lemma 5.7 *If chaos condition 2 (transitivity) holds as defined with the projection g, then this condition 2 also holds when g and $V \subset \mathcal{C}^M$ are replaced with g_1 and $V \subset \mathcal{C}$ respectively in the definition.*

Proof:

Suppose chaos condition 2 holds as defined with g. Then choose any pair of open sets $U_1 \subset \mathbb{R}^N \times \mathcal{C}^M$, $V_1 \subset \mathcal{C}$. Now choose any sets $V_2 \subset \mathcal{C}^M$, $V_3 \subset \mathbb{R}^N \times \mathcal{C}^M$, such that $g(V_3) = V_2$ and $g_1(V_3) = V_1$. From the definition, we have equivalently that $\exists$ $\vec{x}_0 \in U_1$ and $k = k_1 > 0$ such that $g \circ f^{k_1}(\vec{x}_0) \in V_2 = g(V_3)$. This then implies that $g_1 \circ f^{k_1}(\vec{x}_0) \in g_1(V_3) = V_1$. Thus chaos condition 2 holds when g and $V \in \mathcal{C}^M$ are replaced by g_1 and $V \in \mathcal{C}$ respectively in the definition. ■

The importance of these lemmas is that, in terms of the first two chaos conditions, they show that the dynamics on the M error variables $(\varepsilon_{n-1}, \ldots, \varepsilon_{n-M})$ are to some extent interchangeable with the dynamics on the first error variable ε_{n-1}. This is not surprising, since the M error variables of $\vec{x}_n$ simply appear as the first error variable $g_1(\vec{x}_i)$ of a previous $\vec{x}_i$ (i.e. ε_{n-k} is the first error variable of $\vec{x}_{n-k+1}$ for $k = 1, \ldots, M$, $n \geq k-1$). These results mean that, for some cases, when proving results regarding the first two chaos conditions, we need only focus on one variable and a mapping to the one dimensional $\mathcal{C}$ rather than the M dimensional $\mathcal{C}^M$. This, rather analogously to the stability theorem proofs of Chapter 3, allows us to look at the associated difference equation (to the dynamical system) solution as a simple expression of what the dynamics are on one variable.

Examples:

In this chapter, various examples of Σ-Δ modulators will be presented that illustrate some of the conditions for chaos. For simplicity, examples 3, 4 and 5, along with those presented in the proofs of Propositions 5.16, 5.25, 5.28, 5.31 and 5.32 will have feedforward elements only, so that $b_j = 0$, $j = 1, \ldots, N$. From (1.2), such systems may be expressed as follows:

$$\varepsilon_n = \frac{\Delta}{2} - [x_n - (a_1\varepsilon_{n-1} + \ldots + a_M\varepsilon_{n-M})] \bmod \Delta, \quad n \geq 0.$$

To simplify the form of this expression, we make the following change of variables: $\varepsilon_n = \frac{\Delta}{2} - v_n\Delta$, $x_n = d_n\Delta + \frac{\Delta}{2}(a_1 + \ldots + a_M)$. This leads to the following form of such systems:

$$v_n = [a_1 v_{n-1} + \ldots + a_M v_{n-M} + d_n] \bmod 1, \quad n \geq 0, \tag{5.1}$$

with $0 \leq v_n < 1$, for all $n \geq -M$. The initial conditions are $v_{-1}, \ldots, v_{-M}$. This change of variables constitutes a translation and scaling of the quantities involved. As such, there will be a one-to-one relationship between the topological properties (i.e. properties pertaining to sensitivity, transitivity, density of periodic points) of the nontransformed and transformed form of a given system. We denote the zeros of $p(z)$ by μ_i.

5.2 Continuous Model

The basic theorems concerning the three conditions for chaos will now be presented. For all the theorems, corollaries, propositions and work to follow in this section, we assume the Σ-Δ modulator has the simplified form presented in Chapter 4, where the a_i and b_j satisfy condition (R) (or the continuity of Theorem 5.15). The results and discussion of Subsection 5.2.1 will focus on systems that are nonminimum phase in general. Subsection 5.2.2 will focus on systems with at least one minimum-phase zero (i.e. nonexpansive), particularly those that are fully minimum or marginally minimum phase.

5.2.1 Nonminimum-Phase Results

Theorem 5.8 *Suppose the a_i and b_j satisfy condition (R). Suppose also that $p(z)$ has a zero with either magnitude greater than 1, or else magnitude equal to 1 with multiplicity greater than 1. Then chaos condition 1 (sensitivity to initial conditions) will hold.*

Proof:

Let δ be a constant satisfying $0 < \delta < \dfrac{\Delta}{2}$. Choose any $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and some neighbourhood $\tilde{N}$ of $\hat{x}$. Now we choose a $\hat{y}_0 \in \tilde{N}_x$ as given in Lemma 4.10. Then, for $\Delta\vec{x}_0$ we have, from the nature of the zeros of $p(z)$ and (2.3), that $\limsup_{n\to\infty} |\Delta\varepsilon_n| = \infty$. Thus $\exists\ n_1 > 0$ such that $|\Delta\varepsilon_{n_1}| > \dfrac{\Delta}{2}$. Now if we replace the $\hat{y}_0$ with $\hat{y}_{\alpha 0} = \hat{x}_0 - \alpha(\hat{x}_0 - \hat{y}_0)$ in $\Delta\vec{x}_0$, with $0 < \alpha < 1$, then, by the properties of linear difference equations, the constants in (2.3) will be scaled down in magnitude by a factor of α, and hence so will $|\Delta\varepsilon_{n_1}|$. We may then choose an $\alpha = \hat{\alpha}$, with $0 < \hat{\alpha} \leq 1$, such that $\delta < |\Delta\varepsilon_{n_1}| < \dfrac{\Delta}{2}$. This then implies that $\delta < \|g_1 \circ f^{n_1+1}(\hat{x}_0) - g_1 \circ f^{n_1+1}(\hat{y}_{\hat{\alpha}0})\|$, with $n_1 > 0$ and $\hat{y}_{\hat{\alpha}0} \in \tilde{N}$. Now applying Lemma 5.6, we have the result that sensitivity to initial conditions holds. ■

Theorem 5.9 *Suppose the a_i and b_j satisfy condition (R). Suppose also that $p(z)$ has at least M zeros, where each zero has either magnitude greater than 1, or else magnitude equal to 1 with multiplicity greater than 1 [counted (multiplicity − 1) times]. Then chaos condition 2 (topological transitivity) will hold.*

Proof:

Let U_1 be any open set with $U_1 \subset \mathbb{R}^N \times \mathcal{C}^M$. Choose an $\hat{x}_0 \in U_1$ such that $g_{k_i}(\hat{x}_0) \neq \Delta/2$, $\forall\ k_i \in \{2, \ldots, M\}$. Now define $\vec{y}_0 = \hat{x}_0 + \Delta\vec{x}_0$, where we assume that $\Delta\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ is small enough in magnitude so that $\vec{y}_0 \in U_1$ and $g_{k_i}(\hat{x}_0 + \alpha\Delta\vec{x}_0) \neq \Delta/2$, for $0 \leq \alpha \leq 1$. Let $\Delta\vec{x}_0 = (\Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-N}; \Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-M})$, where the $\max(N, M)$ $\Delta\varepsilon_i$ serve as the ε and r coordinates of $\Delta\vec{x}_0$ in this manner. From Theorem 4.4, Theorem 4.5 and its proof, it then follows that the solution to (4.1), with initial conditions $\Delta\vec{\varepsilon}_0 = (\Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-\max(N,M)})$, satisfies $g_1(\Delta\vec{\varepsilon}_n) \cong g_1 \circ f^n(\vec{y}_0) - g_1 \circ f^n(\hat{x}_0)$, for $n \geq 0$, where the subscripts from $\Delta\vec{x}_0$ above are decreased by N.

Now choose any point $\hat{z}_0 \in \mathcal{C}^M$. We want to find an $n_1 \geq 0$ such that $g \circ f^{n_1}(\hat{y}_0) = \hat{z}_0$, for a $\hat{y}_0$ satisfying our conditions on $\vec{y}_0$ above. Thus we will require $g(\Delta\hat{\varepsilon}_{n_1}) = \hat{z}_0 - g \circ f^{n_1}(\hat{x}_0)$ to be satisfied for some $\Delta\hat{\varepsilon}_0$, and hence $\Delta\hat{x}_0$ of sufficiently small magnitude.

With the above equation, we then may construct the following using the matrix notation given at the end of Chapter 4.

$$(v(k)_i)^T = [R_k](A_{ij}, B_{ij}, C_{ij})^T = [R_k] \cdot [R_0]^{-1}(\Delta\varepsilon_i)^T,$$

which leads to $(\Delta\varepsilon_i)^T = [R_0] \cdot [R_k]^{-1}(v(k)_i)^T$. For this we define $v(k)_i = g_i(\hat{z}_0) - g_i \circ f^k(\hat{x}_0)$, for $i = 1, \ldots, M$, and $v(k)_i$ are arbitrary, for $i = M+1, \ldots, \max(N, M)$. We set the constraint $|v(k)_i| \leq \Delta$, for $i = 1, \ldots, M$. The matrix $[R_k]$ is invertible, and hence solutions for $(\Delta\varepsilon_i)^T$ exist.

Suppose there are q zeros of $p(z)$ with magnitude greater than 1, or magnitude 1 with extra multiplicity. Let $\tilde{q} = \max(N, M)^* - q$. Thus without loss of generality we take $q = M$ here. We define $(v(k)_i)_q^T$ and $(v(k)_i)_{\tilde{q}}^T$ to be the corresponding vectors of the first q and last $\tilde{q}$ entries of $(v(k)_i)^T$ respectively.

Now we set $\vec{\alpha}_{\tilde{q}} = 0$. This leads to the equation $\vec{\alpha}_q = [R_{kq}]^{-1}(v(k)_i)_q^T$, for $\vec{\alpha}_q$ in terms of the fixed $(v(k)_q)^T$. We also then get $(v(k)_i)_{\tilde{q}}^T = [R_{k\tilde{q}}] \cdot [R_{kq}]^{-1}(v(k)_i)_q^T$ for the arbitrary $(v(k)_i)_{\tilde{q}}^T$.

Now we have that the lim inf of the magnitudes of the entries of the matrix $[R_{kq}]^{-1}$ will go to zero (uniformly) as $k \to \infty$. Thus $\liminf \vec{\alpha}_q \to 0$ as $k \to \infty$. Since $(\Delta\varepsilon_i)^T = [R_0]\vec{\alpha}$, it follows that $\liminf(\Delta\varepsilon_i)^T \to 0$ as $k \to \infty$. Thus we may find an n_1 such that $g \circ f^{n_1}(\hat{y}_0) = \hat{z}_0$ for some $\hat{y}_0 \in U_1$, and $g_{k_i}(\hat{x}_0 + \alpha\Delta\hat{x}_0) \neq \Delta/2$, for $0 \leq \alpha \leq 1$. The latter condition is needed for the validity of the method used. Since $\hat{z}_0$ is arbitrary, it may be an element of any set $V_1 \in \mathcal{C}^M$, so that $g \circ f^{n_1}(U_1) \cap V_1 \neq \emptyset$. With U_1 and V_1 arbitrary, we have the result that topological transitivity holds under the conditions of the theorem. ■

Theorem 5.10 [2] *Suppose the a_i and b_j satisfy condition (R). Suppose also that $p(z)$ has $\max(N, M)$ zeros, where each zero has magnitude greater than 1. Then chaos condition 3 (density of periodic points) will hold.*

Proof:

Suppose that the system has a periodic point $\hat{x}_0^* \in \mathcal{C}^M$. Then $\hat{x}_0^* = g \circ f^{kp}(\hat{x}_0) = g(\hat{x}_0)$ for some period $p \geq 1$ and $k \in \mathbb{Z}^+$, and for some $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$. Now we choose a $\vec{y}_0$ so that the perturbation $\Delta\vec{x}_0 = \vec{y}_0 - \hat{x}_0 = (\Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-N}; \Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-M})$.

To account for the $M > 1$ case when $g_{k_i}(\hat{x}_0) = \Delta/2$ for some $k_i \in \{2, \ldots, M\}$; we may extend the definition of system (1.2) backward by $M - 1$ iterations. For this, we choose inputs x_i, and initial $\varepsilon_{i-M} \neq \Delta/2$, r_{i-N}, for $\hat{x}_i$, $i = -1, \ldots, -(M-1)$, such that $f^{n+M-1}(\hat{x}_{-(M-1)}) = f^n(\hat{x}_0)$, $\forall\, n \geq 0$. We would now choose $\vec{y}_{-(M-1)}$ (instead of $\vec{y}_0$) and the corresponding perturbation $\Delta\vec{x}_{-(M-1)} = \vec{y}_{-(M-1)} - \hat{x}_{-(M-1)} = (\Delta\varepsilon_{-1-(M-1)}, \ldots, \Delta\varepsilon_{-N-(M-1)}; \Delta\varepsilon_{-1-(M-1)}, \ldots, \Delta\varepsilon_{-M-(M-1)})$. From Theorem 4.4, Theorem 4.5 and its proof, it follows that the solution to (4.1) with initial conditions $\Delta\vec{\varepsilon}_{-a} = (\Delta\varepsilon_{-1-a}, \ldots, \Delta\varepsilon_{-\max(N,M)-a})$, satisfies $g_1(\Delta\vec{\varepsilon}_n) \cong g_1 \circ f^{n+a}(\vec{y}_{-a}) - g_1 \circ f^{n+a}(\hat{x}_{-a})$ for $n \geq -a$, where the subscripts from $\Delta\vec{x}_{-a}$ above are decreased by $N + a$, and $a = 0$ or $M - 1$, $(M > 1)$.

We will now require that $\Delta\vec{x}_0$ be an M dimensional periodic point in $\mathcal{C}^M$ of (4.1), defined by $g(\Delta\vec{x}_{\hat{k}p_1}) = g(\Delta\vec{x}_0)$, $\forall\ \hat{k} \in \mathbb{Z}^+$, with some period $p_1 \geq 1$. We set $p_1 = lp$ for some $l \in \mathbb{Z}^+$. This then requires that $\Delta\varepsilon_{lp+i} = \Delta\varepsilon_i + m_i\Delta$, for some $m_i \in \mathbb{Z}$, not all zero, and $i = -1, \ldots, -M$. This holds when system (4.1) cycles back to produce the same periodic point from the initial condition modulo Δ, every lp iterations. Extending the range of i to $-\max(N, M)$ in this equation, and to $-lp$ in the definition of the m_i, we

[2] Using the standard definition of density, we conjecture that this result will hold when $N = 0, 1$, with no conclusion otherwise. If existence of periodic points is added to the density definition, this result must revert to Corollary 5.22, with no conclusion otherwise.

have the following. From the iterative relationship developed in the proof of Theorem 4.4, we have that

$$\Delta\varepsilon_{lp+i} = \Delta r_{lp+i} + \sum_{j=1}^{lp+i} \tilde{r}_j m_{i-j}\Delta, \quad \text{where } i = -1, \ldots, -N,$$

and the Δr_{lp+i} are the corresponding internal variable components from $\Delta\vec{x}_{lp}$. We can now set up the following scheme to define the m_i, $i = -1, \ldots, -\max(N, M)$, not all zero, for i up to $-M$, in terms of $\max(1, M - N + 1)$ arbitrary integers. We set

$$m_i = \sum_{j=1}^{M+i} \tilde{r}_j m_{i-j} + \tilde{r}_{M+i+1}\hat{m}, \quad \text{for } i = -1, \ldots, -N,$$

with $\hat{m} \in \mathbb{Z}$, and m_i arbitrary for $i = -N - 1, \ldots, -M$, if $N < M$. Substituting into the $\Delta\varepsilon_i$ equation above, this leads to

$$\Delta r_{lp+i} = \Delta\varepsilon_i - [\sum_{j=M+i+1}^{lp+i} \tilde{r}_j m_{i-j} - \tilde{r}_{M+i+1}\hat{m}]\Delta, \quad \text{for } i = -1, \ldots, -N.$$

With these Δr_{lp+i}, the effect of the bracketed terms will cancel out modulo Δ, via condition (R), and thus $g(\Delta\vec{x}_{lp+j}) = g(\Delta\vec{x}_j)$, $\forall\ j \geq 0$, as is sufficient for a periodic point.

We now apply the matrix notation given at the end of Chapter 4, with $(\Delta\varepsilon_{i-a})^T$ corresponding to the initial conditions $\Delta\vec{\varepsilon}_{-a}$. Also, we define $[I_a] = [R_a][R_0]^{-1}$ if $a = M - 1$, and $[I_a] = I$, the identity matrix, if $a = 0$. With the above equation, we then may construct the following.

$$[I_a] \cdot (\Delta\varepsilon_{i-a})^T + (\vec{\beta}_i)^T = [R_k]\vec{\alpha} = [R_k] \cdot [R_0]^{-1} \cdot [I_a] \cdot (\Delta\varepsilon_{i-a})^T.$$

For this we define $(\vec{\beta}_i)^T$ to be the $\max(N, M) \times 1$ vector with entries $m_{-i}\Delta$, for $i = 1, \ldots, \max(N, M)$, as designated above. Without loss of generality, we can replace $[I_a] \cdot (\Delta\varepsilon_{i-a})^T$ with $(\Delta\varepsilon_i)^T$ in this expression.

The equation above leads to $(\vec{\beta}_i)^T = ([R_k] - [R_0])\vec{\alpha}$. This leads to the equation $\vec{\alpha} = ([R_k] - [R_0])^{-1}(\vec{\beta}_i)^T$ for $\vec{\alpha}$ in terms of the Δ multiples in $(\vec{\beta}_i)^T$, which we take to be fixed. Clearly $([R_k] - [R_0])$ is invertible. Now we have that the lim inf of the magnitudes of the entries of the matrix $([R_k] - [R_0])^{-1}$ will go to zero (uniformly) as $k \to \infty$. Thus $\liminf \vec{\alpha} \to 0$ as $k \to \infty$. Since $(\Delta\varepsilon_i)^T = [R_0]\vec{\alpha}$, it follows that $\liminf(\Delta\varepsilon_i)^T \to 0$ as $k \to \infty$.

Now we choose k so that $k = lp$, where $l \in \mathbb{Z}^+$. With this, it may be concluded that for any specific values of m_i, $(\Delta\varepsilon_i)^T$ will become arbitrarily close to zero in its entries and magnitude as l becomes arbitrarily large. Thus, for specific m_i, there must exist an l such that the solution for $\Delta\vec{\varepsilon}_0$ yields an associated $\Delta\hat{x}_0$, with $\hat{y}_0 = \hat{x}_0 + \Delta\hat{x}_0$ satisfying $g(\hat{y}_0) \in \tilde{N}$, for some neighbourhood $\tilde{N} \subset \mathcal{C}^M$ about $g(\hat{x}_0)$. Let $\Delta\hat{\varepsilon}_0$, with corresponding $\hat{l}$, be such a solution. From our construction in the equations, $\Delta\hat{\varepsilon}_0$ repeats as an effective initial condition of (4.1), such that $\Delta\hat{\varepsilon}_0$ and hence $g(\Delta\hat{\varepsilon}_0)$ is cyclic and periodic with period $\hat{l}p$. Thus we must have that

$$\begin{aligned} g(\Delta\hat{x}_{\hat{k}\hat{l}p}) &= g \circ f^{\hat{k}\hat{l}p}(\hat{y}_0) - g \circ f^{\hat{k}\hat{l}p}(\hat{x}_0) + (\gamma(1), \ldots, \gamma(M))\Delta = g(\Delta\hat{x}_0) \\ &= g(\hat{y}_0) - g(\hat{x}_0) + (\gamma(1), \ldots, \gamma(M))\Delta, \qquad \text{for } \hat{k} \in \mathbb{Z}^+. \end{aligned}$$

For this, with $\Delta\hat{x}_0$ small, we define $\gamma(i) = 1$ if $g_i(\hat{x}_0) = \Delta/2$ and $\mathrm{sgn}(g_i(\Delta\hat{x}_0)) > 0$, and $\gamma(i) = 0$ otherwise, $i = 1, \ldots, M$. Thus

$$\begin{aligned} g \circ f^{\hat{k}(\hat{l}p)}(\hat{y}_0) &= g \circ f^{\hat{k}\hat{l}p}(\hat{y}_0) - g \circ f^{\hat{k}\hat{l}p}(\hat{x}_0) + g \circ f^{\hat{k}\hat{l}p}(\hat{x}_0) \\ &= g(\hat{y}_0) - g(\hat{x}_0) + g \circ f^{(\hat{k}\hat{l})p}(\hat{x}_0) = g(\hat{y}_0), \end{aligned}$$

from the above and the fact that $g(\hat{x}_0)$ is a periodic point with period p. Thus $g(\hat{y}_0)$ is a periodic point with period $\hat{l}p$. With $\hat{y}_0$ as chosen, and $g(\hat{x}_0)$ any periodic point, it then follows that the result of density of periodic points holds. ■

Corollary 5.11 [3] *Suppose the a_i and b_j satisfy condition (R). Suppose also that $p(z)$ has q zeros, where each zero has either magnitude greater than 1, or else magnitude equal to 1 with multiplicity greater than 1 [counted at most (multiplicity − 1) times], and $M \leq q < \max(N, M)$. Suppose further that the a_i, $\tilde{b}_j$ of the reduced form of (1.2) associated with the lower order factor of $p(z)$ containing strictly the q zeros, satisfy condition (R). Then chaos condition 3 (density of periodic points) will hold. Moreover, if $p(z) = 1$, then chaos condition 3 will hold if and only if the unique orbit does not allow a periodic point.*

Proof:

The method of the proof of Theorem 5.10 is followed, with the matrix notation from the end of Chapter 4. We have $(\Delta\varepsilon_i)_{\tilde{q}}^T = [R_{0\tilde{q}}] \cdot [R_{0q}]^{-1}(\Delta\varepsilon_i)_q^T$, where $\Delta\vec{x}_0 = ((\Delta\varepsilon_i)_q, (\Delta\varepsilon_i)_{\tilde{q}}, \vec{0};\ \Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-M})$. We consider the reduced form of (1.2) associated with the lower order factor of $p(z)$, denoted $\tilde{p}(z)$, containing strictly the q nonminimum-phase zeros. We denote the b_j of the reduced form by $\tilde{b}_j$. Let $[b_j]_q$, $[b_j]_{\tilde{q}}$ denote the vectors $(b_1, \ldots, b_q)$, $(b_{q+1}, \ldots, N)$ respectively. Then we have, from Proposition 4.11, the definition $[\tilde{b}_j]_q \equiv [b_j]_q + [b_j]_{\tilde{q}}[R_{0\tilde{q}}][R_{0q}]^{-1}$. The characteristic polynomial of the reduced system is then $\tilde{p}(z) = z^q + \sum_{k=1}^{q}(\tilde{b}_k - a_k)z^{q-k}$. From the iterative relationship developed in the proof of Theorem 4.4 applied to the reduced system, we have that $\Delta\varepsilon_{lp+i} = \Delta r_{lp+i} + \sum_{j=1}^{lp+i} \tilde{r}_j m_{i-j}\Delta$, where $i = -1, \ldots, -q$, and the $\tilde{r}_j$ are defined from condition (R) applied to the reduced system, with the a_i, $\tilde{b}_j$. We then define m_i and $\hat{m}$ as in the proof of Theorem 5.10, for $i = -1, \ldots, -q$. The remainder of the proof of Theorem 5.10 may then be applied, with $[R_k]$, $[R_0]$, $[R_a]$, $\vec{\alpha}$ replaced by $[R_{kq}]$, $[R_{0q}]$, $[R_{aq}]$, $\vec{\alpha}_q$ respectively, and the corresponding vectors reduced to length q.

[3] Using the standard definition of density, this result does not apply, and we make no such conclusion.

Now suppose that $p(z) = 1$. From the degenerate form of (1.2), we have that $\Delta\varepsilon_n = 0$, $\forall\ n \geq 0$. This implies that, for any initial condition $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$, there exists a unique orbit $\{\hat{e}_{M+n} \in \mathcal{C}^M \mid n \geq 0\}$ such that $g(\vec{x}_{M+n}) = \hat{e}_{M+n},\ \forall\ n \geq 0$. Clearly chaos condition 3 holds if the unique orbit cannot allow a periodic point; that is if there does not exist an $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and $p \in \mathbb{Z}^+$ such that $\hat{e}_{kp} = g(\vec{x}_0)$, $\forall\ k \in \mathbb{Z}^+$ such that $kp \geq M$. Now suppose that there exists such an $\vec{x}_0$ and p. Suppose that $\vec{y}_0$ is another periodic point with period t. It then follows that $g(\vec{x}_0) = g(\vec{x}_{lpt}) = \hat{e}_{lpt} = g(\vec{y}_{lpt}) = g(\vec{y}_0),\ \forall\ l \in \mathbb{Z}^+$ such that $lpt \geq M$. This implies that the periodic point is unique, and hence chaos condition 3 does not hold. ■

The results of these theorems show that all three chaos conditions are satisfied when the system has a noise transfer function with all zeros strictly nonminimum phase. Thus our version of Devaney chaos holds for these cases (with condition (R) assumed). In Subsection 5.2.2, we find that chaos condition 1 fails for the minimum or marginally minimum-phase (with magnitude 1 zeros of multiplicity 1) cases examined (i.e. with condition (R) or the continuity of Theorem 5.15 assumed), and so this Devaney chaos fails for these cases as well. These results, while generally consistent with the vaguer "chaos" claims of previous research (i.e. [56], [62], [53], [50]), provide a much clearer picture of the manner in which the chaos conditions come into play in the nonminimum-phase cases. In these papers, the existence of chaos is suggested for the nonminimum-phase case only, but no rigorous chaos definition or proof of when this chaos does or does not exist is given.

Here we see that sensitivity to initial conditions holds automatically whenever any zero has either magnitude greater than 1 or else magnitude 1, multiplicity greater than 1. Density of periodic points generally requires that all zeros (of which there must be at least M) have magnitude greater than 1, although some broad conditions are presented in

Corollary 5.11, and further conditions in Theorem 5.12 and the subsequent Corollary 5.13, whereby this condition will hold for at least M such zeros, or fewer such zeros, respectively. Topological transitivity holds under the less stringent requirement of at least M nonminimum-phase zeros, but, as shall be seen, would seem to depart from this rule only for more restrictive classes of cases.

The manner in which these nonminimum-phase cases break into the different possibilities of chaos conditions being satisfied is not too difficult to understand. For strictly nonminimum-phase zeros, the expansivity contribution to the error differences $\Delta\varepsilon_n$ is unbounded. When there is a magnitude 1 zero with multiplicity greater than 1, this gives a resonance like result where the expansivity contribution from the (multiplicity $-$ 1) repeated zeros is also unbounded. Sensitivity to initial conditions will hold with expansivity only in a given error coordinate projection $g_i(\vec{x}_n)$, and therefore requires only one zero to be expansive. Topological transitivity requires expansivity in all M directions of $\mathcal{C}^M$ to be guaranteed, since we are trying to map the neighbourhood of a given point to any other neighbourhood. This requires M expansive zeros. Density of periodic points, in the most general case, has even more constraining properties, and we require that all zeros ($\geq M$) be expansive for it to be shown. This follows from the fact that periodic points must not only map back to themselves, but must do so repeatedly. The requirement may be reduced to as few as M expansive zeros, however, if condition (R) is satisfied, in parallel, on the "reduced" lower order system associated with these fewer zeros (assuming this is mathematically possible), (see Proposition 4.11 for a general definition of this, and Corollary 5.11 for a precise definition of the $\tilde{b}_j$). Under appropriate conditions, even fewer than M such zeros may be sufficient, since the dense periodic points in $\mathcal{C}^M$ may be dense over manifolds of dimension less than M in $\mathcal{C}^M$.

It is interesting to find that, while the system has order $\max(N, M)$, our requirements for chaos condition 2 are only for M nonminimum-phase zeros, even if $N > M$. We only require M degrees of freedom offered by some M zeros to attain these chaos conditions, and hence the remaining $N - M$ zeros and their values play no role, by virtue of the linear structure of the system, in these determinations. When the multiplicities of the magnitude 1 zeros are 1, or greater than 1 and counted once, we have boundedness in the expansivity contribution to the error differences $\Delta\varepsilon_n$, as for strictly minimum-phase zeros. Thus it is easy to see why we have no contributions to general results for any chaos conditions being satisfied in these cases, and indeed have the converse. Thus in summary, for the case of filter gains satisfying condition (R) or of the continuity conditions of Theorem 5.15 holding, we have broad and conclusive statements concerning the existence of chaos.

From the developments used in the proofs of the above theorems, we note that the nominally $\max(N, M)$ dimensional system (4.1) reduces to an analogous lower $\hat{q}$ dimensional system when either the dimensionality of the state space of initial conditions is reduced to $\hat{q}$, or $\hat{q} = \max(N, M)^* < \max(N, M)$. In the latter case, the $\max(N, M)$ dimensional space of initial conditions $\Delta\vec{\varepsilon}_0$ is projected onto a $\max(N, M)^*$ dimensional general solution subspace for $\Delta\vec{\varepsilon}_n$, $n \geq \max(N, M) - \max(N, M)^*$. In particular, if $p(z) = 1$ ($\max(N, M)^* = 0$), this projection corresponds to a single unique orbit for $\varepsilon_n = f^{n+1}(\vec{x}_0)$ over all initial conditions $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$.

A theorem and subsequent corollary which extend chaos condition 3 to more general nonminimum-phase cases are now presented.

Theorem 5.12 [4] *Suppose the a_i and b_j satisfy condition (R) and one of the following applies.*

1. *Suppose $p(z)$ has at least q zeros with $1 \le q \le \max(N,M)^*$, where each zero μ_i has magnitude greater than 1 and is counted once if the multiplicity is greater than 1; and that the argument $\theta_i = 2\pi\tilde{\mu}_i$ for some rational number $\tilde{\mu}_i$ if the zero μ_i is complex. Suppose also that there exist q numbers $\gamma_j \in \mathbb{R}$, $j = 1,\ldots,q$, not all zero, such that*

$$\sum_{j=1}^{q} \gamma_j z_{ij} = \bar{\gamma}_i \in \mathbb{Z}, \qquad \begin{array}{l} \text{for } i = 1,\ldots,q_1, \text{ with } \bar{\gamma}_i,\ i = 1,\ldots,M, \text{ not all zero,} \\ \text{and} \quad \max(M,q) \le q_1 \le \max(N,M); \end{array}$$

where z_{ij} is the (i,j)th element of the matrix $[R_0]$ specified with respect to the q zeros, as analogously defined in Chapter 4, and extended at the bottom to include q_1 rows. Suppose further that either:

(a) The a_i, $\tilde{b}_j$ of the reduced form of (1.2) associated with a reduction in the number of recursions from $\max(N,M)$ to q_1, satisfy condition (R). Also, $\bar{\gamma}_i = m_{-i}$, $i = 1,\ldots,q_1$, for some $\hat{m}$, m_{-k}, $k = N+1,\ldots,M$, from the definition in the proof of Theorem 5.10 applied to the reduced system;

(b) $M = 1$, with (R) holding on the reduced system; or

(c) a_i, $\tilde{b}_j \in \mathbb{Z}$ for $i = 1,\ldots,M$, $j = 1,\ldots,\min(N,q_1)$.

2. *In particular, suppose $q = 1$ or $q = 2$ above, and that the zero(s) have rational magnitude $|\mu|$ and is real (with value μ), or are complex with $\theta = \pm\pi/4$, respectively. Suppose further that either:*

(a) There exist nonzero numbers $\hat{\gamma}_1$, $\hat{\gamma}_{-1} \in \mathbb{R}$, such that $\hat{\gamma}_1(\mu)^{-i} = m_{-i}$, in the $q = 1$ case; and $\hat{\gamma}_{(-1)^i}|\mu|^{-i}(-1)^{A(i)} = m_{-i}$, where $A(i) = -i/2 \mod 1$, in the $q = 2$ case; with $i = 1,\ldots,q_1 = \max(N,M)$, and m_{-i} as in 1(a); or

[4] Using the standard definition of density, this result will generally not hold.

one of 1*(b), (c), with* $q_1 = \max(N, M)$*, holds.*

Then chaos condition 3 *(density of periodic points) will hold.*

Proof:

1. For the proof of this theorem, we follow the same definitions and development as for the proof of Theorem 5.10 and its corollary, leading to the equation $(\vec{\beta}_i)^T = ([R_k] - [R_0])\vec{\alpha}$. Now we apply the matrix notation given at the end of Chapter 4 with respect to the q zeros in question, and introduce the following additional notation. Let $[R_{kq_1}]_{(q)}$ and $[R_{0q_1}]_{(q)}$ be matrices that comprise the first q columns (corresponding to the q zeros) of $[R_k]$ and $[R_0]$ respectively, where the number of rows has been extended from $\max(N, M)^*$ to q_1 at the bottom.

Now set $\vec{\alpha}_{\tilde{q}} = 0$. Let $k = \hat{k} m_\theta$, for $\hat{k} \in \mathbb{Z}^+$, where

$m_\theta = LCM(2, 2\pi/\theta_i, \forall$ complex arguments $\theta_i \neq 0$ corresponding to the q nonminimum -phase zeros satisfying the theorem).

Then the (i, j) element of $([R_{kq_1}]_{(q)} - [R_{0q_1}]_{(q)})$ will be $z_{ij}(|\mu_j|^k - 1)$, where $1 \leq i \leq q_1$, $1 \leq j \leq q$, and z_{ij} is the (i, j) element of $[R_{0q}]_{(q)}$. The RHS of the equation above then gives the vector $(\sum_{j=1}^{q} z_{ij}(|\mu_j|^k - 1)\alpha_j)^T$ with entries over $1 \leq i \leq q_1$, where α_j is the jth entry of $\vec{\alpha}_q$. Now we set $\alpha_j = \dfrac{\gamma_j \Delta}{|\mu_j|^k - 1}$, where the γ_j satisfy the theorem requirements. Then the RHS vector becomes $(\sum_{j=1}^{q} \gamma_j z_{ij} \Delta)^T = (\bar{\gamma}_i \Delta)^T$.

(a) This is the required form of the LHS of the equation if case (a) holds, and thus our $\vec{\alpha}$ have been correctly chosen. Let $[b_j]_{q_1}$, $[b_j]_{\tilde{q}_1}$ denote the vectors $(b_1, \ldots, b_{q_1})$, $(b_{q_1+1}, \ldots, N)$ respectively. Then, using Proposition 4.11, we have the definition $[\tilde{b}_j]_{q_1} \equiv [b_j]_{q_1} + [b_j]_{\tilde{q}_1}[R_{0\tilde{q}_1}]_{(q)}[R_{0q}]^{-1}$, where $[R_{0\tilde{q}_1}]_{(q)}$ is defined analogously (consistent with earlier

definitions) with respect to $[R_{0q_1}]_{(q)}$. Similarly, we have $(\Delta\varepsilon_i)^T_{\tilde{q}_1} = [R_{0\tilde{q}_1}]_{(q)} \cdot [R_{0q}]^{-1}(\Delta\varepsilon_i)^T_q$. For case (a), we now have the iterative relationship form of Corollary 5.11 applied to the reduced system satisfying condition (R) here.

(b) Suppose case (b) holds. Then we scale the α_j as defined above by an integer factor $\hat{m}_\gamma$, replacing α_j by $\tilde{\alpha}_j = \hat{m}_\gamma \alpha_j$. The vector of the RHS of the equation above then becomes $(\hat{m}_\gamma \bar{\gamma}_i \Delta)^T$. We now set $\hat{m}_\gamma \bar{\gamma}_i = \sum_{s=1}^{q_1} \tilde{r}_{M-i+s} \hat{m}_s$, for $i = 1, \ldots, q_1$, for some $\hat{m}_s$ that we wish to choose in $\mathbb{Z}$, with $s = 1, \ldots, q_1$, and where the $\tilde{r}_{M-i+s}$ are defined with respect to the reduced system. From the structure of the vectors $(\tilde{r}_{M-i+s})^T \equiv (\vec{0}, \tilde{r}_1, \ldots, \tilde{r}_{M-i+q_1})^T$, $i = 1, \ldots, q_1$, and the conditions on M and q_1, it follows from (R) that these vectors are linearly independent on Z^{q_1}. Thus $\exists\ \hat{m}_\gamma,\ \hat{m}_s \in \mathbb{Z},\ s = 1, \ldots, q_1$, such that these $\hat{m}_\gamma \bar{\gamma}_i$ equations are satisfied for $i = 1, \ldots, q_1$. From the arguments in the proof of Theorem 5.10, this choice of $m_{-i} = \hat{m}_\gamma \bar{\gamma}_i$, $i = 1, \ldots, q_1$, will then cancel out modulo Δ in the internal Δr coordinate mappings, in the reduced system, and the requirements for a periodic point will be met.

(c) Suppose case (c) holds. Then any integer choice of the $m_{-i} = \bar{\gamma}_i$ will cancel out modulo Δ in the internal Δr coordinate mappings, in the reduced system, as occurs with cases (a) and (b).

It is now clear that $\vec{\alpha}_q \to 0$ as $\hat{k} \to \infty$, where $k = \hat{k} m_\theta$. With $\hat{k} = lp$, it may be concluded that with these specific values of m_{-i}, there exists large l in $\mathbb{Z}^+$ to make $(\Delta\varepsilon_i)^T$ arbitrarily close to zero in its entries and magnitude. The rest of the proof then follows the proof of Corollary 5.11 to the end, with $[R_{kq_1}]$, $[R_{0q_1}]$, $[R_{aq_1}]$, and $\vec{\alpha}_{q_1} \equiv (\vec{\alpha}_q, \vec{0})$.

2. If $q = 1$ and the zero is real, or $q = 2$ and the zero is complex with $\theta = \pm\pi/4$, then the entries z_{ij} have magnitude $|\mu|^{-1}$, $i = 1, \ldots, \max(N, M)$, $j = 1, \ldots, q$, and will be rational if and only if the corresponding zeros have rational magnitude. Under these

conditions, we may choose nonzero integers γ_j such that $\gamma_1 z_{i1} = \gamma_1(\mu)^{-i} \in \mathbb{Z}$ if $q = 1$, and $\gamma_1 z_{i1} + \gamma_2 z_{i2} = \gamma_j|\mu|^{-i}(-1)^{A(i)} \in \mathbb{Z}$, where $A(i) = -i/2 \mod 1$, $j = 1, 2$ for i odd/even respectively, if $q = 2$. Thus if 2(a) holds, 1(a) is satisfied; and if 1(a), (b) or (c) hold, the condition in the theorem will then be satisfied. ■

Corollary 5.13 [5] *Suppose the a_i and b_j satisfy condition (R). Suppose also that $p(z)$ has at least q zeros, with $1 \leq q \leq \max(N, M)^*$, where each zero has either magnitude greater than 1, or else magnitude equal to 1 with multiplicity greater than 1 [counted (multiplicity − 1) times]. Suppose further that there exists an infinite sequence of numbers $\gamma_j(k_m) \in \mathbb{R}$, with $j = 1, \ldots, q$, and $k_m \in \mathbb{Z}^+$, $m \geq 1$, such that $k_{m+1} > k_m$ for all such m, and that the $\gamma_j(k_m)$ are not all zero at any $m \geq 1$. Now suppose that the following hold:*

$$\sum_{j=1}^{q} \gamma_j(k_m) z_{ij}(k_m) = \bar{\gamma}_i(k_m) \in \mathbb{Z}, \quad \text{for } i = 1, \ldots, q_1, \text{ with } \gamma_i(\bar{k}_m),\ i = 1, \ldots, M,$$
$$\text{not all zero at any } m \geq 1, \text{ and } M \leq q_1 \leq \max(N, M);$$

there exists $K > 0$ such that $|z_{ij}(k_m)| > K$, for all $m \geq 1$, where $z_{ij}(k_m)|\mu_j|^{k_m}$ is the (i, j)th element of the matrix $[R_{k_m}]$ specified with respect to the q zeros, as analogously defined in Chapter 4, and extended at the bottom to include q_1 rows; and the $\gamma_i(k_m)$ satisfy $|\gamma_i(k_m)| < \hat{K}$ for some $\hat{K} > 0$, for all $m \geq 1$. In addition, suppose that either: 1(a) with $\bar{\gamma}_i(k_m)$, for all $m \geq 1$; or one of 1(b), (c) from Theorem 5.12 hold. Then chaos condition 3 (density of periodic points) will hold.

Proof:

The proof of this corollary follows the proof of Theorem 5.12, with $k = k_m$, $m \in \mathbb{Z}^+$, and the following changes. The (i, j) element of $([R_{kq_1}]_{(q)} - [R_{0q_1}]_{(q)})$ will be

[5] Using the standard definition of density, this result will generally not hold.

$(z_{ij}(k)|\mu_j|^k - z_{ij}(0)) = z_{ij}(k)(|\mu_j|^k - \frac{z_{ij}(0)}{z_{ij}(k)})$, for $m \in \mathbb{Z}^+$, where $1 \leq i \leq q_1$, $1 \leq j \leq q$, with $z_{ij}(k)|\mu_j|^k$ and z_{ij} the corresponding elements of $[R_{kq_1}]_{(q)}$ and $[R_{0q_1}]_{(q)}$ respectively. For this, we use $|z_{ij}| > K$. The RHS of the equation then gives the vector $(\sum_{j=1}^{q} z_{ij}(k)(|\mu_j|^k - \frac{z_{ij}(0)}{z_{ij}(k)})\alpha_j)^T$ with entries over $1 \leq i \leq q_1$. Now we set $\alpha_j = \frac{\gamma_j(k)\Delta}{|\mu_j|^k - \frac{z_{ij}(0)}{z_{ij}(k)}}$, where the $\gamma_j(k)$ satisfy the corollary. The rest of the proof follows that of Theorem 5.12. It is clear that $\vec{\alpha}_q \to 0$ as $m \to \infty$, since the resulting k are unbounded, and the corresponding $\gamma_j(k)$ and $z_{ij}(k)$ are bounded above and below in magnitude respectively. Thus, from the proof of Theorem 5.12, we have the required result. ■

Theorem 5.12 provides general conditions on the zeros under which density of periodic points may be obtained when fewer than $\max(N, M)$ of these zeros satisfy the requirements of Theorem 5.10. These conditions are most easily described when all of the coefficients a_i, b_j are integers, or when $M = 1$. If only one such real zero, or one complex conjugate pair of zeros with argument $\pm\pi/4$ is considered, the condition of a rational magnitude for the zero is sufficient to obtain chaos condition 3 in these special cases, and to a lesser extent more generally. If more zeros of a restrictive nature are considered, then less restrictive sets of such zeros that meet a generalization of this condition are sufficient for the analogous cases. For a given number of such zeros, there is also trade off between the level of restrictiveness of the sets, and the order of the reduced system that must satisfy (R), as part of the conditions (if $q_1 = \max(N, M)$, then the parallel (R) condition drops altogether).

Corollary 5.13 extends the result of Theorem 5.12 to allow for all nonminimum-phase zeros of the form considered in Theorem 5.10, that is, to include complex zeros with irrational arguments and all corresponding zeros with multiplicity greater than 1. The condition that such zeros must meet, however, is far stricter, since the existence of a certain

infinite sequence of bounded numbers is necessary. Theorem 5.12 and Corollary 5.13 work as relaxations of Theorem 5.10 because we have used certain rational number related symmetries that exist in the zeros (and relate to the structure of the mapping f^n) to show density of periodic points on manifolds of dimension less than M in $\mathcal{C}^M$.

Certain types of more general nonminimum-phase cases may satisfy both chaos conditions 2 and 3, and be fully chaotic. This will occur when the system (1.2) describes a mapping that is a hyperbolic toral automorphism. In [7], Devaney defines this type dynamical system and shows that it is chaotic. Such a dynamical system, in two dimensions, is given by the mapping $L : [0,1)^2 \to [0,1)^2$, with $L(v_{n-1}, v_{n-2}) = A \cdot (v_{n-1}, v_{n-2})^T$ on $[0,1)^2$, for all $n \geq 0$, where A is a 2×2 matrix with all entries integer, $\det(A) = \pm 1$, and A is hyperbolic. The following proposition considers a more specific form of this, arising from a Σ-Δ modulator system, and asserts that it is chaotic.

Proposition 5.14 [6] *Suppose that the system is second order with $M = 2$, $N = 0$, and has a constant input given by $x_n = \frac{\Delta}{2}(a_1 + a_2)$, for all $n \geq 0$. If $a_1 \in \mathbb{Z}$, $a_2 = \pm 1$, and the zeros of $p(z)$ are real with one of magnitude less than one, and the other of magnitude greater than one, then both chaos conditions 2 and 3 will hold, and the system will be chaotic.*

Proof:

The second sentence in the proposition is the definition of a hyperbolic toral automorphism in [7] corresponding to the dynamical system given from (1.2) and (2.1) for a system satisfying the conditions in the first sentence, and transformed to the form (5.1), (with $d_n = 0, \forall\ n \geq 0$). [7] proves such systems are chaotic. The adapted requirements for chaos used by Devaney for this are sufficient to imply chaos under the adapted requirements de-

[6] Using the standard definition of density, with existence of periodic points added as well, this result will still hold.

fined at the beginning of Chapter 5 in this thesis. Chaos condition 1 follows automatically from Theorem 5.8 as well. Thus the system will be chaotic. ■

An example satisfying this proposition would be $a_1 = 2$, $a_2 = 1$, so that $r_{1,2} = 1 \pm \frac{\sqrt{2}}{2}$. More general classes of hyperbolic toral automorphisms (or chaotic mappings) would exist for higher order systems, and systems, we expect, with $N \geq 1$ and/or more general input x_n.

For topological transitivity then, we have some lesser flexibility (compared to density of periodic points) beyond our relative requirements. The following is a nonminimum-phase example with rational zeros which fails the conditions of Theorem 5.10 and is indeed shown not to be topologically transitive:

Example 1:

Consider the second-order system with $a_1 = 2$, $a_2 = -1$, $b_1 = -\frac{1}{2}$, $x_n = \Delta/2$, for all $n \geq 0$; so that $M = 2$ and $N = 1$. Then $p(z) = z^2 - \frac{5}{2}z + 1$, with zeros $\mu_1 = 2$ and $\mu_2 = \frac{1}{2}$. This system satisfies condition (R), since $\tilde{r}_1 = 2$, $\tilde{r}_i = 0$, for $i \geq 2$. The system is also internally stable with $|b_1| < 1$. Now choose $\hat{x}_0 = 0 \in \mathbb{R} \times \mathcal{C}^2$ and an open set $V_1 \in \mathcal{C}^2$, such that $z_1 - 2z_2 \ni \mathbb{Z}$, for any $\vec{z} = (z_1, z_2) \in V_1$. Now choose any $\vec{v} = (v_1, v_2) \in V_1$. Using the approach of the proof of Theorem 5.9, we have $\hat{x}_k = 0$, and $(v(k)_1, v(k)_2) = (v_1 + m(k)_1\Delta, v_2 + m(k)_2\Delta)$, $\forall\ k \geq 0$, with $m(k)_1$, $m(k)_2 \in \mathbb{Z}$. Here we have extended the definition of (v_1, v_2) to allow for all valid forms of this quantity for the solutions of (4.1).

Now we have

$$[R_0] \cdot [R_k]^{-1} = (\frac{2}{3}) \begin{bmatrix} 2^{1-k} - 2^{k-1} & -2^{-k} + 2^k \\ 2^{-k} - 2^k & -2^{-k-1} + 2^{k+1} \end{bmatrix}.$$

Multiplying this matrix by $(v(k)_i)^T$ and equating the second line to $\Delta\varepsilon_{-2}$ (the second initial condition), we arrive at the following.

$$\Delta\varepsilon_{-2} = (2^k(-v_1 + 2v_2 - m(k)_1 + 2m(k)_2) + 2^{-k}(v_1 - \frac{1}{2}v_2 + m(k)_1 - \frac{1}{2}m(k)_2))(\frac{2}{3}).$$

We cannot choose integer $m(k)_1$, $m(k)_2$ to make the first bracket on the right zero. Letting $m(k)_1 = 2m(k)_2$ and equating the first line to the first initial condition as well, we have

$$\begin{aligned}\Delta\varepsilon_{-2} &= (2^k(-v_1 + 2v_2) + 2^{-k}(\tfrac{3}{2}m(k)_2) + 2^{-k}(v_1 - \tfrac{1}{2}v_2))(\tfrac{2}{3}),\\ \Delta\varepsilon_{-1} &= (2^{k-1}(-v_1 + 2v_2) + 2^{1-k}(\tfrac{3}{2}m(k)_2) + 2^{-k}(v_1 - \tfrac{1}{2}v_2))(\tfrac{2}{3}).\end{aligned}$$

The third terms on the right above go to zero as k goes to infinity. If we choose $m(k)_1$ to make $\Delta\varepsilon_{-2}$ go to zero as k goes to infinity, then $\Delta\varepsilon_{-1}$ will be $O(k)$. Similarly, $\Delta\varepsilon_{-2}$ will be $O(k)$ if $m(k)_1$ is chosen to make $\Delta\varepsilon_{-1}$ go to zero as k goes to infinity. Thus we cannot find arbitrarily small $\Delta\varepsilon_{-1}$, $\Delta\varepsilon_{-2}$ that will map to $\vec{v}$. This implies, from the formation of the initial conditions of (4.1) (with subscripts increased by 1) from (1.2), that no initial condition $\Delta\hat{\varepsilon}_0$ of the system with arbitrarily small coordinates $\Delta\varepsilon_{-2}$, $\Delta\varepsilon_{-1}$, Δr_{-1}, exists that will map to $\vec{v}$. Hence there is no such $\Delta\hat{\varepsilon}_0$ in an arbitrarily small neighbourhood about $\hat{x}_0 = 0$ that will map into V_1. Thus this example is not topologically transitive. ■

Example 1 will, nevertheless, satisfy density of periodic points by Theorem 5.12[7]. Example 2 below with an irrational nonminimum-phase zero fails the conditions of Theorem 5.12 and is shown not to satisfy density of periodic points. Note that it is also unstable and thus of theoretical interest only.

Example 2:

Consider the second-order system with $a_1 = 1$, $a_2 = -\eta$, $b_1 = -\eta$, $x_n = \Delta/2$, for all $n \geq 0$; so that $M = 2$ and $N = 1$. We set η to be an irrational number greater than 1. Then $p(z) = z^2 - (1+\eta)z + \eta$, with zeros $\mu_1 = \eta$ and $\mu_2 = 1$. This system satisfies condition

[7]Using the standard definition of density (with periodic points existing here), we make no such conclusion.

(R), since $\tilde{r}_1 = 1$, $\tilde{r}_i = 0$, for $i \geq 2$. The system is not internally stable since $|b_1| > 1$. Clearly $\hat{x}_0^* = 0$ is a periodic point with period $p = 1$.

Using the approach of the proof of Theorem 5.10, we have $(\vec{\beta}_i)^T = ([R_k] - [R_0])\vec{\alpha}$. From the example here, this gives

$$(\vec{\beta}_i)^T = \begin{bmatrix} \eta(\eta^k - 1) & 0 \\ \eta^k - 1 & 0 \end{bmatrix} \vec{\alpha}.$$

This gives $(m_1, m_2)^T \Delta = (\eta, 1)^T (\vec{\alpha}_q(\eta^k - 1))$, for m_1, $m_2 \in \mathbb{Z}$. The LHS vector has entries that are integers, while the RHS vector has entries that are of an irrational ratio. Thus a solution for $\vec{\alpha}_q$ and $\vec{\alpha}$ is not possible. Therefore, by the construction in the proof of Theorem 5.10, no periodic point lies in a small neighbourhood of the periodic point 0 and, in fact, no other periodic point exists at all. Thus this example does not have density of periodic points. ■

General cases that do not satisfy Theorems 5.10, 5.12 or their corollaries, or Theorem 5.9, and are not hyperbolic toral automorphisms in either situation, may be shown not to satisfy chaos condition 3 or 2 respectively by similar methods. An exception is for the case when some zeros are on the unit circle, when Theorem 5.12 and Corollary 5.13 extend to give Theorem 5.17 in the next subsection.

Thus we conjecture, without rigorous proof, that Theorem 5.9, or the existence of a hyperbolic toral automorphism, constitute necessary and sufficient conditions for chaos condition 2 to be satisfied when the system has an input x_n which satisfies certain "basic" (in some sense) conditions. We extend this conjecture of necessity to the combination of Theorems 5.10, 5.12 and 5.17, with Corollaries 5.11, 5.13 (which satisfy sufficiency), when the number of expansive zeros condition is adjusted to M from $\max(N, M)$ as in Theorem 5.9, for chaos condition 3. We will refrain from a further postulation of a precise

definition of these "basic" conditions, except to say that we take them to hold when the system has at least one periodic point, such as the point 0 in Examples 1 and 2. If the input is arbitrary, then these theorems represent sufficient conditions only. It is of particular interest to see Theorem 5.9 accomplish a role as a prospective necessary condition, since this theorem proves a somewhat stronger result than topological transitivity, namely that every neighbourhood in $\mathcal{C}^M$ contains a point that maps to any other point in $\mathcal{C}^M$. The existence of chaos via hyperbolic toral automorphisms suggests an added complexity in the nature by which chaos may arise, thereby posing a challenge in establishing the validity of the conjectures above.

Construction of higher-order examples of arbitrary form that satisfy condition (R) and are stable is difficult. There are natural restrictions that arise for simple examples. Thus we will not emphasize this in the rest of the thesis.

5.2.2 Minimum-Phase Results

It would seem reasonable to assume that condition (R) cannot be satisfied by a system which is strictly minimum phase (and perhaps not if marginally minimum phase with at least 1 zero inside the unit circle and no zeros of multiplicity greater than 1 on the unit circle). Such systems have mappings which are contractive on $\mathcal{C}$ and would thus cause some points in $\mathcal{C}$ to move apart as the length of the entire interval on the circle $\mathcal{C}$ shrinks. No proof of this will be presented here although the nature of such a proof can be surmised from the other work in this section. To allow for something to be said about these minimum-phase cases here, we suppose that the a_i and b_j do not satisfy condition (R) but that the mappings are otherwise continuous on some subset of $\mathbb{R}^N \times \mathcal{C}^M$. Allowing for this possibility when all zeros have magnitude 1, multiplicity 1 as well, we then have the following additional theorem:

Theorem 5.15 *Suppose the a_i and b_j satisfy condition (R), or that the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ is continuous on $\mathcal{C}$ at any $\vec{x}_0 \in U$, where $U \subset \mathbb{R}^N \times \mathcal{C}^M$, for any $n \geq 0$. Suppose also that none of the zeros of $p(z)$ have either magnitude greater than 1, or else magnitude equal to 1 with multiplicity greater than 1. Then chaos condition 1 (sensitivity to initial conditions) will not hold.*

Proof:

Let δ be a constant satisfying $0 < \delta \leq \dfrac{\Delta}{2}$. Choose any $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ such that $g_k(\hat{x}_0) \neq \Delta/2$, $k = 1, \ldots, M$, if (R) holds, and any $\hat{x}_0 \in U$ otherwise. Choose some neighbourhood $\tilde{N}$ of $\hat{x}_0$ with $\tilde{N} \subset U$ if (R) does not hold. Now we choose $\vec{y}_0 \in \tilde{N}$ and apply Theorem 4.4 (if (R) holds) and Theorem 4.5. Then, for $\Delta\vec{x}_0 = \vec{y}_0 - \hat{x}_0$, we have, from the nature of the zeros of $p(z)$ and (2.3), that $|\Delta\varepsilon_n| < K_{\vec{y}_0}$, $\forall\ n \geq -M$ for some bound $K_{\vec{y}_0} > 0$ ($\Delta\varepsilon_n = 0$, $\forall\ n \geq 0$ if $p(z) = 1$ as well). Now choose K such that $K > K_{\vec{y}_0}$, $\forall\ \vec{y}_0 \in \tilde{N}$. Such a bound K clearly exists. We may define a new neighbourhood about $\hat{x}_0$ in $\tilde{N}$ by $\tilde{N}_\alpha = \{\vec{y}_{\alpha 0} \in \mathbb{R}^N \times \mathcal{C}^M \mid \vec{y}_{\alpha 0} = \hat{x}_0 - \alpha(\hat{x}_0 - \vec{y}_0),\ \forall\ \vec{y}_0 \in \tilde{N}\}$, with $0 < \alpha < 1$. Now, with $\Delta\vec{x}_0 = \vec{y}_{\alpha 0} - \hat{x}_0$, by the properties of linear difference equations, the constants in (2.3) will be scaled down in magnitude by a factor of α, and hence so will $|\Delta\varepsilon_n|$. Thus $\alpha K > K_{\vec{y}_0}$, $\forall\ \vec{y}_0 \in \tilde{N}_\alpha$. Now we pick an $\alpha = \hat{\alpha}$ such that $\hat{\alpha} < \dfrac{\delta}{K}$ and $0 < \hat{\alpha} < 1$, which leads to the result that $|\Delta\varepsilon_n| < \delta$, $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$ and $\forall\ n \geq -M$. This then implies that $\|g_1 \circ f^n(\hat{x}_0) - g_1 \circ f^n(\vec{y}_0)\| < \delta$ for any $\vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$ and all $n \geq 0$. The δ was chosen arbitrarily. Now applying Lemma 5.6 we have the result that sensitivity to initial conditions does not hold. ■

Such a condition for the Σ-Δ modulator is reasonable. For example, if $p(z)$ has positive real zeros all of magnitude less than 1 and the input $x_n = \Delta/2$, then the mapping

for ε_n is continuous on $\mathcal{C}$ at any $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ satisfying $\varepsilon_{-1}, \varepsilon_{-2}, \ldots, \varepsilon_{-M} \neq \Delta/2$, for any $n \geq 0$. In such cases, the contractive nature of the mappings dominates the dynamics so as to ensure the relative continuity (and in this case nontransitivity as well). In general, for a given minimum or marginally minimum-phase system, there will exist simple general constraints on the input x_n to guarantee continuity of the mapping over $\mathbb{R}^N \times (\mathcal{C}^M | \varepsilon_{-1}, \varepsilon_{-2}, \ldots, \varepsilon_{-M} \neq \Delta/2)$ or some subset thereof, for all $n \geq 0$. Theorem 5.15 essentially implies that Theorem 5.8 provides necessary and sufficient conditions for chaos condition 1 to be satisfied, when (R) holds.

A proposition and some additional theorems pertaining to cases with marginally minimum-phase zeros, with coefficients satisfying condition (R), will now be given in this subsection.

Proposition 5.16 *Suppose the a_i and b_j satisfy condition (R). If all the zeros of $p(z)$ have magnitude 1 and multiplicity 1, and if the input x_n is constant, then chaos condition 2 (topological transitivity) is neither excluded nor assured.*

Proof:

(i) Consider the systems with the following maps: $v_n = v_{n-1} \mod 1$, with $\mu_1 = 1$; $v_n = -v_{n-1} \mod 1$, with $\mu_1 = -1$; $v_n = v_{n-2} \mod 1$, with $\mu_1 = 1$, $\mu_2 = -1$. The first two systems map points on $[0, 1)$ only to themselves or themselves and their negatives. The third system maps points in $[0, 1)^2$ only to themselves and their reflections about the line $v_n = v_{n-1}$. Thus none of these systems possess topological transitivity. A system with any number of magnitude 1, multiplicity 1 zeros that is nontransitive, may be constructed in this manner, e.g. with the map $v_n = v_{n-M} \mod 1$. These results show an application of Theorem 5.29 as well.

(ii) Now consider the system with the map $v_n = v_{n-1} + d \mod 1$, with $\mu_1 = 1$, where d is an irrational real number. The orbit of any initial condition in $[0, 1)$ will be quasiperiodic on $[0, 1)$. This implies that the orbit is densely distributed in $[0, 1)$, and that the mapping of v_{-1} can hence be made arbitrarily close to any other point in $[0, 1)$, for some sufficiently large n, and for any $v_{-1} \in [0, 1)$. Thus this system satisfies topological transitivity.

For a given irrational "input" d, examples of nontransitivity may generally be found for a system with magnitude 1, multiplicity 1 zeros only, of any order M, with $N = 0$, when $\mu_1 = 1$ is not a root of $p(z)$. For these cases considered, the initial condition $v_i = \dfrac{d \mod 1}{1 - \sum_{j=1}^{M} a_j}$, $i = -1, \ldots, -M$, (if $\sum a_j < 1$), and $v_i = \dfrac{d \mod 1 - 1}{1 - \sum_{j=1}^{M} a_j}$, (if $\sum a_j > 1$), is a fixed point, so that, by Theorem 5.18, transitivity does not hold. In all systems considered the a_i are integers thus satisfying (R). The proof is thus complete. ■

Theorem 5.17 [8] *Suppose the a_i and b_j satisfy condition (R). Suppose also that some of the zeros of $p(z)$ have magnitude 1 and multiplicity 1. Then*

1. *If $p(z)$ has at least q of these zeros μ_i with $q \geq 1$, such that the zeros are real or complex with argument $\theta_i = 2\pi\tilde{\mu}_i$ for some rational number $\tilde{\mu}_i$, then chaos condition 3 (density of periodic points) will hold. Moreover, if $q \geq M$, then either all points in $\mathcal{C}^M$ are periodic or none are.*

2. *If $p(z)$ has $\max(N, M)^* \geq 1$ such zeros that are all complex and do not satisfy condition 1 above, i.e. $q = 0$, then chaos condition 3 will hold if and only if there are no periodic points.*

[8] Using the standard definition of density, this result will hold if $q \geq M$, and will generally not hold otherwise. There is no conclusion regarding the existence of periodic points.

Proof:

1. The proof of this theorem follows the proof of Theorem 5.12, where the matrices, vectors, and m_θ involving q pertain to the q zeros as defined in Theorem 5.17, analogously to definitions in Theorem 5.12. We have the equation $(\vec{\beta}_i)^T = ([R_k] - [R_0])\vec{\alpha}$. Set $\vec{\alpha}_{\tilde{q}} = 0$. From the form of the zeros and $k = \hat{k}m_\theta$, we have $[R_{kq_1}]_{(q)} - [R_{0q_1}]_{(q)} = 0, \ \forall\ \hat{k} \geq 0$, and $M \leq q_1 \leq \max(N, M)$. This is the required form of the LHS of the equation with $m_{-i} = 0$, $i = 1, \ldots, q_1$. Condition 1(a) from Theorem 5.12 is also satisfied (by taking $\hat{m}$, $m_{-i} = 0$, in the proof of Theorem 5.10). Thus $\vec{\alpha}_q$ is a solution for any $\vec{\alpha}_q$ in the span of $[R_{0q}]^{-1}(\Delta\varepsilon_i)^T$, $(\Delta\varepsilon_i)^T \in \mathcal{C}^q$ (i.e. the first q of the $\Delta\varepsilon_i$), $\forall\ \hat{k} \in \mathbb{Z}^+$. Thus $\vec{\alpha}$ and $(\Delta\varepsilon_i)^T$ may be arbitrarily small in magnitude. The rest of the proof follows the proof of Theorem 5.10 to the end.

If $q \geq M$, then, without loss of generality, we may consider the first M of the q zeros and set $q = M$, so that $\vec{\alpha}_M$ is a solution for any $\vec{\alpha}_M$ in the span of $[R_{0M}](\Delta\varepsilon_i)^T$, $(\Delta\varepsilon_i)^T \in \mathcal{C}^M$, $\forall\ \hat{k} \in \mathbb{Z}^+$. Thus every point in $\mathcal{C}^M$ is a periodic point, given that the existence of a periodic point is first assumed. Thus the second result follows.

2. We follow the development in part 1. Let $M^* = \min(M, \max(N, M)^*)$. We denote $[M]^{(i)}$ to be the first i rows of a matrix $[M]$. Now suppose there exists a nonzero $\vec{\alpha} \in \mathbb{R}^{\max(N,M)^*}$, and $p \in \mathbb{Z}^+$, such that $(\vec{\beta}_i)^T_{M^*} = ([R_p]^{(M^*)} - [R_0]^{(M^*)})\vec{\alpha} = (m_i\Delta)^T_{M^*}$, for $m_{-i} \in \mathbb{Z}$, $i = 1, \ldots, M^*$ (not possible if $N \leq M$, since matrix in brackets is then square and nonsingular). From the form of the zeros, the entries of $[R_k]$ will behave quasiperiodically as k increases, such that $([R_{\hat{k}p}]^{(M^*)} - [R_0]^{(M^*)})\vec{\alpha} \neq (m_{\hat{k},i}\Delta)^T_{M^*}$ for some $m_{\hat{k},-i} \in \mathbb{Z}$, $i = 1, \ldots, M^*$, $\forall\ \hat{k} \geq 1$. Thus we cannot have permissible forms of $(\vec{\beta}_i)^T$ as needed for a nearby periodic point. Thus we conclude that density of periodic points cannot hold if a periodic point exists. This chaos condition then holds only if no periodic points exist. ■

This proposition and theorem expand the chaos condition analysis for cases (marginally minimum phase) which violate the sensitivity condition and are therefore not chaotic. From Proposition 5.16 we see that topological transitivity may or may not hold, with the nature of the system input x_n being a simple factor in determining which is true. Specifically, we have examples of transitivity/nontransitivity for periodic irrational/periodic rational inputs respectively (the inputs are in fact constant). Note that small perturbations in the irrational/rational input may maintain transitivity/nontransitivity. These general results make sense since these cases are transitional between nonminimum phase where transitivity holds, and strictly minimum phase where transitivity will be seen to be problematic.

The higher-order solution form for ε_n in (4.2) suggests that the quasiperiodic behaviour of the first order transitive case with irrational input in the proof of Proposition 5.16 may extend to higher-order cases with periodic input forms. A higher-order extension will in fact be shown when considering long run error behaviour (on one error coordinate) in Theorems 7.13 and 7.14 of Chapter 7. From (4.2), it is also apparent that this behaviour is unlikely to occur independently over $M > 1$ dimensions to bring about topological transitivity without a more irregular input.

We see with Theorem 5.17 that density of periodic points is guaranteed when there are some zeros on the unit circle with multiplicity one, unless these are all complex with phase angles that are not rational fractions of 2π. In this case, condition 3 may hold only by default, if there are no periodic points. The existence of just one periodic point here (for a given filter) is always possible with an appropriate input x_n. If at least M phase angles (of the zeros) are rational fractions of 2π, then periodic points are not only dense, but if they exist, they make up all points in $\mathcal{C}^M$. Theorem 5.17 is essentially a more powerful extension of Theorem 5.12 to cases with at least one zero on the unit circle (with multiplicity one). The unitary magnitude nature of the zeros allows the cyclic

component of the error solutions to automatically form additional periodic points when one is known to exist, without any further restrictions on the zeros. Interestingly, a path to chaos exists — a system with $N > M$, M nonminimum-phase zeros, and another zero satisfying Theorem 5.17 part 1, will be chaotic from the results in this section. The results of both parts of Theorem 5.17 show, rather counterintuitively, that density of periodic points tends to become more prevalent when we go from the weakly nonminimum to marginally minimum-phase case under certain circumstances.

We now present a theorem that will permit a further identification of conditions for chaos in marginally minimum-phase cases.

Theorem 5.18 *Suppose the a_i and b_j satisfy condition (R). Suppose also that all the zeros of $p(z)$ have magnitude 1 and multiplicity 1. Suppose further that there exists an $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ such that the set $\{g_M(\hat{x}_n),\ \forall\ n \geq 0\}$ is finite. Then chaos condition 2 (topological transitivity) will not hold.*

Proof:

Let $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ be as given in the theorem, with the number of elements in the set $\{g_M(\hat{x}_n),\ \forall\ n \geq 0\}$ given by p. Let $\tilde{N}$ be some neighbourhood of $\hat{x}_0$. Now we choose a $\hat{y}_0 \in \tilde{N}_x$ as given in Lemma 4.10. Then for $\Delta\vec{x}_0$ we have, from the nature of the zeros of $p(z)$ and (2.3), that $|\Delta\varepsilon_n| < K_{\hat{y}_0},\ \forall\ n \geq -M$ for some bound $K_{\hat{y}_0} > 0$. Now choose K such that $K > K_{\hat{y}_0},\ \forall\ \vec{y}_0 \in \tilde{N}_x$. Such a bound K clearly exists. We may define a new neighbourhood about $\hat{x}_0$ in $\tilde{N}_x$ by $\tilde{N}_\alpha = \{\vec{y}_{\alpha 0} \in \mathbb{R}^N \times \mathcal{C}^M \mid \vec{y}_{\alpha 0} = \hat{x}_0 - \alpha(\hat{x}_0 - \vec{y}_0),\ \forall\ \vec{y}_0 \in \tilde{N}_x\}$, with $0 < \alpha < 1$. Now, with $\Delta\vec{x}_0 = \vec{y}_{\alpha 0} - \hat{x}_0$, by the properties of linear difference equations, the constants in (2.3) will be scaled down in magnitude by a factor of α, and hence so will $|\Delta\varepsilon_n|$. Thus $\alpha K > K_{\vec{y}_0},\ \forall\ \vec{y}_0 \in \tilde{N}_\alpha$. Now we pick $\alpha = \hat{\alpha}$ such that $\hat{\alpha} < \dfrac{\Delta}{4Kp}$ and $0 < \hat{\alpha} < 1$, which leads to the result that $|\Delta\varepsilon_n| < \dfrac{\Delta}{4p},\ \forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$ and $\forall\ n \geq -M$. Thus the mapping

$g_1 \circ f^n(\tilde{N}_{\hat{\alpha}}) \subset [g_1 \circ f^n(\hat{x}_0) - \frac{\Delta}{4p},\ g_1 \circ f^n(\hat{x}_0) + \frac{\Delta}{4p}]$ holds on $\mathcal{C}\ \forall\ n \geq 0$. From the finite size of the orbit set for $\hat{x}_0$ above, we have that $g_1 \circ f^n(\tilde{N}_{\hat{\alpha}}) \subset V_1 \subseteq \mathcal{C}$, where V_1 is the union of p closed intervals on $\mathcal{C}$, each of length $\Delta/2p$, $\forall\ n \geq 0$. Then V_1 covers at most one half of $\mathcal{C}$. Thus there must exist a $\hat{z} \in \mathcal{C}$ with some neighbourhood $\tilde{N}_z \subset \mathcal{C}$ about $\hat{z}$ such that $g_1 \circ f^n(\tilde{N}_{\hat{\alpha}}) \cap \tilde{N}_z = \emptyset$. Now applying Lemma 5.7, we have the result that topological transitivity does not hold.

If $N = 0$, the input x_n is periodic with period q, and the system has an arbitrary periodic point with period p_1, then, from the properties of the difference equations (1.2), the requirements of the theorem are satisfied with the cardinality p satisfying $p \leq p_1 q$.

■

With this theorem we have, for the marginally minimum-phase cases of Proposition 5.16 and Theorem 5.17, that if the input x_n may be chosen such that the orbit of a point $\hat{x}_0$ covers finitely many points in its projection on $\mathcal{C}^M$, then with this input topological transitivity fails. Notably, this yields the consequence, for these phase cases, that topological transitivity and existence of a periodic point cannot coexist in some cases. Specifically, this will be true when a limit cycle orbit exists in $\mathcal{C}^M$ so that every point in the orbit is a periodic point. If there are no feedforward elements in the filter ($N = 0$), and the input x_n is periodic with period q, then from the form of the difference equations (1.2), we see that any periodic point that exists with period p_1 must lie on a limit cycle orbit with period $p = p_1 q$.

It would seem reasonable to conjecture that this result may be extended to the case where $M \geq N$. In such cases, if a periodic point exists with periodic input, and we assume that the point is not on a limit cycle orbit, we then require a different solution for each p to a non-underdetermined linear system in N variables r_i, $i = -1 + p, \ldots, -N + p$,

$p \geq 0$, (satisfying M constraints) up to mod Δ. Since we must also have the infinite sets of N solutions r_i bounded for stability, this suggests the general nonexistence of any other solution, implying that the point is on a limit cycle. Conversely, if $M < N$, it seems possible to have infinite sets of N r_i lying on a bounded $N - M$ dimensional manifold in $\mathbb{R}^N$ that give rise in (1.2) to a periodic point that is not on a limit cycle orbit when the input is periodic. For the $M < N$ case, the input for example would require some stricter constraint for Theorem 5.18 to be applicable via a finite set.

Thus the examples (with $N = 0$) used in the proof of Proposition 5.16 where topological transitivity holds, have no periodic points and thereby trivially satisfy chaos condition 3 as well. The marginally minimum-phase cases from Theorem 5.17 with periodic points on limit cycle orbits via some periodic or other form of input fail topological transitivity.

A summary of the classification of the various cases and sub-cases presented in this section, in terms of whether or not the three conditions for chaos and overall chaos hold, is given in Tables 5.1 and 5.2 (the first two tables) at the beginning of Section 5.4 at the end of this chapter.

5.3 General Model

To continue with the analysis for chaos conditions, we now turn to the general Σ-Δ modulator form with no constraints on the feedback and feedforward gains in the filter, that is $a_i \in \mathbb{R}$ and $b_j \in \mathbb{R}$ for $i = 1, 2, \ldots, M$; $j = 1, 2, \ldots, N$. With this context, we no longer have recourse to Theorem 4.4 or Theorem 4.5 and Corollary 4.6. When condition (R) does not hold, continuity in the mappings for all $n \geq 0$ does not hold over any simply connected subset of $\mathbb{R}^N \times \mathcal{C}^M$, except possibly over a strict such subset for minimum or marginally minimum-phase cases. Thus we face more complications in the analysis. Ultimately, broad

results may be arrived at, although our picture is of a more qualified and less comprehensive nature compared to the previous context arising with condition (R). With this complexity, come new cases of interest and investigation.

To begin, we first discuss the framework by which we approach the limitations of applying Proposition 4.1 and what happens when it breaks down. Considering two nearby initial conditions $\vec{x}_0,\ \vec{y}_0 \in \mathbb{R}^N \times \mathcal{C}^M$, we have that, for any $n_1 \geq N$, there must exist a finite sequence of points $\vec{y}_{\alpha_{n_1,i}} \in \mathbb{R}^N \times \mathcal{C}^M$, of the form $\vec{y}_{\alpha_{n_1,i}} = \vec{x}_0 + \alpha_{n_1,i}(\vec{y}_0 - \vec{x}_0)$ with $0 \leq \alpha_{n_1,i} \leq 1,\ \alpha_{n_1,i} < \alpha_{n_1,(i+1)},\ i = 1, 2, \ldots,$ such that Proposition 4.1 holds (i.e. $\Delta\varepsilon_n$ satisfies (4.1) for $N \leq n \leq n_1$) for $\vec{x}_{1,0} = \vec{y}_{\hat{\alpha}_{1,0}},\ \vec{x}_{2,0} = \vec{y}_{\hat{\alpha}_{2,0}}$, where $\alpha_{n_1,i} < \hat{\alpha}_{1,0} < \hat{\alpha}_{2,0} < \alpha_{n_1,(i+1)}$ for all such i. This follows from having equality of the quantizers holding in the second line of the proposition for the respective cases. Taking the limit as the points $\vec{x}_{1,0}$ and $\vec{x}_{2,0}$ approach $\vec{y}_{\alpha_{n_1,i}}$ and $\vec{y}_{\alpha_{n_1,(i+1)}}$ respectively, for each i, we have the result that Proposition 4.1 essentially holds for each "subinterval" of a subdivision of the "interval" $[\vec{x}_0, \vec{y}_0]$ in $\mathbb{R}^N \times \mathcal{C}^M$. The error difference interval $\Delta\varepsilon_n$ on $\mathcal{C}$ for each initial subinterval is interpreted not to contain the point $\Delta/2$ for $N \leq n \leq n_1$ except possibly at endpoints, as would normally be true when the proposition holds. As n_1 is allowed to increase, the initial subintervals of $[\vec{x}_0, \vec{y}_0]$ will become smaller and more numerous. The error difference interval $\Delta\varepsilon_n$ on $\mathcal{C}$ for $\vec{x}_0$ and $\vec{y}_0$ is thus interpreted as "split" into at least two error difference subintervals when Proposition 4.1 breaks down for the smallest such n. These error difference subintervals then split into further subintervals as n increases and the proposition breaks down for the respective difference intervals, and so on. We have no general means of relating the position or orientation on $\mathcal{C}$ of the split error difference intervals relative to each other as n increases (unlike the case when (R) holds in which the "split" error difference interval remains joined).

5.3.1 General Results

The basic theorems concerning the three conditions of chaos in the most general Σ-Δ modulator context will now be presented.

Theorem 5.19 *Suppose $p(z)$ has a zero with either magnitude greater than 1, or else magnitude equal to 1 with multiplicity greater than 1. Then chaos condition 1 (sensitivity to initial conditions) will hold.*

Proof:

Suppose condition (R) does not hold. Let $d = \min(\tilde{r}_t\Delta \mod \Delta,\ \Delta - \tilde{r}_t\Delta \mod \Delta)$, where $t = \min\{k \mid \tilde{r}_k \text{ is not an element of } \mathbb{Z}, \text{ where the } \tilde{r}_k \text{ are from (R)}\}$. Let δ be a constant satisfying $0 < \delta < \frac{d}{2}$. Choose any $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and some neighbourhood $\tilde{N}$ of $\hat{x}_0$. The results of Lemma 4.10 clearly hold when $\tilde{N}_x$ is replaced by $\tilde{N}$. We thus choose a $\hat{y}_0 \in \tilde{N}$ to satisfy part 2 of Lemma 4.10.

Suppose that the result of Proposition 4.1 holds for $N \leq n \leq n_1$, for some n_1. Then, for $\Delta\vec{x}_0$ we have, from the nature of the zeros of $p(z)$ and (2.3), that $\limsup_{n\to\infty} |\Delta\varepsilon_n| = \infty$ as $n_1 \to \infty$. Thus $\exists\ n_2 = \min\{n \geq 0 \mid |\Delta\varepsilon_n| > \delta\}$ with $n_2 \leq n_1$, if n_1 is sufficiently large. If we replace the $\hat{y}_0$ with $\hat{y}_{\alpha 0} = \hat{x}_0 - \alpha(\hat{x}_0 - \hat{y}_0)$ in $\Delta\vec{x}_0$, with $0 \leq \alpha \leq 1$, then, by the properties of linear difference equations, the constants in (2.3) will be scaled down in magnitude by a factor of α, and hence so will the applicable $|\Delta\varepsilon_n|$. Suppose there exists $n_3 = \min\{k \mid Q(x_k - r_{y,k}) - Q(x_k - r_{x,k}) \neq 0,\ k \geq 0\}$, where the variables with subscripts x and y correspond to the system with initial conditions $\hat{x}_0$ and $\hat{y}_0$ respectively. Then, from propagating this result through (1.2) (as with the proof of Theorem 4.4), the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ will have one or more jump discontinuities of magnitude $|\tilde{r}_t|\Delta$, when $n = n_3 + t$ at the finite points $\{z_i\}$ (where $(x_{n_3} - r_{z_i,n_3}) = m_i\Delta$, $m_i \in \mathbb{Z}$) over the interval defined by $I_{\vec{x}_0} = \{\hat{y}_{\alpha 0},\ 0 \leq \alpha \leq 1\}$. The function $g_1 \circ f^{n+1}(\vec{x}_0)$ will be continuous on $I_{\vec{x}_0}$

for $0 \leq n \leq n_3 + t - 1$, and on $I_{\vec{x}_0} - \{z_i\}$ for $n = n_3 + t$. Thus, from Theorem 4.5 and the above, with $\alpha = 1$, $\Delta\varepsilon_n$ from (2.3) is valid for $N \leq n \leq n_3 + t - 1$, and this $\Delta\varepsilon_n$ summed with the jump discontinuities for $n = n_3 + t$. Clearly there exists either n_2 or n_3. Let $n_4 = \min(n_2, n_3 + t)$, ($n_4 = n_3 + t$ if n_2 does not exist). We may choose an $\alpha = \hat{\alpha}$ with $0 < \hat{\alpha} \leq 1$, such that $\delta < |\Delta\varepsilon_{n_4}| < \dfrac{\Delta}{2}$ if $n_4 = n_2 < n_3 + t$. If $n_4 = n_3 + t < n_2$, then $0 < |\Delta\varepsilon_{n_4}| < \delta$, and there can be only one jump discontinuity and z_i on the interval. Thus, in this case, we leave $\hat{\alpha} = 1$. If $n_4 = n_2 = n_3 + t$, then we choose $\hat{\alpha}$ such that $\delta < |\Delta\varepsilon_{n_4}| < \dfrac{d}{2}$. If a jump discontinuity remains on $\{\hat{y}_{\alpha 0},\, 0 \leq \alpha \leq \hat{\alpha}\}$, then there can be only one and one z_i. In all cases this then implies that $\delta < \|g_1 \circ f^{n_4+1}(\hat{x}_0) - g_1 \circ f^{n_4+1}(\hat{y}_{\hat{\alpha}0})\|$, with $n_4 + 1 > 0$ and $\hat{y}_{\hat{\alpha}0} \in \tilde{N}$. Now applying Lemma 5.6, we have the result that sensitivity to initial conditions holds. If condition (R) holds we have this result from Theorem 5.8.

■

Theorem 5.20 *Suppose $p(z)$ has at least M zeros, where each zero has either magnitude greater than 2, or else magnitude equal to 2 with multiplicity greater than 1 [counted (multiplicity − 1) times]. Then chaos condition 2 (topological transitivity) will hold.*

Proof:

For this proof, we apply the matrix notation given at the end of Chapter 4. We shall take $q = M$ here. Let U_1 be any open set with $U_1 \subset \mathbb{R}^N \times \mathcal{C}^M$. Choose an $\hat{x}_0 \in U_1$, such that $g_{k_i}(\hat{x}_0) \neq \Delta/2$, $\forall\ k_i \in \{1, \ldots, M\}$. Let $\tilde{D}$ be a small q dimensional subset of U_1 defined by

$$\tilde{D} = \{\vec{y}_0 \in U_1 \mid \vec{y}_0 - \hat{x}_0 = \Delta\vec{x}_0 \text{ has the form } \Delta\vec{x}_0 = (\Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-N};\, \Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-M});$$
$$g_{k_i}(\vec{y}_0) \neq \Delta/2,\ \forall\ k_i \in \{1, \ldots, M\};\ \text{and } \vec{\alpha}_{\tilde{q}} = 0\}.$$

Let $\tilde{D}_Z = \{\vec{z}_0 \mid \vec{z}_0 = \vec{y}_0 - \hat{x}_0,\ \forall\ \vec{y}_0 \in \tilde{D}\}$.

Now we define the sets D_k and E_k, where $k = \hat{k}M$, $\hat{k} \in \mathbb{Z}^+ + \{0\}$, with the following. First, we define an aligned M-d square to be a closed rectangular region of M dimensional volume in $\mathbb{R}^M$, with all edges parallel to the ε_i coordinate axes, and of equal length. Now let C_{D_k} be the set of all so defined aligned M-d squares $C_i \subseteq D_k$ satisfying $g_1 \circ f^j(\hat{x}_0 + \vec{z}_0) \neq \Delta/2$, $\forall\ \vec{z}_0 \in \tilde{D}_Z$, such that $[R_{kq}]\vec{\alpha}_q \in C_i \subseteq D_k$, for $j = k-1, \ldots, k-M$, $\hat{k} \geq 1$. Then we define $E_k = \{C_l \in C_{D_k} \mid V(C_l) \geq V(C_i), \forall\ C_i \in C_{D_k}\}$, for $\hat{k} \geq 1$, where $V(*)$ denotes the M dimensional volume of the set in its argument. The choice of C_l need not be unique. Now we apply the mapping of system (4.1) to $\tilde{D}_Z$, that is, over initial conditions in $\tilde{D}_Z$. From this we define the set $D_k \subset \mathbb{R}^M$ by

$$D_k = \{\Delta\vec{\varepsilon}_{q,k} = (\Delta\varepsilon_{k-1}, \ldots, \Delta\varepsilon_{k-q}) \mid \Delta\vec{\varepsilon}_{q,k} = [R_{kq}]\vec{\alpha}_q, \text{ where } \vec{\alpha} = [R_0]^{-1} g^*(\vec{z}_0),$$
$$\forall\ \vec{z}_0 \in \tilde{D}_Z, \text{ such that } [R_{(k-M)q}]\vec{\alpha}_q \in E_{k-M}\},$$

for $\hat{k} \geq 1$; where g^* is defined to be the projection from $\mathbb{R}^N \times \mathcal{C}^M$ onto $\mathbb{R}^N$ if $N > M$, and g otherwise. We also define $D_0 = g(\tilde{D}_Z)$. These D_k, for $\hat{k} \geq 1$, will be M dimensional convex regions in $\mathbb{R}^M$.

The magnitude of each of the M zeros of $p(z)$ is greater than 2. Therefore, by the properties of the linear mapping defined above with eigenvalues represented by the M zeros, the M-d volume of the mapping of any region in $\tilde{D}_Z$ will increase by a factor of greater than 2 (product of the eigenvalue magnitudes) in the successive mapping. From the recursive form of the mapping, state space stretching will occur in the successive $\varepsilon_{-M}, \ldots, \varepsilon_{-1}$ coordinate directions with successive mappings, and this process will cycle through to repeat every M mappings. With every eigenvalue magnitude greater than 2, there will then be an expansivity of greater than 2 in every state space direction over the cumulative effect of M successive mappings. It then follows that the largest aligned M-d square contained in the Mth mapping of a previous aligned M-d square will have an M-d volume of greater than 2^M times that of the original aligned M-d square.

For some $\epsilon > 0$, satisfying $\epsilon < |\mu_i| - 2$, $i = 1, \ldots, M$, where $|\mu_i|$ are the magnitudes of the M zeros; we then have the following:

$$V(\hat{D}_{k+M}) > (2(1+\epsilon))^M V(E_k), \text{ and } V(E_k) \geq \min(2^{-M} V(\hat{D}_k), V(\mathcal{C}^M)), \quad \hat{k} \geq 1,$$

where 2^{-M} is the maximum reduction of volume due to the region splitting continuity condition used in defining the C_i earlier (if original volume is less than $V(\mathcal{C}^M)$), and $\hat{D}_k$ is the largest aligned M-d square contained in D_k. Combining the results, we have $V(E_{k+M}) > \min((1+\epsilon)^M V(E_k), V(\mathcal{C}^M))$ for $\hat{k} \geq 1$. Clearly $\exists$ a $\hat{k}$ and corresponding K such that $V(D_K) > V(\mathcal{C}^M)$ and $V(E_K) = V(\mathcal{C}^M)$, so that $E_K \equiv \mathcal{C}^M$, when considering $\mathcal{C}^M$ as the projection of $\mathbb{R}^N$.

Now we define $\tilde{E}_k = \{\vec{y}_0 \in \tilde{D} \mid [R_{kq}]\vec{\alpha}_q \in E_k\}$ for $k \in \mathbb{Z}^+$. Here $\vec{\alpha}_q$ is as given in the definition of $\tilde{D}$. Next, choose $\hat{y}_{1,0}$, $\hat{y}_{2,0} \in \tilde{E}_K$. Then, from the construction of the E_i, we have that the conditions of Proposition 4.1 hold for $0 \leq n \leq K-1$, and hence the continuity of Theorem 4.5 holds for $0 \leq n \leq K$. Thus Theorem 4.5, with its initial condition formulation, may be applied. It then follows that the solution to (4.1), with initial conditions $\Delta\vec{\varepsilon}_0 = (\Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-\max(N,M)})$, satisfies $g_1(\Delta\vec{x}_n) \cong g_1 \circ f^n(\hat{y}_{2,0}) - g_1 \circ f^n(\hat{y}_{1,0})$ for $0 \leq n \leq K$, where the subscripts from $\Delta\vec{x}_0$ above are decreased by N. Now we have $\Delta\hat{\varepsilon}_0 = I_{x,2} - I_{x,1}$, where $I_{x,i}$ is the initial condition for (4.1) associated with $\hat{y}_{i,0} - \hat{x}_0$, $i = 1, 2$. From the respective definitions of $\vec{\alpha}_{i,q}$ with $q = M$, and the solution of (4.1) in (2.3), this leads to $(g \circ f^n(\hat{y}_{2,0}) - g \circ f^n(\hat{y}_{1,0}))^T \cong [R_{nq}](\vec{\alpha}_{2,q} - \vec{\alpha}_{1,q})$, and

$$(g \circ f^n(\hat{y}_{2,0}) - g \circ f^n(\hat{y}_{1,0}))^T = [R_{nq}](\vec{\alpha}_{2,q} - \vec{\alpha}_{1,q}) + (m_i)^T\Delta, \quad 0 \leq n \leq K,$$

where $(m_i)^T$ is a vector in $\mathbb{Z}^M$ depending on $\hat{y}_{1,0}$ and $\hat{y}_{2,0}$. Rearranging and fixing $\hat{y}_{1,0}$ gives

$$(g \circ f^K(\hat{y}_{2,0}))^T = [R_{Kq}]\vec{\alpha}_{2,q} + \tilde{c}(\hat{y}_{2,0}),$$

where $\tilde{c}(\hat{y}_{2,0}) = \vec{c} + (m_i)^T\Delta$, $\vec{c}$ is a constant vector in $\mathbb{R}^M$, $(m_i)^T\Delta$ depends on $\hat{y}_{2,0}$, and we set $n = K$.

Considering all $\hat{y}_{2,0} \in \tilde{E}_K$, it is clear from this, the definition of $\tilde{E}_K$, and the nature of the congruency property relating $g \circ f^n(\hat{y}_{2,0})$ above, that $g \circ f^K(\tilde{E}_K) = \hat{P}_{\mathcal{C}^1}(E_K + \vec{c})$, (i.e. projection of RHS onto $\mathcal{C}^M$). Thus $g \circ f^K(\tilde{E}_K) = \mathcal{C}^M$ from the previous volume result for D_K. This gives $g \circ f^K(U_1) = g \circ f^K(\tilde{D}) = \mathcal{C}^M$. Thus we have the result that topological transitivity holds. ■

Theorem 5.21 [9] *Suppose $p(z)$ has $\max(N, M)$ zeros, and that each zero has magnitude greater than 2. Suppose also that the input x_n is periodic. Then chaos condition 3 (density of periodic points) will hold and periodic points will exist.*

Proof:

We shall first provide a proof for the case when $N = 0$, so as to make the nature of the methods used more clear. The proof will then be extended to the case when N is arbitrary in value. In both cases, the proof follows as a continuation of the proof of Theorem 5.20, where we now take $q = \max(N, M)$.

1. With $N = 0$, we have $q = M$. Suppose that the input x_n is periodic with period p. It is clear from the previous development, that $E_k \equiv \mathcal{C}^M$, and $V(E_k) = V(\mathcal{C}^M)$ as deduced, will hold $\forall\ \hat{k}$ such that $k \geq K$. Thus we choose $\hat{k}$ so that $k = Kp$ and label this K_p.

Extending the results at the end of the proof for this, we have $(g \circ f^{K_p}(\hat{y}_{2,0}))^T \cong [R_{(K_p)q}]\vec{\alpha}_{2,q} + \vec{c}$, (LHS equals projection of RHS on $\mathcal{C}^M$) for some $\vec{c} \in \mathbb{R}^M$. From this we focus on a subset of the domain where there is a one-to-one mapping to $\mathcal{C}^M$, so as to define the inverse mapping $h : \hat{\mathcal{C}}_1^M \to g(\tilde{E}_{K_p}) \subset \hat{\mathcal{C}}_1^M$ given by

$$h(\vec{y}) = [R_{0q}] \cdot [R_{(K_p)q}]^{-1}(\vec{y} - \hat{c}) + g(\hat{x}_0).$$

[9] Using the standard definition of density, with existence of periodic points added as well, this result will still hold.

For this, E_{K_p} is taken to be a region of volume Δ^M that is equivalent to $\mathcal{C}^M$ (mod Δ). $\mathcal{C}_1^M$ is defined to be an aligned M-d square of side length Δ containing the origin of $\mathbb{R}^M$, representing $\mathcal{C}^M$, so that $g(\tilde{E}_{K_p}) \subset \mathcal{C}^M$ (in $\mathbb{R}^M$) $= \mathcal{C}_1^M$ is satisfied. We interpret $\mathcal{C}_1^M$ as a translation of E_{K_p}. This translation may be carried out in such a manner that, for a finite collection of disjoint subsets $\{H_i\} \subset \mathbb{R}^M$ satisfying $\bigcup_i H_i = E_{K_p}$, every point in H_i is translated a distance that is a multiple of Δ in each vector coordinate. We denote this translation by $\vec{e}(H_i)$, where the components of $\vec{e}(H_i)$ will vary by multiples of Δ over $H_i \subseteq E_{K_p}$. From the geometric properties of E_{K_p}, and the fact that $\tilde{E}_{K_p}$ is small, we can hence choose $H_1 \in \{H_i\}$ such that $g(\tilde{E}_{K_p}) \subset H_1 + \vec{e}(H_1) \equiv \hat{\mathcal{C}}_1^M$. We let $\hat{c} = \vec{c} + \vec{e}(H_1) \in \mathbb{R}^M$. Note that, if we let $\Delta\varepsilon_i$, $i = -1, \ldots, -M$, be the independent variables for points in $\tilde{E}_{K_p} - \hat{x}_0$, this effectively allows $g(\tilde{E}_{K_p} - \hat{x}_0)$, as an M dimensional subspace, to define the M dimensional subspace $\tilde{E}_{K_p} - \hat{x}_0$.

Clearly h is a contraction mapping on $\hat{\mathcal{C}}_1^M$, since the eigenvalues of the matrix $[R_{(K_p)q}] \cdot [R_{0q}]^{-1}$ are all of magnitude greater than 1 (by the zeros of $p(z)$), and the inverse $[R_{0q}] \cdot [R_{(K_p)q}]^{-1}$ thus has entries that are all of magnitude less than 1. Thus, from the contraction mapping theorem, h has a unique fixed point in $g(\tilde{E}_{K_p})$. This implies that the solution

$$\vec{y} = \vec{y}_{3,0} = [R_{0q}] \cdot ([R_{(K_p)q}] - [R_{0q}])^{-1}(g(\hat{x}_0) - \hat{c}) + g(\hat{x}_0) \in g(\tilde{E}_{Kp}),$$

since it is a fixed point of the inverse mapping h.

We may thus apply the $(g \circ f^{K_p}(\hat{y}_{2,0}))^T$ equation above to $\hat{y}_{3,0}$, where $g(\hat{y}_{3,0}) = \vec{y}_{3,0}$, and this extension of $\vec{y}_{3,0}$ to $\hat{y}_{3,0} \in \mathbb{R}^N \times \mathcal{C}^M$ is the unique one satisfying $\hat{y}_{3,0} \in \tilde{E}_{K_p}$, as is implicit from the broader interpretation of how we defined the inverse mapping h above. Letting $\vec{\alpha}_{3,q} = [R_{0q}]^{-1}(\vec{y}_{3,0} - g(\hat{x}_0))$, substituting this into the RHS, and using the formula for $\vec{y}_{3,0}$ to substitute and eliminate $\vec{c}$; the equation reduces, after simplification, to $(g \circ f^{K_p}(\hat{y}_{3,0}))^T = \hat{P}_{\mathcal{C}^1}(\vec{y}_{3,0} - \vec{e}(H_1))$. Since the $\vec{e}$ term is a multiple of Δ, we have

the required result that $g \circ f^{K_p}(\hat{y}_{3,0}) = g(\hat{y}_{3,0})$ on $\mathcal{C}^M$. Since $N = 0$, this implies that $f^{K_p}(\hat{y}_{3,0}) = \hat{y}_{3,0}$. Since the system input x_n is periodic with period p, this mapping rule of f will be repeated, with initial conditions $\hat{y}_{3,K_p} = \hat{x}_0 + \Delta\vec{x}_{3,K_p}$, every p iterations. Thus we have $g \circ f^{K_p m}(\hat{y}_{3,0}) = g(\hat{y}_{3,0}), \ \forall\ m \geq 0$. Thus $\hat{y}_{3,0}$ is a periodic point. We also have $\hat{y}_{3,0} \in U_1$. It thus follows that periodic points exist and are dense in $\mathcal{C}^M$. The chaos result of density of periodic points thus holds.

2. Now we assume no restriction on the value of N. We follow the method of the proof of part 1, as a continuation of the proof of Theorem 5.20. For this, we now set $\tilde{D} = \{\vec{y}_0 \in U_1 \mid g_{k_i}(\vec{y}_0) \neq \Delta/2, \forall\ k_i \in \{1, \ldots, M\}\}$ here. Since the conditions of Proposition 4.1 hold (i.e. for $0 \leq n \leq K_p - 1$) over $\tilde{E}_{K_p}$, we have $\Delta r_{K_p - i} = \Delta\varepsilon_{K_p - i}$, for $i = 1, \ldots, \max(N, M)$, for the coordinates of $\Delta x_{K_p} = \vec{y}_{K_p} - \hat{x}_{K_p}$, when $\vec{y}_0 \in \tilde{E}_{K_p}$. The equation in the proof of Theorem 5.20 may then be extended to give

$$(f^n(\hat{y}_{2,0}) - f^n(\hat{y}_{1,0}))^T \cong [\frac{R_n(M)}{R_n(N)}](\vec{\alpha}_{2,q} - \vec{\alpha}_{1,q}), \quad 0 \leq n \leq K, \quad \hat{y}_{1,0}, \hat{y}_{2,0} \in \tilde{E}_{K_p},$$

where $[R_k(M)]$, $[R_k(N)]$ are matrices with the first M and N rows of $[R_k]$ respectively, and $q = \max(N, M)$. This leads to the analogous equation

$$(f^{K_p}(\hat{y}_{2,0}))^T \cong [\frac{R_{K_p - \max(N,M)}(M)}{R_{K_p - \max(N,M)}(N)}]\vec{\alpha}_2 + \vec{c},$$

for some $\vec{c} \in \mathbb{R}^{M+N}$, and dropping the q from α. We choose $k = K_p$ large enough so that $K_p - \max(N, M) > 0$.

Now we provide the relationship between $\vec{\alpha}_2$ and the associated initial conditions $\Delta\varepsilon_i$, Δr_i. First we have $\vec{\alpha}_2 = [R_0]^{-1}(\Delta\vec{\varepsilon}_0)^T$, where $(\Delta\vec{\varepsilon}_0) = (\Delta\varepsilon_{\max(N,M)-1}, \ldots, \Delta\varepsilon_0)$ here. Now, from iteration of the difference equations of (1.2), with the results from Theorem 4.5, we have the following:

$$\Delta\varepsilon_k = \sum_{i=1}^{k}(a_i - b_i)\Delta\varepsilon_{k-i} + \sum_{i=k+1}^{\max(N,M)}(a_i\Delta\varepsilon_{k-i} - b_i\Delta r_{k-i}), \quad k = 0, \ldots, \max(N, M) - 1.$$

With the appropriately defined $\max(N,M) \times \max(N,M)$ matrix $[A]$, and $\max(N,M) \times (N+M)$ matrix $[B]$, this leads to the matrix equation $[A](\Delta\vec{\varepsilon}_0)^T = [B]\Delta\hat{z}$, where $\Delta\hat{z} \in \mathbb{R}^N \times \mathcal{C}^M$ is the associated initial condition. It can be seen that $[A]$ is invertible, so that $\vec{\alpha}_2 = [R_0]^{-1} \cdot [A]^{-1} \cdot [B]\Delta\hat{z}$.

We may define the analogous noninverse mapping $h : \breve{E}_{K_p} \to \hat{\mathcal{C}}_1^{N+M}$, where $\breve{E}_{K_p} \subset \hat{\mathcal{C}}_1^{N+M}$, given by

$$h(\vec{y}) = [C](\vec{y} - \hat{x}_0) + \hat{c},$$

where $[C] = [\frac{R_{K_p - \max(N,M)}(M)}{R_{K_p - \max(N,M)}(N)}] \cdot [R_0]^{-1} \cdot [A]^{-1} \cdot [B]$. $\breve{E}_{K_p}$ has the definition corresponding to that of $\tilde{E}_{K_p}$, with respect to the extension of $\tilde{D}$ used here. Note that $[C]$ is not invertible. The specifications of the proof then follow analogously to that of part one, including the following: $f^{K_p}(\breve{E}_{Kp}) = \mathcal{C}_1^{N+M}$; $\mathcal{C}_1^{N+M}$ is the $N+M$ dimensional extension of E_{K_p} (i.e. (ε_{-i} coordinate subspace of E_{K_p}) $\times$ (ε_{-j} coordinate subspace of E_{K_p}); $i, j = 1, \ldots, M/N$ respectively) translated to contain the origin of $\mathbb{R}^{N+M}$, so that $\breve{E}_{K_p} \subseteq \mathcal{C}_1^{N+M}$ is satisfied.

Now we have the solution

$$\hat{y} = \hat{y}_{3,0} = ([C] - [I])^{-1}(\hat{x}_0 - \hat{c}) + \hat{x}_0.$$

This $\hat{y}_{3,0}$ will be of the same order of magnitude as the $\vec{y}_{3,0}$ in part 1, since the components of $[C]$ and $\hat{c}$ will, by the construction, be of the same order of magnitude. Thus we conclude here that $\hat{y}_{3,0} \in \hat{\mathcal{C}}_1^{N+M}$, and thus $\hat{y}_{3,0} \in \breve{E}_{K_p}$. We may then apply the $(f^{K_p}(\hat{y}_{2,0}))^T$ equation above to $\hat{y}_{3,0} \in \breve{E}_{K_p}$. Following the steps as in part one, we arrive at the required result that $f^{K_p}(\hat{y}_{3,0}) = \hat{y}_{3,0}$ on $\mathbb{R}^N \times \mathcal{C}^M$. The rest of the proof follows as in part one. ■

These theorems constitute an attempt to extend the results of Theorems 5.8, 5.9, 5.10 and 5.15 under the limitations introduced by the added complexity of having no constraints on the filter coefficients. They show that all three chaos conditions can be affirmed for the

more limited scenario of when the system has noise transfer function with all zeros of magnitude greater than 2, and when the input x_n is periodic.

The proof of the sensitivity condition under condition (R) extends directly to the general nonminimum and marginally minimum-phase (with magnitude 1 zero of multiplicity greater than 1) cases here, since the introduction of interval splitting only tends to enhance sensitivity (i.e. hasten divergence of nearby orbits). The proof for topological transitivity, however, requires M zeros of magnitude greater than 2 (or 2 and repeated) and the proof of density of periodic points requires not only this corresponding extension (with $\max(N, M)$ such zeros) but a periodic input as well. The natural reasons for these requirements can be seen. Sufficient expansivity (given by a zero of magnitude greater than 2) is needed to overcome the possible contractive effects of interval splitting in the mappings (i.e. a maximum contraction factor of 1/2 on each iteration when an interval, or more generally a region in $\mathcal{C}^M$, is split in half and the one half is mapped onto the other) and maintain transitivity. This is thus needed for our method of drawing on transitivity to show density of periodic points as well. In addition, the ad hoc nature of interval splitting means we cannot assume that any input x_n that allows a periodic point will cause a nearby point that repeats (in $\mathcal{C}^M$) once (in tandem with the periodic point) to continue to repeat so as to be periodic. Having a periodic input, however, allows this property to be preserved.

The central approach used in the proofs of Theorems 5.20 and 5.21, while similar to that of Theorems 5.9 and 5.10, is rather more subtle and abstract. For the latter, we directly applied a linear model equation (following from circle map properties) to solve for the required point in a neighbourhood. For the former here, we must first establish the existence of a mapping (analogous to that arising from the circle map) that maps a required neighbourhood to $\mathcal{C}^M$, construct it, and then, for Theorem 5.21, form the required solution point (from its linear equation) and show that it satisfies a recurrent point (e.g. when

$N = 0$ since the inverse mapping is a contraction). Theorem 5.19 supercedes Theorem 5.8, although we see the proof is technically simpler when (R) holds (Theorem 5.8).

Thus from these results we have only that for general systems, Devaney chaos is guaranteed when the noise transfer function has all zeros of magnitude greater than 2 and the input is periodic. As we shall see from a counterexample, however, the increase in our zero magnitude threshold boundary from 1 to 2 is, in general terms, not particularly conservative. In fact, there are important dynamical reasons that confirm our intuition of why chaos conditions 2 and 3 will not hold for systems with at least M zeros of magnitude greater than 1 but without M zeros of magnitude at least 2. Thus our stronger zero conditions of Theorems 5.20 and 5.21 are a feature of the more complex structure of the system, and not simply a reflection of the added difficulty in proving a clear cut result. The extended (number of) expansive zero requirements for Theorems 5.21 and 5.10 (condition 3) over those of Theorems 5.20 and 5.9 (condition 2) respectively also seem not overly conservative, from previous comments. This provides for a potential delineation between the presence of Li and Yorke versus Devaney chaos, which could be attributed to complex nonlinearities and/or discontinuities in the systems for which this differentiation arises.

Since condition (R) need not hold, we no longer require that $M > N$ to be assured of bounded internal stability. Thus, for Theorems 5.19 and 5.20, as opposed to their condition (R) counterparts, the distinction between M and $\max(N, M)$ in the zeros condition is one of practical as well as theoretical relevance.

From the consequences of Theorem 5.21, we have the following natural extension to the situation when condition (R) holds:

Corollary 5.22 [10] *Suppose the conditions of Theorem 5.10 hold, and $N = 0$. Suppose also that the input x_n is periodic. Then chaos condition 3 (density of periodic points) will hold, and periodic points will exist.*

Proof:

The proof of this corollary follows the proofs of Theorems 5.20 and 5.21, with $q = \max(N, M)$, and the following changes. Here, we define E_k to be simply the aligned q-d square of greatest q dimensional volume to be contained in D_k, for $\hat{k} \geq 1$ (may be nonunique). The magnitude of each of the q zeros of $p(z)$ is greater than 1. From analogous arguments to those in the proof of Theorem 5.20, we arrive at the following: $V(E_{k+q}) > (1+\epsilon)^q V(E_k)$, for $\hat{k} \geq 1$. We proceed and choose $\hat{y}_{1,0}$, $\hat{y}_{2,0} \in \tilde{E}_K$. Then, from condition (R), we have that Theorem 4.4 holds, so that Theorem 4.5 may be applied for $0 \leq n \leq K$. The remainder of the proof of Theorem 5.20, and that of Theorem 5.21, part 1, then follow, and lead to the required result. ■

This result and the analogous one of Theorem 5.21 essentially show that a periodic input will tend to bring about periodic points in a system with zeros that are all expansive. It also shows, under the pertinent circumstances, that if the condition of topological transitivity is achieved by the nonexistence of a periodic point, then either the input was not periodic, or these expansive zero conditions were not met. These results are consistent with what we expect from the relationships between periodic aspects of the system, and also point the way towards an analysis of more complicated sub-cases. Note that Corollary 5.22 does not extend more generally to cases when $N > 0$, because the proof relies on the continuity

[10] Using the standard definition of density, with existence of periodic points added as well, this result will still hold.

from condition (R), and not Proposition 4.1, as in the proof of part 2 of Theorem 5.21. This does not necessarily imply that a full extension would not hold, however.

We may now build upon these basic broad results in a rather piecemeal way to arrive at an expanded analysis of cases where chaos and the three chaos conditions do or do not hold, and hence a broader understanding of the nature and conditions of possible chaotic or nonchaotic behaviour in the Σ-Δ modulator.

First, we consider the case of a random input x_n, which allows us to take a more general dynamical perspective, and formulate some fairly broad results with the following two theorems:

Theorem 5.23 [11] *Suppose that either*

(a) for any $n_1 \geq 0$, there exists an $n_2 \geq n_1$ such that the input x_{n_2} is random and its projection $\hat{P}_{\mathcal{C}^1}$ is described by a discrete probability mass function defined over $\mathcal{C}$, with $\text{Prob}(\hat{P}_{\mathcal{C}^1}(x_{n_2}) = c \mid x_i,\ i = 0, \ldots, n_2 - 1) \leq K$ *for all $c \in \mathcal{C}$, for some K independent of n_1, with $0 < K < 1$; or*

(b) there exists an $n_2 \geq 0$ such that the input x_{n_2} is random and its projection $\hat{P}_{\mathcal{C}^1}$ is described by a piecewise continuous probability density function defined over $\mathcal{C}$.

Then the system has no periodic points and thus satisfies chaos condition 3 *(density of periodic points) with probability* 1*.*

Proof:

From the conditions of the theorem, there exists an infinite sequence of the form $\tilde{n}_k$, with $\tilde{n}_{k+1} > \tilde{n}_k$, $k \geq 1$, $\tilde{n}_1 \geq 0$, such that the input $x_{\tilde{n}_k}$ is random as described in the theorem. Let $h_k(x)$ be the probability mass/density function for the $x_{\tilde{n}_k}$ defined over $\mathcal{C}$,

[11] Using the standard definition of density, this result will still hold. If existence of periodic points is added to the density definition, then chaos condition 3 is satisfied with probability zero.

which is conditional upon the values of x_i, $i = 0, \ldots, \tilde{n}_k - 1$. Suppose $\exists\ \vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ such that $g \circ f^{mp}(\vec{x}_0) = g(\vec{x}_0)$, $\forall\ m \geq 1$, for some $p \in \mathbb{Z}$. Then, for any input x_{mp+i}, $i = 0, \ldots, p - M - 1$, $p > M$, $m \geq 0$, we have from (1.2) that the remaining input for $i = p - M, \ldots, p - 1$ is determined. Thus the input x_{mp+i}, $i = 0, \ldots, p - 1$, must exist on the same $\max(p - M, 0)$ dimensional manifold in $\mathcal{C}^p$, $\forall\ m \in \mathbb{Z}^+$. This manifold must not be parallel to one of the coordinate axes in $\mathbb{R}^p$, since, for a given m, the perturbation of one x_{mp+i} value requires the perturbation of others in order for the set to remain on the manifold. Now there exists an infinite sequence of $m_j \in \mathbb{Z}^+$, where $m_{j+1} > m_j$, such that $\tilde{n}_{k_j} = m_j p + i_j$ for some $i_j \in \{0, \ldots, p - 1\}$ with $\tilde{n}_{k_j} \in \{\tilde{n}_k\}$, $k \geq 1$, and $k_{j+1} > k_j$, $j \geq 1$. We drop k from the subscript of h_{k_j} above, and now let $h_{j,l}(x)$ be the PMFs and PDFs for the random inputs $x_{m_j p+l}$, $l \in \{0, \ldots, p - 1\}$.

(a) Suppose condition (a) holds and that $h_{j,d_j}(x)$ are the PMFs for the discrete random inputs $x_{m_j p+d_j}$, $d_j \in \{0, \ldots, p - 1\}$, and are defined for $\hat{P}_{\mathcal{C}^1}(x) = c_{i,j,d_j}$, $i = 1, \ldots, n_{j,d_j}$, $\forall$ d_j and $\forall\ j \geq 1$. Then we have the following:
Let A_j be the event that $\{x_{m_j p+i},\ i = 0, \ldots, p - 1\}$ and $\{x_i,\ i = 0, \ldots, p - 1\}$ exist on the same $\max(p - M, 0)$ dimensional manifold in $\mathcal{C}^p$, for $j \geq 1$. Suppose that $\text{Prob}(\bigcup_{j=1}^{k-1} A_j) > 0$ for some $k \geq 2$. Then we have

$$\begin{aligned} \text{Prob}(A_k \mid \bigcup_{j=1}^{k-1} A_j) &\leq \text{Prob}(A_k | x_i,\ i = 0, \ldots, m_k p - 1) \\ &\leq \max\{h_{k,d_k}(c_{i,k,d_k}),\ i = 1, \ldots, n_{k,d_k},\ \forall\ d_k\} \ = \ q_k, \end{aligned}$$

where $0 < q_k \leq K$. For this we take x_i, $i = 0, \ldots, m_k p - 1$, to be a particular realization that gives the event $\bigcup_{j=1}^{k-1} A_j$, with the maximum probability for event A_k. The final inequality follows since, at best, a strict subset of the possible random outcomes of $x_{m_k p+i}$, $i = 0, \ldots, p - 1$, is required for this set to lie on a given manifold, and the PMF of each random

$x_{m_k p+d_k}$ is bounded by K, independent of the outcomes of all previous inputs. (If any $h_{k,l}(x)$ is a PDF, then the LHS probability is zero, from (b) below.) Similarly we have, $\text{Prob}(A_1) \leq q_1$. From these results it follows that for $P_n = \text{Prob}(\bigcup_{j=1}^{n} A_j)$, $n \geq 1$, either $P_n = 0$, or else $P_n = \prod_{k=2}^{n} \text{Prob}(A_k \mid \bigcup_{j=1}^{k-1} A_j)\text{Prob}(A_1) \leq \prod_{k=1}^{n} q_k$. In either case, we have $\lim_{n\to\infty} P_n \leq \lim_{n\to\infty} K^n = 0$.

(b) Now suppose condition (b) holds. This implies at least one piecewise continuous PDF over $\mathcal{C}$, from the set $\{h_{j,l}(x)\}$, for some $j \geq 1$. It then follows that $\{x_{m_j p+i},\ i = 0, \ldots, p-1\}$ will be described by a joint PDF that is piecewise continuous over a manifold that is parallel to the coordinate axes in $\mathbb{R}^p$ corresponding to the continuous random inputs $x_{m_j p+l}$. This manifold will then either be of dimension greater than $\max(p-M, 0)$ (possible if there are greater than $\max(p-M, 0)$ continuously random $x_{m_j p+l}$), or else will intersect the required $\max(p-M, 0)$ manifold on which $\{x_i,\ i = 0, \ldots, p-1\}$ exists to form a set of measure zero. The latter property follows from the fact that the two manifolds, in this case, must not be parallel. In either case it then follows that

$$\text{Prob}(\{x_{m_j p+i},\ i = 0, \ldots, p-1\} \text{ and } \{x_i,\ i = 0, \ldots, p-1\} \text{ exist on the same } \max(p-M, 0) \text{ dimensional manifold in } \mathcal{C}^p) \ = \ 0.$$

Thus if either conditions (a) or (b) hold, the probability that $\vec{x}_0$ is a periodic point with period p as defined, is zero. If no such $\vec{x}_0$ exist for any p, then there are no periodic points. Since the choice of p was arbitrary, we conclude, with probability 1, that there are no periodic points and density of periodic points thus holds, when (a) or (b) are satisfied.

■

Theorem 5.24 *Suppose that, for any $n_1 \geq 0$, there exists an $n_2 \geq n_1$ such that the input x_{n_2+k}, $k = 0, \ldots, M-1$, is random and its projection $\hat{P}_{\mathcal{C}}$ is described by a piecewise continuous probability density function $h_{n_2}(\vec{x})$ defined over $\mathcal{C}^M$, and satisfying $h_{n_2}(\vec{x}|x_i, i = 0, \ldots, n_2 - 1) \geq K$, for all $\vec{x} \in \mathcal{C}^M$, for some K independent of n_1, with $K > 0$. Then chaos condition 2 (topological transitivity) will hold with probability 1.*

Proof:

Let U_1, V_1, be any open sets with $U_1 \subset \mathbb{R}^N \times \mathcal{C}^M$, $V_1 \in \mathcal{C}^M$. Choose an $\hat{x}_0 \in U_1$ and a $\hat{z} \in V_1$. Now let V_2 be some neighbourhood of $\hat{z}$ in V_1 defined by $V_2 = \{\vec{z} \in \mathcal{C}^M \mid \|g_i(\vec{z}) - g_i(\hat{z})\| < \delta,\ i = 1, \ldots, M\} \subset V_1$ for some $\delta > 0$. From the conditions of the theorem, there exists an infinite sequence of the form $\tilde{n}_k$, with $\tilde{n}_{k+1} > \tilde{n}_k + M - 1$, $k \geq 1$, $\tilde{n}_1 \geq 0$, such that the random input $x_{\tilde{n}_k+i}$, $i = 0, 1, \ldots, M-1$, is described by the piecewise continuous PDF $h_k(\xi_1, \ldots, \xi_M)$ defined over $\mathcal{C}^M$, which is conditional upon the values of x_i, $i = 0, \ldots, \tilde{n}_k - 1$. From the given conditions we have

$$q \equiv K\delta^M \leq \min\{\int_{z_1}^{z_1+\delta} \ldots \int_{z_M}^{z_M+\delta} h_k(\xi_1, \ldots, \xi_M) d\xi_1 \ldots d\xi_M \mid (z_1, \ldots, z_M) \in \mathcal{C}^M\},$$

$\forall\ k \geq 1$. This gives that, for some previous given input x_i, $i = 0, \ldots, \tilde{n}_k - 1$ (if $\tilde{n}_k > 0$);

$$\text{Prob}(g \circ f^{\tilde{n}_k+M}(\hat{x}_0) \in V_2 \mid x_i,\ i = 0, \ldots, \tilde{n}_k - 1, \tilde{n}_k > 0) \ \geq \ q,$$

where $0 < q < 1$, $\forall\ k \geq 1$. Thus $\text{Prob}(g \circ f^n(\hat{x}_0) \ni V_2,\ 0 \leq n \leq \tilde{n}_k + M) \leq (1-q)^k$, and $\lim_{k\to\infty} \text{Prob}(g \circ f^n(\hat{x}_0) \ni V_2,\ 0 \leq n \leq \tilde{n}_k + M) = 0$. Thus we conclude that with probability 1, $\exists\ \hat{n} > 0$ such that $g \circ f^{\hat{n}}(\hat{x}_0) \in V_2$, and hence such that $g \circ f^{\hat{n}}(U_1) \cap V_1 \neq \emptyset$. Thus we have the result that topological transitivity holds. ■

Theorem 5.23 shows that when the input possesses any sort of persistent randomness at all in its projection $\hat{P}_{\mathcal{C}^1}$ onto $\mathcal{C}$, chaos condition 3 (density of periodic points) is automatically satisfied regardless of the coefficients of the noise transfer function. The reason

is that the randomness ensures that no point in $\mathcal{C}^M$ can cycle back to itself with the same period, for ever, and thus no periodic points exist. Thus condition 3 is trivially satisfied. In fact, from part (b), a repeat cycle cannot happen even once following any iteration of input that is piecewise continuously random in its projection. Thus the persistence property is not required for the existence of any continuously random inputs to yield chaos condition 3.

Theorem 5.24 shows that, if in addition, the persistently random sequences x_{n_2+k}, $k = 0, \ldots, M-1$, may take on values that cover $\mathcal{C}^M$ in their projections $\hat{P}_{\mathcal{C}}$ onto $\mathcal{C}^M$, then chaos condition 2 (topological transitivity) will hold. The reason is that the randomness throughout (at least) $\mathcal{C}^M$ in the resulting quantizer input guarantees that the orbit of $(\varepsilon_{n-1}, \ldots, \varepsilon_{n-M})$ in $\mathcal{C}^M$ will be dense for any initializations and filter coefficients, thus giving transitivity. If the conditionality constraint on the probability density/mass functions in these theorems is relaxed, then independence over all n_1 must be imposed to maintain the full theorem results. Otherwise, it is possible to have cases that yield the respective chaos conditions with probabilities that are less than one.

Therefore in conclusion, when the randomness condition of Theorem 5.24 is satisfied by the input, Theorems 5.23, 5.24 and 5.19 allow us to extend the assertion of Devaney chaos beyond the cases covered under condition (R) for the gains in the filter, to those simply satisfying sensitivity — that is when the system is nonminimum phase or marginally minimum phase with a magnitude 1 zero of the noise transfer function having multiplicity greater than 1. Theorem 5.26 in turn will extend the applicability of this chaos condition 1 to minimum-phase limits.

The extension of chaos to non-strictly minimum-phase cases with an appropriate random input may suggest that all these cases (defined by the conditions for the zeros of $p(z)$) are indeed chaotic, when condition (R) does not hold, without any additional conditions. The

following proposition and counterexample proves this not to be true:

Example 3:

Consider the system with the map $v_n = [(1.5)v_{n-1} - 0.2] \mod 1$, with $\mu_1 = 1.5$, and where (R) does not hold. This system maps the interval $I = [(0.8), 1) \cup [0, (0.3)]$ to itself. This interval is also a trapping region for the mappings of all initial conditions in $\mathcal{C}$. This system has an isolated periodic point at $v_{-1} = 0.4$. All initial conditions in a small neighbourhood of 0.4 will, after a sufficient number of mappings, lie in the trapping region. ■

Proposition 5.25 *Suppose the a_i and b_j do not satisfy condition (R). If the largest magnitude zero of $p(z)$ has magnitude greater than 1 and less than 2, then chaos condition 2 (topological transitivity) and chaos condition 3 (density of periodic points) are neither excluded*[12] *nor assured.*

Proof:

(i) Consider Example 3. If $v_{-1} \in I$, then $v_n \in I$, $\forall\ n \geq 0$, so that v_n is not in an arbitrarily small neighbourhood of the point 0.4, for any $n \geq -1$. Thus topological transitivity does not hold. If v_{-1} is in a small neighbourhood of the periodic point 0.4, such that $v_{-1} \ni I$, then $v_n \in I$, $\forall\ n \geq n_1$ for some $n_1 \geq 0$, and v_{-1} is not a periodic point. Thus density of periodic points does not hold.

(ii) Now consider Example 4 given below. We define $\tilde{f} : [0, 1) \to [0, 1)$ as the mapping function here, where $\tilde{f}(x) = \frac{1}{2} - \frac{f(\Delta/2 - x\Delta)}{\Delta}$. Let $J = [c_{-1}, e_{-1}]$ be any interval on

[12] Using the standard definition of density, with existence of periodic points added as well, this result will still hold.

$[0,1)$ such that $J \subset [0,1)$. If we let $\Delta v_{-1} = e_{-1} - c_{-1}$, then in (2.3), we have $|\Delta v_n| \to \infty$ as $n \to \infty$. Thus $\exists\ n_1 > 0$ such that $0 \ni \tilde{f}^n(J)$, for $0 \le n < n_1$, and $0 \in \tilde{f}^{n_1}(J)$. We may express the last mapped interval as $\tilde{f}^{n_1}(J) = [z_a, 1) \cup [0, z_b]$, where $z_a = \tilde{f}^{n_1}(c_{-1})$, $z_b = \tilde{f}^{n_1}(e_{-1})$. Then clearly, $\tilde{f}^q([0, (\beta)^{-q})) = [0,1)$, where q is chosen so that $(\beta)^{-q} \in [0, z_b]$. Thus $\tilde{f}^{n_1+q}(J) = [0,1)$, and in particular, $\tilde{f}^{n_1+q}(z_{-1}, z_{-1} + (e_{-1} - z_{-1})\dfrac{(\beta)^{-q}}{z_b}) = [0,1)$ identically, where $z_{-1} \in J$, with $\tilde{f}^{n_1}(z_{-1}) = 0$. Since the interval J was arbitrary, this implies that topological transitivity holds.

From the corresponding scaling of $\tilde{f}^n(x)$ as x is scaled, we also have that $\tilde{f}^{n_1+q}(z_c) = z_c$, where $z_c = \dfrac{z_{-1}}{1 - (e_{-1} - z_{-1})\dfrac{(\beta)^{-q}}{z_b}}$, and $z_c \in J$. Since the input is constant, we then have $\tilde{f}^{(n_1+q)m}(z_c) = z_c$, $\forall\ m \in \mathbb{Z}$. Thus $\exists$ the periodic point $z_c \in J$. Since the interval J is arbitrary, it follows that periodic points are densely distributed throughout $[0,1)$. Thus density of periodic points holds. ■

Thus with a periodic (constant) input x_n and a zero of the noise transfer function of magnitude greater than 1 but less than 2, Example 3 is nonchaotic. The reason that transitivity fails is that, with interval splitting in the mappings, the input acts so as to induce a trapping region for the orbits in $\mathcal{C}$. The mapping has a repulsive fixed point (periodic point). Points on one side of the fixed point, where the backward displacement due to the input dominates, are drawn to the trapping region from one side, while points on the other side of the fixed point, where the forward displacement due to the expansivity arising from the nonminimum-phase property dominates, are drawn to the trapping region from the other side. Condition 3 (density of periodic points) also fails since the fixed point is isolated.

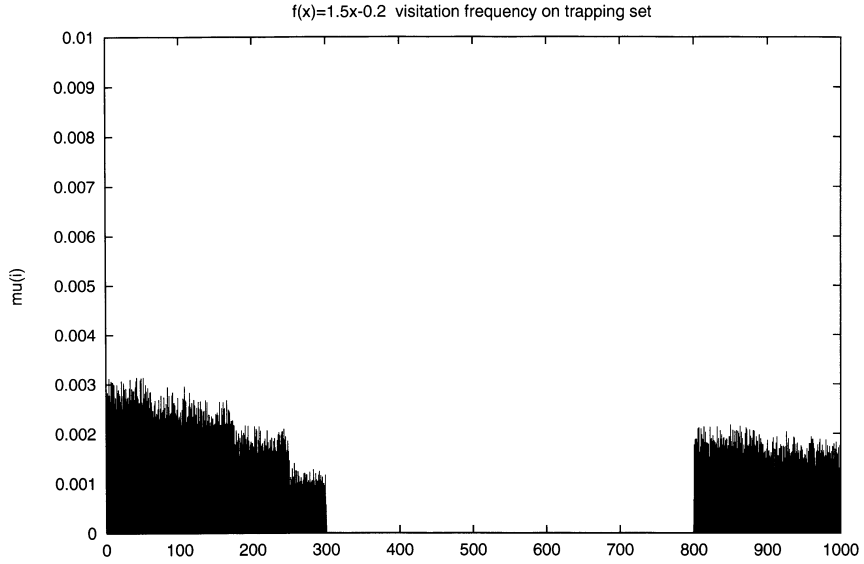


Figure 5.1: Visitation frequency for simulation of Example 3

Figure 5.1 gives a computer generated plot provided by [61] of the relative visitation frequency of the mapping on $\mathcal{C}$ in Example 3 for a simulation (with a particular initial condition) of many mapping iterations. The numerical values on the axes have been scaled. The horizontal axis represents the range $[0, 1)$ scaled by a factor of 1000, and the vertical axis the relative frequency of occurrence of each point in the range. The required trapping region is clearly evident on this plot. The plot shows the long run probability of the orbit visiting outside of the trapping region to be zero, as expected. The plot also shows the orbit visitation to be comparatively recurrent throughout the trapping region, although with a somewhat uneven probability distribution, at least over the finite number of iterations carried out. It may well be that conditions 2 and 3 hold strictly on the trapping region. The dynamics here certainly appear complex and suggestive of a chaotic nature. This, however, is not sufficient to meet the definition of chaos of our overall analysis and so we

classify Example 3 as nonchaotic. Taking the constant input $x_n = c$ and gain factor a_1, $1 < a_1 < 2$, as bifurcation parameters for the first-order system, one may clearly form conditions on these parameters under which the system remains topologically equivalent to Example 3, with a trapping region and isolated periodic point. We shall see in Example 4 such conditions violated by x_n.

Although the input used above was constant, clearly this periodicity is not crucial to the nonchaos, since we may perturb each x_n slightly while retaining a system with a sufficiently similar basic dynamical structure (at least in terms of transitivity). As well, it seems reasonable to assume that such nonchaotic counterexamples exist in systems of higher dimensionality, since the added complexity with this would tend to increase the possible range of dynamic possibilities. It may be conjectured that similar counterexamples that are nonchaotic exist for the marginally minimum-phase case with a zero of magnitude 1, multiplicity greater than 1, since such systems possess expansivity in the mappings that is even less dominant than that of any nonminimum-phase case.

To reinforce the notion of topological transitivity and density of periodic points for general nonminimum-phase systems that do not satisfy condition (R) and have periodic input, we have the following simple example:

Example 4:

Consider the system with the map $v_n = \beta v_{n-1} \mod 1$, with $\mu_1 = \beta$, and where $1 < \beta < 2$, so that (R) does not hold. ■

From the proof of Proposition 5.25, chaos conditions 2 and 3 clearly hold here when the system has 1 zero of magnitude greater than 1 and less than 2, with the constant input of $x_n = \beta\dfrac{\Delta}{2}$ (condition (R) does not hold). Note that from Theorem 5.8, chaos condition 1 holds as well so that Example 4 is chaotic.

If we consider extending the complexity of Example 4 to allow any form of periodic input and any order of system with its M zero(s) of greatest magnitude having either magnitude greater than 1 and less than 2, or else magnitude 1, multiplicity greater than 1, then we would expect that the more complicated dynamical structure in many such examples would only further ensure the prevalence of topological transitivity in these general cases. Similarly, it seems reasonable to believe that this very complexity would also lead to a prevalence of density of periodic points in such general cases with periodic inputs (including the case of example 4 above). In fact, we should not discount such prevalence when the M zeros condition is relaxed as well, whether (R) holds or not. It is this very complexity, however, that surpasses our intuition about it in terms of being able to construct more general examples or propositions to show this.

To further this discussion, we may conjecture, with consideration of Theorems 5.23 and 5.24, that specific deterministic (not necessarily periodic) inputs exist that give rise to topological transitivity and/or density of periodic points in all general nonminimum or marginally minimum-phase (with zeros of multiplicity greater than 1 on the unit circle) cases that do not satisfy condition (R). We may conjecture further that this result holds without the requirement of the M zeros, or those of Theorem 5.12 and Corollary 5.13, with (R) either holding or not. Such inputs, for example, would function dynamically in essentially the same manner as the random input of Theorems 5.23 and 5.24. This conjecture is also supported by the role of periodic and deterministic inputs that we will study for marginally minimum and minimum-phase systems where evidence for an analogous result will be shown for topological transitivity.

In summary, these conjectures taken together suggest that chaos is ubiquitous to all nonminimum and marginally minimum-phase (with zeros on the unit circle of multiplicity greater than 1) cases with different forms of input, analogously in particular to the proven

chaos, when condition (R) is satisfied, in spite of the counterexample which shows that chaos does not automatically hold if (R) does not hold. The "pseudo" chaos of the counterexample also hints at this ubiquitous nature of chaos here.

5.3.2 Minimum-Phase Results

We now investigate the marginally minimum and minimum-phase cases. To begin, we return to the case of random input x_n as a broad starting point, and have the following two theorems:

Theorem 5.26 *Suppose the a_i and b_j do not satisfy condition (R). Suppose also that $p(z)$ has a zero of magnitude 1 or greater. Suppose further that, for any $n_1 \geq 0$, there exists an $n_2 \geq n_1$ such that the input x_{n_2} is random and its projection $\hat{P}_{\mathcal{C}^1}$ is described by a piecewise continuous probability density function $h_{n_2}(x)$ defined over $\mathcal{C}$, and satisfying $h_{n_2}(x|x_i,\ i = 0,\ldots,n_2-1) \geq K$, for all $x \in \mathcal{C}$, for some K independent of n_1, with $K > 0$. Then chaos condition 1 (sensitivity to initial conditions) will hold with probability 1.*

Proof:

From the conditions of the theorem, there exists an infinite sequence of the form $\tilde{n}_k$, with $\tilde{n}_{k+1} > \tilde{n}_k$, $k \geq 1$, $\tilde{n}_1 \geq 0$, such that the input $x_{\tilde{n}_k}$ is random and described by the PDF $h_k(x)$ defined over $\mathcal{C}$ which is conditional upon the values of x_i, $i = 0,\ldots,n-1$.

Suppose that the largest magnitude zeros of $p(z)$ have magnitude 1, with multiplicity 1. Let $d = \min(\tilde{r}_t\Delta \mod \Delta, \Delta - \tilde{r}_t\Delta \mod \Delta)$, where $t = \min\{k \mid \tilde{r}_k$ is not an element of $\mathbb{Z}$, where the $\tilde{r}_k$ are from (R)$\}$. Let δ be a constant satisfying $0 < \delta < \frac{d}{2}$. Choose any $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and some neighbourhood $\tilde{N}$ of $\hat{x}_0$. The range of initial conditions $\Delta\vec{\varepsilon}_0 = \{\Delta\varepsilon_n$ for $n = N-1,\ldots,N-\max(N,M)^*\}$ of (4.1) that lead to the constants A_{ij}, B_{ij}, C_{ij} (corresponding to zeros of magnitude 1, multiplicity 1) all equalling zero constitutes

a $\tilde{q}$ dimensional subspace of initial conditions $C^{\tilde{q}}$, where $\tilde{q}$ is the number of zeros of $p(z)$ with magnitude less than 1, and $0 \leq \tilde{q} < \max(N, M)^*$. Thus there exists a $\max(N, M)^*$ dimensional subspace of initial conditions, excluding the set contained in the $\tilde{q}$ dimensional subspace, $\mathcal{C}^{\max(N,M)^*} - C^{\tilde{q}}$, over which not all of these constants are zero. Thus there must exist $\vec{y}_0 \in \tilde{N}$, such that the initial condition difference $\Delta\vec{x}_0 = \hat{x}_0 - \vec{y}_0$ will give rise to such an initial condition $\Delta\vec{\varepsilon}_0$ of (4.1), in the $\max(N, M)^*$ dimensional space, that is not in the $\tilde{q}$ dimensional subspace.

Further to this, $\exists$ a sequence $\{\hat{y}_{0,k}\}$, $k \geq 1$, of such $\vec{y}_0$ with the following additional property. Suppose that, when taking $\vec{y}_0 = \hat{y}_{0,k}$, that the conditions of Proposition 4.1 hold, for $0 \leq n \leq n_{1,k}$, for some $n_{1,k} \geq 0$. Then, for the $\Delta\vec{x}_0$ associated with $\hat{y}_{0,k}$, we have, from the nature of the zeros of $p(z)$ and (2.3), that $K_1 < |\Delta\varepsilon_{\tilde{n}_k}|$ and $|\Delta\varepsilon_n| < K_2$, $\forall\ 0 \leq n \leq n_{1,k}$, where $K_1 > 0$, $K_2 > 0$, are some constants independent of k. The existence of such a sequence of $\vec{y}_0$ in $\tilde{N}$ satisfying the lower bound requirement for any complex zeros scenarios, in particular, follows from the fact that, when considering (2.3), $|\min(\cos(\theta), \sin(\theta))| = \dfrac{1}{\sqrt{2}} > 0$ over all $\theta \in \mathbb{R}$; and from the fact that the subspace $C^{\tilde{q}}$ will allow choices of the B_{ij}, C_{ij}, associated with a given complex conjugate zero pair that span a two dimensional space.

Without loss of generality, we assume $\tilde{N}$ is small enough so that $K_2 < \delta$. From the condition of the theorem, we have $q \equiv KK_1 \leq \min\{\displaystyle\int_z^{z+K_1} h_k(x)dx \mid z \in \mathcal{C}\}$. Then

$$\text{Prob}[Q(x_{\tilde{n}_k} - r_{y,k,\tilde{n}_k}) - Q(x_{\tilde{n}_k} - r_{x,\tilde{n}_k}) \neq 0 \mid Q(x_i - r_{y,k,i}) - Q(x_i - r_{x,i}) = 0,$$
$$i = 0, \ldots, \tilde{n}_k - 1] \ \geq\ q, \quad \text{where } 0 < q < 1, \forall\ k \geq 1,$$

and where the variables with subscripts x and y, k correspond to the system with initial conditions $\hat{x}_0$ and $\hat{y}_{0,k}$ respectively. That this is true follows from the fact that we have inequality with the quantized values above, when the interval between the quantizer inputs on $\mathcal{C}$ contains the point 0. This interval has magnitude lying between K_1 and K_2. Thus

$$P_k \equiv \text{Prob}[Q(x_i - r_{y,k,i}) - Q(x_i - r_{x,i}) = 0,\, i = 0, \ldots, \tilde{n}_k] \leq (1-q)^k,$$

and $\lim_{k\to\infty} P_k = 0$. Thus, with probability 1, there must exist n_2 and $\hat{k}$, where $n_2 = \min\{i \mid Q(x_i - r_{y,\hat{k},i}) - Q(x_i - r_{x,i}) \neq 0\}$.

Following the method of the proof of Theorem 5.19, the function $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ will have one jump discontinuity of magnitude $|\tilde{r}_t|\Delta$, when $n = n_2 + t$ at the finite point $\hat{z}_0$ where $(x_{n_2} - r_{z,n_2}) = \Delta$, over the interval defined by $I_{\vec{x}_0} = \{\hat{y}_{\alpha 0,\hat{k}},\, 0 \leq \alpha \leq 1\}$. We then have that $|\Delta\varepsilon_n| < K_2$ for $0 \leq n \leq n_2 + t - 1$, from (2.3). At $n = n_2 + t$, the valid value of $\Delta\varepsilon_n$ is that given from (2.3) (magnitude bounded by K_2) summed with the jump discontinuity. This then implies that $\delta < \|g_1 \circ f^{n_2+t+1}(\hat{x}_0) - g_1 \circ f^{n_2+t+1}(\hat{y}_0)\|$, with $n_2 + t + 1 > 0$ and $\hat{y}_0 \in \tilde{N}$. Now applying Lemma 5.6, we have the result that sensitivity to initial conditions holds with probability 1. If there are zeros of $p(z)$ with magnitude greater than 1, or magnitude 1, multiplicity greater than 1, then this result holds from Theorem 5.19. ■

This result shows that when condition (R) on the a_i and b_j is not satisfied, then chaos condition 1 (sensitivity to initial conditions) may be extended to the marginally minimum-phase cases with zeros of multiplicity 1 only on the unit circle, if there is a persistently random input x_n that may take on values that cover $\mathcal{C}$ in their projections $\hat{P}_{\mathcal{C}^1}$ onto $\mathcal{C}$. Thus if the input is in the form of persistently random sequences as specified in Theorem 5.24, then we may apply Theorems 5.23, 5.24 and 5.26 to establish that chaos holds for these marginally minimum-phase cases when coefficient condition (R) does not hold. The role of random input for the Σ-Δ modulator is then clearly one that facilitates the conditions for chaos in a broader array of circumstances. This role does not extend, however, to the analogous strictly minimum-phase case. The following theorem shows, in fact, that random inputs will still give rise to strict nonchaos in this case.

Theorem 5.27 *Suppose all the zeros of $p(z)$ have magnitude less than 1. Suppose also that, for all $n \geq 0$, the input x_n is random and its projection $\hat{P}_{\mathcal{C}^1}$ is described by a piecewise continuous probability density function $h_n(x)$ defined over $\mathcal{C}$, and satisfying $h_n(x|x_i,\ i = 0,\ldots,n-1) \leq K_U$, for all $x \in \mathcal{C}$, for some K_U independent of n, with $K_U > 0$. Then chaos condition 1 (sensitivity to initial conditions) will hold with zero probability.*

Proof:

Let δ be a constant satisfying $0 < \delta \leq \min(\frac{\Delta}{2}, \frac{\Delta}{K_U})$. Choose any $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ and some neighbourhood $\tilde{N}$ of $\hat{x}_0$. Suppose that Proposition 4.1 holds for $N \leq n \leq n_1$, for some n_1, and $\forall\ \vec{y}_0 \in \tilde{N}$. Then, for $\Delta\vec{x}_0 = \vec{y}_0 - \hat{x}_0$, $\vec{y}_0 \in \tilde{N}$, we have, from the nature of the zeros of $p(z)$ and (2.3), that $|\Delta\varepsilon_n| < K_{\vec{y}_0}$, $\forall\ -M \leq n \leq n_1$, for some bound $K_{\vec{y}_0} > 0$. Now choose K such that $K_{\vec{y}_0} < K$, $\forall\ \vec{y}_0 \in \tilde{N}$. Such a bound K clearly exists. We may define a new neighbourhood about $\hat{x}_0$ in $\tilde{N}$ by $\tilde{N}_\alpha = \{\vec{y}_{\alpha 0} \in \mathbb{R}^N \times \mathcal{C}^M \mid \vec{y}_{\alpha 0} = \hat{x}_0 - \alpha(\hat{x}_0 - \vec{y}_0),\ \forall\ \vec{y}_0 \in \tilde{N}\}$, with $0 < \alpha < 1$. Now, with $\Delta\vec{x}_0 = \vec{y}_{\alpha 0} - \hat{x}_0$, by the properties of linear difference equations, the constants in (2.3) will be scaled down in magnitude by a factor of α, and hence so will $|\Delta\varepsilon_n|$. Thus $K_{\vec{y}_0} < \alpha K$, $\forall\ \vec{y}_0 \in \tilde{N}_\alpha$. Now we pick an $\alpha = \hat{\alpha}$, such that $\hat{\alpha} < \frac{\delta}{K}$, and $0 < \hat{\alpha} < 1$, which leads to the result that $|\Delta\varepsilon_n| < \delta$, $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$, and for $-M \leq n \leq n_1$. Now we also have, from the nature of the zeros of $p(z)$ and (2.3), that $\exists$ constants $\tilde{K}$, $\tilde{\mu}$, independent of n_1, with $\tilde{K} > 0$, and $0 < \tilde{\mu} < 1$, such that $|\Delta\varepsilon_n| < \tilde{K}\tilde{\mu}^n$, $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$, and for $0 \leq n \leq n_1$.

We choose a $\hat{y}_0 \in \tilde{N}_{\hat{\alpha}}$ for the following. There $\exists\ k_1 \geq 0$ such that $K_U\tilde{K}\tilde{\mu}^{k_1} < 1$. We define $q_k \equiv K_U|\Delta\varepsilon_k|$, $k = 0,\ldots,k_1 - 1$; $q_k \equiv K_U\tilde{K}\tilde{\mu}^k$, $\forall\ k \geq k_1$. Let $h_n(x)$ be the PDF for the input x_n defined over $\mathcal{C}$ which is conditional upon the values of x_i, $i = 0,\ldots,n-1$. Using the condition involving K_U in the theorem, we have

$$q_k \geq \max\{\int_z^{z+\frac{q_k}{K_U}} h_k(x)dx \mid z \in \mathcal{C}\},\quad \forall\ k \geq 0.$$

Then, using Proposition 4.1, we have

$$\text{Prob}[Q(x_k - r_{y,k}) - Q(x_k - r_{x,k}) \neq 0 \mid Q(x_i - r_{y,i}) - Q(x_i - r_{x,i}) = 0, \\ i = 0, \ldots, k-1,\, k \geq 1] \;\leq\; q_k, \quad \text{where } 0 < q_k < 1,\, \forall\, k \geq 0,$$

and where the variables with subscripts x and y correspond to the system with initial conditions $\hat{x}_0$ and $\hat{y}_0$ respectively. That this is true follows from the fact that we have inequality with the quantized values above when the interval between the quantizer inputs contains the point 0 (mod Δ). This interval has magnitude less than or equal to $\tilde{K}\tilde{\mu}^k$. Thus

$$P_n \;\equiv\; \text{Prob}[Q(x_i - r_{y,i}) - Q(x_i - r_{x,i}) = 0,\, i = 0, \ldots, n] \;\geq\; q_p \prod_{k=k_1}^{n} (1 - q_k),$$

where $q_p \equiv \prod_{k=0}^{k_1-1} (1 - q_k)$. It then follows that $P_n \geq q_p \prod_{k=0}^{n} (1 - \hat{K}\tilde{\mu}^k)$, where $\hat{K} = \tilde{K}\tilde{\mu}^{k_1}$. If we define $\bar{P} = \lim_{n\to\infty} P_n$, then

$$\ln(\bar{P}) \geq \sum_{k=0}^{\infty} \ln(1 - \hat{K}\tilde{\mu}^k) + \ln q_p = -\sum_{k=1}^{\infty} \frac{\hat{K}^k}{k} \left(\frac{1}{1 - \tilde{\mu}^n}\right) + \ln q_p.$$

This sum, by the ratio test, converges to a negative value, and thus $0 < \bar{P} < 1$. Thus, with probability $\bar{P}$, we have that $Q(x_n - r_{y,n}) - Q(x_n - r_{x,n}) = 0$, and hence Proposition 4.1 holds $\forall\, n \geq N$.

Since $\hat{y}_0$ was chosen arbitrarily, this gives $|\Delta\varepsilon_n| < \delta$, $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$, and $\forall\ n \geq -M$. This then implies that, with probability $\bar{P}$, $\|g_1 \circ f^n(\hat{x}_0) - g_1 \circ f^n(\vec{y}_0)\| < \delta$ for any $\vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$, and all $n \geq 0$. If we scale down the value of $\hat{\alpha}$ in $\tilde{N}_{\hat{\alpha}}$, then the values of $\tilde{K}$ and $|\Delta\varepsilon_k|$, $k = 0, \ldots, k_1$, will scale down correspondingly, so that $\lim_{\hat{\alpha}\to 0} \prod_{k=0}^{n} (1 - \tilde{K}\tilde{\mu}^k) = 1$, and $\lim_{\hat{\alpha}\to 0} \bar{P} = 1$. Thus it must be concluded that the statement $\|g_1 \circ f^n(\hat{x}_0) - g_1 \circ f^n(\vec{y}_0)\| < \delta$ for any $\vec{y}_0 \in \tilde{N}_{\alpha}$ and all $n \geq 0$, with $0 < \alpha < \hat{\alpha}$, holds with probability that is arbitrarily close to 1; hence taken to be 1. The δ was chosen arbitrarily. Now applying Lemma 5.6, we have the result that sensitivity to initial conditions holds with probability 0. ■

It is not clear that the results of Theorem 5.27 will hold if the input x_n is allowed to be random with discrete distributions and one may, in fact, conceive that counterexamples exist in light of Proposition 5.28 below. The same consequence holds if we allow only persistently random rather than always random input over each $n \geq 0$. As with Theorems 5.23 and 5.24, if the conditionality constraints in Theorems 5.26 and 5.27 are relaxed and independence over all n_1 or n respectively is not imposed, then the probability margin of the chaos result goes from one to less than or equal to one as supported by cases.

Clearly, Theorems 5.23 and 5.24 will hold with the random x_n of Theorem 5.27, and so the lack of sensitivity to initial conditions remains a barrier to chaos. Moreover, it is reasonable to conjecture that in this situation of stochastic input then, a clear bifurcation point exists for the system delineating chaos from nonchaos. Specifically, the system becomes chaotic if and only if either (a) both condition (R) fails and a zero of $p(z)$ attains 1 in magnitude, or (b) a zero of $p(z)$ attains either magnitude greater than 1, or else magnitude 1 with multiplicity greater than 1. This bifurcation result serves to affirm the view of chaos as some intrinsic deterministic dynamical property of the system whose onset characterizes a change in structural stability. The reason sensitivity is more difficult to obtain than transitivity with random input is because we require more than simply dense orbits. Arbitrary intervals in $\mathcal{C}$ must include a specific point in $\mathcal{C}$ in their mappings, and the contractive nature of these mappings in the strictly minimum-phase case counteracts the dense nature of the mappings arising from random input in accomplishing this.

As an alternative approach, we may consider the following proposition, which assures sensitivity to initial conditions and topological transitivity for the first-order minimum-phase system with a particular input x_n:

Proposition 5.28 *Suppose the system is strictly first order, and that* $M = 1$, $N = 0$. *Then we may construct an input* x_n *such that chaos condition* 2 *(topological transitivity) holds. If the zero of* $p(z)$ *has magnitude less than* 1, *then we may construct an input* x_n *such that chaos condition* 1 *(sensitivity to initial conditions) holds or (with a more complex input) such that both chaos conditions* 1 *and* 2 *hold.*

Proof:

Consider the system with the map $v_n = \beta v_{n-1} + d_n \mod 1$, with $|\beta| > 0$, and $\mu_1 = \beta$. For a given set of input $d_i \in \mathbb{R}$, $i = 0, \ldots, k$, $(k \geq 0)$, one can choose a $d_{k+1} \in \mathbb{R}$ such that $v_{k+1} = e$, for any $e \in [0, 1)$, and any initial condition $v_{-1} \in [0, 1)$, $(k \geq -1)$. From this, it follows inductively that for a given set of initial conditions $v_{-1(i)} \in [0, 1)$, and constants $e_i \in \mathbb{R}$, there exists a set of input d_i such that $v_{i(i)} = e_i$, for $i = 0, \ldots, k$, and any $k \geq 0$. Now consider the infinite sequence p_i given by $\{0, 0, \frac{1}{2}, 0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 0, \frac{1}{8}, \frac{2}{8}, \ldots\}$. Then the points in the sequence are defined by $p_{2^k+i} = \frac{i}{2^k}$, for $i = 0, \ldots, 2^k - 1$, $k \geq 0$. We wish to map all the points p_{2^k+i}, $i = 0, \ldots, 2^k - 1$, to each other in an ordered manner, and then in succession as k increases. To start, we define $\alpha(0) = 1$, $\alpha(k+1) = \alpha(k) + 2^{2k}$, for $k \geq 0$. Now, from the result above, we may choose the input d_i to satisfy the following. If the initial condition of the above system is $v_{-1} = p_{2^k+i}$, then $v_{\alpha(k)+(i)2^k+j} = p_{2^k+j}$, for $j = 0, \ldots, 2^k - 1$, $i = 0, \ldots, 2^k - 1$, $k \geq 0$.

Now choose any two open intervals I_1, $I_2 \subset [0, 1)$. There exist points $p_{2^{k_1}+i_1} \in I_1$, $p_{2^{k_1}+i_2} \in I_2$, for some $k_1 \geq 0$, and i_1, i_2, between 0 and $2^{k_1} - 1$. If $v_{-1} = p_{2^{k_1}+i_1}$, then $v_{\alpha(k_1)+(i_1)2^{k_1}+i_2} = p_{2^{k_1}+i_2}$. Thus under the mapping $\tilde{f}$ (following the notation of the proof of Proposition 5.25), $\tilde{f}^{n_1+1}(I_1) \cap I_2 \neq \emptyset$, where $n_1 = \alpha(k_1) + (i_1)2^{k_1} + i_2$. Since the intervals I_1, I_2 were chosen arbitrarily, we conclude that topological transitivity holds.

Suppose $|\beta| < 1$. Let $\hat{d} = \min(|\beta|, 1 - |\beta|)$. Let δ be a constant satisfying $0 < \delta < \frac{\hat{d}}{2}$. Now choose a small interval $I = [c_{-1}, e_{-1}] \subset [0, 1)$ such that $|e_{-1} - c_{-1}| < \frac{\hat{d}}{2}$. There exists a point $p_{2^{k_1}+i_1} \in I$, for some $k_1 \geq 0$, and i_1 between 0 and $2^{k_1} - 1$. If $v_{-1} = p_{2^{k_1}+i_1}$, then $v_{\alpha(k_1)+(i_1)2^{k_1}} = p_{2^{k_1}} = 0$. Thus $\exists$ an $n_2 \geq 0$ such that $\tilde{f}^n(z) \neq 0$, $\forall\ z \in I$, with $0 \leq n \leq n_2$, and $\tilde{f}^{n_2+1}(z_{-1}) = z_{n_2} = 0$, for some $z_{-1} \in I$. We have that $\tilde{f}^{n_2}$ is continuous over I so that, from (2.3), $|e_{n_2-1} - c_{n_2-1}| = |e_{-1} - c_{-1}||\beta|^{n_2}$, and $\|e_{n_2} - c_{n_2}\| = |e_{-1} - c_{-1}||\beta|^{n_2+1}$ as well. Now we have that $\tilde{f}^{n_2+2}(I) = I_3 \cup I_4$, where I_3, $I_4 \subset [0, 1)$ are intervals with closed endpoints at c_{n_2+1} and e_{n_2+1} respectively. One interval has another closed endpoint at z_{n_2+1} and the other has an open end. The sum of the lengths of the two intervals is $|e_{-1} - c_{-1}||\beta|^{n_2+2}$, and the displacement between the second set of ends is $1 - |\beta|$. Thus $\|e_{n_2+1} - c_{n_2+1}\| \geq \hat{d} - |e_{-1} - c_{-1}||\beta|^{n_2+2} > \frac{\hat{d}}{2} > \delta$. Since the interval I was chosen arbitrarily, we conclude that sensitivity to initial conditions holds.

When $|\beta| < 1$, we may choose the input d_i so that the system will satisfy sensitivity to initial conditions, but not necessarily topological transitivity, as follows. If the initial condition of the above system is $v_{-1} = p_{2^k+i}$, then $v_{2^k+i} = 0$, for $i = 0, \ldots, 2^k - 1$, $k \geq 0$. The proof of sensitivity to initial conditions then follows the same method as given above.

■

It would seem reasonable to expect that this proposition could be extended to systems of any order, that is with any values of M and N, since the structure of the proof would appear amenable to such an extension. Similarly, we might expect it to be possible to construct an input of simpler structure than that in the proof which yields transitivity only, but not sensitivity when the system is minimum or marginally minimum phase (or sensitivity but not transitivity when marginally minimum phase). (If the system has all

its zeros of magnitude 1, multiplicity 1, then this input (e.g. constant) follows from that in Proposition 5.16 when condition (R) holds.) The point of this proposition is that it provides examples of how sensitivity may be satisfied for the strictly minimum-phase case, and of how transitivity may be satisfied for an arbitrary filter form (with some fixed M and N).

To construct the appropriate input, we have relied on the interval splitting property in the mappings that occurs when $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$ is discontinuous, to force sensitivity to initial conditions as was essentially the case in Theorem 5.26 for the marginally minimum-phase case, and used a particular deterministic input to remove any chance of missing sensitivity for some interval as may occur with random input. Deterministic inputs may then extend further the prevalence of chaos condition 1 compared to random ones.

We have no means of showing that a constructed input that achieves this may also make the system chaotic by satisfying condition 3. Considering the random-like complexity of the constructed inputs however, it would seem plausible to believe that nonexistence and hence density of periodic points would hold for any filter form (with the given M, N), as with appropriate random input. Thus we may well have a means of extending chaos to hold for all filter forms with deterministic inputs, notably the strictly minimum-phase case where random input failed to give this. In any event, we have shown how chaos can occur for all filter forms that are not strictly minimum phase. Note that in applying the proposition to the minimum-phase case, we require that the mappings of the system not satisfy the continuity of Theorem 5.15. Otherwise chaos condition 1 would not hold.

Some additional theorems and propositions pertaining to the marginally minimum-phase Σ-Δ system will now be considered. Generalizing Theorem 5.18 to encompass the case where condition (R) does not hold and where the system has zeros inside the unit circle, we bring in the issue of sensitivity to initial conditions as well to give the following:

Theorem 5.29 *Suppose the largest magnitude zero(s) of $p(z)$ have either magnitude less than 1, or else magnitude 1 and multiplicity 1. Suppose also that there exists an $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ such that the set $\{g_M(\hat{x}_n),\ \forall\ n \geq 0\}$ is finite. Suppose further that $g_M(\hat{x}_n) \neq \frac{\Delta}{2}$ holds for all $n \geq 0$. Then chaos condition 1 (sensitivity to initial conditions) and chaos condition 2 (topological transitivity) will not hold.*

Proof:

Let $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ be as given in the theorem, with the number of elements in the set $\{g_M(\hat{x}_n),\ \forall\ n \geq 0\}$ given by p. Now we choose a δ such that $0 < \delta \leq \frac{\Delta}{2}$. Let $\tilde{N}$ be some neighbourhood of $\hat{x}_0$, where $\tilde{N}$ is sufficiently small that, $\forall\ \vec{y}_0 \in \tilde{N}$, $\Delta/2$ does not lie in the smallest closed interval between $g_i(\vec{y}_0)$ and $g_i(\hat{x}_0)$ on $\mathcal{C}$, for $i = 1, \ldots, M$; and $Q(x_k - r_{y,k}) - Q(x_k - r_{x,k}) = 0,\ \ k = 0, \ldots, N$. The variables here with subscripts x and y correspond to the system with initial conditions $\hat{x}_0$ and $\vec{y}_0$ respectively. Thus, from the conditions in the proof of Theorem 4.4, continuity on the line connecting $\hat{x}_0$ and $\vec{y}_0$ in $\mathbb{R}^N \times \mathcal{C}^M$ holds, and thus (4.1) and (2.3) from Theorem 4.5 hold for $n = N$. Now choose a $\hat{y}_0 \in \tilde{N}$. Then, for $\Delta\vec{x}_0 = \hat{y}_0 - \hat{x}_0$, we have, from the nature of the zeros of $p(z)$ and (2.3); that $|\Delta\varepsilon_n| < K_{\hat{y}_0}$, with $-M \leq n \leq \hat{n}$, for some bound $K_{\hat{y}_0} > 0$, if Theorem 4.5 holds, for $N \leq n \leq \hat{n}$, for some $\hat{n} \geq N$. For now, we have $\hat{n} = N$. Now choose K such that $K > K_{\vec{y}_0}$, $\forall\ \vec{y}_0 \in \tilde{N}$. Such a bound K clearly exists. We may define a new neighbourhood about $\hat{x}_0$ in $\tilde{N}$ by $\ \tilde{N}_\alpha = \{\vec{y}_{\alpha 0} \in \mathbb{R}^N \times \mathcal{C}^M \mid \vec{y}_{\alpha 0} = \hat{x}_0 - \alpha(\hat{x}_0 - \vec{y}_0),\ \forall\ \vec{y}_0 \in \tilde{N}\},\ $ with $0 < \alpha < 1$. Now, with $\Delta\vec{x}_0 = \vec{y}_{\alpha 0} - \hat{x}_0$, by the properties of linear difference equations, the constants in (2.3) will be scaled down in magnitude by a factor of α, and hence so will $|\Delta\varepsilon_n|$. Thus $\alpha K > K_{\vec{y}_0},\ \forall\ \vec{y}_0 \in \tilde{N}_\alpha$.

Now we pick $\alpha = \hat{\alpha}$ such that $\hat{\alpha} < \min(\frac{K_1}{K}, \frac{\Delta}{4Kp}, \frac{\delta}{K})$, and $0 < \hat{\alpha} < 1$, where $K_1 > 0$ is such that $\ Q(x_k - r_{x,k} \pm K_1) - Q(x_k - r_{x,k}) = 0,\ \ k = 0, \ldots, n_1$. We label this condition

(#). For this we choose n_1 sufficiently large to satisfy $\{g_1(\hat{x}_n),\, 0 < n \leq n_1\} = \{g_1(\hat{x}_n),\, \forall\ n > 0\}$. With $\hat{\alpha} < K_1/K$, we have $|\Delta\varepsilon_n| < K_1$, for $-M \leq n \leq N$. From the form of (1.2), it holds that the applicability of (2.3), for $-M \leq n \leq N$, extends to $n = N+1$, if $\Delta\varepsilon_{N+1}$ is relabelled Δr_{N+1} on the LHS of (2.3). Thus we have $|\Delta r_{N+1}| < K_1$, which implies that $Q(x_{N+1} - r_{y,(N+1)}) - Q(x_{N+1} - r_{x,(N+1)}) = 0$, if $n_1 > N$, from (#). If $n_1 \leq N$, we have

$$\begin{aligned} Q(x_{N+1} - r_{y,(N+1)}) - Q(x_{N+1} - r_{x,(N+1)}) & \\ &= Q(x_{N+1} - r_{x,(N+1)} - \Delta r_{N+1}) - Q(x_{N+1} - r_{x,(N+1)}) \\ &= Q(x_{n_2} - r_{x,(n_2)} + m\Delta - \Delta r_{n_1+1}) - Q(x_{n_2} - r_{x,(n_2)} + m\Delta) \\ &= 0, \qquad \text{for some } n_2 \text{ satisfying } 0 \leq n_2 \leq n_1. \end{aligned}$$

We have used $g_1(\hat{x}_{N+2}) = g_1(\hat{x}_{n_2+1})$, which also implies from (1.2) that $x_{N+1} - r_{x,(N+1)} = x_{n_2} - r_{x,(n_2)} + m\Delta$, for some $m \in \mathbb{Z}$. Then (#) was applied, with the $m\Delta$ cancelling out. Thus, extending the conclusions of above, (4.1) and (2.3) hold for $-M \leq n \leq N+1$. We may thus continue this process inductively, breaking into $n_1 > N+k$, $n_1 \leq N+k$, cases analogously, to show that (4.1) and (2.3) hold for n up to $N+k+1$, when (4.1) and (2.3) hold for n up to $N+k$, $k \geq 0$. We then conclude that (4.1) and (2.3) hold $\forall\ n \geq N$, and $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$.

Now, with this applicability of (4.1) and (2.3), we have $\hat{n} \to \infty$, and the result that $|\Delta\varepsilon_n| < \dfrac{\Delta}{4p}$, $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$, and $\forall\ n \geq -M$. Thus the mapping $g_1 \circ f^n(\tilde{N}_{\hat{\alpha}}) \subset [g_1 \circ f^n(\hat{x}_0) - \dfrac{\Delta}{4p},\ g_1 \circ f^n(\hat{x}_0) + \dfrac{\Delta}{4p}]$ holds on $\mathcal{C}$ $\forall\ n \geq 0$. From the finite size of the orbit set for $\hat{x}_0$ above, we have that $g_1 \circ f^n(\tilde{N}_{\hat{\alpha}}) \subset V_1 \subseteq \mathcal{C}$, where V_1 is the union of p closed intervals on $\mathcal{C}$, each of length $\Delta/2p$, $\forall\ n \geq 0$. Thus there must exist a $\hat{z} \in \mathcal{C}$ with some neighbourhood $\tilde{N}_z \subset \mathcal{C}$ about $\hat{z}$, such that $g_1 \circ f^n(\tilde{N}_{\hat{\alpha}}) \cap \tilde{N}_z = \emptyset$, $\forall\ n \geq 0$. Now applying Lemma 5.7, we have the result that topological transitivity does not hold. From above, we also have the result that $|\Delta\varepsilon_n| < \delta$, $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$ and $\forall\ n \geq -M$. This then implies that $\|g_1 \circ f^n(\hat{x}_0) - g_1 \circ f^n(\vec{y}_0)\| < \delta$ for any $\vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$, and all $n \geq 0$. The δ was chosen arbitrarily. Now applying Lemma 5.6 we

have the result that sensitivity to initial conditions does not hold. As with Theorem 5.18, if $N = 0$, the input x_n is periodic with period q, and the system has an arbitrary periodic point with period p_1, then, from the properties of the difference equations (1.2), the requirements of the theorem are satisfied with the cardinality p satisfying $p \leq p_1 q$. ■

This theorem extends Theorem 5.18 to include all minimum and marginally minimum-phase cases when (R) does not hold, under the constraint that no coordinate ε_n in the orbit of the point $\hat{x}_0$ ever takes on the boundary value $\Delta/2$. Chaos condition 1 also emerges directly in this version, compared to its separate treatment in Theorem 5.15 previously when (R) holds. The implications of this theorem for establishing chaos are more suggestive compared to that of disproving chaos. Applying the results of the discussion following Theorem 5.18, we can argue as follows. With the theorem above, to show chaos we need only show that conditions 1 and 2 hold in the nonminimum or marginally minimum-phase (zeros on the unit circle with multiplicity 1) cases that have the property that all periodic points must lie on a limit cycle orbit in $\mathcal{C}^M$. This will imply the nonexistence of a periodic point and hence satisfy chaos condition 3 trivially. The limit cycle orbit property will hold, for example, if $N = 0$ with periodic input, but we must also have $\Delta/2$ not on the prospective limit cycle to apply Theorem 5.29.

It was seen how chaos condition 3 (density of periodic points) may arise trivially when no periodic point exists. From the following theorem, we see that for certain minimum-phase systems this is the only way condition 3 may be satisfied:

Theorem 5.30 *Suppose all the zeros of $p(z)$ have magnitude less than 1. Suppose also that there exists an $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$ such that the set $\{g_M(\hat{x}_n),\ \forall\ n \geq 0\}$ is finite. Suppose further that $g_M(\hat{x}_n) \neq \dfrac{\Delta}{2}$ holds for all $n \geq 0$. Then there are no periodic points in a sufficiently small neighbourhood of $g(\hat{x}_0)$.*

Proof:

The first part of the proof of this theorem follows the first paragraph of the proof to Theorem 5.29. With the applicability of (4.1) and (2.3), we have the result that $\lim_{n\to\infty} \Delta\varepsilon_n = 0$. This then implies that $\lim_{n\to\infty} g_1 \circ f^n(\tilde{N}_{\hat{\alpha}}) \subseteq \{g_M(\hat{x}_n), \forall\ n \geq 0\}$. We may assume, without loss of generality, that $\tilde{N}$ and $\tilde{N}_{\hat{\alpha}}$ are small enough so that $g(\vec{y}_0) \ni \{g(\hat{x}_n), \forall\ n \geq 0\}$, $\forall\ \vec{y}_0 \in \tilde{N}_{\hat{\alpha}} - \{\hat{x}_0\}$. It then follows, from the limit above, that the condition $g \circ f^{p_2 k}(\vec{y}_0) = g(\vec{y}_0)$, $\forall\ k \in \mathbb{Z}^+$ and some $p_2 > 0$, cannot hold for any $\vec{y}_0 \in \tilde{N}_{\hat{\alpha}}$. Hence there can be no periodic point in the neighbourhood $\tilde{N}_{\hat{\alpha}}$ of $\hat{x}_0$. As with Theorem 5.29, if $N = 0$, the input x_n is periodic with period q, and the system has an arbitrary periodic point with period p_1, then, from the properties of the difference equations (1.2), the requirements of the theorem are satisfied with the cardinality p satisfying $p \leq p_1 q$. In this case, this implies that density of periodic points does not hold. ■

This result suggests a relative limit on the nature of the dynamics for the minimum-phase case. Not only chaos, but density of periodic points occurs only in the rather uninteresting scenario of having no periodic points, at least for cases where any prospective periodic point must exist on the appropriate limit cycle orbit in $\mathcal{C}^M$. This result takes the same form as that of part 2 of Theorem 5.17 with $M = \max(N, M)$, where a periodic point implies non-density of periodic points.

In passing, we may note that it is possible to relax the set condition in Theorems 5.18, 5.29 and 5.30 to that of $\{g_M(\hat{x}_n), \forall\ n \geq 0\} \cap \tilde{N}_0 = \emptyset$, where $\tilde{N}_0$ is some neighbourhood in $\mathcal{C}^M$ about any point in $\mathcal{C}^M$ for Theorem 5.18, and the point $\Delta/2$ for Theorems 5.29 and 5.30. This shows that these theorems apply to any system that stops short of generally pure quasiperiodic or dense orbit behaviour, although no added simple applications are obvious with this relaxation.

Drawing on the results of Theorems 5.29 and 5.30, we have a corresponding extension of Proposition 5.16 to the converse case when (R) does not hold, as follows:

Proposition 5.31 *Suppose the a_i and b_j do not satisfy condition (R). If the largest magnitude zero(s) have magnitude* 1 *and multiplicity* 1*, and if the input x_n is constant, then chaos condition* 1 *(sensitivity to initial conditions) and chaos condition* 2 *(topological transitivity) are not assured.*

Proof:

(i) Consider the system with the map $v_n = (\frac{1}{2}v_{n-1} - v_{n-2} + d) \mod 1, \ \ 0 < d < \frac{3}{2}$, with $\mu_{1,2} = \frac{1}{4} \pm \hat{i}\frac{\sqrt{15}}{4}, \ |\mu_{1,2}| = 1$, and where (R) does not hold. This map has a periodic point at $(v_{-1}, v_{-2}) = (\frac{2d}{3}, \frac{2d}{3})$, and thus the untransformed system has a periodic point at $(\varepsilon_{-1}, \varepsilon_{-2}) = (\frac{\Delta}{2} - \frac{2d\Delta}{3}, \frac{\Delta}{2} - \frac{2d\Delta}{3})$. Thus, by Theorem 5.29, chaos conditions 1 and 2 do not hold.

(ii) Now consider the system with the map $v_n = (\frac{3}{2}v_{n-1} - \frac{1}{2}v_{n-2}) \mod 1$, with $\mu_1 = 1, \ \mu_2 = \frac{1}{2}$, and where (R) does not hold. This map has a periodic point at all points of the form $(v_{-1}, v_{-2}) = (\alpha, \alpha)$, and thus the untransformed system has a periodic point at $(\varepsilon_{-1}, \varepsilon_{-2}) = (\frac{\Delta}{2} - \alpha\Delta, \frac{\Delta}{2} - \alpha\Delta)$, for $0 \leq \alpha < 1$. Thus, by Theorem 5.29, chaos conditions 1 and 2 do not hold.

(iii) Consider the system with the map $v_n = (\frac{1}{2}v_{n-1} - v_{n-2} + \frac{1}{2}v_{n-3} + d) \mod 1$, $0 < d < 1$, with $\mu_{1,2} = \pm\hat{i}, \ \mu_3 = \frac{1}{2}$, and where (R) does not hold. This map has a periodic point at $(v_{-1}, v_{-2}, v_{-3}) = (d, d, d)$, and thus the untransformed system has a periodic point at $(\varepsilon_{-1}, \varepsilon_{-2}, \varepsilon_{-3}) = (\frac{\Delta}{2} - d\Delta, \frac{\Delta}{2} - d\Delta, \frac{\Delta}{2} - d\Delta)$. Thus, by Theorem 5.29, chaos conditions 1 and 2 do not hold. ■

Thus the nonchaos consequences of Proposition 5.16 naturally extend to the question of initial conditions when the lack of condition (R) raises this as an undetermined factor. Unlike this previous proposition, having a periodic input that is irrational did not provide a means for showing that either chaos condition in question may be satisfied. We may nonetheless conjecture that such examples exist (and indeed may be chaotic by the discussion following Theorem 5.29). In this event, we would have, as was the case of Proposition 5.16 vs. Theorem 5.24 for topological transitivity, that a form of irrational periodic input may lead to sensitivity to initial conditions or topological transitivity (or indeed chaos) with much the same mechanism as that of the random input in Theorem 5.26 or Theorem 5.24 (or these plus Theorem 5.23) respectively. Small perturbations of the input for each of the examples in the proof of the proposition (or related examples satisfying some chaos conditions) may maintain the respective sensitivity/transitivity or nonsensitivity/nontransitivity.

It is easy to see examples with periodic rational input x_n where sensitivity to initial conditions and topological transitivity do not hold for the strictly minimum-phase case. The following example shows this property and motivates the following proposition:

Example 5:

Consider the system with the map $v_n = \beta v_{n-1} + d \mod 1$, with $\mu_1 = \beta$, where $0 < \beta < 1$, and $0 \leq d < (1 - \beta)$. ■

Proposition 5.32 *Suppose all the zeros of $p(z)$ have magnitude less than 1, and that the input x_n is constant. Then chaos condition 1 (sensitivity to initial conditions), chaos condition 2 (topological transitivity), and chaos condition 3 (density of periodic points) are not assured.*

Proof:

Consider Example 5. First suppose that $0 < d < (1-\beta)$. This map has a periodic point at $v_{-1} = \dfrac{d}{1-\beta}$, and thus the untransformed system has a periodic point at $\varepsilon_{-1} = \dfrac{\Delta}{2} - \dfrac{d\Delta}{1-\beta}$. Thus, by Theorem 5.29, chaos conditions 1 and 2 do not hold, and by Theorem 5.30, chaos condition 3 does not hold.

Now consider the case when $d = 0$. The mapping $\tilde{f}$ (following the notation of the proof of Proposition 5.25) is clearly continuous on $[0, 1)$ so that, by Theorem 5.15, sensitivity does not hold. Let $I = [0, \beta] \subset [0, 1)$. Clearly $\tilde{f}^n(I) \subset I, \forall\ n \geq 0$, so that $\tilde{f}^n(I) \cap (\beta, 1) = \emptyset, \ \forall\ n \geq 0$. Thus transitivity does not hold. We have $\lim_{n\to\infty} \tilde{f}^n(v_{-1}) = 0, \ \forall\ v_{-1} \in [0, 1)$, which implies that 0 can be the only periodic point. Thus density of periodic points does not hold. ■

From Example 5 and Proposition 5.32, we see how all three chaos conditions may fail for a general minimum-phase case when the input x_n is appropriately constrained relative to the contractility of the mappings (determined by the magnitude of the zeros of the NTF), so that the mappings do not exhibit interval splitting. We have no proposition or examples to show the possible existence of sensitivity or transitivity in the minimum-phase case when the input x_n is periodic. This is because of the difficulty in analyzing the behaviour of mappings with ever shrinking intervals. We may conjecture, as was analogously done regarding Proposition 5.31 however, that with irrational periodic input, topological transitivity would hold in some cases. Extending this conjecture to sensitivity to initial conditions, as well, would seem dubious considering that, from Theorem 5.27, a purely random input is insufficient to bring this about. Applying Theorem 5.29, however, will extend this to the conclusion that such conjectured cases (with the property that all periodic points must exist on a limit cycle in $\mathcal{C}^M$) satisfy chaos condition 3 and, in the event that

sensitivity holds, are chaotic. It would seem reasonable to expect that such irrational periodic input chaotic examples (minimum and marginally minimum phase) could be extended to higher-order examples.

No extension of the density of periodic points result of Theorem 5.17 could be constructed to apply when condition (R) does not hold. This is a consequence of the fact that, while the method of proof of Theorem 5.10 was easily adaptable to the nonexpansive case, the method of proof of Theorem 5.21 relies on sufficient expansivity and hence is not even adaptable to less expansive cases. The more complex, interval splitting nature of the mappings indeed makes it difficult to conjecture that relatively simple extensions would exist.

We have seen how density of periodic points (i.e. no periodic points) may arise for some minimum and marginally minimum-phase (zeros on the unit circle having multiplicity 1, some zeros inside the unit circle) cases when chaos conditions 1 and 2 are made to hold with a periodic input, or when the input is random. No propositions or examples yielding this chaos condition were constructed for these cases when the continuity of Theorem 5.15 holds together with a periodic input, although it seems reasonable to expect that such examples exist for at least the marginally minimum-phase case with irrational periodic input, given the complexity and asymmetry of the structure involved. Note that there exists an infinite number of certain general perturbations of the periodic input, modulo $\mathcal{C}$, in the examples considered in the proofs of Propositions 5.31 and 5.32 that maintain the periodic point. Theorems 5.29 and 5.30, however, will not apply if the input, modulo $\mathcal{C}$, does not exist in a certain finite set, for each $\hat{x}_0$.

5.4 Summary

A summary of the classification of the various cases and sub-cases in terms of whether or not the three conditions for chaos and overall chaos hold is now presented in Tables 5.1 to 5.5 that follow.

The basic content of these tables follows directly from the theorems, propositions and examples/counterexamples of this thesis. In a few cases, as indicated, specific classifications follow in part from conjectures resulting from the discussion or extrapolation of results in this chapter. By "persistent random input", we mean input x_n satisfying the conditions of Theorem 5.23.

Summary of Results:

To summarize the results, chaos was shown to hold when the system is nonminimum phase with $\max(N, M)$ zeros outside the unit circle, for any input x_n, provided the filter coefficients a_i and b_j satisfy condition (R) as presented. Chaos was also shown to hold when the system is nonminimum phase with $\max(N, M)$ zeros outside the circle of radius 2, with any such filter coefficients, provided the input x_n is periodic; or with each of at least M zeros of magnitude greater than 2, or equal to 2 and repeated [counted (multiplicity $-$ 1) times], provided the input (projected on $\mathcal{C}$) is persistently random. Chaos was shown to hold as well, when the system is not strictly minimum phase, with any such filter coefficients, provided the system input has a more restrictive random structure.

Chaos was shown not to hold when the system is marginally minimum phase and the filter coefficients satisfy condition (R); and when the system is minimum/marginally minimum phase (with a zero inside the unit circle) and the continuity of Theorem 5.15 holds. Examples of chaos was shown for nonminimum-phase cases (with or without M nonminimum-phase zeros) with constant input, and examples of nonchaos were shown in all general phase

Condition (R) Holds					
NTF Zeros	*Chaos Conditions*				
	1	2	3	Chaos	C*+ S
(a) $\max(N, M)$ nonminimum-phase zeros	$\surd$	$\surd$	$\surd$	$\surd$	E
(b) $M < N$ nonminimum-phase zeros	$\surd$	$\surd$	$\mathrm{E}_1^{\surd}$	$\mathrm{E}_1^{\surd}$	E
(c) 1 to $M-1$ nonminimum-phase zeros	$\surd$	E	E	E	E
(d) $\max(N, M)^* \geq 1$ marginally minimum-phase zeros	$\times$	E	E	$\times$	

Table 5.1: Chaos condition classifications when (R) holds

Condition (R) Holds										
	Periodic Input					*Persistent Random Input on* $\mathcal{C}$				
NTF Zeros	*Chaos Conditions*					*Chaos Conditions*				
(as above)	1	2	3	Chaos	C*+ S	1	2	3	Chaos	C*+ S
(a)	$\surd$	$\surd$	$\surd$	$\surd$	$\times_1$	$\surd$	$\surd$	$\surd$	$\surd$	E
(b)	$\surd$	$\surd$	$\mathrm{E}_1^{\surd}$	$\mathrm{E}_1^{\surd}$	$\times_1$	$\surd$	$\surd$	$\surd$	$\surd$	E
(c)	$\surd$	E	E	E	$(\times)$	$\surd$	$(\surd)$	$\surd$	$(\surd)$	E
(d)	$\times$	E	$(\surd)$	$\times$		$\times$	$(\surd)$	$\surd$	$\times$	

Table 5.2: Chaos condition classifications according to input when (R) holds

Condition (R) Fails to Hold					
NTF Zeros	*Chaos Conditions*				
	1	2	3	Chaos	C*+ S
(a) $\max(N,M)$ zeros of magnitude > 2	$\surd$	$\surd$	$E_1^{\surd}$	$E_1^{\surd}$	E
(b) $M < N$ zeros of magnitude > 2	$\surd$	$\surd$	E_1	E_1	E
(c) $< M$ zeros of magnitude > 2, ≥ 1 NMP zeros	$\surd$	E	E	E	E
(d) $\max(N,M)^* \geq 1$ MMP zeros	E	E	$(\surd)$	E	E
(e) 1 to $\max(N,M)^*-1$ MMP zeros, no NMP zeros	E	E	$(\surd)$	E	E
(f) $\max(N,M)^* \geq 1$ minimum-phase zeros	E	E	E	$(\times)$	

Table 5.3: Chaos condition classifications when (R) fails to hold

Condition (R) Fails to Hold										
	Periodic Input					*Persistent Random Input on* $\mathcal{C}$				
NTF Zeros	*Chaos Conditions*					*Chaos Conditions*				
(as above)	1	2	3	Chaos	C*+ S	1	2	3	Chaos	C*+ S
(a)	$\surd$	$\surd$	$\surd$	$\surd$	E	$\surd$	$\surd$	$\surd$	$\surd$	E
(b)	$\surd$	$\surd$	E_1	E_1	E_1	$\surd$	$\surd$	$\surd$	$\surd$	E
(c)	$\surd$	E	E	E	E_1	$\surd$	$(\surd)$	$\surd$	$(\surd)$	E
(d)	$(\times)$	$(\times)$		$(\times)$		$(\surd)$	$(\surd)$	$\surd$	$(\surd)$	E
(e)	$(\times)$	$(\times)$		$(\times)$		$(\surd)$	$(\surd)$	$\surd$	$(\surd)$	E
(f)	$(\times)$	$(\times)$	$(\times)$	$(\times)$		$(\times)$	$(\surd)$	$\surd$	$(\times)$	

Table 5.4: Chaos condition classifications according to input when (R) tails to hold

Glossary	
Symbol	*Meaning*
C*+ S	Stability holds when Chaos holds, and $N \geq 1$
$\surd$	condition always satisfied
$(\surd)$	condition satisfied in some cases
$\times$	condition never satisfied
$(\times)$	condition not satisfied in some cases
E	condition satisfied or not satisfied depending on the case
$\mathrm{E}_1^{\surd}$	$(\surd)$ holds, $(\times)$ conjectured to hold
E_1	$(\surd)$ and $(\times)$ both conjectured to hold
$\times_1$	$\times$ conjectured to hold

Table 5.5: Glossary for chaos condition classifications

cases. Many cases were found in which some of the chaos conditions held or did not hold, but where no overall chaos results were obtained. No clear conclusions were drawn in many cases, particularly when condition (R) and the continuity of Theorem 5.15 do not hold, or the input is neither definably random or periodic. Some of the nonminimum-phase results obtained were conjectured to be necessary and sufficient for chaos when condition (R) holds. A prevalence of chaos and "pseudo-chaotic" nonchaos was further conjectured over the general nonminimum-phase and, to some extent, minimum/marginally minimum-phase cases, when condition (R) or the continuity of Theorem 5.15 do not hold in the relevant cases.

Density of Periodic Points:

If the standard (stricter) definition of density of periodic points is used in chaos condition 3, then our conclusions in the last two columns of line (a) in Table 5.1, and the periodic input segment of table 5.2, are weaker — condition 3 and chaos are satisfied in some cases (i.e. when $N = 0, 1$, by conjecture) rather than all cases. Conclusions are weaker in the other lines as well (see footnotes in this chapter), although the classification symbols remain the same. If existence of periodic points is added to the condition 3 definition as well, then conclusions are generally weaker yet (see footnotes), although the only fundamental change in the tables shows up in the persistent random input segment of Tables 5.2 and 5.4, where, as expected, condition 3 and chaos fail in all cases — hence the symbols not involving "$\times$" in the last two columns of Tables 5.1 and 5.3 would convert to "E" throughout.

We also observe that if the strictest interpretation of condition 3 in Devaney's chaos (i.e. including existence of periodic points) is adopted, we have the following for the condition (R) results: Sensitivity always holds for the cases where we show transitivity and density of periodic points both holding, and does not hold for the cases when we show at least one of conditions 2 or 3 failing. This is particularly notable in the minimum-phase situation (no sensitivity), where creating quasiperiodicity to give transitivity eliminates periodic points, and creating enough regularity to give dense periodic points makes transitivity unattainable. These consequences are consistent with the property of Devaney chaos that transitivity and density of periodic points taken together imply sensitivity for continuous maps, since the mappings here are nearly continuous (continuous, except on boundaries) under condition (R). We see some carryover of these consequences when (R) does not hold, as well. Using our definition of condition 3, however, we have under (R) that conditions 2 and 3 may both hold, but not 1, when the system is minimum-phase with random input. Thus removing the existence condition for density of periodic points, as was deemed

meaningful for nonautonomous systems, has some effect in making the sensitivity condition less redundant for characterizing chaos in general classes of mappings.

Chaos with Stability:

It is of more practical importance to understand what conditions will bring about chaos together with particular types of stability for the Σ-Δ modulator. For the discussion here, we shall assume that this is bounded internal stability as studied in Chapter 3, and that it is taken to hold over all initial conditions in the state space $\mathbb{R}^N \times \mathcal{C}^M$. This is, of course, the most fundamental type of stability of concern. A summary of the classifications of the cases, according to whether or not stability holds when chaos holds, under the nontrivial condition of $N \geq 1$, is provided as well in Tables 5.1 to 5.5. We discuss the arguments behind these classifications with the following.

For a simplified system with no feedback elements ($N = 0$), bounded stability is automatic. Such systems are contained within the treatment of the general results of this chapter, and thus provide trivial examples of chaos with stability, when the pertinent chaos conditions are met (e.g. conditions on the feedback gains a_i).

For more general systems, it is straightforward to obtain stability when satisfying conditions for chaos via Theorem 5.21, or for that matter, the pertinent chaos conditions of the general results of Section 5.3 overall. One simply chooses the N feedforward gains b_j to give stability, and then, independently, the a_i gains so that the $a_k - b_k$ coefficients of $p(z)$ satisfy any required conditions (e.g. the zeros of greater than magnitude two conditions on $p(z)$ in Theorems 5.20 or 5.21). Similarly, the input requirements are satisfied independently in the theorems pertaining to a persistent random input. In general, from the work of Section 5.3, we would conjecture, as well, the existence of systems having all zeros of $p(z)$ with magnitude less than two (but typically all greater than 1), $N \geq 1$, and a deterministic input, that are both chaotic and stable.

When $N \geq 1$, and condition (R) is required to hold, the situation is more complicated. As mentioned in Chapter 4, and alluded to in this chapter, it is difficult to ensure that all the zeros of $p_r(z)$ lie inside the unit circle (i.e. bounded stability) while requiring that some zeros of $p(z)$ lie outside the unit circle (to make the system nonminimum phase) at the same time. In fact, our analytical evidence would seem to suggest that a necessary condition[13] for the existence of q nonminimum-phase zeros of $p(z)$ and stability, where $1 \leq q \leq \max(N, M)^*$, is that $M \geq N + q$. Adopting this conjecture leads to the conclusion of stability with the chaos condition of sensitivity in Theorem 5.8 only when $M > N$, and stability with transitivity and density of periodic points in Theorems 5.9 and 5.10 respectively, only when $N = 0$. In short, we have no examples under which full chaos holds with stability when both $N \geq 1$ and (R) hold, and when the input is arbitrary and deterministic — indeed we appear to have general results in Section 5.1 for the existence of chaos with nonstability for such systems. Other results in this section allow for cases, or conjectures, under which some of the chaos conditions will hold with stability for such systems (although not conditions 1 and 2 together). At best, we might speculate that more general, fully chaotic forms of the hyperbolic toral automorphism mappings (which satisfy (R)) of Proposition 5.14, arising with $N \geq 1$, might exist, where fewer than M nonminimum-phase zeros are present.

Defining Chaos:

The counterexample of Example 3 provides an interesting case of dynamical behaviour where our Devaney chaos may hold on the trapping region but not on the whole set $\mathcal{C}$, as sufficient for our chaos requirement. This raises the issue of whether a more appropriate, or at least a more viable adaptation of Devaney's definition of chaos to apply to a system such

[13]Such a bound on M may, in general, be sufficient, since we have shown stability in the particular case of Example 1, where $M = 2$, $N = 1$, and one zero of $p(z)$ is nonminimum phase.

as the Σ-Δ modulator here, would involve a relaxation of the adapted definition we have used, to assert that chaotic behaviour on a subset of $\mathcal{C}^M$ (i.e. such as a trapping region) would be sufficient to warrant an overall classification of chaos for the system. Devaney's basic definition of chaos is one that is particularly well suited for the characterization of chaos on such subsets of state space.

Such a relaxed definition would allow a convenient classification as chaotic for many such expansive mapping systems that demonstrate complex, intuitively chaotic like behaviour, but elude our stricter chaos designation based on perhaps less relevant details inherent in the complexity of the dynamics. Such a way of characterizing chaos may also be more meaningful as a practical way of distinguishing or bifurcating between two qualitatively different types of behaviour in the functioning Σ-Δ modulator. In addition, such notions of chaos come closer to the ideal of viewing all cases where each zero is nonminimum phase, as being chaotic, and other cases as generally nonchaotic. This is an extrapolation of the circumstance when condition (R) is satisfied, and is consistent with the unsubstantiated claims of other research work. This view has some intuitive support from the minimum-phase results as well, where it seems reasonable to argue that nonchaos (resulting from nonsensitivity to initial conditions) is the prevalent generic condition, and chaos arises only under very specific circumstances (i.e. on the input).

However we define chaos, it may well be that, in further investigations of Σ-Δ modulator systems with zeros of magnitude between 1 and 2, the establishment of Devaney chaos on a trapping region is the best that we can do in terms of asserting the possible existence of overall Devaney chaos. One possible tool in such a study involves Devaney's concept of a nonwandering set [7], which is essentially a set that is internally topologically transitive. Some work on nonwandering sets for monotonic modular functions has been done by Hofbauer [23]. The mapping f of the first-order Σ-Δ modulator is of this general

form. Studies involving an application of this work to expansive maps would be a relevant starting point for further investigation.

Such investigations, or a study involving a change of the chaos definitions used, are thus of further interest but will not be pursued in this thesis due to the added complexity and work involved for such an analysis to be as thorough as that given for the definitions of chaos already chosen. The counterexample of Example 3 also provides motivation for a framework for further study of the complex dynamics arising from such maps, but this will not be pursued within the scope of this thesis.

Modulator Implications:

In this chapter, an adapted version of Devaney's definition of chaos has been applied to the general multi-bit Σ-Δ modulator with the goal of determining conditions under which chaos does or does not hold. The approach of the analysis was to consider general mathematical variations in the nature of the parameters involved (i.e. filter coefficients a_i, b_j and input x_n) either directly (i.e. with condition (R)) or indirectly (i.e. with the position of the zeros of the noise transfer function), so as to arrive at cases that are as simple as possible and for which as broad conclusions as possible could be derived.

The practical implications of varying the zeros of the noise transfer function, maintaining stability and meeting any filter coefficient conditions obviously follow directly when choosing the filter coefficients of the Σ-Δ modulator. Other factors in design may of course make some coefficient class scenarios more prevalent than others. The implications of the different input cases (involved for questions of both stability and chaos) are less obvious or controllable in practice. The relevance of considering a particular deterministic input, or a type of periodic or random input, may be little if the particular input signal desired for processing by the Σ-Δ modulator conforms to none of these simplified classifications, as may generally be the case. Nevertheless, the analysis undertaken for chaotic behaviour

at the more abstract level in this chapter serves to provide a broad yet thorough overview of the general mathematical nature of the dynamics of the Σ-Δ modulator, using the most standard concept of chaos available. The same assertion holds for the analysis of stability in Chapter 3. Such an underpinning is both theoretically necessary and practically useful in providing a structure with which to understand the behaviour and approach the design of Σ-Δ modulators.

Quantizer Implications:

The analysis and results for chaos in this chapter were derived for a Σ-Δ modulator with multi-bit quantizer of an arbitrary number of bits. Thus the results we have obtained readily apply to the b-bit quantizer case under the condition of no overload (see Section 1.3 for explanation). A broader study of the low-bit case might make no assumption of the no overload condition. Under these circumstances, the mappings inherent in the Σ-Δ modulator dynamical system become much more complicated to analyze, since we would be effectively "turning off" the quantizer in the system for inputs beyond a certain magnitude but not otherwise. As noted in the Introduction, such a study at the level of analysis conducted in this chapter concerning chaos would be much more complicated with the possibility of full tractability and conclusions unclear. The work of others, e.g. [56], [62], [53], hints at how to mathematically tackle mappings under the 2-bit case with no assumption of the no overload condition, but these ideas do not extend in an obvious way to dealing with the complexity inherent in analyzing our Devaney chaos conditions, particularly when extending the Σ-Δ modulator to general higher-order form (i.e. $M > 2$ or $N > 0$) from a first or second-order form (i.e. $M = 1,\ 2,\ N = 0$).

The no overload condition is a reasonable requirement to make, since Σ-Δ modulators will generally operate with a design and input that yield this, or may be readily controlled to yield this, and since the stability requirement of a bounded filter output tends to make a

no overload quantizer requirement a practical natural extension. Therefore the analysis of chaos for the arbitrarily large multi-bit case in this chapter provides a meaningful framework with which to assert relevant general results on chaos for the low-bit cases of practical interest, along with the extensions to analogous multi-bit cases for $b > 2$. Similarly, the work of Chapters 6, 7 and 8 applied to this model will accomplish this for the respective issues to be investigated ahead.

Chapter 6

Dithered Model and Chaos

In this chapter we extend the dynamical system model of the Σ-Δ modulator in Chapter 2 to include an i.i.d. dither signal, and discuss model formulation issues from this more general perspective. The study of chaos is then extended to the case when such a dither is applied to the system.

6.1 Dithered Model

When the dynamical system model of (1.2) is considered with dither incorporated, the dynamical system (2.1) takes the general form

$$\begin{aligned} \vec{x}_{n+1} &= \mathcal{F}(\vec{x}_n, x_n, \nu_n) \equiv f_n(\vec{x}_n) \\ y_n &= \mathcal{Q}(\vec{x}_n, x_n, \nu_n) \equiv Q_n(\vec{x}_n), \end{aligned} \tag{6.1}$$

for $n \geq 0$, where

$$\vec{x}_n = (r_{n-1}, \ldots, r_{n-N};\ \varepsilon_{n-1}, \ldots, \varepsilon_{n-M}) \in \mathbb{R}^N \times \mathbb{R}^M,$$

$x_n \in \mathbb{R}, \quad \nu_n \in \mathbb{R}, \quad y_n \in (\mathbb{Z} \cdot \Delta + \Delta/2),$

$$\begin{aligned} f_n : (\mathbb{R}^N \times \mathbb{R}^M) &\rightarrow (\mathbb{R}^N \times \mathbb{R}^M) \\ Q_n : (\mathbb{R}^N \times \mathbb{R}^M) &\rightarrow (\mathbb{Z} \cdot \Delta + \Delta/2), \qquad n \geq 0, \end{aligned}$$

with, in particular,

$$\begin{aligned} f_{\varepsilon(1),n}(\vec{x}_n) &= \varepsilon_n = Q(x_n - r_n + \nu_n) - (x_n - r_n) \\ Q_n(\vec{x}_n) &= Q(x_n - r_n + \nu_n), \end{aligned}$$

for $n \geq 0$, and where the remaining notation and definitions from Chapter 2 are unchanged. The dither ν_n, in addition to the input x_n, is now removed as an independent variable and incorporated into the functional form when forming f_n and Q_n from $\mathcal{F}$ and $\mathcal{Q}$.

Error State Space:

From this, we find that the range of the error coordinate $g_1(\vec{x}_n)$, $n > 0$, is now extended from the interval $(-\Delta/2, \Delta/2]$ to the interval $I_D = (a_L - \Delta/2, a_R + \Delta/2]$, where the dither ν_n lies in the interval $[a_L, a_R] \in \mathbb{R}$. While the initial coordinates $g(\vec{x}_0)$ may be chosen to remain bounded on $(-\Delta/2, \Delta/2]$ in practice, we extend their range in the formulation as well, to maintain consistency in the generality of the dynamical system model, (e.g. to allow flexibility in shifting the time step n). This situation raises the issue of what the appropriate definition of the state space should be for $g_1(\vec{x}_n) = \varepsilon_{n-1}$ for the purpose of studying the dynamics, at least for extending our chaos analysis. Clearly, we wish to define a state space that will preserve the symmetry properties of the circle map as much as possible, which represented $(-\Delta/2, \Delta/2]$ as a circle $\mathcal{C}$ for the range in the nondithered case. The central aspect of this symmetry, relating to system (1.2), is the fact that $g_1(\vec{x}_n)$ on $\mathcal{C}$ is continuous as a function of the remaining coordinates of $\vec{x}_n$, and x_n. Clearly this is not the case with either I_D as a subset of R_N, or with the enlarged circle of circumference

$a_R - a_L + \Delta$ formed by joining the ends of I_D, for the dithered case ($a_R > a_L$ generally assumed).

Continuity can be obtained with the following two projections of I_D onto $\mathcal{C}$. First we have the internal quantizer error $q_n = \varepsilon_n - \nu_n$ on $\mathcal{C}$, and second, the value of ε_n itself on $\mathcal{C}$. The latter is given from the circle map projection of a point x from $\mathbb{R}$ to its point modulo $\mathcal{C}$ on $\mathcal{C}$ by $\hat{P}_{\mathcal{C}^1}$. Our dithered case state space should also have the property that the value of $g_1(\vec{x}_n)$ on the state space should be equivalent, if possible, to its usual value on $\mathcal{C}$ when the dither ν_n is insufficient to shift it outside the interval $(-\Delta/2, \Delta/2]$. Only the second of the two continuity cases has this property. Therefore, from the point of view of our definitions of chaos, we will choose the modulo value of $g(\vec{x}_n)$ on the Cartesian product of circles $\mathcal{C}^M$ as the sensible state space of definition. It is believed this approach will thus continue to capture the essential dynamical properties of the error behaviour, while taking advantage of symmetry to allow simplicity.

Analysis and Transformed Variables:

The immediate effect of adding a dither ν_p is to bring about a possible perturbation in the value of the error ε_n, for $n = p$, by strictly integer multiples of Δ. From the circle map projection used, it is obvious that such perturbations do not show up when considering the projection of the error ε_p on $\mathcal{C}$. Thus an immediate property of the state space that we are considering for ε_n is the initial suppression of any randomizing effect of the dither on the error. Under the coefficient condition (R), we will see that this result extends to all $n \geq p$.

Our approach to analyzing the dithered form of system (1.2) for the Σ-Δ modulator is to first transform (1.2), by a simple change of variables, into an equivalent system expressed in terms of the value of ε_n on $\mathcal{C}$ (i.e. modulo $\mathcal{C}$ as defined above). To begin, we let $\breve{\varepsilon}_n = \varepsilon_n - m_n\Delta$, $n \geq -M$, with $m_i = 0$ for $i = -1, \ldots, -M$. For this, we define $m_n \in \mathbb{Z}$ such that $m_n\Delta = Q(x_n + \nu_n - r_n) - Q(x_n - r_n)$ so that $\breve{\varepsilon}_n = Q(x_n - r_n) - (x_n - r_n)$,

where $\varepsilon_n = Q(x_n + \nu_n - r_n) - (x_n - r_n)$, and $n \geq 0$, from (1.2). Thus $\breve{\varepsilon}_n$ is the value of ε_n on $\mathcal{C}$ for our state space definition, and the m_i are the displacement effects, in multiples of Δ caused by the dither ν_n. Substituting the $\breve{\varepsilon}_i$ into the difference equation in (1.2), we then find that we need to transform the r_i to preserve the general form of this equation. Specifically, we let $\breve{r}_n = r_n - \sum_{i=1}^{n} \tilde{r}_i m_{n-i}\Delta$ for $n \geq 1$, with $\breve{r}_n = r_n$ for $n = 0, \ldots, -N$, where the $\tilde{r}_i$ are from coefficient condition (R). Substituting $\breve{r}_n$ into the second quantizer equation of (1.2), we then transform x_n to preserve this equation form. Thus, we let

$$\breve{c}_n = x_n - \sum_{i=1}^{n} \tilde{r}_i m_{n-i}\Delta, \quad n \geq 1, \quad \breve{c}_0 = x_0.$$

The quantizer equation is now $\breve{\varepsilon}_n = Q(\breve{c}_n - \breve{r}_n) - (\breve{c}_n - \breve{r}_n)$. Now the required state space $\breve{\varepsilon}_n$ is described by this transformed system which has the same form as the undithered (1.2). Thus we may apply the analysis and results developed in Chapter 5 to this system to arrive at chaos condition results for the equivalent dithered system under our state space definition.

The first immediate result we have for the new state space dynamics concerns the case when condition (R) holds. For this we have the following:

Theorem 6.1 *Suppose the a_i and b_j satisfy condition (R). Then the value of the error $\varepsilon_n = g_1 \circ f^{n+1}(\vec{x}_0)$, $n \geq 0$, on $\mathcal{C}$ is independent of the dither ν_n, $n \geq 0$. Hence the presence of a dither signal has no net effect on the internal dynamics on $\mathcal{C}^M$ of the Σ-Δ modulator.*

Proof:

With condition (R) holding, we have that $\breve{c}_n = x_n + K_n\Delta$, where $K_n \in \mathbb{Z}$, $\forall\, n \geq 0$. This implies that $Q(\breve{c}_n - \breve{r}_n) - (\breve{c}_n - \breve{r}_n) = Q(x_n - \breve{r}_n) - (x_n - \breve{r}_n)$, $\forall\, n \geq 0$. Thus system (1.2) in the transformed variables is equivalent to (1.2) in the untransformed variables with the dither ν_n set to zero. ■

This theorem demonstrates that, while a dither signal introduces a random element to the dynamics of the Σ-Δ modulator, the structure of the effects of this randomness is such that the randomness will be assimilated out of the error dynamics when condition (R) holds. This is because, unlike the case of a random input from Chapter 5, the effects here are only discrete value perturbations in the value of the quantizer level, and hence perturbations in the gross error value ε_n by a integer multiple of Δ. Condition (R) thus cancels out the net perturbation effects in the same manner as continuity (no interval splitting) is assured as shown in Chapter 5. This type of property does not extend in general to the projections of the initial error conditions $g(\vec{x}_0)$, on $\mathcal{C}^M$, unless the a_i and b_j are all integers, however. For example, the error dynamics of $g(\vec{x}_n)$ on $\mathcal{C}^M$, for $n \geq 0$, when $\vec{x}_0 = (\hat{y}, \hat{z}_1)$ for some $\hat{y} \in \mathbb{R}^N$, $\hat{z}_1 \in \mathbb{R}^M$, will generally be different from that when $\vec{x}_0 = (\hat{y}, \hat{z}_2)$, where $\hat{z}_2 = \hat{z}_1 + (\Delta, \ldots, \Delta) \in \mathbb{R}^M$, even though $\hat{P}_{\mathcal{C}}(\hat{z}_1) = \hat{P}_{\mathcal{C}}(\hat{z}_2)$.

Switching System Formulation:

The fact that $\breve{c}_n$ in the formulation above is dependent on the initial condition $\vec{x}_0$ limits the applicability of this approach. To examine the more general case, when (R) may not hold, we introduce the following formulation of the dithered system (1.2) as a switching system:

$$\varepsilon_n = \sum_{k=1}^{\max(N,M)^*} (a_k - b_k)\varepsilon_{n-k} + \sum_{j=0}^{N} b_j(y_{n-j} - x_{n-j}), \tag{6.2}$$

for $n \geq \max(N, M)$, where

$$y_n = Q(x_n + \nu_n - \sum_{k=1}^{\max(N,M)} ((a_k - b_k)\varepsilon_{n-k} + b_k(y_{n-k} - x_{n-k}))), \quad b_0 \equiv 1,$$

for $n \geq \max(N, M)$. These equations are obtained simply by eliminating the r_n variables from the equations of (1.2). This system then applies to the state space quantity $g(\vec{x}_n) \in \mathcal{C}^M$, for $n \geq \max(N, M)$. The initial state coordinates $g_1(\vec{x}_{\max(N,M)-k})$, for $0 \leq k \leq$

$\max(N,M)^* - 1$, $(\max(N,M)^* \geq 1)$, along with the initial quantizations $y_k = Q(x_k + \nu_k - r_k)$, for $\max(N,M) - N \leq k \leq \max(N,M) - 1$, $(N \geq 1)$, are obtained by iterating through system (1.2) in the usual fashion, from the initial condition $\vec{x}_0$. The nonlinear part of the difference equation for ε_n is given by the vector $(y_n, \ldots, y_{n-N})$. We identify the subsystem "modes" to be the set of vectors of length $N+1$, with entries integer multiples of $\Delta/2$, that this vector may equate to. The system switches to (or remains at) a particular effective "input" mode given by this vector at iteration n according to the values of ν_n, x_n, ε_{n-k}, x_{n-j}, y_{n-j}, for $k = 1, \ldots, \max(N,M)^*$, $j = 1, \ldots, N$, $n \geq \max(N,M)$. If no dither is present and the input is not random, the switching rule is deterministic, and based on the current state of these quantities. If dither ν_n or random input x_n is present, there is then a stochastic as well as deterministic component to the switching rule. In a more practical context, we would typically model the input as fixed, and only the dither as random. Therefore the formulation (6.2) serves to distill the dynamical role of dither, as the most meaningful stochastic element in the system, via a randomization of the switching rule.

Whenever the switching rule, for any iteration n, is initial condition dependent (i.e. dependent on $\vec{x}_0$) via the dependency of y_n on the current state of the deterministic quantities mentioned, the switching system formulation provides a meaningful and structurally non-redundant conversion from one nonlinear system to a set of linear subsystems. This is generally expected to hold for any system that is nonminimum phase, has expansive NTF zeros, or has sufficiently random input or dither; where condition (R) does not hold. When this dependence does not hold, the switching rule will depend only on the iteration n, regardless of the orbit or initial condition, and the overall system reduces to a single linear difference equation system for ε_n. More generally, and when condition (R) holds, this dependence will disappear modulo Δ (i.e. the only dependence will involve additive mul-

tiples of Δ on the RHS of the difference equation), since the system is continuous (linear) except on one "boundary" in $\mathcal{C}^M$.

The switching system we have in (6.2) represents a formulation of the dynamical system of (6.1) where the state space is given entirely in terms of the error coordinates ε_n. Strictly speaking, the error coordinates constitute the part of the state space defined on continuously on subregions of the circles $\mathcal{C}$, while the output coordinates y_n are the part defined on discrete multiples of $\Delta/2$, and are incorporated into the input structure. This serves to emphasize how the external quantities of interest, the errors, may function as the essential and sufficient states describing the dynamics of the Σ-Δ modulator system.

The extension to $\max(N, M)^*$ dimensions is analogous to what we have with the interval difference equation system (4.1), and gives perhaps a less consistent state space structure then the M dimensions associated with $g(\vec{x}_0)$ used for our chaos definitions and analysis (i.e. more/less restrictive if $\max(N, M)^* > / < M$); and one diluted by the form of the b_j coefficients of the r_n. This formulation, however, may open up new methods for confirming or advancing the Σ-Δ modulator dynamics studies in this thesis, beyond the specific treatment of this and the previous chapter. Notice also that a linear combination of the total or "shaped" errors $y_n - x_n$ of the Σ-Δ modulator constitutes the "effective input" in the difference equation for the unshaped error ε_n.

6.2 Chaos with Dither

An obvious result when condition (R) holds, and a result for the more general case when (R) may not hold, follow from the following proposition:

Proposition 6.2 [1] *Suppose there exists a dither ν_n for the system.*

1. *Then Theorems 5.19, 5.20, and 5.23 - 5.27 hold. Theorems 5.29 and 5.30 also hold, with $\mathcal{C}$ and g replaced by $\mathbb{R}$ and $\hat{P}_{\mathcal{C}^1} \circ g$ respectively in the theorem statements.*

2. *Suppose also that the a_i and b_j satisfy condition (R). Then Theorems 5.8 - 5.18, Corollary 5.13 and Proposition 5.16 also hold, with $\mathcal{C}$ and g replaced by $\mathbb{R}$ and $\hat{P}_{\mathcal{C}^1} \circ g$ respectively in the statements of Theorems 5.15 and 5.18.*

Proof:

1. The addition of dither ν_n in the switching system formulation (6.2) has the effect of leading to an affine shift in the boundaries on the domain $\mathbb{R}^N \times \mathcal{C}^M$ of initial conditions at which the system switches between given modes at a given iteration n. The proofs given for Theorems 5.19 and 5.20, did not make any assumptions as to how such a switching rule that preserves these (boundary) properties, and gives rise to discontinuities, is determined. Therefore the approach of these proofs may be extended to the case when dither is present. These theorems, and Theorems 5.29, 5.30, also place no explicit conditions on the form of the external input $\breve{c}_n$ in the dithered system (1.2) under the transformed variables. Thus all of these theorems hold as required.

Considering Theorems 5.23 - 5.27 applied to the dithered (1.2) under the transformed variables, we have $\breve{c}_n = x_n - \sum_{i=1}^{n} \tilde{r}_i m_{n-i} \Delta$, for $n \geq 1$, $\breve{c}_0 = x_0$. The term x_n is statistically

[1] The footnotes to these results in Chapter 5, regarding chaos condition 3, will still hold as well.

independent of $\sum_{i=1}^{n} \tilde{r}_i m_{n-i}\Delta$ in $\breve{c}_n$. This implies that the upper/lower bound on the PDF or PMF, as defined in the theorems, applied to $\breve{c}_n$, will be less/greater than that applied to x_n, and that the corresponding PDF or PMF forms will be the same. It follows that the input $\breve{c}_n$ satisfies the requirements of each of Theorems 5.23 - 5.27 whenever the input x_n satisfies these requirements.

The proofs of all these theorems may each be extended for the more general case when $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$. These theorems thus hold as required.

2. Applying Theorem 6.1, we then have that these results hold for systems with $\vec{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$. The proofs of these results may each be extended for the more general case when $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$. Thus we have the required result. ■

Thus all the chaos results derived in Section 4.1 when condition (R) holds extend simply to the more general case of the presence of some arbitrary dither. From Proposition 6.2, this is true as well of many of the results derived more generally, with (R) not required, where no external input restriction was imposed on the examples used in the proofs. The density of periodic points result of Theorem 5.21 (requiring periodic input) is a notable exception. When condition (R) does not hold, the addition of dither may well alter the error dynamics of the Σ-Δ modulator, and this will generally be true when the probability density of the dither signal is defined over the full interval $(-\Delta/2, \Delta/2]$. The extension of the initial condition domain of $g(\vec{x}_0)$ from $\mathcal{C}^M$ to $\mathbb{R}^M$ has no effect on the applicability of the analogous chaos results, since this change has no net topological effect on the quantities dealt with in the proofs of the results.

We now explore what the nature of the input $\breve{c}_n$ may allow us to conclude about the chaotic properties of the system when condition (R) does not hold. For this, we have the

following results:

Theorem 6.3 [2] *Suppose the a_i and b_j do not satisfy condition (R). Suppose also that there exists a dither ν_n for the system that is described by a piecewise continuous and nonzero probability density function over the interval $[\tilde{a}_L, \tilde{a}_R]$, where $\tilde{a}_R - \tilde{a}_L > \Delta$. Then the system has no periodic points and satisfies chaos condition* 3 *(density of periodic points) with probability* 1.

Proof:

Let $t = \min\{k \mid \tilde{r}_k \ni \mathbb{Z}$, where the $\tilde{r}_k$ are from (R)$\}$. Let $h_n(x|*)$ denote a discrete conditional PMF of $\hat{P}_{\mathcal{C}^1}(\breve{c}_n)$ over $\mathcal{C}$, where $\breve{c}_n = x_n - \sum_{i=1}^{n} \tilde{r}_i m_{n-i}\Delta$, as defined earlier, and $n \geq 0$. For any values of x_i, m_j, $i = 0, \ldots, n$, $j = 0, \ldots, n-t-1$, we have, with the dither PDF of ν_n as given in the theorem, that m_{n-t} takes on at least two values in $\mathbb{Z}$ with nonzero probability, $\forall\ n \geq t$. Thus, with $\tilde{r}_t \ni \mathbb{Z}$ and $\tilde{r}_i \in \mathbb{Z}$, $i = 1, \ldots, t-1$, we have that $h_n(x|x_i, m_j,\ i = 0, \ldots, n,\ j = 0, \ldots, n-t-1)$ describes a discrete distribution over $\mathcal{C}$. Furthermore, $\exists\ K < 1$ such that $h_n(x|x_i, m_j,\ i = 0, \ldots, n,\ j = 0, \ldots, n-t-1) < K$, $\forall\ x \in \mathcal{C}$, $\forall\ n \geq t$, since the area under the PDF of ν_n over an interval of length Δ must always be less than some value $K < 1$. It then follows that $h_n(x|x_i, m_j,\ i = 0, \ldots, k,\ j = 0, \ldots, k-t-1) < K$, $\forall\ x \in \mathcal{C}$, $k = 0, \ldots, n-1$, $n \geq t$. This implies that $h_n(x|\breve{c}_k,\ k = 1, \ldots, n-1) < K$, $\forall\ x \in \mathcal{C}$, $n \geq t$. Thus the transformed input $\breve{c}_n$ satisfies the requirements of Theorem 5.23(a). We may then apply this theorem to obtain the result.

■

[2] Using the standard definition of density, this result will still hold. If existence of periodic points is added to the density definition, then chaos condition 3 is satisfied with probability zero.

Corollary 6.4 [3] *Suppose the conditions of Theorem 6.3 hold with* $\tilde{a}_R - \tilde{a}_L = \Delta$. *Then, with probability* 1, *the system both has no periodic points and satisfies chaos condition* 3, *if and only if there is no periodic point* $\hat{x}_0$ *satisfying* $g_i(\hat{x}_0) = \hat{P}_{\mathcal{C}^1}(\frac{\tilde{a}_R + \tilde{a}_L}{2})$ *for* $i = 1, \ldots, M$.

Proof:

Suppose there exists a periodic point $\hat{x}_0$ satisfying $\varepsilon_{-k} = g_k(\hat{x}_0) \neq \hat{P}_{\mathcal{C}^1}(\frac{\tilde{a}_R + \tilde{a}_L}{2})$, for some $k \in \{1, \ldots, M\}$. Then this implies that $\varepsilon_{\hat{m}p-k} = g_k \circ f^{\hat{m}p}(\hat{x}_0) \neq \hat{P}_{\mathcal{C}^1}(\frac{\tilde{a}_R + \tilde{a}_L}{2})$, $\forall\ \hat{m} \geq 0$, for some $p \in \mathbb{Z}^+$. It follows here that, with the dither PDF of ν_n as given in the corollary, that m_{n-t} (from the definition of $\breve{c}_n$) takes on at least two values in $\mathbb{Z}$ with nonzero probability, when $n = n_{\hat{m}} = \hat{m}p - k + t$, $\forall\ \hat{m} \geq K_m$, for some integer $K_m \geq 0$.

Following the proof of Theorem 6.3, with this infinite sequence $\{n_{\hat{m}}\}$, we may apply Theorem 5.23(a) to obtain the conclusion that $\hat{x}_0$ is a periodic point with probability zero. This contradicts the initial assumption. Therefore it is necessary to have $g_i(\hat{x}_0) = \hat{P}_{\mathcal{C}^1}(\frac{\tilde{a}_R + \tilde{a}_L}{2})$, for $i = 1, \ldots, M$, $n = 0$, for $\hat{x}_0$ to be a periodic point. If $\hat{x}_0$ is such a periodic point, then $\hat{x}_0$ is an isolated periodic point and chaos condition 3 fails. Note that if $\varepsilon_k = \hat{P}_{\mathcal{C}^1}(\frac{\tilde{a}_R + \tilde{a}_L}{2})$, $\forall\ k \geq -M$, then $m_i = 0$, $\forall\ i \geq 0$, so that $\hat{x}_0$ is a periodic point with probability 1. ■

Thus we see that for a piecewise continuous nonzero dither over an interval of length greater than Δ, the random dither achieves the same effect as a persistent random input x_n of the form in Theorem 5.23 in terms of preventing periodic points from existing, when (R) does not hold. Our transformation of the dithered system (1.2) by change of coordinates makes the relationship between the randomizing effect of dither and that of a random input

[3] Using the standard definition of density, this result will still hold. If existence of periodic points is added to the density definition, then chaos condition 3 is satisfied with probability zero, regardless of whether the periodic point $\hat{x}_0$ exists or not.

more explicit, by shifting the dither dependent terms from the quantizer Q to the input $\breve{c}_n$. If the interval of the dither above is only of length Δ, such as an RPDF, then chaos condition 3 holds as well, unless the input x_n is chosen to force a periodic point at the one value in $\mathcal{C}^M$ for which the dither can have no effect on the subsequent dynamics, for $n \geq 0$. In this case, the periodic point is unique, and condition 3 fails.

To comment on this situation further, suppose $\hat{x}_0$ is a periodic point satisfying the condition $g_i(\hat{x}_n) = \hat{P}_{\mathcal{C}^1}(\frac{\tilde{a}_R + \tilde{a}_L}{2})$, for $i = 1, \ldots, M$, $n = 0$, and $n \geq n_1$, for some $n_1 > 0$, when no dither is present. If the dither of Corollary 6.4 is introduced, then we can say that in general, $\hat{x}_0$ will remain a periodic point with probability 1 if $n_1 = 1$, with some probability $\hat{p}$ with $0 \leq \hat{p} \leq 1$ if $n_1 > 1$, and with probability 0 if no such n_1 exists. These results follow from the methods of the proof of the Corollary 6.4. If the length of the dither interval decreases further, then clearly condition 3 becomes harder to ensure, and may remain only over certain sub-cases. This is to say, over a certain subset of $\mathcal{C}^M$, or for certain forms of $p(z)$, or certain forms of input x_n, for example. Condition 3 may not hold at all, or may always hold for some systems with some form of dither. Specific answers are speculative, although the general consequences are clearer.

Theorems 5.24, 5.26 and 5.27 all require stricter conditions on the input $\breve{c}_n$ to apply, than those easily afforded by the dither dependent effects in $\breve{c}_n$. Thus we have no further clear results on the contributions to chaos from the dither when (R) does not hold. It seems reasonable to conjecture, however, that the addition of dither will drive some cases of systems to satisfy sensitivity to initial conditions and/or topological transitivity that did not satisfy some these conditions without dither, and that the prevalence of this would increase with a longer interval of definition for the dither. This conjecture is of particular value in the case of topological transitivity, where there is more room for this condition to be met, and where, for nonminimum-phase systems, such a condition, with Theorems 5.19

and 6.3, would bring about chaos.

Discontinuous Case:

We now briefly examine the dithered case when condition (R) does not hold, from the perspective of considering the situation, as with the nondithered case, of what happens when the application of Proposition 4.1 breaks down. In this case, the discussion follows exactly that of the second paragraph in the beginning of Section 5.3, with one difference. The error difference interval $\Delta\varepsilon_n$ on $\mathcal{C}$ for each subinterval is now interpreted not to contain the point $\frac{\Delta}{2} + \nu_n$ if $k\Delta \leq \nu_n \leq (2k+1)\frac{\Delta}{2}$, and $(2k+1)\frac{\Delta}{2} + \nu_n$ if $(2k+1)\frac{\Delta}{2} \leq \nu_n \leq (2k+2)\frac{\Delta}{2}$, for $k = 0, 1, \ldots$, except possibly at the endpoints. The reason is as follows. With no dither, the normal point $\Delta/2$ of error discontinuity occurs when the input to the quantizer $x_n - r_n$ crosses some multiple of Δ. With dither ν_n added, the input to the quantizer $x_n - r_n + \nu_n$ crosses a multiple of Δ when the predithered quantizer input $x_n - r_n$ crosses $-\nu_n$. This effectively shifts the point of error discontinuity away from $\Delta/2$ as given in the expressions above. Proposition 4.1 then holds over these intervals. These dither induced shifts are replicated in shifts in the switching rule thresholds for the switching system formulation. The stochastic effects of dither, in general, appear to play a greater role in randomizing the switching rules, than in randomizing the effective external input $\breve{c}_n$.

From this discussion, we see that the effect of adding dither to systems where condition (R) does not hold has the effect of creating a form of randomized perturbation of the discontinuities in the interval mappings, and of the locations of the error difference intervals upon which the proposition holds for some n from above. Such behaviour provides additional support for the contention that dither may bring some non-topologically transitive systems to transitivity (and chaos), particularly those with M zeros of magnitude between 1 and 2, for example.

The natural state space on which the points of error discontinuity are shifted back to $\Delta/2$ for all n in the dithered system is given by the internal quantizer error $q_n = \varepsilon_n - \nu_n$. If we make a change of coordinates analogous to that done for $\breve{\varepsilon}_n$, we get $\breve{c}_n = x_n + \nu_n - \sum_{i=1}^{n} \tilde{r}_i \nu_{n-i}$, for $n \geq 1$, $\breve{c}_0 = x_0$. For such an analogous form of the dithered system (1.2), $\breve{c}_n$ will meet the conditions of Theorem 5.23, and hence chaos condition 3 will hold, when any nonzero dither exists. Theorems 5.24 and 5.26 may be applied as well when the dither interval is greater than Δ if (R) holds, and for a sufficiently wide interval otherwise, to give that q_n on $\mathcal{C}^M$ is chaotic, if the zeros of $p(z)$ are not all strictly minimum phase. Although the symmetry given above, and the more clear cut chaos result make q_n look like an appealing state space for the dynamics, this is clearly not appropriate, since we have the onset of an arbitrarily small interval for dither invoking chaos condition 3 in cases where such dither has little or no effect on the actual dynamics of ε_n.

Error Dependencies on Dither:

To comment and compare further, we note that adding a dither ν_n always perturbs the value of q_n at the given value of n, while it never perturbs the value of $\breve{\varepsilon}_n$ at the given n. To consider the effects of previous dithers, we have the following result:

Proposition 6.5 *The value of the error ε_n on $\mathcal{C}$, and the internal quantizer error q_n, at a given $n \geq 0$, are each independent of the previous dithers ν_k, $0 \leq k \leq n$, if the a_i and b_j satisfy condition (R). Moreover, if there exists a dither ν_n for the system that is described by a piecewise continuous and nonzero probability density function over the interval $[\tilde{a}_L, \tilde{a}_R]$, where $\tilde{a}_R - \tilde{a}_L > \Delta$, then these independence results will hold for some $n \geq K$, with $K > 0$ sufficiently large, only if condition (R) is satisfied.*

Proof:

Suppose that condition (R) holds. From relationships stated earlier, we obtain $q_n = \breve{\varepsilon}_n + m_n\Delta - \nu_n = \hat{P}_{\mathcal{C}^1}(\breve{\varepsilon}_n - \nu_n)$, since $q_n \in \mathcal{C}$ and $m_n \in \mathbb{Z}$. From Theorem 6.1 and the i.i.d. dither form, we have that there is no dependence of $\breve{\varepsilon}_n$ or ν_n upon the previous dither values. Thus this result follows for q_n as well.

Now suppose that condition (R) does not hold. Considering the transformed form of (1.2), we have that the value of $\breve{c}_n$ is dependent upon $m_{n-l} \in \mathbb{Z}$, while $\breve{r}_n$ is independent of m_{n-l}, where $l = \min(k \mid \tilde{r}_k \ni \mathbb{Z})$, and $n \geq l$. Thus the value of $\breve{\varepsilon}_n$ is dependent upon m_{n-l}. From the form of the dither PDF, it follows that m_{n-l} and $\breve{\varepsilon}_n$ are dependent upon ν_{n-l}. Applying the above relation for q_n, it follows that q_n is dependent upon ν_{n-l}. ■

Thus the form of $\breve{c}_n$ has enough structure to bring about the independence result for either quantity when (R) holds. If the dither interval is Δ, as with the RPDF case for example, then the necessity condition of (R) will clearly apply when fixed point orbits having $\breve{\varepsilon}_n = \Delta/2$, $n \geq l$, are excluded. For shorter dither intervals or weaker dither conditions, we expect the independence result to fail with at least as much generally for q_n as it will for $\breve{\varepsilon}_n$.

Summary:

In summary, we have with our defined state space, that most of the chaos results for nondithered systems hold when an arbitrary dither is added, and that the dynamics are unchanged if condition (R) holds. In addition, with the addition of sufficient dither, chaos condition 3 holds, and conditions 2 and 1 are conjectured to hold in more cases than with no dither, when (R) is not satisfied. This suggest that dither provides a moderate form of randomization effect on the dynamics of the Σ-Δ modulator, similar to that of a certain level of random input. Overall, we expect the addition of dither to induce more chaotic or

near chaotic behaviour.

Under the conditions of Theorem 6.3, a classification of cases in terms of whether conditions for chaos hold would take an equivalent form to that of the persistent random input segment in Table 5.4 (the fourth table) presented at the beginning of Section 5.4.

Chapter 7

Stochastically Modelled Dynamics

In this chapter, we study the long run error state space dynamics of the Σ-Δ modulator from the point of view of characterizing this behaviour as a stochastic or random process. This approach is partly inspired by consideration of the randomizing effects that are explicitly introduced when dither is added. In particular, with the theorems to follow, we present here the beginnings of a general theory to address the question of how desirable control of the error variance level may be achieved for the Σ-Δ modulator with RPDF dither — a question which will be dealt with in Chapter 8. In this theory, uniformity in the distribution of long run error behaviour is the underlying principle. Statistical and dynamical implications associated with this will be considered as we proceed. First, we present and discuss the theoretical background for how to characterize the error dynamics as a random process.

7.1 Background and Approach

For the subsequent work in this chapter, we focus on the error value ε_n that arises in the Σ-Δ modulator as a random variable. If the value of n is specified, and the values of the

external input x_i and initial condition $\vec{x}_0$ are fixed and known for $0 \leq i \leq n$, and no dither is present, then the corresponding value of ε_n will be known and deterministic with no random uncertainty. If a dither is present under these conditions, and condition (R) holds, then the value of ε_n on $\mathcal{C}$ would still be known, as noted earlier. If the external input (x_i for some $0 \leq i \leq n$) or initial condition is statistically random with some probability distribution, or if appropriately sufficient dither relative to the system is present, then ε_n will be (in general) statistically random with some probability distribution (i.e. with some PDF/PMF) for the specified value of n. For the rest of Chapter 7 and 8, we will use the term "probability density function/PDF" to refer to either a PDF or PMF, and distinguish the two cases, when necessary, with the adjective "continuous/discrete".

For a system with any characterization of input, dither or initial condition, the phenomenon that is often of practical interest is the long run or steady state behaviour of ε_n. This means the behaviour of ε_n following an arbitrarily or at least sufficiently large number of iterations of n following the initial $n = 0$. This is useful because it represents the steady state behaviour of ε_n that would normally be the prevalent condition during a period of observation and use of the Σ-Δ modulator. This would typically be long enough after initialization for any transient aspects of the error dynamics to be negligible, for example. Mathematically, for a system with any form of x_n, $\vec{x}_0$ and dither ν_n, we may regard the steady state of ε_n as a random variable. For this definition, we take this random variable to be the quantity ε_n where n is some arbitrary unknown integer greater than 0.

To proceed with an analysis of the steady state as a random variable, and for subsequent work, we need to introduce the concept of convergence of a sequence of random variables. The random variable $\bar{X}$ with PDF $\bar{h}(\vec{x})$ may be defined to be the limit of a sequence of random variables X_i with PDFs $h_i(\vec{x})$, if $\bar{h}(\vec{x})$ is the limiting PDF in pointwise convergence of the sequence of PDFs $h_i(\vec{x})$ when such a limit $\bar{h}(\vec{x})$ exists. This is called

convergence in distribution. Stronger forms of convergence exist but will not be considered here. Convergence in distribution is normally applied to a sequence of discrete/continuous random variables converging to a discrete/continuous random variable with the appropriate respective PDF definition. For our purposes, we will consider a more general definition of convergence in distribution given as follows:

Definition 7.1 (Convergence in Distribution) *Let X_i, $i \geq 1$, be a sequence of random variables defined over a set $U \subseteq \mathbb{R}^N \times \mathcal{C}^M$. Let $\bar{X}$ be a unique and well defined random variable over U. Then the sequence X_i is defined to converge (in distribution) to the random variable $\bar{X}$ if, for any Borel set $V \subseteq U$, we have* $\lim_{i\to\infty} \text{Prob}(X_i \in V) = \text{Prob}(\bar{X} \in V) \in [0, 1]$.

This definition is stronger than the usual one for convergence in distribution. The requirement that V be a Borel set is necessary for this definition to yield the PDF interpretation mentioned above. (normal sets of consideration have this property.) In subsequent work, we shall describe the PDFs $h_i(\vec{x})$ as converging to the PDF $\bar{h}(\vec{x})$ when the sequence X_i converges to $\bar{X}$ by the definition above, and vice versa.

There is no obvious unique way to precisely define the steady state random variable, as we have conceived it, in the broad context of a general random process. One approach is to look at limiting distributions of random variables, each involving a finite set of random variables $\{X_i\}$ for $i = 0, \ldots, k$, for some $k \geq 0$, and then to consider average distributions. This approach, and the associated Definitions 7.2 and 7.4, are specially constructed to integrate and apply to the work and goals of this thesis. We define the first average to be the limiting distribution or stable distribution $\bar{X}_{(1)} = \lim_{n\to\infty} X_n$, if it exists, where X_n is the random variable with n specified, and the set $\{X_n,\ n \geq 0\}$ constitutes the random process. The second average is defined be the random variable $\bar{X}_{(2)} = \lim_{n\to\infty} X_{p_n(2)}$, if it

exists, where $X_{p_n(2)}$ is a random variable and $p_n(2)$ is chosen, with equal probability, from among the integers in the set $\{0, \ldots, n\}$, with n specified. The PDF of $\bar{X}_{(2)}$ corresponds essentially to the function that a convergent histogram tends to as the sample sequence taken from the respective random variables X_0, $X_1, \ldots$, goes to infinity (see discussion of time series analysis in Section 7.3). If X_n represents the error ε_n, such samples would represent observations from simulation or realizations. In general, the mth average with $m > 1$ is defined as the random variable $\bar{X}_{(m)} = \lim_{n\to\infty} X_{p_n(m)}$, if it exists, where $X_{p_n(m)}$ is the random variable for the value found by choosing, with equal probability, from one of the distributions in the set $\{X_{p_i(m-1)}\}$, with $i = 0, \ldots, n$, with n specified. Note that $p_i(1) = i$, and $p_i(m)$ is itself a discrete random variable for $m > 1$ here. These descriptions lead to the following formal definition:

Definition 7.2 (Average Distributions) *Let $X_{(1),k}$, $k \geq 1$, be a sequence of random variables defined over a set $U \subseteq \mathbb{R}^N \times C^M$, with PDFs $h_{(1),k}(\vec{x})$. The first average distribution is given by $\bar{X}_{(1)} = \lim_{k\to\infty} X_{(1),k}$, if this limit exists. Let $X_{(m+1),k}$ be the random variable with PDF given by $h_{(m+1),k}(\vec{x}) \equiv \frac{1}{k}\sum_{i=1}^{k} h_{(m),i}(\vec{x})$, for $k \geq 1$, $m \geq 1$. Then the "$m+1$"th average distribution is given by $\bar{X}_{(m+1)} = \lim_{k\to\infty} X_{(m+1),k}$, if this limit exists.*

Lemma 7.3 *If the mth average distribution of a sequence of random variables X_i, $i \geq 1$, exists for some $m \geq 1$, then the kth-average of the sequence of X_i will exist and will equal the mth average, for any $k > m$.*

Proof:

We extend the notation to X_i. The random variable $\bar{X}_{(m+1)}$ given by the limit is unchanged if $X_{p_n(m+1)}$, $n \geq n_1$, arises from choosing with respect to the set $\{X_{p_i(m)}\}$, with $i = n_1, \ldots, n$, for some $n_1 \geq 0$. This is because the effect of the contribution of the elements

with $i < n_1$ will go to zero as $n \to \infty$. Now if we let $n_1 \to \infty$, the elements in the set $\{X_{p_i(m)}\}$ will tend to $\bar{X}_{(m)}$ uniformly. Thus $\bar{X}_{(m+1)} = \bar{X}_{(m)}$. By induction, we then get the result. ■

From these definitions, it is clear, with the lemma above, that if a given average distribution exists, then all higher averages will exist and have the same distribution. Furthermore, any such average that exists will satisfy our definition of what a steady state distribution for X_n should be. This is fairly obvious for the first and second averages. Consider the following example that illustrates this for the third average. Let the values of X_i be defined by 0 if $2^{2k} - 1 \leq i \leq 2^{2k+1} - 2$, and by 1 if $2^{2k+1} - 1 \leq i \leq 2^{2k+2} - 2$, for $k \geq 0$. The first-average limit for this X_n jumps between sequences of 0s and 1s and does not converge. The second-average limit will oscillate between $\text{Prob}(0) = \frac{1}{3}$, $\text{Prob}(1) = \frac{2}{3}$, and $\text{Prob}(0) = \frac{2}{3}$, $\text{Prob}(1) = \frac{1}{3}$, every 2^k iterations as $k \to \infty$, but does not converge. Thus a sample histogram will not converge to a steady state PDF profile. The third-average limit converges to $\text{Prob}(0) = \text{Prob}(1) = \frac{1}{2}$. Thus we see that an inherent steady state distribution in the sense of how we defined it exists for this X_n, but requires the third average distribution to uncover it from the pattern of the X_i.

In light of these average distribution properties, we will formally define the steady state of X_n to be the random variable of the mth average distribution defined above, if such an average exists for some $m \geq 1$:

Definition 7.4 (Steady State) *Suppose the mth average distribution $\bar{X}_{(m)}$ of a sequence of random variables X_i, $i \geq 1$, exists for some $m \geq 1$. Then the steady state random variable of the sequence of X_i is uniquely defined to be $\bar{X}_{(m)}$.*

We use the word "average" in average distribution because the mth average PDF is essentially the average of the PDFs of the set of $\{X_{p_i(m-1)}\}$, with $i = 0, \ldots, n$, as $n \to \infty$,

as follows from the definitions. A valid steady state X_n in our conception should always exist. We will not be concerned about a specific mathematical description if no average distribution exists for any $m \geq 1$, because we have no clear way to mathematically define this or its existence, although the subsequent use of the concept of a steady state X_n (or error ε_n) in this thesis would apply to any such form.

Clearly the first average will not yield a steady state ε_n if the quantizer system is deterministic unless ε_n happens to converge to a fixed value. Thus we expect the second average to be significantly more likely to converge to a steady state ε_n. Higher averages require more effort to determine and are less likely to reveal a steady state that is not already present in the second average, since these cases are less generic. Thus, for practical purposes, the second average distribution should generally be considered the most useful means for uncovering a steady state ε_n, and could serve as a more generic, though more restrictive definition of what we mean by steady state ε_n. In this thesis, we utilize some of these second-average advantages.

Under some cases of a particular arbitrary or chosen input and initial conditions, it may not be possible to physically establish the notion of a steady state error ε_n. This would correspond mathematically to the case where no steady state distribution of ε_n exists, or at least the case where no average distribution exists. Observations and simulations of Σ-Δ behaviour show that an effective steady state statistical behaviour of the error ε_n, for all practical purposes, commonly exists. This concept therefore forms a useful framework with which to analyze Σ-Δ behaviour and performance. We will conjecture that a steady state probability distribution for ε_n will generally exist for any initial condition $\vec{x}_0$, if the input x_n is either periodic, or random with some given PDF that is independent and periodic in n. Here it is assumed that the fixed structure of such an input will tend to bring about a fixed structure in the long term behaviour of ε_n in the form of a steady state distribution.

Such inputs are also more characteristic of those used in simulation to study steady state behaviour.

Statistical studies of Σ-Δ modulators in previous research has generally not dealt with the definition or specific concept of the steady state distribution of Σ-Δ modulator quantities. In such work, e.g. [63] by Wannamaker and Lipshitz, the analysis of particular statistical quantities of interest, such as error moments, has been done directly, without considering general random variable convergence issues. For this, the assumption of stationary processes provides sufficient steady state criteria to define and analyze the moments of the stochastic processes of interest. If assumptions of ergodicity are also made, then time series averages over a single simulation may be used to form estimators of moments associated with an ensemble of simulations at a given iteration point. More generally, the existence of a power spectrum implicitly assumes the existence of steady state properties for the data correlations. In this thesis, the study of Σ-Δ modulator random variables will be approached from the steady state perspective outlined above.

To continue, we outline several properties of the steady state distribution concept. From the definition of the random variable, it follows that the PDF of the steady state X_n as the mth average distribution may be expressed by $h_S(\vec{x}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} h_i(\vec{x})$, if this converges to a function $h_S(\vec{x})$ defined over the same domain as the $h_i(\vec{x})$. Here, the $h_i(x)$ are the PDFs of the random variables $X_{p_i(m-1)}$, $i \geq 0$, defined earlier, and $m > 1$. This result generalizes for our general convergence definition.

PDF Mapping Properties:

Now suppose the random variable $X_n \equiv \vec{x}_n$, for all $n \geq 0$, from (2.1), with PDF $h_n(\vec{x})$ over $\mathbb{R}^N \times \mathcal{C}^M$. We denote the mapping of $h_0(\vec{x})$ to $h_n(\vec{x})$ by $h_n(\vec{x}) = f^n(h_0(\vec{x}))$. If the input x_n is not random (i.e. fixed), and no dither ν_n is present, then this mapping

is defined by $h_n(\hat{x}) = \alpha_n \sum_{i=1}^{n_q} h_0(q_i)$, $\hat{x} \in \mathbb{R}^N \times \mathcal{C}^M$, where the q_i are the n_q points in $\mathbb{R}^N \times \mathcal{C}^M$ satisfying $f^n(q_i) = \hat{x}$ if $n_q > 0$, and $h_n(\hat{x}) = 0$ if $n_q = 0$, for $n \geq 0$. We also set $\alpha_n = 1$ if $h_0(\vec{x})$ is discrete, and $\alpha_n = \lim_{r\to\infty} \frac{V(B_r \times \mathcal{C}^M)}{V(f_E^n(B_r \times \mathcal{C}^M))}$ if $h_0(\vec{x})$ is continuous. Here B_r is denotes a ball of radius r about 0 in $\mathbb{R}^N$. The mapping f_E^n in the denominator denotes the mapped region from f^n, without reduction due to overlapping on $\mathcal{C}^M$ (i.e. as an extended mapping on $\mathbb{R}^N \times \mathbb{R}^M$ that is one-to-one from $\mathcal{C}^M$ to $\mathbb{R}^M$). $V(*)$ denotes the volume of the region in its argument. Thus α_n is a scaling normalization parameter in the continuous PDF case, to account for the scaling in the relative PDF values as the volume of its domain changes. Note also that from the properties of (2.1), well defined piecewise continuous PDFs will be mapped by f^n to well defined piecewise continuous PDFs. The mapping definition above will hold, more generally, if the input x_n is independent of the initial condition $\vec{x}_0$, and the dither is i.i.d., since this allows a fixed realization interpretation of these quantities relative to X_0.

Suppose the system input is constant so that $x_n = c$ for all $n \geq 0$. Then each mapping f_i from the system (2.1) is equivalent so that $f_i \equiv \tilde{f}$, $i \geq 0$. If $h_i(\vec{x})$ are the PDFs for the respective random variables $\vec{x}_i$, it follows that $\tilde{f}(h_i(\vec{x})) = h_{i+1}(\vec{x})$, for $i \geq 0$. Suppose also that the steady state $h_S(\vec{x})$ exists at the second average. Using these facts, the summation formula above, and the associative property for adding the PDF contributions from mappings of PDF domains, one can arrive at the result $\tilde{f}(h_S(\vec{x})) = h_S(\vec{x})$. This is deduced as follows:

$$\tilde{f}(h_S(\vec{x})) = \tilde{f}(\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} h_i(\vec{x})) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} \tilde{f}(h_i(\vec{x})) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} h_{i+1}(\vec{x}) = h_S(\vec{x}).$$

We believe this holds regardless of the domain of $h_S(\vec{x})$. This says that, if we regard the dynamical system (2.1), with the given constraints, as a functional that maps the function space of classes of PDFs with domain $\mathbb{R}^N \times \mathcal{C}^M$ to itself, then the PDF of the steady state

$\vec{x}_n$, if it exists in the second average, is a fixed point of the mapping. A PDF $h(\vec{x})$ over $\mathbb{R}^N \times \mathcal{C}^M$ is defined to be a fixed point in general if $\tilde{f}^n(h(\vec{x})) = h(\vec{x})$ holds for all $n \geq 0$. This consequence is consistent with standard functional analysis results.

The easiest and most standard piecewise continuous fixed point PDF to imagine is the uniform PDF over the bounded coordinate space of definition. If a piecewise continuous fixed point of some other form exists, and it is well behaved in the sense of having a finite number of maxima and minima over $\mathcal{C}^M$, then this would imply that the orbit of $g(\vec{x}_0)$, for any initial condition, would lie in a finite element subset of $\mathcal{C}^M$ if $\tilde{f}$ is one-to-one. This is rather special dynamical behaviour. Clearly a fixed point $\hat{h}(\vec{x})$ is a necessary but not sufficient condition for a steady state $\vec{x}_n$ PDF to exist (having PDF $\hat{h}$) in the second average, when the input x_n is constant. If the input x_n of (2.1) is not constant, then the f_i are not all constant, and it is more difficult to interpret (2.1) as a simple mapping functional with simple properties. In general, we would not expect the steady state $\vec{x}_n$ PDF, if it exists, to be a fixed point of such a system with arbitrary initial condition $\vec{x}_0$. Of course if a fixed point PDF is used to characterize the initial conditions randomly, then it is also the steady state PDF in this case.

The steady state distribution for ε_n over $\mathcal{C}$, in general, is given by the marginal distribution $g_M(X_S)$ of the steady state distribution X_S of $\vec{x}_n$ over $\mathbb{R}^N \times \mathcal{C}^M$. This follows, for $M > 1$, because $g_M(\vec{x}_n)$, the earliest mapping iterate, is the independent random variable, and $g_i(\vec{x}_n)$, $i = 1, \ldots, M-1$, are dependent random variables upon this as later iterates. It follows directly that if the fixed point PDF is also the steady state $\vec{x}_n$ PDF, then the steady state PDF over $\mathcal{C}$ for ε_n is the marginal distribution of any of the M ε_i coordinates associated with the steady state $\vec{x}_n$ PDF over $\mathbb{R}^N \times \mathcal{C}^M$. Hence this $\vec{x}_n$ PDF must be symmetric in its ε_i coordinates.

Extensions of the definitions, formulations and discussion about PDF mapping prop-

erties above may be analogously developed for the consideration of the random variable $X_n \equiv \vec{x}_n$, for all $n \geq 0$, from (6.1), with PDF $h_n(\vec{x})$ over $\mathbb{R}^N \times \mathbb{R}^M$; as arises from the more general dithered model formally introduced in Section 6.1. Even for a system with some general dither added, it is relevant and useful however to consider the random variable $X_n \equiv \vec{x}_n$ for the value of $\vec{x}_n$ on $\mathbb{R}^N \times \mathcal{C}^M$, that is with the "value" of the $g(\vec{x}_n)$ part of $\vec{x}_n$ corresponding to the projection of its actual value (on $\mathbb{R}^M$) onto $\mathcal{C}^M$. This projection is given by $\hat{P}_{\mathcal{C}}(x)$, and the PDF for this $\vec{x}_n$ value is then defined over $\mathbb{R}^N \times \mathcal{C}^M$. This random variable definition is consistent with the state space definition on $\mathcal{C}^M$ for the analysis of chaos in Section 6.2. It also provides a formulation that allows for broad results in the next section to be derived — results that in turn may be used to answer statistical questions for dithered systems in Chapter 8.

Now the subsequent theorems and results of this chapter may be presented and discussed.

7.2 Uniform Steady State Results

All of the theoretical results in this section will be concerned with establishing conditions under which the steady state error state space dynamics of $(\varepsilon_{n-1}, \ldots, \varepsilon_{n-q})$ projected on $\mathcal{C}^q$, with $1 \leq q \leq M$, are described by a uniform probability distribution over $\mathcal{C}^q$ (generally $q = M$ or $q = 1$). The results will basically apply to systems satisfying condition (R) which, from Chapter 6, have dithered and nondithered dynamics that are equivalent on $\mathcal{C}^q$. Only Propositions 7.9, 7.12 and 7.15, apply more generally. Thus, for the most part, the nondithered dynamical systems model (2.1) in Chapter 2 will be the formulation applied. Application of continuity relationships from Chapter 4, and their derived implications, will also follow in this context.

Theorem 7.5 *Suppose the a_i and b_j satisfy condition (R). Suppose also that $p(z)$ has at least M zeros, where each zero has either magnitude greater than 1, or else magnitude equal to 1 with multiplicity greater than 1 [counted (multiplicity − 1) times]. Suppose further that either*

(a) the initial condition $\vec{x}_0$ is such that $g(\vec{x}_0)$ is random and described by a piecewise continuous marginal PDF over $\mathbb{R}^M$; or

(b) there exists an input x_{n_1+k}, $k = 0, \ldots, M-1$, that is random and $(x_{n_1}, \ldots, x_{n_1+M-1})$ is described by a piecewise continuous PDF over $\mathbb{R}^M$, for some $n_1 \geq 0$;

where x_n is independent of $\vec{x}_0$, for $n \geq 0$, in (a), and $(x_{n_1}, \ldots, x_{n_1+M-1})$, for $n \geq n_1+M$, in (b). Then the PDF of the long run steady state behaviour of $g(\vec{x}_n)$ will be uniformly distributed over $\mathcal{C}^M$ in the first average distribution.

Proof:

To begin, if a dither ν_n is present, we simply drop the dither from the system. The error and state space dynamics of the undithered system will correspond to the respective dynamics of the dithered system on $\mathcal{C}$ and $\mathcal{C}^M$, by Theorem 6.1. We follow the notation of the proof of Theorem 5.9.

Suppose condition (b) holds. Then, without loss of generality, we relable the state $\vec{x}_{n_1+M+k}$ as $\vec{x}_k$, and the input x_{n_1+M+k} as x_k, $\forall\ k \geq 0$. From the properties of the mapping f, the new $g(\vec{x}_0)$ will be random and satisfy condition (a). Moreover, the long run steady state behaviour of the new shifted system will be equivalent to that of the original system, since the first $n_1 + M$ states in the orbit (of the original system) are finite in number and can be neglected.

Now for the rest of the proof, we suppose condition (a) holds. Choose an $\hat{x}_0 \in \mathbb{R}^N \times \mathcal{C}^M$, and some neighbourhood $\tilde{N}_x$ of $\hat{x}_0$ as given in Lemma 4.10. We define

$$\tilde{D} = \{\vec{y}_0 \in \tilde{N}_x \mid \vec{y}_0 - \hat{x}_0 = \Delta\vec{x}_0 \text{ has the form } \Delta\vec{x}_0 = (\Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-N}; \Delta\varepsilon_{-1}, \ldots, \Delta\varepsilon_{-M}),$$
$$\text{and } \vec{\alpha}_{\tilde{q}} = 0\},$$

and choose a $\vec{y}_0 \in \tilde{D}$. We apply the matrix notation given at the end of Chapter 4. Let $q = M$. We choose $(\Delta\varepsilon_i)_{\tilde{q}}^T = [R_{0\tilde{q}}] \cdot [R_{0q}]^{-1}(\Delta\varepsilon_i)_q^T$, for the first q given initial conditions $(\Delta\varepsilon_i)_q^T$. We denote the projection onto $\mathcal{C}^q$ by $g_{(q)}(*) = (g_1(*), \ldots, g_q(*))$. With the proof of Lemma 4.10, we now have that $g_{(q)}(\tilde{D})$ covers a full $\mathcal{C}^q$ dimensional set in $\mathbb{R}^N \times \mathcal{C}^M$.

Let $\Delta V_{\hat{x}_0} \in \tilde{D}$ be some small rectangular volume element in $\mathbb{R}^N \times \mathcal{C}^M$ (of dimension q), with vertices given by $\hat{x}_0$ and $\vec{y}_{0,i}$, for such choices of $\vec{y}_{0,i} \in \tilde{D}$, which may then be constructed. We have that $g_{(q)}(\Delta V_{\hat{x}_0})$ is a small convex volume element in $\mathcal{C}^q$. Now define the mapping $F^n : \tilde{D} \to \mathbb{R}^N \times \mathcal{C}^M$ by $F^n(\vec{\alpha}_q) = [R_{nq}]\vec{\alpha}_q$, for the $\vec{\alpha}_q$ of $\tilde{D}$. We have, from the nature of the q zeros and (2.3), the following. The magnitude of the projection on each coordinate axis for each side of the rectangle mapping $g_{(q)} \circ F^n(\Delta V_{\hat{x}_0})$ will go to infinity, and thus each side length of $g_{(q)} \circ F^n(\Delta V_{\hat{x}_0})$ will go to infinity, as $n \to \infty$. Since each of the q zeros of $p(z)$ and its eigendirection is expansive, this expansion will scale in a manner so that the volume $g_{(q)} \circ F^n(\Delta V_{\hat{x}_0})$ will go to infinity as $n \to \infty$, with the components of this expansion on $\mathcal{C}^q$ wrapping around $\mathcal{C}^q$ continuously. With the continuity implications arising from condition (R) via Theorem 4.4 and Theorem 4.5, Corollary 4.6 and Proposition 4.7, we have that $g_{(q)} \circ f^n(\Delta V_{\hat{x}_0})$ will be this "wrapping" of $g_{(q)} \circ F^n(\Delta V_{\hat{x}_0})$ about $\mathcal{C}^q$ (i.e. its projection). With this, and the continuity of the mapping f^n, we have the following. For any point $\hat{p} \in \mathcal{C}^q$, there exists, for a sufficiently large n_1, a point $\hat{z}_0 \in \Delta V_{\hat{x}_0}$, such that $g_{(q)} \circ f^{n_1}(\hat{z}_0) = \hat{p}$. This follows from the proof of Theorem 5.9 as well.

For the subsequent treatment, we shall extend the boundaries of the neighbourhood $\tilde{D} \in \mathbb{R}^{N+M}$ so that it constitutes this full domain, where $g_{(q)}(\tilde{D}) = \mathcal{C}^q$. Now we want to characterize the set of points $\vec{y}_0 \in \tilde{D}$ such that $g_{(q)} \circ f^k(\vec{y}_0) = \hat{p}$, for any point $\hat{p} \in \mathcal{C}^q$, and any $k > 0$. From the method of the proof of Theorem 5.9, we define $v(k)_i = \hat{p} - g_i \circ f^k(\hat{x}_0)$,

for $i = 1, \ldots, q$. We then solve the equation

$$(v(k)_i)_q^T + (m_i\Delta)_q^T = [R_{kq}][R_{0q}]^{-1}(\Delta\varepsilon_i)_q^T, \quad \text{for } (\Delta\varepsilon_i)_q^T \in \mathcal{C}^q,$$

where $(m_i\Delta)_q^T$ is a vector with entries that are some integer multiples m_i of Δ. This equation has the solution

$$(\Delta\varepsilon_i)_q^T = [R_{0q}][R_{kq}]^{-1}(v(k)_i)_q^T + [R_{0q}][R_{kq}]^{-1}(m_i\Delta)_q^T.$$

The second term is a linear combination of integer multiples of vectors that span $\mathbb{R}^q$. From the nature of the q nonminimum-phase zeros of $p(z)$, and the resulting form of $[R_{kq}]$, each of these vectors will go to zero in magnitude as $k \to \infty$. We may choose the q integers m_i in any manner such that the elements of the solution vector $(\Delta\varepsilon_i)_q^T$ are in the interval $(-\Delta/2, \Delta/2]$. From these results, it is clear that the solutions for $(\Delta\varepsilon_i)_q^T$, and hence $g_{(q)}(\vec{y}_0) = g_{(q)}(\hat{x}_0) + (\Delta\varepsilon_i)_q$, will be uniformly distributed in a grid structure throughout $\mathcal{C}^q$; with the number of such solutions being roughly of order $|\tilde{\mu}|^k$, for $k \geq 0$, where $|\tilde{\mu}|$ is the product of the magnitudes of the q nonminimum-phase zeros.

Let X_0 be a random variable defined over $\mathbb{R}^N \times \mathcal{C}^M$, such that the marginal distribution of $g_{(q)}(X_0)$ is piecewise continuous as defined over $\mathcal{C}^q$. Now suppose that $h(\vec{x})$ is a piecewise continuous probability density function defined over $\mathcal{C}^q$ of the (marginal) random variable $g_{(q)}(X_0)$. We set

$$\begin{aligned}(\varepsilon_i)_{\tilde{q}}^T &= [R_{0\tilde{q}}] \cdot [R_{0q}]^{-1}(\varepsilon_i)_q^T + (d_i)_{\tilde{q}}^T, \; (r_i)_q^T = (\varepsilon_i)_{\tilde{q}}^T + (e_i)_q^T, \\ (r_i)_{\tilde{q}}^T &= [R_{0\tilde{q}}] \cdot [R_{0q}]^{-1}(\varepsilon_i)_q^T + (e_i)_{\tilde{q}}^T,\end{aligned}$$

for the components of X_0, where $(d_i)_{\tilde{q}}^T \in \hat{\mathcal{C}}^{\tilde{q}}$ (a set) $\cong \mathcal{C}^{\tilde{q}}$ and $(e_i)^T \in \mathbb{R}^N$ are arbitrary and fixed. Thus

$$(\Delta\varepsilon_i)_{\tilde{q}}^T = [R_{0\tilde{q}}] \cdot [R_{0q}]^{-1}(\Delta\varepsilon_i)_q^T, \; (\Delta r_i)^T = (\Delta\varepsilon_i)^T;$$

and we can use $F^n(\vec{\alpha}_q)$, with some arbitrary $\hat{x}_0 = ((e_i), \vec{0}, (d_i)_{\tilde{q}})$, to describe the mappings of the domain of the conditional PDF of $g_{(q)}(X_0)$, given this $\hat{x}_0$, denoted $h(\vec{x}|\hat{x}_0)$. From the piecewise continuity of $h(\vec{x})$, it follows that $h(\vec{x}|\hat{x}_0)$ must be piecewise continuous over $\mathcal{C}^q$, for any $(d_i)_{\tilde{q}}^T$, $(e_i)^T$ in a set S, and hence any $\hat{x}_0$ in a corresponding set $\hat{S}$ satisfying $\text{Prob}(\hat{x}_0 \in \hat{S}) = 1$.

We define the mapping $f^n(h(\vec{x}|\hat{x}_0))$ to be the PDF of the random variable $g_{(q)} \circ f^n(X_0|\hat{x}_0)$, for $n \geq 0$, with a fixed realization of the input x_n assumed as follows from (c). Thus, for some $\hat{q} \in \mathcal{C}^q$, $f^n(h(\hat{q}|\hat{x}_0)) = \alpha_n \sum_{j=1}^{n_q} h(q_j|\hat{x}_0)$, where the q_j are the n_q points in $\mathcal{C}^q$ satisfying $g_{(q)} \circ f^n(X_0|\hat{x}_0) = \hat{q}$, and $g_{(q)}(X_0|\hat{x}_0) = q_j$, for $n \geq 0$, $j = 1, \ldots, n_q$; and $f^n(h(\hat{q})) = 0$, if $n_q = 0$. We define the scaling renormalization α_n by $\alpha_n = \dfrac{V(\tilde{D})}{V(F^n(\tilde{D}))}$, using the notation defined earlier.

From the linear structure of the mapping f, the discussion about volume elements above, and the characterization of points $\vec{y}_0$ with mappings on $\mathcal{C}^q$ to a point $\hat{p}$ given above; it can be concluded that $\forall\ n \geq 0$, the region $\mathcal{C}^q$ can be partitioned into a uniform grid of $n_q(n)$ q dimensional convex volume elements ΔV_j, with $\vec{\alpha}_{\tilde{q}} = 0$, such that $g_{(q)} \circ f^n(\Delta V_j) = \mathcal{C}^q$, and each ΔV_j contains a point q_j. As $n \to \infty$, we have $n_q(n) \to \infty$, and the volume of these elements ΔV_j will go to zero. Thus the uniformly distributed points q_j on $\mathcal{C}^q$ become dense as $n \to \infty$. Let n_p be the largest integer such that $n_p^q \leq n_q$. Then, from the uniform grid structure, we may assume that $\dfrac{1}{(n_p+1)^q} \leq \alpha_n \leq \dfrac{1}{n_p^q}$. This gives $\lim_{n\to\infty} \alpha_n n_q(n) = 1$. From the summation formula above, these results imply that, for any $\hat{q} \in \mathcal{C}^q$, the value of $f^n(h(\hat{q}|\hat{x}_0))$ approaches the average value of the function $f^n(h(\vec{x}|\hat{x}_0))$ over $\mathcal{C}^q$, as $n \to \infty$. Thus $\lim_{n\to\infty} f^n(h(\hat{q}|\hat{x}_0)) = \dfrac{1}{\Delta^q}$. Thus we conclude that the mapping by the system (2.1) of $h(\vec{x}|\hat{x}_0)$ converges to the uniform distribution over $\mathcal{C}^q$ as the number of successive mappings goes to infinity, for any choice of such $\hat{x}_0$ (i.e. any choice of $(d_i)_{\tilde{q}}^T \in \hat{\mathcal{C}}^{\tilde{q}}$, $(e_i)^T \in \mathbb{R}^N$). Since this conditional PDF convergence result holds $\forall\ \hat{x}_0 \in \hat{S}$, we thus have that it holds

equivalently for the marginal PDF $h(\vec{x})$. Since this result holds for any realization of the input x_n, it will thus hold for any random input satisfying the theorem conditions. Thus we have the required result when the PDF of $g(\vec{x}_0)$ is strictly defined over $\mathcal{C}^M$.

Now suppose that the PDF of $g(\vec{x}_0)$ is defined over $\mathbb{R}^M$. We partition a minimal domain in $\mathbb{R}^M$ over which the PDF of $g(\vec{x}_0)$ is defined, into a uniform grid defined by the union of simply connected sets $C_i \in \mathbb{R}^M$ satisfying the following: $\hat{P}_{\mathcal{C}}(C_i) = \mathcal{C}^M$, $\forall\ i$; $C_i \cap C_j = \emptyset$, $\forall\ i \neq j$; $\text{Prob}(g(\vec{x}_0) \in C_i) = \tilde{p}_i > 0$, $\forall\ i$; and $\text{Prob}(g(\vec{x}_0) \in \bigcup_i C_i) = 1$. From (1.2), the dynamics of $g(\vec{x}_n)$ for $g(\vec{x}_0) \in C_l$ will be equivalent to the dynamics of $g(\vec{y}_n)$ for $g(\vec{y}_0) = \hat{P}_{\mathcal{C}}(g(\vec{x}_0)) \in \mathcal{C}^M$, (where $\vec{x}_0$ and $\vec{y}_0$ have the same r_i values), with the first M input values of x_n in the system shifted to account for the translation of $\mathcal{C}^M$ to C_l. With $g(\vec{x}_0)$ satisfying (a) over C_l, and the input shift having no effect on maintaining the requirements, it then follows from applying the proof above, that we have the required result conditionally when $g(\vec{x}_0) \in C_l$. Since C_l was chosen arbitrarily, it follows that the required result holds when $g(\vec{x}_0)$ is defined over $\mathbb{R}^M$. ■

Corollary 7.6 *Suppose the conditions of Theorem 7.5 are satisfied for at least one zero of $p(z)$, and that (c) with either (a) or (b) hold, where $g(\vec{x}_0)$ in (a) and $(x_{n_1}, \ldots, x_{n_1+M-1})$ in (b) may be described by a piecewise continuous marginal PDF over some L dimensional manifold of $\mathcal{C}^M$, with $1 \leq L \leq M$, $L \in \mathbb{Z}^+$. Then the PDF of the long run steady state behaviour of the system error ε_n will be uniformly distributed over $\mathcal{C}$ in the first average distribution.*

Proof:

Without loss of generality, we relable the state $\vec{x}_M$ as $\vec{x}_0$, and the input x_{k+M} as x_k, $\forall\ k \geq 0$. From the properties of the mapping f, the new $g_1(\vec{x}_0)$ will be random and satisfy

condition (a). The rest of the proof then follows that of the proof of Theorem 7.5, with $q = 1$. ■

Theorem 7.7 *Suppose the a_i and b_j satisfy condition (R), and that the conditions on the zeros of $p(z)$ given in Theorem 7.5 hold. Suppose also that the input x_n is either periodic, or is random and described by the random variables X_{x_n} for each x_n respectively, where the X_{x_n} are independent (jointly over all sets) and periodic in n. Suppose further that the initial condition $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$ satisfies* $\text{Prob}(\lim_{t\to\infty} \frac{\tilde{P}_t}{t} = 0) = 1$, *for all $\tilde{P}$, where*

(i) $\tilde{P}_t$ is defined to be the number of elements of a set defined by $\{k \mid g \circ f^k(\vec{x}_0) \in \tilde{P}$ for some $k \in \{0, 1, \ldots, t\}\}$, for the nondithered system; and

(ii) $\tilde{P}$ is any set contained in $\mathcal{C}^M$ for which there is no closed set $\tilde{P}_M \subset \mathcal{C}^M$ of dimension M such that $\tilde{P}_M \subset \tilde{P}$.

Then the PDF of the long run steady state behaviour of $g(\vec{x}_n)$ will be uniformly distributed over $\mathcal{C}^M$.

Proof:

To begin, if a dither ν_n is present, we simply drop the dither from the system. The error and state space dynamics of the undithered system will correspond to the respective dynamics of the dithered system on $\mathcal{C}$ and $\mathcal{C}^M$, by Theorem 6.1.

The system with $g(\vec{x}_0) \in \mathbb{R}^M$ may be adjusted to an equivalent system with $g(\vec{x}_0)$ shifted to $\hat{P}_{\mathcal{C}}(g(\vec{x}_0))$ in $\vec{x}_0$, and the values of x_i, $i = 0, \ldots, M-1$ shifted by an associated amount. This has no effect on the long run behaviour of $\vec{x_n}$, and thus, without loss of generality, we simply take $x_0 \in \mathbb{R}^N \times \mathcal{C}^M$ for the proof that follows.

Suppose that the input x_n, or the input random variable X_{x_n} is periodic with period p. We let $q = M$, and use $g_{(q)}$ as defined in the proof of Theorem 7.5. We also denote

the marginal random variable over $\mathcal{C}^q$, of a random variable X over $\mathbb{R}^N \times \mathcal{C}^M$, by $g_{(q)}(X)$. For the orbit of system (2.1) with the given initial condition $\vec{x}_0$, we define $X_{n,i}$ to be the random variable corresponding to a sample in $\mathbb{R}^N \times \mathcal{C}^M$ from the set $\{\vec{x}_{mp+i},\, 0 \leq m \leq n\}$, with $i = 0, \ldots, p-1$, and where each element is selected with equal probability.

Then we now have the following conditions on the behaviour of $X_{n,i}$, as $n \to \infty$. From the condition on $\vec{x}_0$ with *(i)* and *(ii)*, it follows that no submanifold set in $\mathcal{C}^q$, with a q dimensional volume of zero, may support a nonzero probability that $g_{(q)}(X_{n,i})$ on $\mathcal{C}^q$ will take on a value in the set. This also means that $g_{(q)}(X_{n,i})$ must not take on a value in $\mathcal{C}^q$ with nonzero probability. These results imply that the steady state error coordinate of $g_{(q)}(\vec{x}_{mp+i})$, m arbitrary, if it exists via the second average with $\lim_{n\to\infty} g_{(q)}(X_{n,i})$, has a probability density function that we may define as piecewise continuous over $\mathcal{C}^q$. These results also imply, analogously, that the PDF of the random variable $g_{(q)}(X_{n,i})$ will increasingly tend towards a piecewise continuous PDF as $n \to \infty$; that is $\lim_{n\to\infty}(g_{(q)}(X_{n,i}) - Y_{n,i}) = 0$, for some sequence of random variables $\{Y_{k,i},\, k \geq 0\} \in S_{Y_i}$, where S_{Y_i} is defined to be a set of random variables with piecewise continuous PDFs defined over $\mathcal{C}^q$. We form S_{Y_i} as a set of elements with PDFs having a given bound $K_i > 0$, as allowed by the following: the converse (of such an allowance) necessarily implies the existence of some sequence $\{n_j\}$, with $n_{j+1} > n_j$, $j > 0$, such that the PDF of $g_{(q)}(X_{n_j,i})$ becomes unbounded as $j \to \infty$. This is not possible with the condition on $\vec{x}_0$ holding.

Now we show that the first average distribution $\bar{Z}_{(1)}$, (i.e. the limiting distribution), of the sequence $Z_{(1),m,i} = g_{(q)} \circ \hat{f}^{mp+i}(\lim_{k\to\infty} X_{k,i})$, $m \geq 0$, converges to a random variable $\bar{Y}_{(1)}$ which is uniform, and that $\lim_{n\to\infty} g_{(q)}(X_{n,i}) = \bar{Y}_{(1)}$. For this, we define the mapping $\hat{f}^{mp+i} \equiv f_{(m+1)p+i-1} \circ \ldots \circ f_i$, for $i = 0, \ldots, p-1$. We may extend the set S_{Y_i} to a set of random variables $\tilde{S}_{Y_i}$ defined over $\mathbb{R}^N \times \mathcal{C}^M$, such that $S_{Y_i} = g_{(q)}(\tilde{S}_{Y_i})$. This is sufficient to represent the convergent nature of $X_{n,i}$: we have again that $\lim_{n\to\infty} g_{(q)}(X_{n,i})$ converges to S_{Y_i}.

Now consider the value of $g(\vec{x}_{m_1p+i})$ in this system for some arbitrary, unknown $m_1 \geq 0$. From our conception of a steady state distribution, we say that the random behaviour of $g(\vec{x}_{m_1p+i})$ is characterized by S_{Y_i}, meaning that the notion of a steady state PDF for $g(\vec{x}_{m_1p+i})$ is constituted randomly from a PDF defined over the function space of PDFs corresponding to elements of S_{Y_i}. Therefore we are justified to form average distributions from S_{Y_i} to try to converge the domain (given by this space of PDFs) to a unique PDF defined over $\mathcal{C}^q$, existing with probability 1.

We find that the first average distribution $\bar{Y}_{(1)}$ of the sequence $Y_{(1),m,i} = g_{(q)} \circ \hat{f}^{mp+i}(Y)$, $m \geq 0$, is a random variable with the uniform distribution over $\mathcal{C}^q$, $\forall\ Y \in \tilde{S}_{Y_i}$. This follows from applying Theorem 7.5 on the set $\tilde{S}_{Y_i}$. Part (c) of this theorem is satisfied via the independent form of the input x_n and X_{x_n} here. This convergence is uniform over all elements of S_{Y_i}, since the rate of convergence will be an increasing function of the supremum values of the elements of S_{Y_i}, and these elements are bounded by K_i.

From the properties and arguments given above, we thus have that $\bar{Z}_{(1)} = \bar{Y}_{(1)}$. By the associative property that arises from our definition of PDF mappings; $Z_{(1),m,i} = \lim_{k\to\infty} g_{(q)} \circ \hat{f}^{mp+i}(X_{k,i})$. We now apply the cyclic property of the mappings f: i.e. $f_l = f_{l+mp}$, $\forall\ l$, $m \geq 0$. The result then reduces to

$$Z_{(1),m,i} = \lim_{k\to\infty} g_{(q)}(X(m+1)_{k,i}) = \lim_{k\to\infty} g_{(q)}(X_{k,i}),$$

where $X(m+1)_{k,i}$ is the random variable corresponding to a sample in $\mathbb{R}^N \times \mathcal{C}^M$ from the set $\{\vec{x}_{jp+i},\ m+1 \leq j \leq k\}$, with $i = 0, \ldots, p-1$, and where each element is selected with equal probability. Thus $\bar{Y}_{(1)} = \bar{Z}_{(1)} = \lim_{n\to\infty} g_{(q)}(X_{n,i})$ has the uniform PDF $h_{\bar{Y}}$ over $\mathcal{C}^q$. The marginal PDF for the steady state of $g_{(q)}(\vec{x}_n)$ is now given (in the second average) by $\frac{1}{p}\sum_{i=0}^{p-1} h_{\bar{Y}} = h_{\bar{Y}}$, over $\mathcal{C}^q$. Thus we have the required result. ■

Corollary 7.8 *Suppose the conditions of Theorem 7.7 are satisfied for at least one zero of* $p(z)$*, and that the conditions on the input* x_n *and initial condition* $\vec{x}_0$ *hold with* g *and* M *replaced by* g_1 *and* 1 *respectively. Then the PDF of the long run steady state behaviour of the system error* ε_n *will be uniformly distributed over* $\mathcal{C}$*.*

Proof:

The proof follows the proof of Theorem 7.7, with $q = 1$, and with the application of Theorem 7.5 replaced by Corollary 7.6. ■

Theorems 7.5 and 7.7 link the existence of a uniform steady state error coordinate distribution over $\mathcal{C}^M$ with nonminimum-phase systems that satisfy the equivalent conditions of Theorem 5.9 for topological transitivity. Their corollaries analogously link a uniform steady state error ε_n to the equivalent conditions of Theorem 5.8 for sensitivity to initial conditions, which only require one nonminimum-phase zero.

Theorem 7.5 essentially shows that piecewise continuous probability distributions of $g(\vec{x}_n)$ converge through successive mappings towards the uniform distribution over $\mathcal{C}^M$, and its corollary analogously for ε_n and $\mathcal{C}$. The expansive nature of the mappings tend to average and hence smooth out the mapped PDFs, leading to a uniform steady state.

Theorem 7.7 and Corollary 7.8 present conditions under which the results of Theorem 7.5 and Corollary 7.6 may be extended to more general probabilistic cases: most generically when the initial condition $\vec{x}_0$ is fixed or random with a discrete distribution, and/or there is some discrete randomness in the input x_n. As such, Theorem 7.7 and its corollary are of more practical use, since they focus on behaviour with a specific realized initial condition.

The condition on $\vec{x}_0$ in Theorem 7.7 is constructed so that the probability density function of the steady state error coordinate $g(\vec{x}_n)$ on $\mathcal{C}^M$ can be properly defined as a

piecewise continuous function over $\mathcal{C}^M$. This condition, with *(i)* and *(ii)*, essentially say that no submanifold set in $\mathcal{C}^M$ with an M dimensional volume of zero may support a nonzero probability that the steady state $g(\vec{x}_n)$ over $\mathcal{C}^M$ will take on a value in the set. This is a generalization of saying that the steady state random variable $g(\vec{x}_n)$ over $\mathcal{C}^M$ must not take on any value in $\mathcal{C}^M$ with a nonzero probability and, more specifically, that $g(\vec{x}_0)$ cannot be a periodic point.

It is expected that in typical topologically transitive systems with sensitivity to initial conditions (particularly those that are chaotic), the initial conditions failing to satisfy this condition form a set of measure zero. Thus one would expect that a random initial condition chosen in $\mathbb{R}^N \times \mathcal{C}^M$ according to a piecewise continuous PDF over $\mathcal{C}^M$ would satisfy the theorem condition with probability one. This is consistent with the prevalence of nonperiodic points in generic topologically transitive systems we see from the study of the dynamics involved with the work of Chapter 5 of this thesis. Thus the implication of a steady state uniform error PDF over $\mathcal{C}$ when the system is fully nonminimum phase (or expansive) from Theorem 7.7 would be of practical relevance. Indeed it is the assumption of this result, under generic initial conditions and the periodic input form, that gives this theorem its real importance. Establishing a proof of the second sentence of this paragraph (if possible) would then fully complete the immediate accomplishments of this theorem. It would, of course, be useful to have results that would enable one to designate classes of systems as having the condition on $\vec{x}_0$ met by all initial conditions, as is the case with the quasiperiodic system in Theorem 7.13 and some extended results in [22].

The periodic conditions on the input x_n in Theorem 7.7 are an important requirement to afford the result. With a fixed initial condition, it is always possible to choose an input that will bring about a densely distributed steady state behaviour (i.e. with piecewise continuous PDF and the theorem condition on $\vec{x}_0$ satisfied), that is not uniformly distributed. (Indeed

any steady state PDF may be so created.) Unlike in Theorem 7.5, additional averaging structure in the input, together with the expansive properties of the mappings, is thus required to bring about a uniform PDF result here. We may speculate that a more general input structure, such as quasiperiodic, or one with an asymptotic steady state behaviour that is periodic or quasiperiodic may be sufficient. It is also possible and intuitively plausible from previous work that, for general input, the set of initial conditions for which the steady state would not be uniform (nonuniform piecewise continuous now included) would still form a set of measure zero.

The proof of Theorem 7.7 essentially involves applying Theorem 7.5, and arguing that the error coordinate, as a random variable over $\mathcal{C}^M$, should converge to a random variable with unique piecewise continuous PDF in the steady state; and that this PDF should then be the actual function that general piecewise continuous PDFs converge to themselves under successive mappings, rather then some other arbitrary function and rather than a condition of nonconvergence.[1]

When the conditions of Theorem 7.10, Proposition 7.9 or the final part of Proposition 7.12 hold, the uniform PDF over $\mathcal{C}^M$ is basically a fixed point of the dynamical system (6.1) viewed as a functional that maps a space of probability density functions for $\hat{P}_{\mathcal{C}}(g(\vec{x}_0))$ to itself, as follows from our discussion in the previous section. With the existence of such a PDF fixed point, the proof of Theorem 7.5 therefore reflects characteristics of the system that we would expect when the contraction mapping theorem applies. In the proofs of Theorems 5.20 and 5.21, we showed that under the conditions of Theorem 5.20

[1]Specifically, the error coordinate ε_n, (or state $\vec{x}_n$ in the full theorem) is shown to converge to a set of random variables described by piecewise continuous functions. The mapping f^n is then applied, with $n \to \infty$, to this process. The convergent sequence of random variables is mapped to itself, via the input structure; while the set of random variables in the they converge to is mapped uniformly to a uniform distribution (1-element set), by Theorem 7.5.

for topological transitivity in the general case, we have "inverse" contraction mappings over neighbourhoods of $\mathcal{C}^M$ as a means to this transitivity. Thus, by analogy to the condition (R) environment, we may speculate that an extension of Theorem 7.5 to meet the transitivity conditions of Theorem 5.20 would also pertain to systems which reflect these contraction mapping characteristics. Hence we would speculate that such extensions of Theorems 7.5, 7.7 and their corollaries (i.e. to require zeros of magnitude greater that 2, but not (R)) would hold for nondithered systems. It is less clear that such an approach could be used to prove analogous results in the case where condition (R) does not hold with a possible dither present.

More generally, we could speculate that, under certain conditions, a uniform steady state PDF for $g(\vec{x}_0)$ or ε_n may exist. As discussed in the previous section, we might expect that the PDF would generally not be a fixed point of the functional. It can be proven, however, that the only piecewise continuous PDF fixed point that any one-to-one topologically transitive system may have is in fact one that is uniform over $\mathcal{C}^M$. If one could apply on the space of piecewise continuous PDFs over $\mathcal{C}^M$, for some class of one-to-one topologically transitive systems, or some class of systems satisfying Theorem 7.10 or the final part of Proposition 7.12; either the convergence proof for Theorem 7.5 or the contraction mapping theorem; one would then arrive at the result of Theorem 7.5, Proposition 7.12 or their corollaries, for this class of nondithered systems (by the same logic).

The following result asserts the fixed point claim:

Proposition 7.9 *Suppose the system satisfies chaos condition* 2 *(topological transitivity). Suppose also that the mapping* $\hat{P}_{\mathcal{C}} \circ g \circ f_n$ *in (6.1), from* $\hat{P}_{\mathcal{C}}(g(\vec{x}_0))$ *to* $\hat{P}_{\mathcal{C}}(g \circ f_n(\vec{x}_0))$, $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$, *is one-to-one over* $\mathcal{C}^M$ *for all* $n \geq 0$. *Suppose further that the input* x_n *is independent of* $\vec{x}_0$, *for* $n \geq 0$. *Then, if a* $\hat{P}_{\mathcal{C}}(g(\vec{x}_n))$ *marginal PDF fixed point of*

the system (6.1) viewed as a functional exists for some $\vec{x}_0$ PDF, with the $\hat{P}_{\mathcal{C}}(g(\vec{x}_n))$ PDF piecewise continuous over $\mathcal{C}^M$, this marginal PDF must be uniformly distributed over $\mathcal{C}^M$.

Proof:

Suppose the given system has a $\hat{P}_{\mathcal{C}}(g(\vec{x}_n))$ PDF fixed point for some given $\vec{x}_0$ PDF, and that the fixed point is also piecewise continuous over its domain $\mathcal{C}^M$. Let $\hat{h}(\vec{x})$ be the PDF of this $\vec{x}_0$, and let $\hat{h}_g(\vec{x})$ denote its $\hat{P}_{\mathcal{C}}(g(\vec{x}_0))$ marginal PDF over $\mathcal{C}^M$. Let $\hat{x} \in \mathbb{R}^N \times \mathbb{R}^M$, $\hat{y} \in \mathcal{C}^M$ be any two points (in particular with $\hat{P}_{\mathcal{C}}(g(\hat{x})) \neq \hat{y}$). Now, from topological transitivity, it follows that there exist two sequences of points $\hat{x}_i \in \mathbb{R}^N \times \mathbb{R}^M$, $\hat{y}_i \in \mathcal{C}^M$; with $\hat{h}_g(\vec{x})$ continuous over some simply connected closed sets $\tilde{X}_i, \tilde{Y}_i \subset \mathcal{C}^M$, containing $\hat{P}_{\mathcal{C}}(g(\hat{x}_i))$, $\hat{P}_{\mathcal{C}}(g(\hat{x}))$, and $\hat{y}_i$, $\hat{y}$, respectively, for $i \geq 0$; and satisfying the following: $\lim_{n\to\infty} \hat{x}_n = \hat{x}$, $\lim_{n\to\infty} \hat{y}_n = \hat{y}$, and $\hat{P}_{\mathcal{C}}(g \circ f^{n_i}(\hat{x}_i)) = \hat{y}_i$, for some $n_i > 0$, $\forall\ i \geq 0$. From the mapping properties of $\hat{h}$, the fixed point property (i.e. $f^n(\hat{h})_g = \hat{h}_g$, $\forall\ n \geq 0$), and the one-to-one property of $\hat{P}_{\mathcal{C}} \circ g \circ f$; it must hold that $\hat{h}_g(\hat{P}_{\mathcal{C}}(g \circ f^n(\vec{x}_0))) = \hat{h}_g(\hat{P}_{\mathcal{C}}(g(\vec{x}_0)))$, $\forall\ n \geq 0$, and $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$. The mapping properties include the assumption of a fixed realization of the input x_n and dither ν_n that follows from x_n, ν_n and $\vec{x}_0$ being independent, for $n \geq 0$. Applying this to the sequences above, we then have that $\hat{h}_g(\hat{P}_{\mathcal{C}}(g(\hat{x}_i))) = \hat{h}_g(\hat{y}_i)$, $\forall\ i \geq 0$. From the continuity of $\hat{h}_g(\vec{x})$ over the sets $\tilde{X}_i$ and $\tilde{Y}_i$, it then follows that $\hat{h}_g(\hat{P}_{\mathcal{C}}(g(\hat{x}))) = \hat{h}_g(\hat{y})$. Since $\hat{x}$ and $\hat{y}$ were chosen arbitrarily, it then follows that $\hat{h}_g(\vec{x})$ is uniform over $\mathcal{C}^M$. Thus we have the result. ■

The requirement of this proposition that the mappings be one-to-one essentially excludes systems that are nonminimum phase or have expansive NTF zeros. It is not obvious whether this condition can be relaxed, and the result extended to include expansive systems which would tend to have PDF mappings that are more contractive.

The following theorem captures the idea of a uniform distribution over $\mathcal{C}^M$ as a fixed point of the system (6.1) as a functional, and shows how this holds whenever a somewhat more restrictive version of condition (R) holds, or when the coefficients of a $\max(N, M)$ order $p(z)$ are integers. No explicit requirements on the zeros of $p(z)$ are made.

Theorem 7.10 *Suppose that one of the following conditions is satisfied:*

(a) the a_i and b_j satisfy condition (R), with $\tilde{r}_{q_1} \neq 0$, and $\tilde{r}_i = 0$ for all $i > q_1$, for some q_1 satisfying $M \leq q_1 \leq \max(N, M)$;

(b) $a_i - b_i \in \mathbb{Z}$, for $i = 1, \ldots, \max(N, M)$, $a_{q_1} - b_{q_1} \neq 0$ if $N = M$, where $q_1 = \max(N, M)$, and the dither $\nu_n = 0$, for all $n \geq 0$.

Suppose also that the "extended" initial error conditions ε_k, $k = -1, \ldots, -q_1$, are random, i.i.d. over $\mathcal{C}$, and jointly uniform, so that $g(\vec{x}_0)$ has a uniform PDF over $\mathcal{C}^M$. Suppose further that the initial internal state conditions r_k, $k = -1, \ldots, -N$, are random variables defined by $r_k = \sum_{i=1}^{q_1+k} \tilde{r}_i \varepsilon_{k-i}$, where the $\tilde{r}_i$ come from condition (R), if (a) holds; and $r_k = \varepsilon_k$ if (b) holds. Suppose as well that the input x_n is independent of x_j and $\vec{x}_0$, for all $0 \leq j < n$ and $n \geq 0$. Then $g(\vec{x}_n)$ will have a PDF that is uniformly distributed over $\mathcal{C}^M$ for all $n \geq 0$. In particular, the system error ε_n will have a PDF that is uniformly distributed over $\mathcal{C}$ for all $n \geq 0$, when (a) or (b) hold for some q_1 satisfying $1 \leq q_1 \leq \max(N, M)$.

Proof:

To begin, if a dither ν_n is present in (a), we simply drop the dither from the system. The error and state space dynamics of the undithered system will correspond to the respective dynamics of the dithered system on $\mathcal{C}$ and $\mathcal{C}^M$, by Theorem 6.1.

Suppose that the errors ε_i, for $i = k - 1, \ldots, k - q_1$, are an i.i.d. independent family (i.e. a jointly independent set) with uniform probability densities over $\mathcal{C}$.

(a) First, suppose condition (a) holds. Suppose that the r_i are random variables defined by $r_i = \sum_{j=1}^{q_1+i} \tilde{r}_j \varepsilon_{i-j}$, for $i = k-1, \ldots, k-N$, for some $k > 0$. By the conditions of the theorem, this holds when $k = 0$. Then, substituting the expressions for the r_i into the difference equation in (1.2), and using the definition of the $\tilde{r}_i$ in (R), we see that this expression holds for $i = k$. By induction, we see the statement is thus valid $\forall\ k \geq 0$.

Now applying the condition $\tilde{r}_i = 0$, for $i > q_1$, we have $r_k = \sum_{j=1}^{q_1} \tilde{r}_j \varepsilon_{k-j}$, for $k \geq 0$. By condition (R), the coefficient of each term in this sum will be an integer. An integer times a uniform distribution over $\mathcal{C}$ simply stretches the distribution and wraps it around $\mathcal{C}$ an integer number of times, thus giving back a uniform distribution over $\mathcal{C}$. Thus each nonzero term in the sum is uniformly distributed over $\mathcal{C}$, and these are not all zero since $\tilde{r}_{q_1} \neq 0$. Since the distribution of each nonzero term in the sum is independent, and these terms form an independent family, the distribution of r_k will be the convolution of up to q_1 uniform or RPDF distributions over $\mathcal{C}$. By Corollary 8.4, we then have that r_k is uniformly distributed over $\mathcal{C}$. From the independence conditions on x_n in the theorem, and (1.2), it follows that x_k and r_k will be independent. The quantizer input $u_k = x_k - r_k$ will then be uniformly distributed over $\mathcal{C}$, since this equation simply inverts (additively) the distribution of r_k, and then shifts it independently by x_k. Since there is a one-to-one mapping between the value of u_k on $\mathcal{C}$ (inverted again) and the value of ε_k on $\mathcal{C}$, we have from the form of this mapping given in the proof of Theorem 8.5, that ε_k is uniformly distributed over $\mathcal{C}$.

(b) Now suppose condition (b) holds. From the arguments above, it is clear that we may write $r_k = \sum_{j=1}^{q_1} (a_j - b_j) T_{j,k}(\varepsilon_{k-j})$, for $k \geq 0$, where the $T_{j,k} : \mathcal{C} \to \mathcal{C}$ are affine mappings with a stretch factor of 1. Each $T_{j,k}(\varepsilon_{k-j})$ will then be uniformly distributed over $\mathcal{C}$, and independent over all such j. With integer coefficients of the terms (not all zero)

given from condition (b), we may then apply the arguments used for the analogous sum from case (a) above to this sum, to give that r_k, and hence ε_k, are uniformly distributed over $\mathcal{C}$ here.

Now let $r_{k,i}$ be the sum of any m terms from the set $\{\hat{r}_j,\, j = 1, \ldots, q_1 - 1\}$, with m any integer between 1 and $q_1 - 1$, and with $\hat{r}_j = \tilde{r}_j \varepsilon_{k-j}$ if (a) holds, or $\hat{r}_j = (a_j - b_j) T_{j,k}(\varepsilon_{k-j})$ if (b) holds. Let $r_{k,a} = r_k - r_{k,i}$. From the above arguments, we have that $r_{k,a}$ has a uniform PDF over $\mathcal{C}$. Since the ε_i, $i = k-1, \ldots, k-M$, are an independent family, it follows that $r_{k,i}$ and $r_{k,a}$ are independent. The PDF for r_k given $r_{k,i}$ then will simply be the PDF of $r_{k,a}$ shifted by some given $r_{k,i}$ value, which is a uniform PDF over $\mathcal{C}$ as well. Since this is the PDF of r_k on its own, r_k is independent of $r_{k,i}$. With this holding over all such $r_{k,i}$, we then have sufficient conditions to say that the ε_i, $i = k, \ldots, k - q_1 + 1$, are an i.i.d. independent family. With the previous result for the PDFs, this then implies that $g_q(\vec{x}_{k+1})$, as well as $g_q(\vec{x}_k)$, has a uniform PDF over $\mathcal{C}^q$, with $1 \leq q \leq q_1$. With this holding for the $k = 0$ case, we then have, by induction, that this holds $\forall\ k \geq 0$. Thus we have the required result of the theorem, when $q = M$. With $q = 1$, it follows that ε_k will have a uniform PDF over $\mathcal{C}$ $\forall\ k \geq 0$. ■

The initial densities of the internal states r_i given in Theorem 7.10 are needed to at least keep the errors ε_i in $g(\vec{x}_n)$ mutually independent, and also help in the recursion and the application of condition (R). It is possible that, with no conditions imposed on these r_i in the theorem, the error ε_n could still be shown to be uniformly distributed over $\mathcal{C}$. It is unclear that the independence of the ε_i in $g(\vec{x}_n)$ can also be shown to then conclude that $g(\vec{x}_n)$ is uniformly distributed over $\mathcal{C}^M$. The more restricted form of condition (R) was required for basically the same purpose with higher iterations. The same speculations may be made about the results when the added restrictions on (R) are relaxed. The restriction

is sufficiently satisfied if $\tilde{r}_i = 0$ for $i = M+1, \ldots, M+N$.

In Theorem 7.10, the definition of a fixed point of system (6.1) as a functional presents this fixed point as the initial condition. It follows immediately that this fixed point is also the steady state distribution of the system when it serves as the initial condition. Note that if Theorem 7.5 is satisfied but Theorem 7.10 is not, then the uniform steady state PDF may not be a fixed point, at least if the input x_n is not constant, even though this PDF must converge back to itself. The continuity results of Chapter 4 generally cannot be applied at the endpoint $\Delta/2$ of $\mathcal{C}$ with condition (R) unless $M = 1$, or all the coefficients a_i, b_j are integers. This provides the difficulty in setting up a uniform PDF fixed point in $\mathcal{C}^M$. Conversely, under these conditions, we might expect that an extension of Theorem 7.10 is readily possible, as partly supported in condition (b) of the theorem.

From a practical point of view, Corollaries 7.6, 7.8 and Theorem 7.10 say that if we choose the initial conditions randomly according to the PDF specifications given in the theorems over an ensemble of p simulations of a given Σ-Δ modulator satisfying (R), then, as $p \to \infty$, the average error ε_n at a given iteration n averaged over the observed errors in the ensemble, will have a PDF that approaches a uniform distribution over $\mathcal{C}$, for any n.

Theorems 7.5, 7.7 and 7.10 have been constructed and proved so that their results are generalizable to any dimension q, and $\mathcal{C}^q$, with $1 \leq q \leq M$. Note that unlike Theorem 7.5; Theorems 7.10, 7.13, and Proposition 7.12 to follow are not exclusive to requiring nonminimum-phase zeros.

With the following, we examine the relationship between steady state error behaviour and the existence of a white error process, as afforded by Theorems 7.5, 7.7, their corollaries, and Theorem 7.10:

Theorem 7.11 *Suppose that one of the following hold:*

1. *The conditions of Corollaries 7.6 or 7.8 are satisfied for at least* 2 *zeros of* $p(z)$*, and with the corresponding adjustments in the other corollary conditions relating to this dimensionality.*

2. *Conditions (a) or (b) of Theorem 7.10 are satisfied, for some* q_1 *satisfying* $2 \leq q_1 \leq \max(N, M)$*. We define the polynomial* $p_R(z) = z^{q_1} + \sum_{k=1}^{q_1} \tilde{r}_k z^{q_1-k}$ *for (a).*

Then the PDF of the long run steady state behaviour of $(\varepsilon_{n+\tau}, \varepsilon_n)$ *will be uniformly distributed over* $\mathcal{C}^2$*, for all* $\tau \geq 1$ *when* 1 *holds; and if and only if the zeros of* $p_R(z)$ *or* $p(z)$ *are not all magnitude one, multiplicity one, with rational complex arguments, when* $2(a)$ *or* $2(b)$ *hold, respectively. Moreover, the spectrum of the error* ε_n *on* $\mathcal{C}$ *(i.e.* $\hat{P}_{\mathcal{C}^1}(\varepsilon_n)$*) will be white in* 1*; and in* 2 *if and only if this zero condition holds as well.*

Proof:

The joint uniform distribution of the steady state behaviour of $(\varepsilon_{n+1}, \varepsilon_n)$ over $\mathcal{C}^2$ follows from the proofs of Corollaries 7.6 and 7.8, and Theorem 7.10, with $q = 2$.

1. We give a generalized extension of the proofs Corollaries 7.6 and 7.8, with the following. The matrix $[R_k]$ is replaced with $[R_k]_\tau$, for $k \geq \tau - 1$, where $[R_k]_\tau$ is the $\max(N, M)^* \times \max(N, M)^*$ matrix, with row 1 equal to row 1 of $[R_k]$, and rows 2 to $\max(N, M)^*$ equal to rows 2 to $\max(N, M)^*$ of $[R_{k-\tau+1}]$, respectively. $[R_{kq}]_\tau$ and $[R_{k\tilde{q}}]_\tau$ are defined analogously, and replace their counterparts. We set $[R_{0q}]_\tau = [R_{0q}]$, $[R_{0\tilde{q}}]_\tau = [R_{0\tilde{q}}]$. The projections g_i are taken to project onto $\mathcal{C}^q$, as spanned by the coordinates $(\varepsilon_{k-1}, \varepsilon_{k-1-\tau}, \ldots, \varepsilon_{k-q+1-\tau})$, when $k \geq \tau - 1$, and $(\varepsilon_{-1}, \ldots, \varepsilon_{-q+1})$, when $k = 0$; where k is the iteration step of the argument being projected. The proofs of Corollaries 7.6 and 7.8, under these modifications, with $q = 2$, then imply a joint uniform distribution of the steady state behaviour of $(\varepsilon_{n+\tau}, \varepsilon_n)$ over $\mathcal{C}^2$, for any $\tau \geq 1$.

2. We apply the arguments and results in the proof of Theorem 7.10. We have that the second-average steady state r_n must satisfy the relationship $r_n = \sum_{j=1}^{q_1} \tilde{r}_j \varepsilon_{n-j}$ in (a), and $r_n = \sum_{j=1}^{q_1} (a_j - b_j) T_{j,n}(\varepsilon_{n-j})$ in (b). These represent difference equations with characteristic polynomials $p_R(z)$ and $p(z)$ respectively. Let $[R_k]$ be as defined at the end of Chapter 4, and corresponding to either difference equation system here. Let $[R_k]^{(1)}$ denote the first row of $[R_k]$. Then, for (a) or (b), we have the relationship that

$$r_{n+\tau} = [R_{\tau+1}]^{(1)} [R_0]^{-1} \cdot (T_{1,n,\tau}(\varepsilon_{n-1}), \ldots, T_{q_1,n,\tau}(\varepsilon_{n-q_1}))^T,$$

where the $T_{j,n,\tau} : \mathcal{C} \to \mathcal{C}$ are some affine mappings with stretch factor 1. If the zeros of $p_R(z)$ or $p(z)$ are not all magnitude one, multiplicity one, with rational complex arguments, then $[R_{\tau+1}]^{(1)}$ and $[R_1]^{(1)}$ will be linearly independent $\forall\ \tau \geq 1$. With $\varepsilon_{n-1}, \ldots, \varepsilon_{n-q_1}$ an independent family of uniform PDFs $\forall\ n \geq 0$, this implies that the pair $(r_{n+\tau}, r_n)$, and hence $(\varepsilon_{n+\tau}, \varepsilon_n)$, are independent and thus uniformly distributed over $\mathcal{C}^2$, for any $\tau \geq 1$, in the second-average steady state.

If an i.i.d. dither is allowed, then the distributions of $(\varepsilon_{n+\tau}, \varepsilon_n)$ over $\mathbb{R}^2$ will, by extension, clearly exist and remain independent in the second average, $\forall\ \tau \geq 1$, in 1 and 2 above. Since the steady state distribution for ε_n exists in the second average, clearly the steady state distribution for $\varepsilon_n \varepsilon_{n+\tau}$ exists in the second average as well, $\forall\ \tau \geq 1$, as a convolution of second-average convergent distributions. Now $<< \varepsilon_n \varepsilon_{n+\tau} >>$ gives the average over all n of the means of the distributions $\varepsilon_n \varepsilon_{n+\tau}$, which is equivalent to the mean of the average over all n of these distributions. From this, we have by definition that $<< \varepsilon_n \varepsilon_{n+\tau} >>$ is the mean of the steady state distribution of $\varepsilon_n \varepsilon_{n+\tau}$, when this steady state exists in the second average. Since the steady states ε_n and $\varepsilon_{n+\tau}$ were shown to be independent in 1 and 2, it follows that $<< \varepsilon_n \varepsilon_{n+\tau} >> \equiv E[\varepsilon_n \varepsilon_{n+\tau}] = 0$, for $\tau \geq 1$. From the methods in

the proof of Proposition 8.1, it then follows that the spectrum of the error ε_n will be white.

Now suppose that the zeros of $p_R(z)$ or $p(z)$ do not satisfy the conditions in part 2 of the theorem. Then there exists a $p \in \mathbb{Z}^+$ such that $[R_{\tau+1}]^{(1)} = [R_1]^{(1)}, \ \forall\ k \geq 0$, when $\tau = kp$. This implies that $r_{n+\tau} = T_a(r_n)$, and hence $\varepsilon_{n+\tau} = T_b(\varepsilon_n)$, for some such affine mappings T_a, T_b, when $\tau = kp$, $k \geq 0$. Thus $(\varepsilon_{n+\tau}, \varepsilon_n)$ are not independent, or thus jointly uniformly distributed over $\mathcal{C}^2$, for these τ. If an i.i.d. dither is allowed, then the values of $(\varepsilon_{n+\tau}, \varepsilon_n)$ over $\mathbb{R}^2$ will, by extension, clearly remain dependent for these τ. Since these pairs are also correlated in second-average steady state, it follows that $<< \varepsilon_n \varepsilon_{n+\tau} >> \neq 0$, when $\tau = kp$, $k \geq 0$. The spectrum of the error ε_n, by definition, is not white under these conditions. ■

These results suggest that a white error spectrum is generally synonymous with having at least two nonminimum-phase zeros. Having a uniform steady state distribution for $(\varepsilon_{n+1}, \varepsilon_n)$ over $\mathcal{C}^2$ does not guarantee whiteness however, as we see for a subclass of marginally minimum-phase systems in part 2, that exhibit a recurrent or cyclic dynamical behaviour. The proof of these results also demonstrates, in particular, that a second-average convergence of steady state error behaviour is associated with any possible white error relationship.

We continue the investigation of this section, beginning with the following proposition which deals with the case of random external input:

Proposition 7.12 *Suppose that a given system with possible dither present has an input x_n with a uniform probability distribution over $\mathcal{C}$ for a given $n \geq 0$, and that x_n is independent of x_k and ν_k, for all $0 \leq k < n$. Then, for any initial condition $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$, the PDF of the system error ε_n will be uniformly distributed over $\mathcal{C}$ at the given n. Moreover, if these conditions hold for all $n \geq 0$, then the PDFs of $g(\vec{x}_n)$ and ε_n will be uniformly distributed*

over $\mathcal{C}^M$ *and* $\mathcal{C}$ *respectively, for all* $n \geq 0$.

Proof:

Suppose that the input x_n has a uniform PDF over $\mathcal{C}$, for a given $n \geq 0$. Supposing that the independence conditions hold as well, it follows that x_n and r_n are independent. Then the PDF of $u_n = x_n - r_n$ will be the PDF of x_n shifted by r_n, and thus, with x_n and r_n independent, will also be uniform over $\mathcal{C}$. Since there is a one-to-one mapping between the value of u_n on $\mathcal{C}$ and the value of ε_n on $\mathcal{C}$ (independent of the dither), we have from the form of this mapping given in the proof of Theorem 8.5, that ε_n is uniformly distributed over $\mathcal{C}$, for the given $n \geq 0$. If the x_n are i.i.d., independent of the corresponding dither, and uniform over $\mathcal{C}$, for $n \geq 0$, it follows that the ε_n are i.i.d. and uniform over $\mathcal{C}$, $\forall\, n \geq 0$. The final result then follows. ■

Here we see, by extension, that a random i.i.d. input x_n with a uniform PDF over $\mathcal{C}$ is sufficient to directly induce a random $g(\vec{x}_n)$ with uniform PDF over $\mathcal{C}^M$ (a PDF fixed point), and hence a random i.i.d. error ε_n with uniform PDF over $\mathcal{C}$, for all $n > 0$. This result is independent of any other conditions, or the need to attain a steady state. Notice that simply having an input x_n that is uniformly distributed over $\mathcal{C}$ in steady state is not, in itself, sufficient to guarantee that u_n and hence ε_n will be uniform over $\mathcal{C}$ in steady state, since x_{n-j}, $j = 1, \ldots, N$, and the feedback r_n will generally not be independent (if the independence requirements on x_n in the proposition hold as well, then this result for ε_n would hold). The question may now arise as to whether, through its quasi-randomizing effects, a constant irrational input can bring about a uniform PDF for ε_n over $\mathcal{C}$ in steady state for any system. The following theorem gives a specific case where this is true:

Theorem 7.13 *Suppose that the first-order system with unity gain, that is $M = 1$, $N = 0$, and $a_1 = 1$, has a constant input $x_n = c$, for all $n \geq 0$, where c is an irrational multiple of Δ. Then, for any initial condition $\varepsilon_{-1} \in \mathbb{R}$, the PDF of the long run steady state behaviour of the system error ε_n will be uniformly distributed over $\mathcal{C}$.*

Proof:

To begin, if a dither ν_n is present, we simply drop the dither from the system. The error dynamics of ε_n in the undithered system will correspond to the error dynamics of the dithered system on $\mathcal{C}$, by Theorem 6.1.

From (1.2), this system may be expressed as $\varepsilon_n = \dfrac{\Delta}{2} - [(c - \varepsilon_{n-1}) \mod \Delta]$, for $n \geq 0$. The orbit of a given initial condition $\varepsilon_{-1} \in \mathbb{R}$ will be quasiperiodic on $\mathcal{C}$. This condition implies a concept of periodicity applied over arbitrarily small intervals. Specifically, this means, for any positive integer p, that if $\mathcal{C}$ is partitioned into p intervals of equal length, then the ratio of the recurrences of ε_n between any two intervals will tend towards 1 as $n \to \infty$. The rate of this convergence will also occur on the same order of magnitude as for the case of a p point limit cycle, such as arises when $c = \Delta/p$. These properties will hold for any such p interval equipartition (with these related via rotation on $\mathcal{C}$), and p may be arbitrarily large, and hence the intervals arbitrarily small. We then have, from this structure, that the long run steady state probability of ε_n lying in an interval of given length is independent of the position of this interval in $\mathcal{C}$. Thus this probability must be the interval length divided by Δ. Allowing the given interval length to tend to zero, these results imply that the steady state PDF of ε_n over $\mathcal{C}$ will be uniform. If the initial condition ε_{-1} is changed, the steady state PDF will simply be shifted on $\mathcal{C}$ by an amount equal to this change. By symmetry, the resulting PDF will still be uniform over $\mathcal{C}$. Thus the result holds for any initial condition $\vec{x}_0 = \varepsilon_{-1}$ on $\mathcal{C}$ or $\mathbb{R}$. ■

For this theorem, the same quasiperiodicity that was used in Theorem 5.16 to show topological transitivity in this case for the study of chaos is used in the proof here.

By considering the first order system of Theorem 7.13 with a particular initial condition $\varepsilon_{-1} \in \mathcal{C}$, and general constant input $x_n \in \mathbb{R}$, we may seek to understand more about the structure of the real numbers on $\mathcal{C}$, and possible relationships between the uniform distribution and the nature of irrational numbers. Let $S_c = \{\varepsilon_n \mid x_n = c, \forall\ n \in \mathbb{Z}\}$ be the set of all points in the orbit of ε_{-1}, iterating in both directions, for a given constant input $c \in \mathbb{R}$. Such sets are invariant sets of the mapping (i.e. $f^n(S_c) = S_c$, $\forall\ n \geq 0$). Let $C_{\mathbb{Z}}$ denote the set of all numbers in $\mathcal{C}$ that are rational multiples of Δ, and $C_{\mathbb{Q}}$ denote the set of all numbers in $\mathcal{C}$ that are irrational multiples of Δ. Clearly we have that $\bigcup_c S_c$ over all c that are rational multiples of Δ, will give $C_{\mathbb{Z}}$, and similarly the union over all c that are irrational multiples of Δ will give $C_{\mathbb{Q}}$. Each set S_c with rational-type c is finite, and each with irrational-type c is infinite and, by quasiperiodicity, uniform on $\mathcal{C}$ and with cardinality on the same order of magnitude as that of $C_{\mathbb{Z}}$ (since common elements α, β, satisfy $m\alpha = \beta + n\Delta$, for some $m, n \in \mathbb{Z}$, such that α, $\beta \in \mathcal{C}$). Any two nonidentical sets S_c are necessarily disjoint. Thus we conclude that $C_{\mathbb{Z}}$ consists of the union of an infinite number of disjoint sets of finite size (limit cycles), and $C_{\mathbb{Q}}$ consists of the union of an infinite number of disjoint quasiperiodic sets of infinite size, and cardinality of each related to that of $C_{\mathbb{Z}}$. It follows that $C_{\mathbb{Q}}$ is uniform on $\mathcal{C}$. $C_{\mathbb{Z}}$ is as well, by quasiperiodic arguments. These lines of argument may perhaps be developed further to give more insight into these properties.

The following theorem gives a more general extension of Theorem 7.13, to higher-order marginally minimum-phase systems that satisfy condition (R).

Theorem 7.14 *Suppose the a_i and b_j satisfy condition (R). Suppose also that the largest magnitude zero(s) have magnitude* 1 *and multiplicity* 1*, and that all such zeros μ_i are all real or complex with argument $\theta_i = 2\pi\tilde{\mu}_i$ for some rational number $\tilde{\mu}_i$. Suppose further that the input x_n is periodic with period p. Suppose, in addition, that the part of the particular solutions of $\varepsilon_{\tilde{p}+k} - \varepsilon_k$ associated with the magnitude* 1 *zeros are irrational multiples of Δ, for $k = 0, \ldots, \tilde{p} - 1$, where $\tilde{p} = LCM(p, \tilde{\mu}_{d,1}, \ldots, \tilde{\mu}_{d,q}, 2)$, and the $\tilde{\mu}_{d,i}$ are the denominators of $\tilde{\mu}_i$ for the respective q complex μ_i. Then, for any initial condition $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$, the PDF of the long run steady state behaviour of the system error ε_n will be uniformly distributed over $\mathcal{C}$.*

Proof:

To begin, if a dither ν_n is present, we simply drop the dither from the system. The error and state space dynamics of the undithered system will correspond to the respective dynamics of the dithered system on $\mathcal{C}$ and $\mathcal{C}^M$, by Theorem 6.1.

The system with $g(\vec{x}_0) \in \mathbb{R}^M$ may be adjusted to an equivalent system with $g(\vec{x}_0)$ shifted to $\hat{P}_{\mathcal{C}}(g(\vec{x}_0))$ in $\vec{x}_0$, and the values of x_i, $i = 0, \ldots, M-1$ shifted by an associated amount. This has no effect on the long run behaviour of $\vec{x_n}$, and thus, without loss of generality, we simply take $x_0 \in \mathbb{R}^N \times \mathcal{C}^M$ for the proof that follows.

With condition (R) holding, the behaviour of ε_n is described by (4.2) in Proposition 4.8. We proceed by considering the overall solution of the difference equation of (4.2) given from (2.3) and (2.4).

From the form of (2.4), the periodicity of the input x_n (and hence $\sum_{j=0}^{N} b_j x_{n-j}$), and the nature of the zeros μ_i, it follows that the terms in the summation portion of (2.4) associated with the magnitude 1 zeros, will repeat in a cycle of period $\tilde{p}$ across the sequence of all such terms going from $k = N + l$ to $k = N + m\tilde{p} + l$, $\forall\ m \in \mathbb{Z}^+$, and any given l with

$0 \leq l \leq \tilde{p} - 1$ (cycle changes with l). For the condition on the corresponding part of the particular solution of $\varepsilon_{\tilde{p}+k} - \varepsilon_k$ given in the theorem to hold, it follows, from (4.2), that the sum of the terms in a cycle will be some irrational multiple of Δ, that we denote e_l. It then follows that the sum of terms from $k = N + l$ to $k = N + m\tilde{p} + l$ will be me_l added to the sum from $k = N + l$ to $k = N + \tilde{p} + l$. Applying the arguments of the proof of Theorem 7.13 with the quantizer relationship of (4.2), it follows that the long run behaviour of the associated part of the particular solution of $\varepsilon_{m\tilde{p}+l}$ will be quasiperiodic and uniform over $\mathcal{C}$.

From the form of (2.4) and periodicity of x_n, it follows that the portion of (2.4) associated with the remaining zeros will converge to a constant value as $n \to \infty$. From the form of (2.3) and the nature of all zeros, it follows that (2.3) will converge to a limit cycle over the period $\tilde{p}$ as $n \to \infty$, and thus a constant value as $m \to \infty$, for $n = m\tilde{p} + l$, with a given l. Combining these results with the quasiperiodic ones above implies that the long run behaviour of the overall solution for $\varepsilon_{m\tilde{p}+l}$ will be uniform over $\mathcal{C}$. Since this holds over all l, the result thus extends to ε_n. ■

For this result, we are using the circle map relationship as applied in Proposition 4.8. To create quasiperiodicity and thus a uniform steady state, we require both unboundedness in the particular solution for $\tilde{w}_n$ in (4.2), and an irrationality property in this solution. Achieving unboundedness of $\tilde{w}_n$ in a marginally minimum-phase system is basically the opposite problem to what was considered for stability in Chapter 3, with the analogous difference equation for r_n. Considering this, it can be shown (with $p_r(z)$ replaced by $p(z)$, and $\tilde{c}_n = \sum_{i=0}^{N} b_j x_{n-j}$) that under the general conditions of Theorem 3.3, with any of the zero conditions violated; or Proposition 3.4 with (a), (b) or (c) violated; the solution $\tilde{w}_n$ will be unbounded. (These general conditions also require the generic assumption that the

nonzero constants in the representation of the particular solution for ε_n are not exclusive to the terms associated with the minimum-phase zeros or non-condition violating marginally minimum-phase zeros). Similarly, Proposition 3.6 provides a way for unboundedness to fail. Unboundedness allows the irrationality component to yield quasiperiodicity. To express the irrationality conditions explicitly in terms of the input, as in Theorem 7.13, would require stronger and more complicated conditions on the magnitude 1 zeros of $p(z)$ in general.

Extensions of Theorems 7.13 or 7.14 to arbitrary marginally minimum-phase systems that satisfy condition (R) (but not Theorem 7.5), for generic initial conditions would seem plausible. In particular, it might be possible to have a more general input structure, such as quasiperiodic, or one that has an asymptotic steady state that is periodic or quasiperiodic, or extend to stochastic input. Extensions to allow any type of nonrepeated magnitude 1 zeros might also be possible. When the irrationality conditions of Theorem 7.14 fail, we typically get cyclic periodic points. This serves as an extension to this phenomenon in the proof of Proposition 5.16.

A particular extension of Theorem 7.13 to the case of M repeated zeros at $\mu = 1$, with $N = 0$, no dither present, and irrational constant input or sinusoidal input (if $M > 2$), was given by He et al. [22]. The general result is that the steady state $g(\vec{x}_n)$ is uniform over $\mathcal{C}^M$. Theorems by Weyl (see [22]), beginning with the assertion that the fractional part of a polynomial $P(n)$, $n \in \mathbb{Z}^+$, with real coefficients, will have essentially a steady state distribution (as a process in n) that is uniform on $[0, 1)$, if at least one coefficient is irrational, are used to prove their results. Their work also asserts that the steady state errors ε_n will be white and input independent, for $M > 1$. The claim that a uniform steady state for $g(\vec{x}_n)$ over $\mathcal{C}^M$ is roughly equivalent to whiteness of the errors when $M \geq 2$ [22] is applied for this. Conditions under which this claim holds are not given, considering that Theorem 7.11 shows limitations to a general interpretation. Their proofs make the

assumption that the initial condition $\vec{x}_0 = 0$, although it would appear that arbitrary initial conditions would suffice, as was shown explicitly in a preceding consideration of the $M = 2$ case with constant irrational input in [21]. Although the systems considered are nominally nonminimum phase, the results of [22] extend somewhat further from what would be obtained from applying Theorem 7.7 or its corollary.

The implied consequence of [21], and we assume [22] (under arbitrary initial conditions), that no periodic points (which would yield discrete histogram orbits) exist, means that their results serve as a counterexample to negate a possible extension of Corollary 5.22 to cases with fewer than M expansive zeros. The pure results also cannot be extended to allow any magnitude one zeros that are not identically one, since our proof of Proposition 5.17 shows that some of these cases yield periodic points.

The relationship between topological transitivity and a steady state error distribution can be clarified and summarized with the following:

Proposition 7.15 *Suppose that for any initial condition $\vec{x}_0 \in \mathbb{R}^N \times \mathbb{R}^M$, the long run steady state $\hat{P}_{\mathcal{C}}(g(\vec{x}_n))$ of the system (6.1) has a piecewise continuous PDF that is nonzero over $\mathcal{C}^M$. Then the system satisfies chaos condition 2 (topological transitivity).*

Proof:

Suppose that topological transitivity is not satisfied. This implies that there exists $\vec{y}_0 \in \mathbb{R}^N \times \mathbb{R}^M$, $\hat{z} \in \mathcal{C}^M$, and a $\delta > 0$, such that $\|\hat{P}_{\mathcal{C}}(g \circ f^n(\vec{y}_0)) - \hat{z}\| > \delta, \ \forall\ n > 0$. Thus $\text{Prob}(\|\hat{P}_{\mathcal{C}}(g \circ f^{n_1}(\vec{y}_0)) - \hat{z})\| < \delta) = 0$, for arbitrary $n_1 > 0$. This implies that the steady state PDF for $\hat{P}_{\mathcal{C}}(g(\vec{x}_n))$, when $\vec{x}_0 = \vec{y}_0$, must be zero over a ball of radius δ about $\hat{z}$ on $\mathcal{C}^M$. This contradicts the conditions of the proposition. Thus chaos condition 2 is satisfied. ■

This result is not too surprising and is supported by the topological transitivity for the systems in Theorems 7.5, 7.7, and 7.10 - 7.13. Applying this proposition to the results of [22] given, we may conclude that a system with M repeated zeros at $\mu = 1$, $N = 0$, and an irrational constant input will be topologically transitive. The conditions for Proposition 7.15 are stronger than simply that of nonzero piecewise continuity of the steady state ε_n PDF, which alone would be insufficient to guarantee topological transitivity if $M > 1$.

The converse of this proposition, at least theoretically, does not hold. Orbits may be dense on $\mathcal{C}^M$, but exist on certain subregions with probability zero. In particular, one may conceive of constructing random inputs x_n that bring about topological transitivity via Theorem 5.24, but a steady state that is zero over subregions of $\mathcal{C}^M$ (as well as being nonuniform). Thus it appears that the nonminimum-phase requirements of Theorem 5.9 (and perhaps 5.20) provide a "stronger" form of transitivity — one that is more central to giving a uniform steady state, than the most general random input conditions of Theorem 5.24. The former are sufficient for Theorem 7.5 and necessary for Theorem 7.7.

The following proposition allows us to interpret the error behaviour of systems satisfying condition (R) and any of Theorems 7.5, 7.7, 7.10, 7.13, their corollaries, or Proposition 7.9, when an arbitrary dither is added:

Proposition 7.16 *Suppose the a_i and b_j satisfy condition (R) and that no dither is present in the system. Suppose also that the system error ε_n (or $g(\vec{x}_n)$) has a uniform PDF over $\mathcal{C}$ (or $\mathcal{C}^M$) (i) for all $n \geq 0$, or (ii) in its long run steady state behaviour. Then, if an arbitrary dither ν_n is added to the system, the PDF of the error ε_n (or $g(\vec{x}_n)$) will be uniformly distributed over $\mathcal{C}$ (or $\mathcal{C}^M$) for cases (i) or (ii) respectively.*

Proof:

For the proof, we shall consider ε_n to be such that either n satisfies *(i)*, or n is sufficiently large that *(ii)* can be taken to apply. Applying Theorem 6.1, the value of ε_n on $\mathcal{C}$ is unchanged when an arbitrary dither ν_n is added. Thus the result follows. ■

The interpretation is thus straightforward. As we saw with Theorem 6.1, the error dynamics on $\mathcal{C}$ are not changed with the addition of dither, and thus, for any system satisfying (R), Theorems 7.5, 7.7, 7.10, 7.13, their corollaries, or Proposition 7.9 simply extend to the case when an arbitrary dither ν_n is added. This holds automatically for Proposition 7.12.

In summary, the theoretical results of this section, apart from their more practical application to follow in Chapter 8, serve to demonstrate the fundamental role of the uniform steady state distribution as a natural description of averaged state space error behaviour for a broad range of Σ-Δ modulator systems characterized by some structure. Furthermore, in establishing these results, we find that persisting questions about stochastically interpreted long run behaviour (in the context of these results), in significant measure, begin to boil down to issues of the properties of dynamical behaviour, chaotic properties, and generally what can be said about the system as a class of dynamical system. This appeals to the dynamical perspective of Chapter 5, in addition to drawing on some of the analysis for conditions of chaos therein.

7.3 Discussion

Time Series Analysis:

When a series of random variables X_n, $n \geq 1$, is used to characterize quantities that arise in a random process, such as a time series of Σ-Δ modulator errors ε_n, the question arises as to whether the general steady state behaviour can be inferred from the behaviour of one particular time series realization (a generalized ergodic property), or whether an ensemble of such realizations is necessary. For consideration of the steady state PDF, it was asserted (without proof) that second-average convergence implies convergence of a sample histogram to the same PDF. This assumes sampling over an ensemble of realized time series, independently at each iteration. The average of a finite number of converged histograms corresponding to a finite number of realizations will approximate this. For $\vec{x}_n$ or ε_n, clearly an ensemble is generally necessary when the initial conditions are random. If the entire Σ-Δ system is deterministic however, there is trivially only one realization to consider.

For cases with fixed initial conditions, and a random input or dither, either situation emerges. Specifically, we assert, without proof, that when condition (R) holds, the input x_n is nonrandom, and an i.i.d. dither ν_n is added, the histogram, for any single realization of $\vec{x}_n$ or ε_n, will converge as $n \to \infty$ to the respective steady state PDF, if it exists in the second average. This assertion follows from recognizing that with (R) the dither has no effect on the state space dynamics, and it acts independently and identically each iteration. Similarly, this assertion could be made about any such dithered system, or any nondithered system at all, when the input x_n is i.i.d. and uniform over $\mathcal{C}$. It is unclear under what greater level of generality single histogram convergence to the steady state PDF will hold with probability 1, for $\vec{x}_n$ or ε_n — one can conceive of the existence of examples where it would not. In theory, only the ensemble interpretation and approach is universally valid.

In light of the practical role of time series for estimating error moments, or representing a specific realization of a process, one could develop the steady state theory of Section 7.1 in an analogous manner for this type of stochastic process. In this case, the sequence of random variables X_i, $i \geq 1$, would not be treated as independent (i.e. from independent ensemble samples). The distribution of X_n would be dependent, in general, on the values of X_i, $i = 1, \ldots n-1$, for all $n \geq 0$. The randomness of a given error ε_n, for example, would arise solely from the randomness of the dither ν_n and input x_n at the same value of n, but its distribution would depend, in general, on all previous outcomes ν_i, x_i, $i = 0, \ldots, n-1$. The definitions for convergence, average distributions and steady state could similarly be applied to the sequence of X_i. A second-average steady state PDF then represents, by definition, the PDF that the associated histogram converges to.

Stationarity/Ergodicity:

We now discuss some of the connections of the results of our theorems to the issues of stationarity and ergodicity, drawing from [1]. A random process x_n will be weakly stationary if $E[x_n]$ and $E[x_n x_{n+\tau}]$ are independent of n, for all $\tau \geq 1$. Stricter forms of stationarity exist as well. The process x_n will also be ergodic if it is stationary, and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_n = K, \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_n x_{n+\tau} = K_\tau,$$

for some real constants K, K_τ, $\tau \geq 1$, with probability 1. The extra ergodic property then requires a form of equivalent long run behaviour over any time series realization. The definitions of stationarity and ergodicity are important tools for analyzing the moments of stochastic processes and forming estimators of these with simulations. They also provide a basic theoretical characterization.

Clearly any process modelled by a sequence of random variables that converge to a steady state distribution in the second average generally satisfies the second condition

for ergodicity, if any realization of the process converges to the same steady state with probability 1. The latter may be thought of as a generalized ergodic property. With this holding, it is generally ergodic if it is stationary. From the preceding discussion on time series analysis, the process $g(\vec{x}_n)$ (and ε_n) in a Σ-Δ system will satisfy the second ergodic condition if condition (R) holds, the initial condition and input are nonrandom, and an i.i.d. dither is added. It will generally not be fully ergodic if undithered (i.e. it will satisfy the second ergodic condition, but generally not be stationary). Conversely, the process $g(\vec{x}_n)$ (and ε_n) on $\mathcal{C}^M$ (or $\mathcal{C}$) will be stationary if the distribution of $g(\vec{x}_0)$ is a fixed point PDF of the system, and the input x_n is constant over all n or i.i.d. Ergodicity may be achieved (on $\mathcal{C}^M/\mathcal{C}$) when both cases hold. This is possible, notably, for the first-order system of Theorem 7.13, where the input c is rational or irrational, that will be analyzed in more detail in Section 8.3. We have that $g(\vec{x}_n)$ and ε_n on $\mathcal{C}^M/\mathcal{C}$ will be ergodic for any fixed initial condition $\vec{x}_0$, if the input x_n is uniformly distributed over $\mathcal{C}$ and i.i.d. for all $n \geq 0$, from Proposition 7.12.

From the error moment results of dither theory, $E[\varepsilon_n|\varepsilon_{n-i}, i = 1, \ldots n+M] = 0$ for all $n \geq 0$ if the dither is a convolution of RPDFs. Thus such dither brings about stationarity, with the mean and autocorrelation moments equating to zero for all $n \geq 0$ and $\tau \geq 1$. The process ε_n will then generally be ergodic if any realization converges with probability 1 to a steady state distribution that exists in the second average as well. Applying Proposition 8.6 and the arguments from the time series analysis discussion with this, we then have the following: with such dither, the process ε_n will generally be ergodic under the conditions of Theorems 8.5, 7.5, 7.7, 7.10 plus (R), 7.13, their corollaries, Propositions 7.9 plus (R), 7.12 for all $n \geq 0$, or 7.16. This result extends to the case of rational input x_n for the first-order system of Theorem 7.13 as will be analyzed in Section 8.3.

PRN Generator Implications:

Our steady state distribution investigation provides a connection with the function of PRN generators. If a Σ-Δ modulator system with fixed initial conditions and nonrandom input has a steady state distribution for $\vec{x}_0$ that is uniform over $\mathcal{C}^M$, such as is provided for when Theorem 7.7, Theorem 7.13 or its Mth order extension from [22] apply, then it follows that for an arbitrary n, elements of the set $\{\varepsilon_{n-1}, \ldots, \varepsilon_{n-M}\}$ will be i.i.d. and jointly uniformly distributed over $\mathcal{C}$. Thus a relatively general class of Σ-Δ modulators with an Mth order feedback filter will function as PRN generators for random sequences of length M or less. If we wish the sequence length to be more arbitrary, or require a lower order feedback filter (such as first order), then we must provide more specification in the design the Σ-Δ system, as discussed in Section 1.7 of the Introduction.

Chaos with PDF State Space:

Returning briefly to the topic of chaos, the generalized probabilistic treatment of the state space coordinate $\vec{x}_n$ that emerges in this chapter raises the question of whether it is possible or sensible to define and investigate chaos for Σ-Δ modulator systems with stochastic or random initial conditions $\vec{x}_0$. For such stochastic systems, it is obvious that we can no longer separate the stochastic elements (i.e. input x_n and/or dither ν_n) from the otherwise "fixed" initial condition $\vec{x}_0$, so as to maintain an easy carry over of the chaos condition definitions. Perhaps no such meaningful carry over exists. Drawing on the developments in this chapter, we have a natural generalization of the chaos conditions however.

Specifically, we apply a "functional" generalization of Devaney's chaos conditions and related definitions (Definitions 2.1, and 5.1 to 5.5). In this context, the state space changes from $\mathbb{R}^N \times \mathcal{C}^M$ to the set of all piecewise continuous/discrete probability density/mass functions defined over the domain $\mathbb{R}^N \times \mathcal{C}^M$ describing the random variable $\vec{x}_0$. The f_n are

taken to be functionals that map the PDF state space to itself according to the definition given in Section 7.1. The projections g/g_i map a PDF of $\vec{x}_k$ to the marginal PDF of the random variable $g(\vec{x}_0)/g_i(\vec{x}_0)$ defined over the domain $\mathcal{C}^M/\mathcal{C}$.

We propose the following metric over the space of PDFs that are defined over the domain $D = \mathbb{R}^N \times \mathcal{C}^M$. For two piecewise continuous PDFs $h_1(\vec{x})$ and $h_2(\vec{x})$, the metric is defined as

$$\tilde{m}_1 = \max \frac{\Delta}{2}\{\| \frac{1}{K_2} h_2(\vec{z}) - \frac{1}{K_1} h_1(\vec{z}) \| \mid \vec{z} \in D\},$$

where $K_i = \max\{h_i(\vec{z}) \mid \vec{z} \in D\}$, $i = 1, 2$. $\| \ \|$ is the metric on D defined in Chapter 2. For two discrete PDFs $h_1(\vec{x})$ and $h_2(\vec{x})$, the metric is defined as

$$\tilde{m}_2 = (1 - \lambda)\tilde{m}_1 + \lambda \min\{\|\vec{z}_2 - \vec{z}_1\|, \forall\ \vec{z}_1, \vec{z}_2 \mid h_1(\vec{z}_1), h_2(\vec{z}_2) \neq 0\},$$

where $\lambda = (K_1 + K_2)/2$, $0 < \lambda \leq 1$, (K_i defined similarly). This particular metric is chosen because it has the properties that the discrete form reduces to the metric defined in Chapter 2 for fixed (single point mass) state space ($\lambda = 1$), and it reduces to the metric for piecewise continuous PDF state space above in the continuous limit of discrete PDFs ($\lambda \to 0$). This metric definition might be refined further, so as to exhibit the continuous limit property on any sequence of PDF pairs, for example, or so as to consider mixed pairs of discrete and continuous PDFs.

Under this formulation, a PDF fixed point, such as those of Theorem 7.10 and Proposition 7.9, is simply a fixed periodic point of the mapping. Under the M "nonminimum-phase" zeros and condition (R) stipulations of Theorem 7.5, we have, from the proof of the theorem, that any initial condition in the piecewise continuous PDF state space will converge to the uniform distribution over $\mathbb{R}^N \times \mathcal{C}^M$. Thus the uniform distribution is the only possible periodic point. If the conditions of Theorem 7.10 also hold, or the contraction mapping theorem, or some other means could be applied to show that this uniform

distribution is a fixed point (or at least a periodic point), then we would have that chaos condition 3 (over the given space of random inputs) would fail. It appears probable that chaos conditions 1 and 2 would generally fail with the conditions above of Theorem 7.10 alone. From these results, we conclude that overall chaos appears to be less prevalent for nonminimum-phase systems with stochastic piecewise continuous initial conditions than for those with fixed initial conditions.

We expect this also for such systems with stochastic discrete PDFs. For such a state space, the possibly chaotic fixed (i.e. single point mass) state space would be a subset of the overall discrete PDF state space. More chaos might be expected on other subsets of the state space as well, i.e. where the positions of the point masses of the PDFs are rationally related to the expansivity factors of the NTF zeros.

If a random input x_n or dither ν_n is added to the system, this would be treated in terms of the randomizing effects on orbits relative to the given initial condition PDF state space in the study of chaos, as was done in this thesis for a fixed state space. Of course the effect of adding these random signals is to convolve their distributions with those of the internal state variables, and this relationship is applied when studying general statistical error behaviour as was explored in this chapter.

A deeper investigation of chaos in Σ-Δ modulators with random initial conditions following this setup and paralleling the chaos analysis of this thesis could be looked at in future work.

Chapter 8

Dithered Error Statistics

In this chapter, we study the statistical properties of the error behaviour of the Σ-Δ modulator when a dither signal is added to the input signal prior to its entry into the quantizer, as shown in Figures 1.5, 1.6 and 1.7. The study of this behaviour, and the underlying error dynamics, from a statistical point of view is consistent with the approach of standard work in dither theory, and will allow an easy application of the results of Chapter 7 to the issues addressed here. In analyzing the statistics of the error ε_n, we will be concerned with the actual value of this error as represented in the Σ-Δ modulator topology and dealt with in practice. The interpretation of ε_n lying on $\mathcal{C}$ in state space will be adopted only when we need to appeal to the approaches and symmetries used in studying the dynamics earlier to help develop our study here. As an important result, we will show how an average error variance level of $\Delta^2/6$ may be achieved for the Σ-Δ modulator with RPDF dither when certain internal dynamical behaviour exists.

The dither theory mentioned in the Introduction, as with much of the previous work on Σ-Δ modulators forming the relevant background to this thesis, was derived using standard frequency domain methods. These methods are standard to the electrical engineering

approach that naturally arises through the study of Σ-Δ modulators from the perspective of digital audio or other traditional applications. The purely dynamical systems approach developed initially for the study of chaos in this thesis provides an alternative point of view for approaching many of the analytical questions and issues concerning Σ-Δ modulator behaviour of both practical and theoretical interest. This general approach will be carried through to the analysis undertaken in this chapter, with the aim of reinforcing and expanding upon the current understanding of dithered Σ-Δ modulator behaviour.

8.1 Dithered Quantizers

To begin, we show, using a simple method grounded in our dynamical systems approach, how the first and second error moment results $E[\varepsilon_n]$ and $E[\varepsilon_n^2]$ may be derived for the case of RPDF and TPDF dither density functions. For these results, we confine the predithered quantizer input u_n to lie on the circle $\mathcal{C}$. This gives no loss of generality, since the quantizer output and hence the moments for a given u_n will be unchanged if any integer multiple of Δ is added to u_n.

Error Moment Results:

RPDF Dither:

Suppose $0 \leq u_n \leq \dfrac{\Delta}{2}$. Then

$$\begin{aligned} E[\varepsilon_n] &= \text{Prob}(-u_n < \nu_n < \tfrac{\Delta}{2})(\tfrac{\Delta}{2} - u_n) + \text{Prob}(-\tfrac{\Delta}{2} < \nu_n < -u_n)(-\tfrac{\Delta}{2} - u_n) \\ &= (\tfrac{1}{2} + \tfrac{u_n}{\Delta})(\tfrac{\Delta}{2} - u_n) + (\tfrac{1}{2} - \tfrac{u_n}{\Delta})(-\tfrac{\Delta}{2} - u_n) \\ &= 0, \end{aligned}$$

where we make use the fact that ν_n has an RPDF distribution, and the effect ν_n may have on displacing the error value by a factor of Δ.

Now suppose $-\dfrac{\Delta}{2} < u_n < 0$. Then we get the same expression for $E[\varepsilon_n]$ as above, with the Δ displaced ε_n, and the non-displaced ε_n terms switched. Thus $E[\varepsilon_n] = 0$ as well.

Extending these results to the second moment, we have

$$\begin{aligned} E[\varepsilon_n^2] &= \text{Prob}(-u_n < \nu_n < \tfrac{\Delta}{2})(\tfrac{\Delta}{2} - u_n)^2 + \text{Prob}(-\tfrac{\Delta}{2} < \nu_n < -u_n)(-\tfrac{\Delta}{2} - u_n)^2 \\ &= (\tfrac{1}{2} + \tfrac{u_n}{\Delta})(\tfrac{\Delta}{2} - u_n)^2 + (\tfrac{1}{2} - \tfrac{u_n}{\Delta})(-\tfrac{\Delta}{2} - u_n)^2 \\ &= \tfrac{1}{\Delta}(\tfrac{\Delta}{2} - u_n)(\tfrac{\Delta}{2} + u_n)(\tfrac{\Delta}{2} + \tfrac{\Delta}{2} + u_n - u_n) \\ &= \tfrac{\Delta^2}{4} - u_n^2. \end{aligned}$$

TPDF Dither:

Suppose $0 \le u_n \le \dfrac{\Delta}{2}$. Then

$$\begin{aligned} E[\varepsilon_n] &= \text{Prob}(-u_n < \nu_n < -u_n + \Delta)(\tfrac{\Delta}{2} - u_n) + \text{Prob}(-\Delta < \nu_n < -u_n)(-\tfrac{\Delta}{2} - u_n) \\ &\quad + \text{Prob}(-u_n + \Delta < \nu_n < \Delta)(\tfrac{3\Delta}{2} - u_n) \\ &= (\tfrac{2\Delta^2 - u_n^2 - (\Delta - u_n)^2}{2\Delta^2})(\tfrac{\Delta}{2} - u_n) + (\tfrac{(\Delta - u_n)^2}{2\Delta^2})(-\tfrac{\Delta}{2} - u_n) + \tfrac{u_n^2}{2\Delta^2}(\tfrac{3\Delta}{2} - u_n). \end{aligned}$$

This follows the same approach as for the RPDF case, where the probabilities are calculated as areas under the triangular TPDF over the given intervals. Collecting terms, we have

$$E[\varepsilon_n] = \tfrac{u_n^3}{\Delta^2}(1 - \tfrac{1}{2} - \tfrac{1}{2}) + \tfrac{u_n^2}{\Delta}(-\tfrac{1}{2} - 1 + 1 - \tfrac{1}{4} + \tfrac{3}{4}) + u_n(-\tfrac{1}{2} + \tfrac{1}{2} - \tfrac{1}{2} + \tfrac{1}{2}) + (\tfrac{1}{4} - \tfrac{1}{4}) = 0.$$

Suppose $-\dfrac{\Delta}{2} < u_n < 0$. Then

$$\begin{aligned} E[\varepsilon_n] &= \text{Prob}(-u_n - \Delta < \nu_n < -u_n)(-\tfrac{\Delta}{2} - u_n) + \text{Prob}(-u_n < \nu_n < \Delta)(\tfrac{\Delta}{2} - u_n) \\ &\quad + \text{Prob}(-\Delta < \nu_n < -u_n - \Delta)(-\tfrac{3\Delta}{2} - u_n) \\ &= (\tfrac{2\Delta^2 - u_n^2 - (\Delta + u_n)^2}{2\Delta^2})(\tfrac{\Delta}{2} - u_n) + (\tfrac{(\Delta + u_n)^2}{2\Delta^2})(\tfrac{\Delta}{2} - u_n) + \tfrac{u_n^2}{2\Delta^2}(-\tfrac{3\Delta}{2} - u_n). \end{aligned}$$

It can be seen that the case of negative u_n simply changes the signs of the values of the three terms in this expression from the situation that exists with positive u_n. Thus we have $E[\varepsilon_n] = 0$ again.

Extending these results to the second moment for $0 \leq u_n \leq \frac{\Delta}{2}$, we have

$$\begin{aligned} E[\varepsilon_n^2] &= \text{Prob}(-u_n < \nu_n < -u_n + \Delta)(\tfrac{\Delta}{2} - u_n)^2 + \text{Prob}(-\Delta < \nu_n < -u_n)(-\tfrac{\Delta}{2} - u_n)^2 \\ &\quad + \text{Prob}(-u_n + \Delta < \nu_n < \Delta)(\tfrac{3\Delta}{2} - u_n)^2 \\ &= (\tfrac{2\Delta^2 - u_n^2 - (\Delta - u_n)^2}{2\Delta^2})(\tfrac{\Delta}{2} - u_n)^2 + (\tfrac{(\Delta - u_n)^2}{2\Delta^2})(-\tfrac{\Delta}{2} - u_n)^2 + \tfrac{u_n^2}{2\Delta^2}(\tfrac{3\Delta}{2} - u_n)^2 \\ &= \tfrac{u_n^4}{\Delta^2}(-1 + \tfrac{1}{2} + \tfrac{1}{2}) + \tfrac{u_n^3}{\Delta}(1 + 1 - 1 + \tfrac{1}{2} - \tfrac{3}{2}) + u_n^2(\tfrac{1}{2} - 1 - \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{2} - 1 + \tfrac{1}{8} + \tfrac{9}{8}) \\ &\quad + u_n\Delta(\tfrac{1}{4} - \tfrac{1}{2} - \tfrac{1}{4} + \tfrac{1}{2}) + \Delta^2(\tfrac{1}{8} + \tfrac{1}{8}) \\ &= \tfrac{\Delta^2}{4}. \end{aligned}$$

If $-\frac{\Delta}{2} < u_n < 0$, then the second moment is

$$E[\varepsilon_n^2] = (\tfrac{2\Delta^2 - u_n^2 - (\Delta + u_n)^2}{2\Delta^2})(\tfrac{\Delta}{2} - u_n)^2 + (\tfrac{(\Delta + u_n)^2}{2\Delta^2})(\tfrac{\Delta}{2} - u_n)^2 + \tfrac{u_n^2}{2\Delta^2}(-\tfrac{3\Delta}{2} - u_n)^2.$$

It can be seen that with negative u_n, the values of the three terms in this expression are the same as in the situation that exists with positive u_n. Thus we have $E[\varepsilon_n^2] = \frac{\Delta^2}{4}$ again.

■

These results are thus consistent with the known results from dither theory. The general formulas involved may be applied in this brute force manner to higher-order RPDF convolution densities and to find higher-order error moments. The determination of the analytic functional form of such higher-order convolutions for application in the formula, and the final analytic determination of moments from the formula (particularly higher-order moments) clearly involve lengthy, tedious algebra. These algebraic steps could be performed symbolically by a computer language such as Maple. Further verification of the dither theory results for moments is then possible. From a comparison with previous work, it is clear that frequency domain methods provide a more analytically effective means for the establishment of the error moment results.

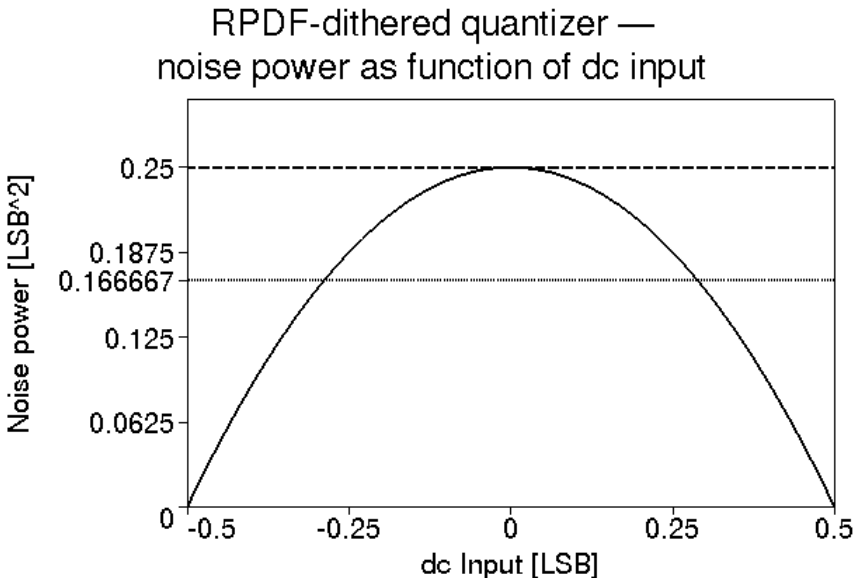


Figure 8.1: $E[\varepsilon_n^2]$ as a function of u_n with RPDF dither

The value of the noise power $E[\varepsilon_n^2]$ under RPDF dither, as a function of u_n over $\mathcal{C}$, is shown in Figure 8.1, with Δ scaled to 1. This corresponds to the more specific case, indicated in the figure, one has with a quantizer having no feedback ($r_n = 0, \forall\ n \geq 0$), and a constant or DC input $u_n = x_n = c \in \mathbb{R}, \forall\ n \geq 0$. The noise power then applies as an average over any number of iterations, and hence as a long run steady state. The graph in Figure 8.1 is just the parabola given from the second error moment result under RPDF dither above.

Now we examine the power spectral density results. Standard dither theory gives that if the dither density function is RPDF or TPDF, then, for a given input sequence u_n, the power spectrum for ε_n is constant and hence white. This condition is equivalent to that of the errors ε_n being statistically uncorrelated in time (i.e. for different n), as will be shown. For RPDF dither, this constant is input dependent ("noise modulation"), whereas for TPDF dither, it is input independent (controlled) [35], [65]. A more general consequence

of the general theory may be stated with the following proposition:

Proposition 8.1 *Suppose that the dither ν_n is chosen with a PDF such that the first error moment (i.e. the error mean) is zero, independent of the predithered quantizer input u_n. Then the power spectrum of the error ε_n will be white.*

Proof:

The power spectral density $P(f)$ is defined as the discrete-time Fourier transform F of the time averaged autocorrelation function $<< \varepsilon_n \varepsilon_{n+\tau} >>$. This gives

$$P(f) = F[\{<< \varepsilon_n \varepsilon_{n+\tau} >>\}](f) = \sum_{\tau=-\infty}^{+\infty} << \varepsilon_n \varepsilon_{n+\tau} >> e^{-2\pi i f \frac{\tau}{f_s}},$$

where f_s is the sampling frequency, $f_s/2$ is the Nyquist frequency, and F acts on the sequence $\{<< \varepsilon_n \varepsilon_{n+\tau} >>, \tau \in \mathbb{Z}\}$. If the errors are statistically uncorrelated in time, then the time averaged autocorrelation function $<< \varepsilon_n \varepsilon_{n+\tau} >> = \lim_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T} < \varepsilon_n \varepsilon_{n+\tau} >$ will be equal to a constant when $\tau = 0$, and equal to zero otherwise. This is equivalent to a Fourier transform and hence power spectrum which is constant or white.

We assume that the error behaviour ε_n may be described either by a discrete or a piecewise continuous probability mass/density function $h(\varepsilon_n)$ over $\mathbb{R}$. The former case implies that the error takes on only a countable number of values. The autocorrelation function may be expressed as follows:

1. If the error has a discrete distribution, then

$$< \varepsilon_n \varepsilon_{n+\tau} > \equiv E[\varepsilon_n \varepsilon_{n+\tau}] = \sum_{\varepsilon_n = e_1}^{e_L} E[\varepsilon_n \varepsilon_{n+\tau} | \varepsilon_n] h(\varepsilon_n) = \sum_{\varepsilon_n = e_1}^{e_L} \varepsilon_n E[\varepsilon_{n+\tau} | \varepsilon_n] h(\varepsilon_n),$$

where $e_1, \ldots, e_L$ are the L discrete values that the error ε_n may take on in $\mathbb{R}$.

2. If the error has a piecewise continuous distribution, then

$$< \varepsilon_n \varepsilon_{n+\tau} > \equiv E[\varepsilon_n \varepsilon_{n+\tau}] = \int_{-\infty}^{+\infty} E[\varepsilon_n \varepsilon_{n+\tau} | \varepsilon_n] h(\varepsilon_n) d\varepsilon_n = \int_{-\infty}^{+\infty} \varepsilon_n E[\varepsilon_{n+\tau} | \varepsilon_n] h(\varepsilon_n) d\varepsilon_n.$$

The first moment or error mean is zero, which gives that $E[\varepsilon_n | u_n] = 0$. Shifting subscripts gives $E[\varepsilon_{n+\tau} | u_{n+\tau}] = 0$. If $\tau > 0$, then $\varepsilon_{n+\tau}$ can only depend on ε_n through the predithered quantizer input $u_{n+\tau}$ via a possible feedback contribution. This implies that $E[\varepsilon_{n+\tau} | \varepsilon_n] = 0$, for $\tau > 0$. Substituting this into the sum or the integral above, we get that the autocorrelation function $< \varepsilon_n \varepsilon_{n+\tau} >= 0$ if $\tau > 0$, and $< \varepsilon_n \varepsilon_{n+\tau} >= D_n$ if $\tau = 0$, where D_n is the error variance. From the formulas above, the time averaged autocorrelation function $<< \varepsilon_n \varepsilon_{n+\tau} >>= 0$ if $\tau > 0$, and $<< \varepsilon_n \varepsilon_{n+\tau} >>= \lim_{T \to \infty} \frac{1}{T} \sum_{n=0}^{T} D_n = D$ if $\tau = 0$, where D is the time averaged or asymptotic steady state error variance, which is assumed to exist. From the discussion at the beginning of the proof, we now conclude that the error ε_n will be white. ■

In [65] by Lipshitz et al., corollaries 1 and 2, one sees that, if the dither is i.i.d., then the existing theory gives the same conclusion as this proposition. From the proof of this proposition, we have that the power spectrum constant is simply the time averaged second error moment. Applying the second error moment results of standard dither theory given in Section 1.6 to this result when Proposition 8.1 is applied to the cases of RPDF and TPDF dither, leads immediately to the relationships between the input and the power spectrum constant stated just before the proposition.

Proposition 8.1 thus provides a simple statement of the relationship between the first error moment and the error power spectrum induced by dither, and thus a specification on a property of the dither that is sufficient to bring about a white error power spectrum. The condition of white errors or white quantization noise implies that all frequency modes

are equally present in the error quantity dynamics over time, so that limit cycles or the dominance of the error pattern by certain periodic modes is not possible. Such conditions improve the quality of the digital representation of sampled analogue signals. Having unshaped errors that are white is therefore one desirable goal for Σ-Δ modulators in practice. It should be noted that having a white error spectrum does not imply that the errors ε_n are statistically independent of each other for different values of n. Such independence constitutes a stronger condition which, conversely, would imply a white spectrum. It is impossible to achieve this independence with nonsubtractive dither.

Strictly speaking, the second average distribution convergence criterion mentioned in the discussion of Theorem 7.11 in Chapter 7 is an implied requirement for the result of Proposition 8.1. Extending the examination of white error behaviour to systems with a more general dither than that of Proposition 8.1 is difficult, and is not pursued here.

For the rest of this chapter, much of our concern will essentially be with the goal of analyzing the second moment or error variance $E[\varepsilon_n^2]$. This quantity is important because it corresponds to the noise power, when the overall errors are regarded as noise. We start by looking at how the quantities associated with the quantizer part of the Σ-Δ modulator relate, and what formulations we can state. We observe from the topology of Figure 1.7 the following: $\varepsilon_n = y_n - u_n$ and $q_n = y_n - (u_n + \nu_n)$. Combining these equations, we arrive at $\varepsilon_n = \nu_n + q_n$. An expression for the error variance is then given as follows:

$$E[\varepsilon_n^2] = E[(\nu_n + q_n)^2] = E[\nu_n^2] + E[q_n^2] + 2E[\nu_n q_n]. \tag{8.1}$$

The third term in this equation will be zero if and only if ν_n and q_n are uncorrelated. From the form of the topology, we see that this will generally not hold. The following proposition provides conditions under which this does hold, along with subsequent results:

Proposition 8.2 *Suppose the predithered quantizer input u_n is random with a uniform PDF over $\mathcal{C}$. Then the internal quantizer error q_n will be independent of the dither ν_n, and q_n will be random with a uniform PDF over $\mathcal{C}$ (RPDF). Also, the modulator error variance will equal the sum of the dither variance and the internal quantizer error variance: $E[\varepsilon_n^2] = E[\nu_n^2] + \Delta^2/12$, where $E[q_n^2] = \Delta^2/12$. Moreover, the PDF of the modulator error value ε_n will be given by $h(\varepsilon) = \dfrac{1}{\Delta}\displaystyle\int_{\varepsilon-\Delta/2}^{\varepsilon+\Delta/2} g(\nu)d\nu$, where $g(\nu)$ is the PDF of the dither ν_n.*

Proof:

For a given value of dither ν_n, the density function for q_n is given by

$$\begin{aligned} h_q(q|\nu_n) &= h_u(\Delta/2 - q - \nu_n), \qquad 0 \le q \le \Delta/2, \\ h_q(q|\nu_n) &= h_u(-\Delta/2 - q - \nu_n), \qquad -\Delta/2 < q < 0, \end{aligned}$$

where $h_u(u)$ is the density function for $\hat{P}_{\mathcal{C}^1}(u_n)$, or u_n on $\mathcal{C}$. These results simply follow from the topology, and the definition of the quantizer and $\mathcal{C}$. Since $h_u(u)$ has a uniform density over $\mathcal{C}$, it then follows in the expressions above that $h_q(q|\nu_n)$ has a uniform density over $\mathcal{C}$. This density has no dependence on the ν_n value, and thus q_n is independent of ν_n, with a uniform density (or RPDF) over $\mathcal{C}$.

With ν_n and q_n independent, we have $E[\nu_n q_n] = E[\nu_n]E[q_n]$. Since q_n has a uniform density over $\mathcal{C}$, this implies $E[q_n] = 0$, so that $E[\nu_n q_n] = 0$, from above. From the probability density, we also have $E[q_n^2] = 2\displaystyle\int_0^{\Delta/2} q^2 dq = \Delta^2/12$. Substituting these results into the error variance formula (8.1), we obtain the desired final result.

From the independence of ν_n and q_n, it follows that the density function for the magnitude of ε_n is given by the convolution of the density functions of ν_n and q_n which make up its sum. Thus

$$h(\varepsilon) = \int_{-\infty}^{+\infty} g(\nu)h_q(\varepsilon - \nu)d\nu = \frac{1}{\Delta}\int_{\varepsilon-\Delta/2}^{\varepsilon+\Delta/2} g(\nu)d\nu,$$

since $h_q(q)$, the density function for q_n, is defined as $1/\Delta$ over $(-\Delta/2, \Delta/2]$, (i.e. over $\mathcal{C}$) and zero elsewhere. ■

Proposition 8.2 therefore provides simple error variance and error probability density results in terms of the statistical properties of the dither, when the predithered quantizer input is uniformly distributed modulo the circle $\mathcal{C}$. Such a form of predithered quantizer input may arise in a closed loop Σ-Δ system (where $u_n = x_n - r_n$) in several ways. One way is to simply have a random external input x_n that is uniformly distributed modulo $\mathcal{C}$. Another way, conversely, is to have an internal feedback r_n that may be interpreted to be random at steady state, with a uniform distribution modulo $\mathcal{C}$. The latter may arise from a topologically transitive dynamic property of the system, as was discussed in Chapter 7. More abstractly, "random" initial conditions (uniform over $\mathcal{C}$) in state space may also bring this about for r_n. Clearly having x_n and r_n independent, and either quantity uniformly distributed over $\mathcal{C}$, will, by arguments analogous to those used in the proof of Proposition 8.2, make $u_n = x_n - r_n$ uniformly distributed over $\mathcal{C}$.

Generalizing from the idea of the last result of Proposition 8.2, we may express the probability density of the error ε_n in terms of the probability densities of the predithered quantizer input u_n and the dither ν_n as follows.

Let $h(\varepsilon)$, $h_u(u)$ and $g(\nu)$ be the probability density functions (probability mass functions if the distribution is discrete) for the error ε_n, predithered quantizer input u_n, and dither ν_n respectively, with each defined over $\mathbb{R}$. Then we have the following:

1. If the dither distribution is discrete, then

$$h(\varepsilon) = \sum_{k=-\infty}^{+\infty} h_u(\frac{\Delta}{2} - \varepsilon + k\Delta) \cdot \sum_{\nu \in S(\varepsilon - \frac{\Delta}{2}, \varepsilon + \frac{\Delta}{2})} g(\nu), \qquad \varepsilon \in \mathbb{R}, \tag{8.2}$$

where $S(a, b)$ designates the set of all dither values ν of nonzero probability lying in the

interval set $[a, b)$ in $\mathbb{R}$.

2. If the dither distribution is piecewise continuous, then

$$h(\varepsilon) = \sum_{k=-\infty}^{+\infty} h_u(\frac{\Delta}{2} - \varepsilon + k\Delta) \cdot \int_{\varepsilon-\Delta/2}^{\varepsilon+\Delta/2} g(\nu)d\nu, \qquad \varepsilon \in \mathbb{R}. \tag{8.3}$$

The derivations of formulas (8.2) and (8.3) follow directly from the mathematical relationships between u_n, ν_n and ε_n in the system topology, and basic probability results. Clearly these formulas can be used to give general expressions and derive formulas for the error moments $E[\varepsilon_n^m]$, $m \geq 1$, in terms of the probability densities of u_n and ν_n, using basic moment definitions involving $h(\varepsilon)$. Our previous moment calculations for the RPDF and TPDF dither cases followed a simple form of this approach.

In the study of this chapter, it becomes both conceptually meaningful and practically relevant to interpret and express the predithered quantizer input in terms of its state position on the unit circle $\mathcal{C}$, that is modulo $\mathcal{C}$, as was done in Proposition 8.2. Conceptually, this is desirable because this is consistent with the circle map approach used for the state space description of the error ε_n in our dynamical system formulation of the Σ-Δ modulator for studying chaos. Practically, it is desirable because we find from the derivation of the formulas above that values of u_n that are equal modulo $\mathcal{C}$ are "mapped" to the same error value ε_n in $\mathbb{R}$ for a given dither ν_n. Hence we see the importance of the circle map symmetry entering into our analysis, even when we are not applying this interpretation to our current treatment of the errors ε_n for which it was first introduced.

If we take $\tilde{u}_n$ to represent the value of u_n on $\mathcal{C}$, that is $\tilde{u}_n = \hat{P}_{\mathcal{C}^1}(u_n)$, and let $h_{\tilde{u}}(\tilde{u})$ denote its PDF/PMF, then we may rewrite the first summation portions of (8.2) and (8.3) as follows:

$$\sum_{k=-\infty}^{+\infty} h_u(\frac{\Delta}{2} - \varepsilon + k\Delta) = h_{\tilde{u}}(\frac{\Delta}{2} - (\varepsilon \bmod \Delta)). \tag{8.4}$$

The probability density functions h_u and $h_{\tilde{u}}$ are defined over $\mathbb{R}$ and $\mathcal{C}$ respectively. Thus (8.4) expresses the fact that the value of the probability density at a point $\tilde{u}_n$ on $\mathcal{C}$ is simply the sum of the probability densities of all the points u_n that are equivalently at the point $\tilde{u}_n$ on $\mathcal{C}$. With this simplification incorporated, we find that if we choose the density to be uniform over $\mathcal{C}$, that is $h_{\tilde{u}}(u) = 1/\Delta$, then (8.3) reduces to the final density result in Proposition 8.2.

8.1.1 RPDF Dither

For the next part of this section, we examine the variance properties of the value of the error ε_n under conditions of RPDF dither, and its higher-order convolutions. By definition, RPDF dither (PDF width Δ) has a uniform probability distribution over $\mathcal{C}$. The following lemma shows that this result also holds for its higher-order convolutions, and indeed for the convolution of any distribution that is uniform over $\mathcal{C}$ with any other independent PDF:

Lemma 8.3 *The probability distribution formed from the convolution of a distribution with a uniform PDF over $\mathcal{C}$, with any other independent probability distribution defined over $\mathbb{R}$, will have a PDF that is uniformly distributed over $\mathcal{C}$.*

Proof:

Let G be a random variable with the PDF of a given distribution over $\mathbb{R}$. Let K be a random variable with a uniform distribution over $\mathcal{C}$. Then the random variable $H = K + G$ will have the PDF corresponding to the convolution of the distributions corresponding to K and G above. A mapping $\hat{P}_{\mathcal{C}^1}$ from $\mathbb{R}$ to the equivalent position on $\mathcal{C}$ may be given by

$$\begin{aligned} \hat{P}_{\mathcal{C}^1}(x) &= \Delta/2 - (Q(x) - x), \qquad m\Delta \le x \le m\Delta + \frac{\Delta}{2}, \\ \hat{P}_{\mathcal{C}^1}(x) &= -\Delta/2 - (Q(x) - x), \qquad m\Delta - \frac{\Delta}{2} < x < m\Delta, \end{aligned}$$

for $x \in \mathbb{R}$, where $m \in \mathbb{Z}$, and $Q(x)$ is the usual Σ-Δ quantizer function. From this mapping form, it is clear that H will be uniformly distributed over $\mathcal{C}$ (i.e. modulo $\mathcal{C}$), if its quantization error $q_H = Q(H) - H$ has a uniform PDF over $\mathcal{C}$. Now applying Proposition 8.2, with $u_n = K$, $\nu_n = G$, and $q_n = q_H$, we have that q_H has a uniform PDF over $\mathcal{C}$. Thus, from the above, the required result follows. ■

Corollary 8.4 *The probability distribution of an nth order convolution of RPDF dither will have a PDF that is uniformly distributed over $\mathcal{C}$, for all $n \geq 1$.*

Proof:

The probability density function of an nth order convolution of RPDF dither can be expressed, by definition, as the convolution of a RPDF with the PDF of an $(n-1)$th order convolution of RPDF dither, if $n > 1$. The RPDF dither has by definition a uniform PDF over $\mathcal{C}$. Applying Lemma 8.3, we then have the required result for the PDF of the nth order convolution of RPDF dither. By definition, this result also holds for $n = 1$. By induction, we have the required result. ■

Now let us consider the case of a system with RPDF dither ν_n and some arbitrary (i.e. unknown or unspecified) predithered quantizer input u_n. We may apply Proposition 8.2 with the roles of ν_n and u_n reversed. By the symmetry of the relationship between u_n, ν_n and q_n, we can conclude, from Proposition 8.2, that the internal quantizer error q_n has a uniform PDF over $\mathcal{C}$ and is independent of u_n. This is a restatement of the result of Schuchman [59]. Both ν_n and q_n are now RPDF with mean zero and variance $\Delta^2/12$. From the modulator error variance formula (8.1), we thus have

$$E[\varepsilon_n^2] = \frac{\Delta^2}{6} + 2E[\nu_n q_n]. \tag{8.5}$$

In the event that ν_n and q_n are uncorrelated, (8.5) reduces to an error variance value of $\Delta^2/6$ (since $E[\nu_n q_n] = 0$ then). This corresponds to the variance of a TPDF distribution of width 2Δ, which is the distribution one obtains for the modulator error when ν_n and q_n are not only uncorrelated but also independent. In this case, the relation $\varepsilon_n = \nu_n + q_n$ corresponds to the distribution of the error as a convolution of the distributions of the RPDF dither and RPDF internal quantizer error, which is hence the TPDF mentioned above. Since (from dither theory) an RPDF dither cannot be guaranteed to control or fix the error variance, then, of course, in general ν_n and q_n will be correlated, and the second term in the variance formula (8.5) will be nonzero. As we shall see, this term may be either positive or negative, with an expected value of zero when u_n is chosen randomly from a uniform distribution over $\mathcal{C}$. Hence the value $\Delta^2/6$ may be considered a generic average value of the error variance when RPDF dither is used[1]. Note that the random processes ν_n and q_n, $n \geq 0$, associated with a closed loop feedback system can never be independent; they are related by the quantization operation Q and the feedback.

We now extend our analysis to the case of a system with dither ν_n that is $p > 1$ convolutions of RPDF. First, we suppose that the predithered quantizer input u_n is uniformly distributed over $\mathcal{C}$ (see Section 8.2 for context). Then, from Proposition 8.2, it follows that the error variance will be the dither variance (p times the RPDF variance) plus the variance of q_n (same as for RPDF) and hence $(p+1)\dfrac{\Delta^2}{12}$. Since, from dither theory, the variance

[1]This result — a consequence of Proposition 8.2 that will be developed further in Section 8.2 — would seem to improve upon the results of Lipshitz et al. which provided for an average variance of $\Delta^2/4$ under general conditions. Lipshitz et al. assume the dither is TPDF and make no other assumptions (see our second moment derivation for TPDF dither, which is this constant hence applies as the average result). We are using RPDF dither, but are making the very strong assumption of a uniformly distributed predithered quantizer input in time over $\mathcal{C}$. It is this special circumstance, which on the surface seems rather general, that allows for our improved average variance result.

with $p > 1$ convolutions of RPDF dither must be constant for any system, it follows that the value of this constant variance is $(p+1)\dfrac{\Delta^2}{12}$. If the predithered quantizer input is now arbitrary, we may apply Proposition 8.2 with the roles of ν_n and u_n reversed, which is a more general form of Schuchman's result. For this we use the fact that the convolution of p RPDF dithers is uniform over $\mathcal{C}$ by Corollary 8.4. Thus q_n is uniform over $\mathcal{C}$. Applying the error variance formula (8.1), we then have that $E[\varepsilon_n^2] = (p+1)\dfrac{\Delta^2}{12} = p\dfrac{\Delta^2}{12} + \dfrac{\Delta^2}{12} + E[\nu_n q_n]$. Thus $E[\nu_n q_n] = 0$, showing that ν_n and q_n are always uncorrelated when $p > 1$. This consequence is consistent with what we might expect, after examining this correlation term in the RPDF dither case earlier.

An important advantage of using RPDF dither over dither of higher RPDF convolutions is the reduction of the error variance level to the general neighbourhood of $\Delta^2/6$ from the higher fixed levels associated with the higher-order PDFs. A disadvantage is that this variance is not fixed, but input dependent. We wish to lessen this disadvantage by understanding more about the dependency of this variance on other aspects of the system. In the work to follow, we will see, for RPDF Σ-Δ modulator systems, how the error variance may vary about the reference level of $\Delta^2/6$ and under what circumstances this variance level may be attained.

8.2 Dithered Σ-Δ Modulators

When extending the statistical analysis of the overall error ε_n from quantizers to a full Σ-Δ modulator system, it is important, at least mathematically, to make the distinction between the statistics of the error ε_n at some fixed value of n, versus a long run steady state error ε_n as characterized in Chapter 7. In a dithered system that is otherwise deterministic (without dither), these quantities generally differ, and it is only the steady state interpretation that

gives a consistent and meaningful result. From a practical point of view, it is the steady state error and its statistics that are of interest, in any event, since this corresponds to what we have with a typical observed error ε_n at arbitrary and large n, or a sample of a long sequence of such errors.

The following results, taken together with the results presented in Section 7.2, provide general conditions under which a variance level of $\Delta^2/6$ may be achieved for the Σ-Δ modulator error, with RPDF dither.

Theorem 8.5 *Suppose the error ε_n has a uniform PDF over $\mathcal{C}$ (i) for a given $n \geq 0$, or (ii) in its long run steady state behaviour. Then the PDF of the predithered quantizer input u_n will be uniformly distributed over $\mathcal{C}$ (iii) for the given $n \geq 0$ if (i) holds, or (iv) in its long run steady state behaviour if (ii) holds.*

Moreover, the conditions and results of Proposition 8.2 will hold for cases (i) or (ii) respectively. More specifically, if the dither is a convolution of p RPDFs, then the PDF of the value of the error ε_n, for the given n in case (i), or in its long run steady state in case (ii), will be a convolution of $(p+1)$ RPDFs for $n \geq 1$, $p \in \mathbb{Z}^+$.

Proof:

For the proof, we shall consider ε_n to be such that either n satisfies *(i)*, or n is chosen arbitrarily so that *(ii)* can be taken to apply. Clearly, there is a one-to-one mapping between the value of u_n on $\mathcal{C}$, and the value of ε_n on $\mathcal{C}$, in the nondithered case, given by $f_\varepsilon(u) = \Delta/2 - u$, if $0 \leq u \leq \Delta/2$; and $f_\varepsilon(u) = -\Delta/2 - u$, if $-\Delta/2 < u < 0$. The addition of dither ν_n may change the value of the quantizer $Q(u_n + \nu_n)$ by some integer multiple of Δ, which does not change the value of ε_n on $\mathcal{C}$. Thus this one-to-one mapping remains the same with any arbitrary dither present. From the form of $f_\varepsilon(u)$, it is clear (considering the inverse of $f_\varepsilon(u)$) that a uniform density function over $\mathcal{C}$ for ε_n can only be

mapped to by a uniform density function for u_n over $\mathcal{C}$. Thus we have the first result.

The conditions for and results of Proposition 8.2 then apply. Now applying Proposition 8.2, we have that the internal quantizer error q_n is independent of the dither ν_n, and has a uniform PDF over $\mathcal{C}$ that is hence RPDF. Applying this to the relation $\varepsilon_n = \nu_n + q_n$, we then have that the PDF of the error value ε_n is the convolution of the PDF of q_n with the PDF of ν_n. Thus if the dither is the convolution of p RPDFs, we get the required result for the PDF of ε_n. ■

We may note that this theorem demonstrates a situation where an error PDF of $(p+1)$ RPDF convolutions arises from a system with dither that is a convolution of p RPDFs. This is not true in general, particularly when considering a given $n \geq 0$ (case *(i)*). It is true here only because of the uniform distribution assumption over $\mathcal{C}$ of ε_n or u_n — an assumption that has a more pertinent role in the context of the steady state. The variance of this error would then be, by independence, $(p+1)$ times the RPDF variance, and hence $(p+1)\dfrac{\Delta^2}{12}$, (as given from Proposition 8.2).

More generally, Theorem 8.5 asserts essentially that if we know that the error value ε_n is uniformly distributed over $\mathcal{C}$, then, for a given dither ν_n with known PDF added, we can determine the PDF of the error value ε_n as a convolution of an RPDF with the dither PDF. In short, we can apply Proposition 8.2. The error value variance $E[\varepsilon_n^2]$ then follows as well. Seeking to utilize Theorem 8.5 for this purpose, Theorems 7.5, 7.7, 7.10, 7.13, their corollaries, and Propositions 7.9 and 7.12 from the work of the last chapter, will provide conditions under which the error value ε_n will be uniformly distributed over $\mathcal{C}$. The following result then follows automatically from Proposition 7.16:

Proposition 8.6 *Suppose that the system satisfies one of Theorems 8.5, 7.5, 7.7, 7.10 plus (R), 7.13, their corollaries, Propositions 7.9 plus (R), 7.12 for all* $n \geq 0$*, or 7.16. Suppose also that the system has a dither* ν_n *that is the convolution of* p *RPDFs. Then, under the conditions of the theorem, corollary or proposition that the system satisfies, the PDF of the value of the system error* ε_n *will be a convolution of* $(p+1)$ *RPDFs in its long run steady state behaviour (or for all* $n \geq 0$*, if Proposition 7.12 is satisfied for all* $n \geq 0$*, or Proposition 7.16 with (i) is satisfied).*

Proof:

This follows from applying Theorems 7.5, 7.7, 7.10, 7.13, their corollaries, and Propositions 7.9 and 7.12; with Proposition 7.16 and Theorem 8.5. ■

Proposition 8.6 now provides conditions under which an error value variance of $\Delta^2/6$ may be achieved in steady state. Specifically, for a system with RPDF dither, if any of Theorems 8.5, 7.5, 7.7, 7.10, 7.13, their corollaries, Propositions 7.9, 7.12 or 7.16 are satisfied, along with condition (R) for 7.10 or 7.9, then the error value ε_n will have a TPDF distribution in the long run (steady state) and hence a variance of $\Delta^2/6$. In short, the attainment of a uniform PDF of ε_n over $\mathcal{C}$ is sufficient for the RPDF dither to yield the nominally ideal value of $\Delta^2/6$ for the error value variance.

This uniform PDF must exist when the dither is present if condition (R) does not hold, (if (R) holds, its existence in the nondithered case is sufficient). Even if we had produced more general results in Section 7.2 concerning the existence of steady state uniform errors over $\mathcal{C}$, we could not, in general, reach the TPDF error distribution conclusions from RPDF dither here if (R) does not hold. Incorporating the effect of dither into the analysis of the error dynamics is difficult under such generality, as was seen in Chapter 6. Therefore

we do not attempt to extend our approach to arriving at error distribution and variance conclusions in thesis to beyond our condition (R) treatment, as such (see the second last paragraph of Section 8.2 for further discussion). Clearly some perturbation of the conditions from those in the theorems might be expected to yield systems whose resulting error variance $E[\varepsilon_n^2]$ is perturbed off the $\Delta^2/6$ result.

To clarify the significance of Proposition 8.6 for Σ-Δ modulator systems, and put it in some overall context, we make the following comments. The results are important in that they provide for a constant (input independent) steady state error variance (i.e. $\Delta^2/6$) over certain classes of input x_n, for various system forms (pertaining to the different theorems and results in Proposition 8.6). For example, when there is a nonminimum-phase zero, the initial condition $\vec{x}_0$ is of generic type, and condition (R) holds (conditions of Theorem 7.5), then the steady state error variance will be constant for any periodic input x_n. Establishing a constant variance result over systems with a range of possible input sequences x_n, in this manner, is an important accomplishment, considering that under RPDF dither, the error variance will generally vary, not only at a given n as the predithered quantizer input u_n varies, but in steady state as the overall structure of u_n varies (e.g. from one periodic form to another). The results also provide for the constant (I.C. independent) steady state variance over general classes of initial conditions $\vec{x}_0$, for various system forms. The absence of these error properties for some general forms of the first-order model of Section 8.3 illustrates the relative nontriviality and uniqueness of our result, and the need for rigorous, theoretical approaches to establish results of this nature.

Also of importance is the fact that the constant steady state variance is at the nominally ideal value of $\Delta^2/6$ for RPDF dither, indicating that the error variance biases, at particular values of n, on average, balance out. This result may not seem overly remarkable, in the sense that this is the mean variance $E[E[\varepsilon_n^2]]$ one would expect at a particular n,

for u_n chosen from a uniform distribution over $\mathcal{C}$ (which is our steady state interpreted result). It is important, however, to assert that this nominal value, and the implied average noncorrelation between q_n and ν_n, hold over the general conditions (e.g. regarding input x_n, initial conditions mentioned above) for the various results in Proposition 8.6, and that the steady state variance is hence below the fixed value of $\Delta^2/4$ given by TPDF dither as well. Note that this steady state variance really pertains to the steady state of $E[\varepsilon_n^2]$ given by $E[E[\varepsilon_n^2]]$, over general n (i.e. "n" is a random variable as discussed and characterized in Section 7.1).

Discussion:

As we noted in the footnote of Subsection 8.1.1, the steady state variance result of $\Delta^2/6$ that follows from Proposition 8.2 requires the strong condition that the predithered quantizer input u_n have a uniform steady state over $\mathcal{C}$. Under the conditions of Theorem 8.5 and satisfied by the conditions of Proposition 8.6, the requirement is still rather strong — namely that the error ε_n have a uniform steady state over $\mathcal{C}$. One might argue that the conditions of the theorems and results of Chapter 7 that satisfy this for Proposition 8.6 are overly restrictive, or of limited practical relevance — they require special filter forms (e.g. condition (R)), an input that is at least periodic, quasiperiodicity in the error, or stochastic initial conditions or input.

We believe that our are assumptions are nevertheless reasonable. As with the case of chaos or stability, the filter conditions at hand may be satisfied via the flexibility one naturally has to choose the filter coefficients. Systems with input that is periodic, constant, or that induces quasiperiodic behaviour are often investigated analytically or via simulation by researchers to learn more about pertinent issues of performance such as stability, limit cycle behaviour, chaos or noise control. General practical input signals may possess periodic components as well, and thus are amenable to some analysis via a periodic input

model. Systems with stochastic initial conditions or input may be studied to help address theoretical questions, and provide insights that are potentially applicable to certain practical systems or issues in the future. Moreover, our conditions for a uniform steady state error over $\mathcal{C}$ are strongly integrated with our conditions for chaos, thus allowing one to often satisfy both properties with little more than the conditions of either one. In important respects, the steady state conditions are weaker, requiring only one nonminimum-phase zero of $p(z)$. In fact, they are weaker than our condition in Chapter 7 for whiteness (which requires two such zeros). And finally, the uniformity property obviously possesses an aspect of symmetry and regularity, which we believe is likely to arise, in steady state, in more general systems with RPDF dither than those provided in Proposition 8.6. Therefore the results we have may constitute an important starting point for extensions, through further analysis, and thus represent the outline of a larger theory with broader direct applications.

Error Dependencies on Dither:

We shall clarify further what we may state about the statistical dependence of the error ε_n on $\mathcal{C}$, and the internal quantizer error q_n, upon a dither ν_n in the Σ-Δ modulator. From the dynamical analysis in Chapter 6, it follows that the value of ε_n on $\mathcal{C}$, i.e. $\breve{\varepsilon}_n$, will always be statistically independent of ν_n at the given value of n, and will be independent of all previous dithers $\nu_k,\ 0 \leq k < n$, as well, when condition (R) holds. From the results at the end of Chapter 6, it follows that q_n will be statistically independent of ν_n at the given n when u_n, at this n, is uniformly distributed over $\mathcal{C}$ (generally not true otherwise — essentially never when the PDF of ν_n is continuous anywhere), and will be independent of all previous dithers when (R) holds. Regardless of these properties however, it is easy to see the truth, asserted earlier, of how the random processes q_n and ν_n can never be independent in a Σ-Δ modulator system with fixed initial conditions and deterministic input:

From the relations discussed in Chapter 6, we have $q_n = \hat{P}_{\mathcal{C}^1}(\breve{\varepsilon}_n - \nu_n)$. Since any

randomness in $\breve{\varepsilon}_n$ depends, at most, only upon the dithers ν_k, $0 \leq k < n$, and the dither is i.i.d., it follows that the value of q_n depends statistically upon all, or some subset of ν_k, $k = 0, \ldots, n$.

The results of Theorem 6.5, as well, extend directly from functional, to statistical dependence of $\breve{\varepsilon}_n$ upon the previous dithers. Note that if (R) holds, then $\breve{\varepsilon}_n$ and u_n are always nonrandom.

The value of u_n, at any n, can be uniformly distributed over $\mathcal{C}$ (and indeed have a PDF that is anywhere continuous) only as a consequence of a random input x_k for some k values between 0 and n, and/or a random initial condition $\vec{x}_0$. This provides a means for q_n to be statistically independent of ν_n at the given value of n, and potentially of all previous dithers (and/or $\breve{\varepsilon}_n$ as well). The steady state concept of $\breve{\varepsilon}_n$, q_n and u_n that we have used also allows for this situation at an arbitrarily interpreted given "n". The notion of the dependence of $\breve{\varepsilon}_n$ or q_n upon previous dithers does not apply in the steady state context, however, because these random variables represent average behaviour, rather than behaviour at a particular value of n.

For a Σ-Δ modulator system with fixed initial conditions and deterministic input, u_n (and $\breve{\varepsilon}_n$), at a given n, will essentially be described by a discrete probability mass function, when (R) does not hold. This is because the dither induces random, discrete perturbations in the value of the error ε_n, as denoted by $m_n\Delta$ in Chapter 6. We may conjecture the possibility that, with certain filters, the discrete distribution for u_n will converge to a uniform distribution over $\mathcal{C}$ in the limit as $n \to \infty$. This might come about, for example, when the a_i and b_j are such that some of the $\tilde{r}_k$ from (R) are irrational, and a quasiperiodic distribution of the perturbation outcomes of u_n is induced over $\mathcal{C}$ as $n \to \infty$. Thus it may be possible to extend the results of Proposition 8.6 to Σ-Δ systems with broad and/or easily constructible classes of filter forms, when (R) does not hold. We might expect these

extensions, in particular, to be synonymous with the existence of chaos. In such results, we would be dealing with a steady state interpretation in the first average distribution. Although generalizations to a full state space analysis on $\mathcal{C}^M$, akin to the more theoretical work of Chapter 7, may be possible, the context put forth here (consistent with that of Chapter 8) is that of the single error coordinate ε_n in one dimension. Exploring these ideas and possibilities would then be a good topic for future research. ■

With these overall results, some basic conditions have been presented under which the introduction of a convolution of p RPDFs of dither to the system will bring about an error value with a PDF that is a convolution of $(p+1)$ RPDFs. This simple consequence represents the practical aims that contributed to motivate the more theoretical process that was undertaken in Chapter 7 to arrive at some understanding of the statistical and dynamical system behaviour here. These theorems and results provide both a diverse range of scenarios under which the resulting error variance of $\Delta^2/6$ arises, and several possible points of reference for further analysis. We shall now show a more specific relevance of these results, with the investigation of the Σ-Δ modulator with RPDF dither, and the variance properties of its error.

8.3 First-Order Model

We continue the analysis by focusing on the simple first-order Σ-Δ modulator system with unity gain and constant input and RPDF dither, so that $M = 1$, $N = 0$, $a_1 = 1$, and $x_n = c \in \mathbb{R}$, for all $n \geq 0$. The difference equation describing this system (nondithered), from (1.2), is as follows:

$$\varepsilon_n \;=\; \frac{\Delta}{2} \;-\; [(c - \varepsilon_{n-1}) \mod \Delta], \quad n \geq 0. \tag{8.6}$$

From Theorem 7.13 and Proposition 8.6, this system will have TPDF error, and hence error variance $\Delta^2/6$ in steady state if c/Δ is irrational. This will hold for any initial condition ε_{-1}. If we allow the gain a_1 in (8.6) to take on integer values different from ± 1 or 0, then, from Theorem 7.7, the system will maintain error variance $\Delta^2/6$ and TPDF error in steady state, for any real value of c/Δ. This is provided the initial condition ε_{-1} is not a periodic point, and does not enter into a limit cycle or fractal attractor after successive mappings. Such initial conditions would be expected to form a set of measure zero on $\mathcal{C}$. Note that for this case and the result, x_n may also vary with n. If, on the other hand, we keep the original conditions as in Theorem 7.13, but specify c/Δ to be strictly rational, we then find that every initial condition ε_{-1} will be a periodic point of the system. The PDF of the steady state error will then be discrete for any initial condition, and the error variance will generally not be $\Delta^2/6$. We will proceed to analyze this case further.

8.3.1 Analysis

We begin this analysis by first characterizing the orbits that arise for a given constant input c that is a rational multiple of Δ, and initial condition ε_{-1}. Let the input be represented by $c = \frac{p}{q}\Delta$, where p and q are coprime integers. If $c = 0$, then $q \equiv 1$. Using the difference equation (8.6) , we have the following:

Orbit Sets of u:

Case (a): If q is odd, or q is a multiple of four (and hence even), then the u_n orbit modulo $\mathcal{C}$ will lie on points in the set $\{\tilde{\varepsilon}^*_{-1} \pm \frac{i}{2k}\Delta,\ \tilde{\varepsilon}^*_{-1} + \frac{\Delta}{2},\ i = 0, \ldots, k-1\}$, where $k = q$ if q is odd, and $k = \frac{q}{2}$ if q is a multiple of four. Here also
$\tilde{\varepsilon}^*_{-1} = -(\varepsilon_{-1} \bmod \frac{\Delta}{2k})$, and $-\frac{\Delta}{2k} < \tilde{\varepsilon}^*_{-1} \leq 0$.

Case (b): If q is even, but not a multiple of four, then the u_n orbit modulo $\mathcal{C}$ lies in the set $\{\tilde{\varepsilon}^*_{-1} \pm \frac{i}{2k-1}\Delta,\ i = 0, \ldots, k-1\}$, where $k = \frac{1}{2}(\frac{q}{2}+1)$. In this case $\tilde{\varepsilon}^*_{-1} = \frac{\Delta}{2(2k-1)} - (\varepsilon_{-1} \bmod \frac{\Delta}{2k-1})$, and $-\frac{\Delta}{2(2k-1)} < \tilde{\varepsilon}^*_{-1} \leq +\frac{\Delta}{2(2k-1)}$. ■

For all cases, the orbit points of ε_n on $\mathcal{C}$ may be obtained from these orbit points for u_n on $\mathcal{C}$ by the relation $\varepsilon_n = \frac{\Delta}{2} - u_n$ if $0 \leq u_n \leq \frac{\Delta}{2}$, and $\varepsilon_n = -\frac{\Delta}{2} - u_n$ if $-\frac{\Delta}{2} < u_n < 0$. From these relations, we see that the orbit sets for ε_n will take on a configuration that is topologically equivalent to that for u_n on $\mathcal{C}$. For the orbit of u_n or ε_n on $\mathcal{C}$, we get either $2k$ or $2k-1$ equally spaced points on the circle, for cases (a) or (b) respectively. The period of each point in the orbit will be $2k$ or $2k-1$ for (a) or (b) respectively as well.

With this information, we want to construct formulas for the error variance as a function of q and the initial condition ε_{-1}, under the regime of RPDF dither. The probability distribution for the error ε_n will be taken essentially to correspond to steady state. For the value of a single input u_n on $\mathcal{C}$, we have seen earlier (beginning of Section 8.1) that the error variance is $\frac{\Delta^2}{4} - u_n^2$. Applying this to the case of an orbit set with a finite number of points, the variance would then be given by the expression $\frac{\Delta^2}{4} - \sum_{i=1}^{n_u} p_i u_{(i)}^2$, where the orbit set $\{u_{(i)}\}$ has n_u elements, and the ith element has (steady state) probability of occurrence (i.e. relative frequency) p_i, with $\sum_{i=1}^{n_u} p_i = 1$. Thus if an element $u_{(i)}$ is periodic with period $\hat{T}_i$, then $p_i = \frac{1}{\hat{T}_i}$. If u_n on $\mathcal{C}$ has a continuous steady state PDF $h_u(u)$, then the error variance is the analogous integral expression $\frac{\Delta^2}{4} - \int_{\frac{-\Delta}{2}}^{\frac{\Delta}{2}} h_u(u) u^2 du$. Applying the finite orbit set formulas to cases (a) and (b) above, we get the following expressions for the error variance:

1. (a) $E[\varepsilon_n^2] = \frac{\Delta^2}{4} - \frac{1}{2k}[2\Delta^2 \sum_{i=1}^{k-1}(\frac{i}{2k})^2 + 2k\tilde{\varepsilon}_{-1}^{*\,2} + \tilde{\varepsilon}_{-1}^{*}\Delta + \frac{1}{4}\Delta^2];$

   (b) $E[\varepsilon_n^2] = \frac{\Delta^2}{4} - \frac{1}{2k-1}[2\Delta^2 \sum_{i=1}^{k-1}(\frac{i}{2k-1})^2 + (2k-1)\tilde{\varepsilon}_{-1}^{*\,2}].$

Using the summation formula $\sum_{i=1}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1)$, the above expressions can be simplified to give the following formulas:

2. (a) $E[\varepsilon_n^2] = \frac{\Delta^2}{6} - [\tilde{\varepsilon}_{-1}^{*\,2} + \frac{\tilde{\varepsilon}_{-1}^{*}}{2k}\Delta + \frac{\Delta^2}{24k^2}];$

   (b) $E[\varepsilon_n^2] = \frac{\Delta^2}{6} - [\tilde{\varepsilon}_{-1}^{*\,2} - \frac{\Delta^2}{12(2k-1)^2}].$

We can combine these two formulas into one single formula, by letting $\tilde{\varepsilon}_{-1} = \tilde{\varepsilon}_{-1}^{*}$, and $s = 2k$ in case (a); and $\tilde{\varepsilon}_{-1} = \tilde{\varepsilon}_{-1}^{*} - \frac{1}{2(2k-1)}$, with $s = 2k-1$ in case (b). With these adjustments, we arrive at the following, incorporating both cases[2]:

2. $E[\varepsilon_n^2] = \frac{\Delta^2}{6} - [\tilde{\varepsilon}_{-1}^2 + \frac{\tilde{\varepsilon}_{-1}}{s}\Delta + \frac{\Delta^2}{6s^2}];$

   where $s = 2q$ if q is odd, $s = q$ if q is a multiple of 4, and $s = \frac{q}{2}$ if q is even, but not a multiple of 4;

   $\tilde{\varepsilon}_{-1} = -(\varepsilon_{-1} \bmod \frac{\Delta}{s}), \quad -\frac{\Delta}{s} < \tilde{\varepsilon}_{-1} \leq 0.$

In these formulas, we see that the bracketed terms represent the deviation or perturbation of the error variance from the reference point ideal of $\frac{\Delta^2}{6}$ discussed earlier. The value of $E[\varepsilon_n^2]$ as a function of $\tilde{\varepsilon}_{-1}$, with s constant, is described by a concave down parabola over the interval of definition of $\tilde{\varepsilon}_{-1}$ as given in formula 2 above, with vertex maximum at the interval midpoint, and minimum at the endpoints. Specifically, the variance maxima

[2]The corresponding error variance formula for the same first-order model with a mid-tread quantizer is given by:

$$E[\varepsilon_n^2] = \frac{\Delta^2}{6} - [\tilde{\varepsilon}_{-1}^2 + \frac{\tilde{\varepsilon}_{-1}}{q}\Delta + \frac{\Delta^2}{6q^2}],$$

where q is as defined at the beginning of this subsection, and $\tilde{\varepsilon}_{-1} = -(\varepsilon_{-1} \bmod \frac{\Delta}{q})$, $-\frac{\Delta}{q} < \tilde{\varepsilon}_{-1} \leq 0$.

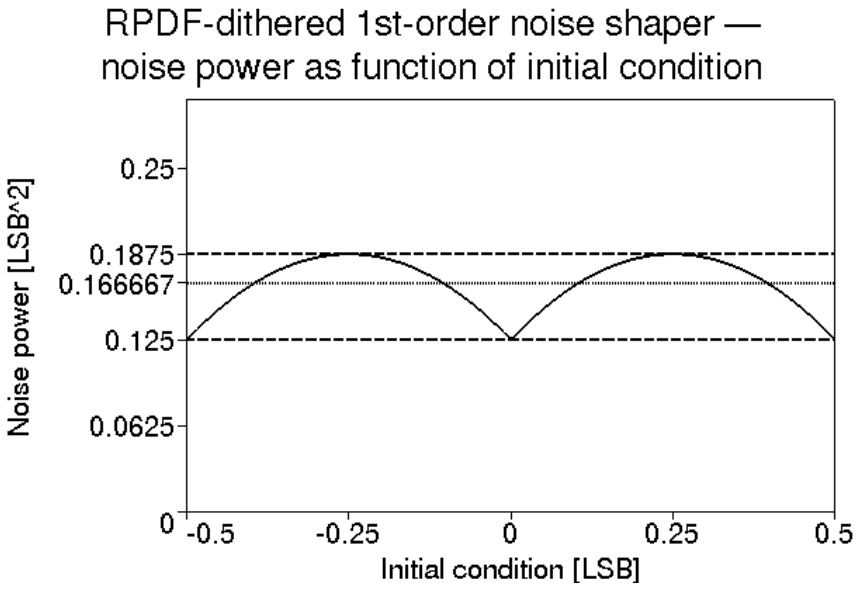


Figure 8.2: $E[\varepsilon_n^2]$ as a function of ε_{-1} when $q = 1$

and minima are given as follows:

3. $E[\varepsilon_n^2]_{max} = \dfrac{\Delta^2}{6} + \dfrac{\Delta^2}{12s^2}$, at $\tilde{\varepsilon}_{-1} = -\dfrac{\Delta}{2s}$;
$E[\varepsilon_n^2]_{min} = \dfrac{\Delta^2}{6} - \dfrac{\Delta^2}{6s^2}$, at $\tilde{\varepsilon}_{-1} = 0,\ -\dfrac{\Delta^+}{s}$.

By extension, it follows that the error variance as a function of ε_{-1} over $\mathcal{C}$ is described by a piecewise continuous union of s equivalent parabolas of this form. The value of $E[\varepsilon_n^2]$, i.e. the noise power, as a function of ε_{-1} for the fairly simple case when $q = 1$, is shown in Figure 8.2, with Δ scaled to 1. This corresponds to an input c that is zero, or more generally an integer multiple of Δ, and yields two parabolas over the domain $\mathcal{C}$.

If we are interested in controlling the error variance to a level below $\dfrac{\Delta^2}{6}$, then we may obtain, from a further analysis of formula 2, the following intervals for $\tilde{\varepsilon}_{-1}$ on which this holds:

4. $E[\varepsilon_n^2] < \dfrac{\Delta^2}{6}$, if $|\tilde{\varepsilon}_{-1} - \dfrac{i}{s}\Delta| < \dfrac{\Delta}{2s}(1 - \dfrac{1}{\sqrt{3}})$, $i = 0, -1$.

Equality holds when the inequality is replaced by equality in the expression. In terms of the value of the initial condition ε_{-1} on $\mathcal{C}$, this result says that the error variance will be a minimum if ε_{-1} is chosen to be a point of the form $\pm\frac{i}{s}\Delta$, for $i = 0, \ldots, \frac{s}{2}$, in case (a), and $i = 0, \ldots, \frac{1}{2}(s+1)-1$, in case (b). The error variance will be less than $\frac{\Delta^2}{6}$ if ε_{-1} is chosen to be within a distance of $\frac{\Delta}{2s}(1-\frac{1}{\sqrt{3}})$ of the respective points. The error variance will be a maximum if ε_{-1} is chosen at the midpoints between these points. From these results it can be seen that intervals on $\mathcal{C}$ over which the initial condition ε_{-1} will give an error variance of less than $\frac{\Delta^2}{6}$ have a combined length of $(1-\frac{1}{\sqrt{3}})\Delta$, or about 42% of the total length of $\mathcal{C}$. Thus if ε_{-1} is chosen at random according to a uniform PDF over $\mathcal{C}$, we would expect the resulting system error variance to be over $\frac{\Delta^2}{6}$ with about 58% probability.

For ε_{-1} chosen randomly on $\mathcal{C}$ according to a PDF $h(x)$, the mean or expected error variance would be given by $\int_{-\frac{\Delta}{2}}^{+\frac{\Delta}{2}} h(x)E[\varepsilon_n^2|\varepsilon_{-1}=x]dx$. If $h(x)$ is uniform so that $h(x) = \frac{1}{\Delta}$ for $x \in \mathcal{C}$, we may calculate the expected error variance as follows:

$$\begin{aligned} 5.\quad E[\varepsilon_n^2] &= \frac{s}{\Delta}\int_{-\frac{\Delta}{s}}^{0}(\frac{\Delta^2}{6} - [x^2 + \frac{x}{s}\Delta + \frac{\Delta^2}{6s^2}])dx \\ &= \frac{\Delta^2}{6} + [\frac{1}{3s^3} - \frac{1}{2s^3} + \frac{1}{6s^3}]s\Delta^2 \qquad = \frac{\Delta^2}{6}. \end{aligned}$$

Thus the expected error variance is always $\frac{\Delta^2}{6}$, for any rational $\frac{c}{\Delta}$. From the functional form of $E[\varepsilon_n^2]$ over $\mathcal{C}$ as a union of parabolas, we have that the deviation of the variance below this level, for ε_{-1} over the corresponding portions of $\mathcal{C}$ (42% of total), is then greater (i.e. carries more weight) than the deviation above on the remaining portions of $\mathcal{C}$. The peak deviations above and below are similarly unequal. Their averages on the corresponding portions of $\mathcal{C}$ are equal and opposite sign and so cancel. Notice now that the system (8.6) under consideration (for any x_n) satisfies the conditions of Theorem 7.10. Our random choice of ε_{-1} with uniform PDF over $\mathcal{C}$ then corresponds to the error PDF fixed point

implied by this theorem. By Theorem 8.5, and Propositions 7.16 and 8.6, the error value ε_n will then have a TPDF and the variance will hence be $\frac{\Delta^2}{6}$. Therefore the results obtained by integration for the expected error variance above is consistent with the prediction from our preceding theory.

The discrete PDF for the steady state error value ε_n may be obtained straightforwardly from the orbit sets for u_n. If the orbit set $\{u_{(i)}\}$ has n_u elements, then the steady state PDF for u_n will be a sequence of n_u equally spaced point masses about $\mathcal{C}$, each with probability value $\frac{1}{n_u}$. The positions of these point masses $\{u_{(i)}\}$ are given specifically by the orbit set descriptions for cases (a) and (b) presented at the beginning of this subsection. The steady state PDF for ε_n will then be a sequence of generally $2n_u$ point masses that are equally spaced along the interval of length 2Δ centred on 0. From consideration of the RPDF dither, and the quantizer Q in the system topology, these point masses will be located at the points $+\frac{\Delta}{2} - u_{(i)}$ and $-\frac{\Delta}{2} - u_{(i)}$, for $i = 1, \ldots, n_u$, from the orbit set $\{u_{(i)}\}$. The corresponding probability values will be $\frac{1}{2} + \frac{u_{(i)}}{\Delta}$ and $\frac{1}{2} - \frac{u_{(i)}}{\Delta}$ respectively, for $i = 1, \ldots, n_u$. From a graphical examination, it can be seen that the point masses of this steady state ε_n PDF trace out an envelope that corresponds to the triangular TPDF with base length 2Δ.

We now consider the situation where we allow s to vary. From the formulas in 3 for the variance maximum and minimum, it is clear that as the number of points n_u in the orbit set of u_n and ε_n on $\mathcal{C}$ increases, and hence s increases, the deviation of these extremum values for $E[\varepsilon_n^2]$ from the value of $\frac{\Delta^2}{6}$ will decrease. Thus the parabolas in the graph of $E[\varepsilon_n^2]$ as a function of ε_{-1} over $\mathcal{C}$ will increase in number, but shrink in vertical length about $\frac{\Delta^2}{6}$ as s increases. The value of s will increase if and only if the denominator q in the representation $c = \frac{p}{q}\Delta$, of the "rational multiple of Δ" input c, increases. This condition corresponds to a rational input c that is becoming more similar in character to an irrational number. Taking this further, let A_q be the set of all rational numbers in

$\mathbb{R}$ with denominator q in its coprime integer fractional representation. Then we expect $\lim_{q\to\infty} A_q$ to be the set of all irrational numbers in $\mathbb{R}$. Thus we expect the limiting case in our variance results, as $q \to \infty$, to correspond to the realization of an irrational value for $\frac{c}{\Delta}$.

In this limiting case, we see from our results above that $k \to \infty$, $\tilde{\varepsilon}_{-1} \to 0$ and $E[\varepsilon_n^2] \to \frac{\Delta^2}{6}$, for all ε_{-1}. The point masses in the steady state PDF for ε_n will also fill up the TPDF envelope, yielding a convergence of this discrete PDF to the continuous TPDF. From Theorem 7.13 and Proposition 8.6, these error variance and PDF results correspond to what is predicted when the input c is an irrational multiple of Δ. Conversely, we may use these limiting properties to give essentially a proof of Theorem 7.13, under the specific conditions of RPDF dither, as follows:

Alternative Proof of Theorem 7.13:

Suppose the system (8.6) is as specified in Theorem 7.13, with input c an irrational multiple of Δ, and, without loss of generality, $\varepsilon_{-1} \in \mathcal{C}$. Let $c_{(n)}$ be a sequence of rational numbers satisfying $c_{(n)} = \frac{e_n}{2^n}$, where $e_n = \{x \in \mathbb{Z} \mid |\frac{x}{2^n} - \frac{c}{\Delta}| \leq |\frac{y}{2^n} - \frac{c}{\Delta}|, \forall\, y \in \mathbb{Z}\}$. Then we have $|c_{(m)} - \frac{c}{\Delta}| < \frac{1}{2^n}$, $\forall\, m \geq n$, and for any $n \geq 1$. Thus the sequence $c_{(n)}$ converges to $\frac{c}{\Delta}$ as $n \to \infty$. From our results above, we have that, for the system (8.6) with input $c_{(n)}$ and RPDF dither, the steady state error value PDF converges to the TPDF as $n \to \infty$. Thus the steady state error value PDF for the system with the given input c and dither will be TPDF and hence uniformly distributed over $\mathcal{C}$. ■

Clearly such a proof can be extended to more general dither cases when a corresponding analysis, showing the expected analogous asymptotic error behaviour for the rational $\frac{c}{\Delta}$, has been established. Thus we have shown how the analysis from our rational input model

yields results corresponding to the irrational input case, in the limit as the rational input approaches an irrational form. In the opposite extreme, we see that an input that is an integer multiple of $\frac{\Delta}{2}$ will yield the maximum deviations of the error variance from $\frac{\Delta^2}{6}$. For this case, if the initial condition ε_{-1} is 0, the variance will be the minimum of 0, and if ε_{-1} is $\frac{\Delta}{2}$, it will be the maximum of $\frac{\Delta^2}{4}$. Note that the shrinking of the variance deviation to zero does not occur uniformly as the value of the denominator q in the rational factor of the input increases — the deviation oscillates within uniformly increasing envelope lines (e.g. showing why $q = 2$ and not $q = 1$ gives the maximum).

The analysis of this subsection proves that the first-order Σ-Δ modulator with unity gain, RPDF dither, and a constant input that is a rational multiple of Δ has a steady state error variance $E[\varepsilon_n^2]$ that is generally not $\frac{\Delta^2}{6}$ (holds only for a countable number of initial conditions). Thus the claim of Reiss in [52] mentioned in Section 1.6 that this variance must always be $\frac{\Delta^2}{6}$ with any constant input is proved incorrect. We have proven that his claim applies only when the input is a constant irrational multiple of Δ. Our results then, more broadly, serve to demonstrate the limits of assuming some input or initial condition independence of the steady state error variance under RPDF dither, as suggested from Proposition 8.6 and discussed in Section 8.2. For the first-order model here, this variance is not constant over the set inputs x_n that are constant, rational multiples of Δ, and varies continuously with the initial condition ε_{-1} as well.

Nevertheless, it is worth summarizing the results for error control that we have established for the first-order model of this section. From Proposition 8.6, we have the constant steady state error variance of $\frac{\Delta^2}{6}$ holding over the set of all inputs x_n that are a constant irrational multiple of Δ, and all initial conditions $\varepsilon_{-1} \in \mathcal{C}$. In addition, equation 2 provides explicit conditions under which the steady state error variance can be made low, via the choice of initial condition and input that is a constant rational multiple of Δ. The lowest

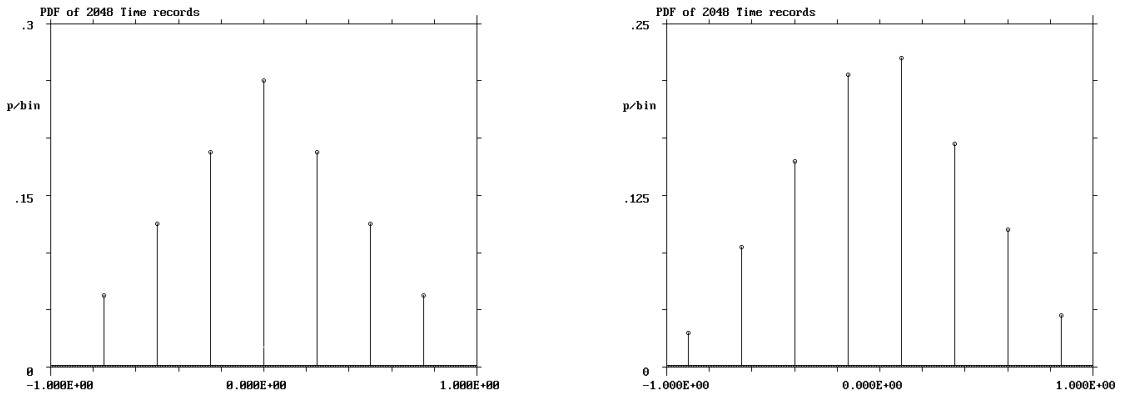
variance values are possible for the lowest values of s.

The methods and approach of this subsection could be similarly applied to higher order systems satisfying (R) with no nonminimum-phase zeros, to give a steady state error variance description. Clearly many classes of such systems will fail to satisfy Theorem 7.14 and Proposition 8.6, and will have steady state error variances that are not generally $\frac{\Delta^2}{6}$ under RPDF dither. Extending the analysis to systems where (R) does not hold would appear to be not much more tractable, however, than for the theoretical studies undertaken previously to establish a uniform steady state error, for example.

8.3.2 Simulations

A number of simulations were carried out in double-precision floating-point arithmetic on a computer to test the viability of the above theory for predicting error variance and PDF results, for the first-order Σ-Δ modulator with unity gain and constant input of (8.6) subjected to an RPDF dither. From the discussion of time series analysis in Section 7.3, it is expected that the asymptotic long run behaviour of any given realization via simulation, for this specific system, will converge to the theoretical steady state form. The simulations were hence conducted and considered with this predictive property in mind.

The noise shaper simulation computer program that was employed, with a given initial condition ε_{-1}, runs a large number of iterations of the difference equations of the dithered system (1.2), and then calculates statistical properties such as error standard deviation and error PDF plots (as appropriately scaled histograms), based on a sample estimation of these properties conducted over the large number of iterations run. Specifically, the program conducts an integer number of "runs", where each run corresponds to $2^{14} = 16384$ iterations or "samples". We present here results from four simulations using a reasonable rational input, and a series of simulation results where a near irrational form of rational input was



(a) (b)

Figure 8.3: Simulation histograms for ε_n with $c = \dfrac{1}{4}$ and (a) $\varepsilon_{-1} = 0$, (b) $\varepsilon_{-1} = 0.1$

used. Note that $\Delta = 1$ here, and that $\dfrac{1}{\sqrt{6}} \approx 0.40824829$. The numerical specifications, predictions, results, and differences between the two are summarized in Tables 8.1 and 8.2 near the end of this subsection.

For the first two simulations, a constant input of $c = \dfrac{1}{4}$ was used. The first case used an initial condition of $\varepsilon_{-1} = 0$, and the second case an initial condition of $\varepsilon_{-1} = 0.1$. For both cases, the simulation went through $2^{11} = 2048$ runs. This gives 33,554,432 ($\approx 3.4 \times 10^7$) samples in the average. For the first case, the error variance formula in 2.(a) predicts a standard deviation (square root of the variance) that is 1.296358×10^{-2} below the base value of $\dfrac{1}{\sqrt{6}}$. The simulation result reports a value that is 1.292263×10^{-2} below $\dfrac{1}{\sqrt{6}}$. The net difference between these two results is 4.096×10^{-5}. In the second case, the error variance formula 2.(a) predicts a standard deviation that is 5.57534×10^{-3} above $\dfrac{1}{\sqrt{6}}$. The simulation result gives a value that is 5.55299×10^{-3} above $\dfrac{1}{\sqrt{6}}$. The net difference between these results is 2.236×10^{-5}. The simulation plots of the PDFs for ε_n in these cases are given, with the $\varepsilon_{-1} = 0$ case in Figure 8.3(a), and the $\varepsilon_{-1} = 0.1$ case in Figure 8.3(b).

```
-------Noise Shaper Simulation------
S)    SamplingFreq 2822400
q)    ADC type     mid-riser Hgn: 1.267678  DCgn: .9998274  Ngn: 1.008873
u)    Q_sat        32767    Q_in(rms): .5631157
v)    Insum_sat    32767    Q_in(avg): .2500432
b)    Bits @ H_out
p)    Prerun smpls 0
F)    Sig1 f, bin# 0  ampl 0 c
h)    Sig2 on bin# 0  ampl 0 s  ACgn2: 0
r)    Ramp rate    0
l)    RPDFinputsig 0
o)    DC offset    .25
M)    Idle bin#    4096
i)    Initial e1#  .1

d)    RPDFinptdith 0

n)    #Rands @ Qin 1  Level(p-p) 1
f)    Q_Dith filt  10000
s)    Bin dither   0  rate 0
R)    Re-seed RANDc2  dith transfrac: .1800127
m)    Meter point: H_in  Avg:-6.818175322077276D-05  rms: .4138012770692862
             max: .85  min:-.9  transfrac: .6224507
A)    Import coeffs   File: Pas1Z
B)    Export coeffs   1bitover: 2473440  1bitunder: 125836
H)    Display and set coeffs
0)    Zero all coeffs

x,c)  Execute anew or continue?     Run#: 2048
```

Figure 8.4: Simulator text output corresponding to Figure 8.3(b) case

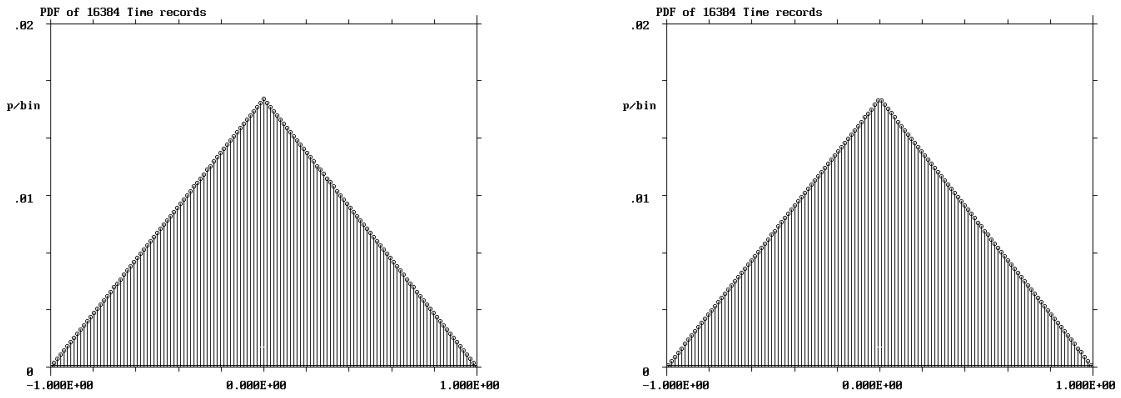
These plots display the predicted point mass PDFs tracing out the TPDF envelope. The position and length of the bars are as given from our expressions, with the bars from Figure 8.3(a) shifted to the right by the initial condition shift of 0.1 in Figure 8.3(b).

Figure 8.4 shows the text output of the simulator corresponding to the second case, and Figure 8.3(b) mentioned above. The information contained includes the simulator set up and the numerical output. We note the following specific lines. The first part of line (q) specifies the quantizer as the mid-riser type. Line (o) sets the value of the constant or DC input c. Line (i) sets the value of the initial condition ε_{-1}. The first part of line (n) specifies that 1 number is to be sampled from a random number generator and then added to the quantizer input at the entry point to the quantizer Q. The second part of (n) defines the PDF for the random number generator as 1 convolution of a uniform PDF of width $\Delta = 1$ ("Level peak to peak"), that is the RPDF itself. Thus line (n) specifies the dither ν_n of the

circuit. Line (m) reports calculated statistics from measurements at the entry point to the filter H, which hence pertain to the error ε_n. The average value of ε_n is reported, followed by its root mean square or standard deviation value, and then other statistics. Line (A) sets the filter coefficients. "Pas1Z" refers to the algebraic form corresponding to the first order noise shaper ($a_1 = 1$, $M = 1$, $N = 0$). Line (x,c) reports the number of runs of the simulator at the time of the given text output, and the associated statistics contained therein.

For the next two simulations, a constant input of $c = \dfrac{17}{64}$ was used. The first case used an initial condition of $\varepsilon_{-1} = 0$, and the second case an initial condition of $\varepsilon_{-1} = \dfrac{1}{128}$. For both cases, the simulation went through $2^{14} = 16384$ runs. This gives 268,435,456 ($\approx 2.7 \times 10^8$) samples in the average. For the first case, the error variance formula in 2.(a) predicts a standard deviation of 4.984×10^{-5} below $\dfrac{1}{\sqrt{6}}$. The simulation result reports a value that is 6.351×10^{-5} below $\dfrac{1}{\sqrt{6}}$. The net difference between these two results is 1.367×10^{-5}. In the second case, the error variance formula 2.(a) predicts a standard deviation that is 2.492×10^{-5} above $\dfrac{1}{\sqrt{6}}$. The simulation result gives a value that is 3.516×10^{-5} above $\dfrac{1}{\sqrt{6}}$. The net difference between these results is 1.024×10^{-5}. The simulation plots of the PDFs for ε_n in these cases are given, with the $\varepsilon_{-1} = 0$ case in Figure 8.5(a), and the $\varepsilon_{-1} = \dfrac{1}{128}$ case in Figure 8.5(b). These plots again display the predicted point masses tracing the TPDF envelope from the theory, with about $4k$ such bars. The bars in Figure 8.5(b) are those in Figure 8.5(a) shifted to the right by the change in initial condition value.

The net difference in the predicted and observed error standard deviation is of the order of 10^{-5} in the four cases. This provides rather good proportional agreement in the first two cases where the variance deviations are larger. It would seem clear, given the very large number of iteration samples performed by the simulation program, that this net difference represents the maximum precision that can be achieved in attaining theoretical results, due



(a) (b)

Figure 8.5: Simulation histograms for ε_n with $c = \dfrac{17}{64}$ and (a) $\varepsilon_{-1} = 0$, (b) $\varepsilon_{-1} = \dfrac{1}{128}$

to round off error and any other numerical method limitations of the computer.

For the fifth simulation, a constant input of $c = \dfrac{1}{4} + 2^{-18} = \dfrac{65537}{262144}$ and initial condition $\varepsilon_{-1} = 0$ was used. This input has a q value of 2^{18}, and was chosen so as to provide as close an approximation to an irrational input as is possible to represent numerically on the computer. The simulation here went through 2^{11} runs. The simulation result gave a standard deviation value that was 2.309×10^{-5} below the prediction of $\dfrac{1}{\sqrt{6}}$. Thus the net difference from the irrational input result of $\dfrac{1}{\sqrt{6}}$ is of the same order of magnitude as the respective net differences in the first four cases. The simulation plot of the PDF for ε_{-1} in Figure 8.6 shows basically as dense a configuration of bars as is possible to plot, that trace out the TPDF, as expected. Figure 8.7 gives the text output of the simulator for the fifth simulation.

The final sequence of nine simulations used the same input and initial condition, but only one to eight and then 33 runs. The simulation plots of the histogram PDFs for ε_n in Figure 8.8 show an interesting pattern. For one run, the probability bars form periodic

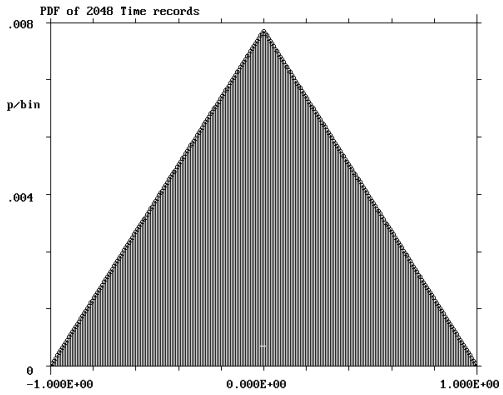
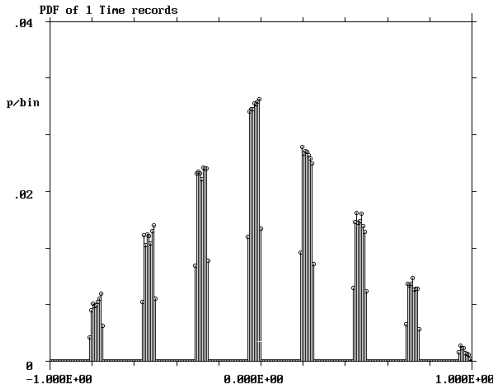


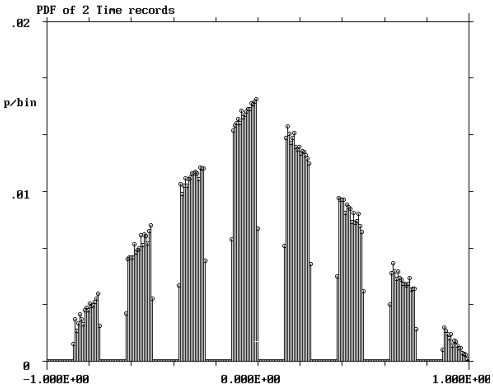
Figure 8.6: Simulation histogram for ε_n with $c = \frac{1}{4} + 2^{-18}$ and $\varepsilon_{-1} = 0$

```
-------Noise Shaper Simulation------
S)    SamplingFreq 2822400
q)    ADC type     mid-riser Hgn: 1.265948  DCgn: 1.00004  Ngn: .9999927
u)    Q_sat        32767    Q_in(rms): .5589705
v)    Insum_sat    32767    Q_in(avg): .2499938
b)    Bits @ H_out
p)    Prerun smpls 0
F)    Sig1 f, bin# 0  ampl 0 c
h)    Sig2 on bin# 0  ampl 0 s  ACgn2: 0
r)    Ramp rate    0
l)    RPDFinputsig 0
o)    DC offset    .2500038146972656
M)    Idle bin#    4095.9375
i)    Initial e1#  0

d)    RPDFinptdith 0

n)    #Rands @ Qin 1  Level(p-p) 1
f)    Q_Dith filt  10000
s)    Bin dither   0  rate 0
R)    Re-seed RANDc2  dith transfrac: .1822852
m)    Meter point: H_in  Avg:-1.5020370483398444D-05  rms: .4082252054115095
            max: .9998283  min:-.9998398  transfrac: .6183935
A)    Import coeffs   File: Pas12
B)    Export coeffs   1bitover: 2358306  1bitunder: 87407
H)    Display and set coeffs
0)    Zero all coeffs

x,c)  Execute anew or continue?     Run#: 2048
```

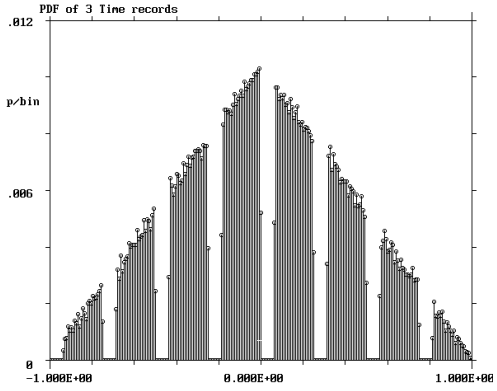
Figure 8.7: Simulator text output corresponding to Figure 8.6 case



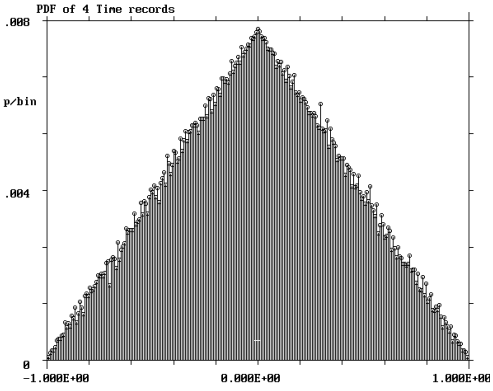
(a) 1 run



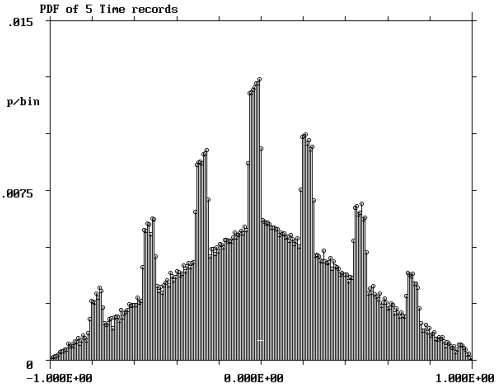
(b) 2 runs



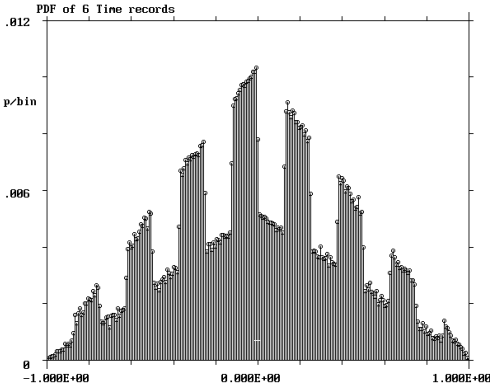
(c) 3 runs



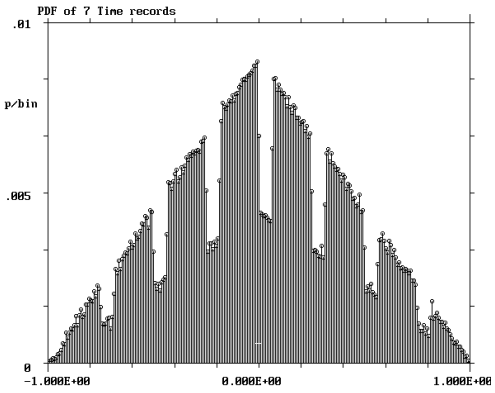
(d) 4 runs



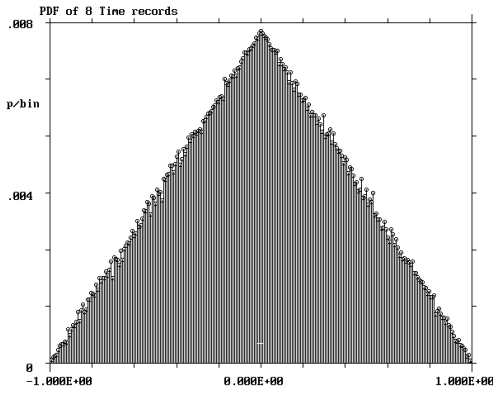
(e) 5 runs



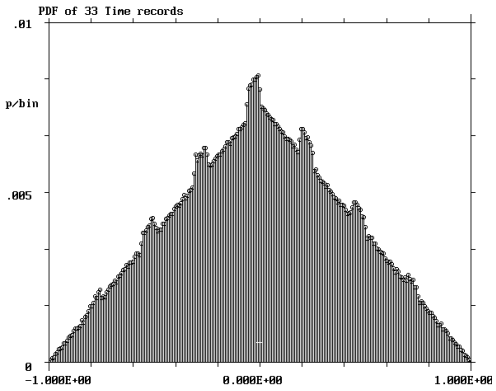
(f) 6 runs



(g) 7 runs



(h) 8 runs



(i) 33 runs

Figure 8.8: Simulation histograms for ε_n with $c = \frac{1}{4} + 2^{-18}$, $\varepsilon_{-1} = 0$ and 1 to 8, 33 runs

```
-------Noise Shaper Simulation------
S)    SamplingFreq 2822400
q)    ADC type     mid-riser Hgn: 1.2663  DCgn: 1.0003  Ngn: 1.001937
u)    Q_sat        32767    Q_in(rms): .560278
v)    Insum_sat    32767    Q_in(avg): .2499403
b)    Bits @ H_out
p)    Prerun smpls 0
F)    Sig1 f, bin# 0  ampl 0 c
h)    Sig2 on bin# 0  ampl 0 s  ACgn2: 0
r)    Ramp rate    0
l)    RPDFinputsig 0
o)    DC offset    .2500038146972656
M)    Idle bin#    4095.9375
i)    Initial e1#  0

d)    RPDFinptdith 0

n)    #Rands @ Qin 1  Level(p-p) 1
f)    Q_Dith filt  10000
s)    Bin dither   0  rate 0
R)    Re-seed RANDc2  dith transfrac: .181366
m)    Meter point: H_in  Avg: 3.032684326171875D-04  rms: .4088708660852532
             max: .9942627  min:-.9971199  transfrac: .6187439
A)    Import coeffs   File: Pas1Z
B)    Export coeffs   1bitover: 4664  1bitunder: 192
H)    Display and set coeffs
0)    Zero all coeffs

x,c)  Execute anew or continue?     Run#: 4
```

Figure 8.9: Simulator text output corresponding to Figure 8.8(d) case

clusters touching the TPDF envelope. As the number of runs increase to four, the clusters grow and fill up the envelope. Proceeding from five to eight runs, this cycle repeats, but with the gaps between clusters partly filled in by a shallower triangular density, the remnant of the first four runs. Even at 33 runs, this phenomenon shows some persistence. These patterns reflect the dynamics of the near quasiperiodic nature of the system with near irrational input. The histograms show short run behaviour that is similar to the finite orbit set behaviour of the rational input case, while displaying longer run behaviour that is increasingly suggestive of the dense orbits in the irrational input case. We must wait for the very long run for the simulations to show unarguable convergence of the PDF to a TPDF. Figure 8.9 provides the text output of the simulator after 4 runs.

Simulations were also carried out of first-order systems where the gain value a_1 was changed from 1 to various larger positive integers. The input, initial condition and number

Simulation Cases and Predicted Standard Deviation					
Case	Plot Figure	c	I.C.	runs	predicted $\sqrt{E[\varepsilon^2]}$
1	8.3(a)	$\frac{1}{4}$	0	2^{11}	$\frac{1}{\sqrt{6}} - 1.296358 \times 10^{-2}$
2	8.3(b)	$\frac{1}{4}$	0.1	2^{11}	$\frac{1}{\sqrt{6}} + 5.57534 \times 10^{-3}$
3	8.5(a)	$\frac{17}{64}$	0	2^{14}	$\frac{1}{\sqrt{6}} - 4.984 \times 10^{-5}$
4	8.5(b)	$\frac{17}{64}$	$\frac{1}{128}$	2^{14}	$\frac{1}{\sqrt{6}} + 2.492 \times 10^{-5}$
5	8.6	$\frac{1}{4} + 2^{-18}$	0	2^{11}	$\approx \frac{1}{\sqrt{6}}$
6 - 14	8.8	$\frac{1}{4} + 2^{-18}$	0	1 - 8, 33	

Table 8.1: Simulation standard deviation predictions

Numerically Estimated Standard Deviation		
Case	estimated $\sqrt{E[\varepsilon^2]}$	|difference|
1	$\frac{1}{\sqrt{6}} - 1.292263 \times 10^{-2}$	4.096×10^{-5}
2	$\frac{1}{\sqrt{6}} + 5.55299 \times 10^{-3}$	2.236×10^{-5}
3	$\frac{1}{\sqrt{6}} - 6.351 \times 10^{-5}$	1.367×10^{-5}
4	$\frac{1}{\sqrt{6}} + 3.516 \times 10^{-5}$	1.024×10^{-5}
5	$\frac{1}{\sqrt{6}} - 2.309 \times 10^{-5}$	$\approx 2.309 \times 10^{-5}$

Table 8.2: Simulation standard deviation results

of runs were as in the fifth simulation above. The standard deviation results generally had net differences from the predicted $\frac{1}{\sqrt{6}}$ value that were of the same order of magnitude as for the previous simulations described. The PDF plots were also dense and traced out the TPDF as predicted from theory. In some cases these triangles showed slight perturbations from a smooth form. This effect was not expected, and perhaps can be attributed to numerical limitations of the computer.

In closing this subsection, we mention the recent paper by Løkken et al. [39], which investigates some noise power issues for Σ-Δ modulators. This work included considering the same first order model form studied in this section, but with a mid-tread quantizer (i.e. quantizer levels at integer multiples of Δ) instead of the mid-riser type used in this thesis. Numerical simulations and analysis of the average error variance were conducted and are presented in the paper. These results showed full agreement with the predictions given from our mid-tread error variance formula presented in the footnote of Subsection 8.3.1. This formula was derived as a simple extension of the corresponding mid-riser formulations, and shows virtually an equivalent (and indeed simpler) description. Note that the subsequent analysis that follows for the mid-riser case in Subsection 8.3.1, will apply analogously to the mid-tread case.

8.4 Discussion

Rational/Irrational Input Implications:

The practical distinction between the rational and irrational input in cases such as those given with the preceding examples is now briefly addressed.

Gray [17] emphasizes the idea that rational inputs are of little interest because they occur with zero probability. The rational/irrational properties of the input are normally

of practical interest and concern when assessing the limit cycle or quasiperiodic behaviour of the Σ-Δ modulator errors they give rise to. To distinguish irrational input from generic cases of rational input in this manner, however, one has to follow an arbitrarily large number of iterations to verify that the errors behave quasiperiodically (implying irrational input) and not periodically with a large period (implying rational input with a typically large denominator).

The practical effects of the errors on the reconstructed audio output are normally governed by the error behaviour over shorter iteration lengths. This is because periodic behaviour with a shorter period corresponds to audio frequencies that are high enough for detection. Furthermore, over these shorter iteration lengths, the irrational input/quasiperiodic, (or rational input/large period) error behaviour will tend to closely approximate the periodic behaviour associated with a nearby rational input with a shorter period. For example, an irrational or rational input very close to zero would yield error behaviour that approximates the period two oscillations with zero input. Therefore, from the point of view of audible detection, the existence of an irrational input generally cannot be distinguished from that of a rational one, and neither can generally be distinguished from that of a rational one associated with a short period.

In a digital Σ-Δ modulator corresponding to a computer simulation, all inputs are rational by default, and so the rational case occurs with probability one. From the practical consideration above, rational inputs with a large denominator may be taken to function equivalently to the irrational inputs they are chosen to approximate, as well.

The situations above counter some of the importance of Gray's emphasis, and suggest that it may be most appropriate to interpret or analyze the modulator behaviour in terms of a rational input x_n. Future work could then involve an extension of the analysis of statistical error properties in this chapter and related issues, from this practically motivated

perspective.

Summary:

In summary, in Chapter 7 we developed a new statistical theoretical approach to handle the concept of the steady state error. Relevant theorems and results were then derived, with direct applicability to the error variance characterization via the uniform PDF property on $\mathcal{C}$. We began Chapter 8 by applying the overall dynamical systems approach of this thesis to dither theory for the Σ-Δ modulator, and the establishment of standard and relevant results. We then explored the issues of the statistics of $E[\varepsilon_n^2]$ by applying the results of Chapter 7. A simple class of examples was then considered, for which simple error variance formulas were developed, analyzed and then successfully tested by simulation.

Chapter 9

Conclusions

This thesis represents an attempt at gaining a broader, deeper and more thorough understanding of important dynamical behaviour in the Σ-Δ modulator system. A dynamical systems approach was taken to initiate a theoretical investigation of chaotic properties of the dynamics, and of statistical dynamical properties in the context of dithered systems. The research was aimed at seeking to resolve previous examinations of chaos, and at extending this philosophy to treat relevant dither issues. The dynamical systems approach was viewed as conceptually important as well as analytically useful, since many diverse physical systems are topologically and hence dynamically equivalent or comparable to Σ-Δ modulator systems under a formal dynamical systems interpretation. The following have been accomplished in the work of this thesis:

The Mth/Nth order feedback/feedforward multi-bit Σ-Δ modulator was modelled as a dynamical system, using the circle map interpretation for the state space of the shaped and dithered quantizer errors ε_n. A general continuity condition on the filter coefficients was established, and continuity properties described. An analysis of conditions on the feedforward filter coefficients and system errors, to establish bounded internal stability

or nonstability, was conducted and some results were obtained, which give an expanded conceptual picture. These results, in particular, concerned extending stability conditions to systems with noise transfer function poles on, as well as inside the unit circle. A viable adaptation of the Devaney definition of chaos was applied to this model. An analysis of conditions on the filter coefficients and system input to satisfy the Devaney conditions for chaos or nonchaos was conducted and extended to the dithered case.

In this analysis, methods and approaches were developed for the verification or rejection of strong chaos conditions for a general mathematical model arising from a specific practical application. As such, the concepts and tools available to study or interpret chaos have been strengthened, both for abstract or theoretical pursuits, and for potential application to other practical systems that can be modelled. Such methods using topological symmetries and model linearities were applied throughout the thesis. Some flexibility arose in the precise manner in which Devaney's conditions for chaos were applied to the model. The approach and choices taken for the adaptation and analysis here are not meant to be the final word, but rather a reasonable first step or prototype for future theoretical studies of chaos in Σ-Δ modulator-like dynamical systems.

The results of the analysis showed that chaos will hold if all the zeros of the noise transfer function lie outside the circle of radius two, provided the input is either periodic or is persistently random modulo Δ. Bounded stability was shown to be readily attainable with such chaos. If the filter coefficients satisfy the continuity condition, then the zeros condition may be relaxed to that of a circle of radius one, and conditions on the input dropped to give chaos. Bounded stability was shown to hold with this chaos, generally if and only if the filter is strictly of feedback form, when the input is periodic. Under the continuity condition, chaos was shown not to hold when all zeros are on the unit circle. In general, chaos is not prevalent for minimum/marginally minimum-phase systems. These results generally extend

to the analogous dithered systems, extending exactly when the continuity condition holds. Other more specific results or examples for particular chaos conditions were determined. These results serve to differentiate important aspects of the dynamical behaviour given by the chaos condition characterizations, (as determined by filter coefficients and input), and may have practical implications in terms of limit cycle behaviour, types of stability or other issues of importance in Σ-Δ modulator performance. The effects of adopting stricter definitions for density of periodic points on the characterizations made of chaos was also examined.

Important error moment and dithered quantizer formulas, relationships and statistical distributions were derived using state space methods. A theoretical approach for describing steady state error distributions and PDF mappings was developed and applied to the Σ-Δ modulator, under certain modulator or dynamic conditions, to obtain insight and conclusions about the steady state error distributions.

The results showed that a generic dense steady state error distribution on the circle will be uniform if one zero of the noise transfer function is nonminimum phase, the input is periodic, and the continuity condition holds. With two such zeros, the errors are also white. Extensions to M dimensions, and general random initial conditions/input were explored. A uniform steady state was also shown if the system is first order with unity gain and irrational constant input (quasiperiodic), or for any system with input that is random, i.i.d. and uniform. Extensions to RPDF dithered systems were made. These showed that a constant average variance of $\Delta^2/6$ followed when most of the uniformity results held, hence extending a measure of input independence from TPDF to RPDF dithered systems. The variance behaviour of the first-order system under general constant input with RPDF dither was analyzed in more detail, and simulations were conducted to support the results.

The bridge between the theory of stochastic processes and the practical analysis of

dithered Σ-Δ modulator behaviour has been broadened by the methods and approaches developed here. The results serve to enhance a general understanding of performance properties of such dithered systems.

In general, this thesis represents the dynamical study of a practical system whose behaviour is strongly characterized by stochastic or stochastic-like attributes. As such, the approach taken has been to often integrate (e.g. in these terms) the work of the separate aspects of the studies (in respective chapters) involved — generally by virtue of the consistent dynamical systems treatment. Subsequent chapters tend to build results successively upon the work of previous ones, and endeavours are made to synthesize earlier treatments, ideas or discussions into later ones. Common threads emerge, as we have from the observation that systems of sufficient structure (e.g. input or filter form) are amenable to satisfying theorems that assert structure in the dynamics. Therefore the strategy and structure of this thesis serve as a role model or philosophy for advancing the dynamical studies further, or for studying other systems whose dynamical aspects possess subtle similarities, and where the accumulation of analytical approaches may be required. Broader implications may be derived in related areas of dynamical systems or signal processing as well. This then serves as an exemplification of the process of applied mathematics at work as theoretical engineering.

9.1 Recommendations

We make the following recommendations for future research, which emerge from the work of this thesis.

Rigorous analyses of conditions for chaos in the higher-order Σ-Δ modulator analogous to those done here could be conducted using other definitions for chaos, such as Lyapunov

exponents or rotational chaos, and then compared with the Devaney results to solidify the conclusions reached here. Further investigation into mathematical methods and approaches could be made in an effort to arrive at more definitive or clear cut theorems and conditions on the existence of Devaney chaos or nonchaos, or particular chaos conditions. This could particularly be so in the case where the continuity condition (R) does not hold, especially regarding density of periodic points. These efforts could be extended to seek more definitive results on how the addition of dither may bring about more chaos, or chaos conditions (e.g. small dither for density of periodic points).

Flexibility in the adaptation of Devaney's chaos conditions to the dynamical model could be exercised to adjust the definitions used here for the analysis. Results from the same analysis with adjustments could then be compared with the results here to provide greater insight into what aspects of the chaos conclusions may be truly adaptation dependent. This could thus aid in future chaos condition formulations to give both the most theoretically and practically meaningful conclusions. In particular, an investigation to classify chaos using stricter definitions for density of periodic points (in the standard manner) could be conducted to give a more comprehensive comparison, or notions of "projected" chaos, considered only on $\mathcal{C}^M$, could be altered with justifications. Overall chaos could be characterized as holding when the required conditions are shown to hold over a subspace or submanifold of the error coordinate state space, or over a pertinent subset of state space, rather than the entire space. This would open up greater classes of cases to satisfy chaos, or foster a deeper analysis of cases that were not otherwise fully chaotic. As a study more separate from that of the Σ-Δ modulator, an investigation into the complex dynamics of such discrete affine modulo mappings that possess chaotic attractors or subsets/submanifolds could be undertaken. Adjustments to the Devaney adaptations could be extended to the dithered system, focusing particularly on the state space definition, where more flexibility

arose in the characterization. The chaos analysis conducted in the thesis could be extended or generalized to a study of systems with stochastic initial conditions.

The switching system formulation of the dynamical system Σ-Δ modulator model could be explored as an alternative approach to study the dynamics. Existing theory on the chaotic properties and stability of deterministic or stochastic discrete subsystem switching systems with infinite input modes could be used, and new theory developed, to apply to studies of chaos or stability in the dithered or nondithered Σ-Δ modulator. More definitive results on bounded stability, or an extension for results on asymptotic forms of stability, could be sought through an investigation of this or other modelling approaches.

The development of the steady state random variable approach could be broadened or made more rigorous, and could hence be linked or incorporated within standard stochastic processes theory. More specifically, the axiomatic theory of probability could be utilized as a theoretical tool for moving the analysis beyond the limits of the more intuitive methods used in this thesis. Further mathematical methods and approaches to apply this theory to dithered Σ-Δ modulators could thus be studied, with the aim of acquiring more clear cut or expanded results about steady state error behaviour, the role of uniformity, and more theoretical relationships involving the dynamics. This could focus, in particular, on nondithered systems when (R) does not hold. Further investigation, with dither added, could then proceed as a separate problem. Conversely, adding dither could be viewed as simplifying the analysis in one dimension when (R) does not hold. This could be an approach for arriving at results, as conjectured in Section 8.2, from which generalizations to $\mathcal{C}^M$, with or without dither, could then be sought. The average variance results under RPDF dither could be developed to reflect and apply any newly derived steady state error results. The error variance formulation and analysis methods of the RPDF dithered first-order case with unity gain could be extended to similarly study more general or complex

cases. A more precise and complete explanation of the behaviour of the simulation results in comparison with the theoretical predictions could be sought. The analysis of statistical error properties could be advanced from the practical perspective that irrational Σ-Δ modulator inputs are essentially perceived by the output observer as having the same effect as rational ones.

In future research, the chaos, stability and statistical error property investigations of the dithered or nondithered Σ-Δ modulator of this thesis could be carried out for the finite bit, and in particular, the 1-bit quantizer case, where the assumption of a no overload condition is dropped. A counterpart theory for the 1-bit quantizer and noise shaper to the existing multi-bit case theory could be pursued in future work. The first approach would be analogous to that of this thesis, in terms of establishing error property statistical results. Efforts could be made at uncovering further pseudorandom number generator properties from Σ-Δ modulators satisfying steady state error distribution theorems arising from future research. Seeking appropriate ways of analyzing dithered Σ-Δ modulators with the aim of obtaining results that can be applied to the study of stochastic resonance system in physics, to obtain insight into their dynamical behaviour, could be the topic of future research.

Bibliography

[1] Bendat J. S., Piersol A. G. 1971. *Random Data: Analysis and Measurement Procedures.* New York N.Y.: Wiley-Interscience.

[2] Benzi R., Sutera A. and Vulpiani A. 1981. The mechanism of stochastic resonance. *J. Phys. A* **14**, L453-L457.

[3] Benzi R., Parisi G., Sutera A. and Vulpiani A. 1982. Stochastic resonance in climate change. *Tellus* **34**, 10-16.

[4] Candy J. C. 1997. An overview of basic concepts. *Delta-Sigma Data Converters, Theory, Design, and Simulation, edited by Norsworthy S. R., Schreier R., Temes G. C.*, 1-43. Piscataway N.J.: IEEE Press.

[5] Casdagli M. 1988. Rotational chaos in dissipative systems. *Physica* **29D**, 365-386.

[6] Cutler C. C. 1960. Transmission system employing quantization. *U.S. Patent No. 2,927,962.*

[7] Devaney R. L. 1989. *An Introduction to Chaotic Dynamical Systems, Second Edition.* Redwood City Ca.: Addison-Wesley.

[8] Dieci L., Van Vleck E. S. 1995. Computation of a few Lyapunov exponents for continuous and discrete dynamical systems. *Applied Numerical Mathematics* **17**, 275-291.

[9] Fauve S. and Heslot F. 1983. Stochastic resonance in a bistable system. *Physics Letters* **97A(1,2)**, 5-7.

[10] Feely O., Chua L. O. 1992. Nonlinear dynamics of a class of analog-to-digital converters. *International Journal of Bifurcation and Chaos* **2(2)**, 325-340.

[11] Friedman V. 1988. The structure of the limit cycles in sigma delta modulation. *IEEE Transactions on Communications* **36(8)**, 972-979.

[12] Gambaudo J. M., Glendinning P. and Tresser C. 1984. The rotation interval as a computable measure of chaos. *Physics Letters* **105A(3)**, 97-100.

[13] Gammaitoni L. 1995. Stochastic resonance and the dithering effect in threshold physical systems. *Physical Review E* **52(5)**, 4691-4698.

[14] Gammaitoni L., Hänggi P., Jung P., Marchesoni F. 1998. Stochastic resonance. *Reviews of Modern Physics* **70(1)**, 223-287.

[15] Garver W. and Moss F. 1995. Detecting signals with noise. *Scientific American* **273(2)**, 100-103.

[16] Gray R. M. 1989. Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input. *IEEE Transactions on Communications* **37(6)**, 588-599.

[17] Gray R. M. 1997. Quantization noise in delta-sigma a/d converters. *Delta-Sigma Data Converters, Theory, Design, and Simulation, edited by Norsworthy S. R., Schreier R., Temes G. C.*, 44-74. Piscataway N.J.: IEEE Press.

[18] Grimmett G. R., Stirzaker D. R. 1992. *Probability and Random Processes, Second Edition*. Oxford U.K.: Oxford University Press.

[19] Güntürk C. S. 2000. Harmonic analysis of two problems in signal quantization and compression. *Ph.D. Thesis, The Faculty of Princeton University (Program in Applied and Computational Mathematics)*, 1-108.

[20] Hao B.-L. 1989. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems*. Singapore: World Scientific.

[21] He N., Kuhlmann F., Buzo A. 1990. Double-loop sigma-delta modulation with dc input. *IEEE Transactions on Communications* **38(4)**, 487-495.

[22] He N., Kuhlmann F., Buzo A. 1992. Multiloop sigma-delta quantization. *IEEE Transactions on Information Theory* **38(3)**, 1015-1028.

[23] Hofbauer F. 1984. Monotonic mod one transformations. *Studia Mathematica, T. LXXX.*, 17-40.

[24] Inose H., Yasuda Y., Murakami J. 1962. A telemetering system by code modulation—delta-sigma modulation. *IRE Trans. Space Electron. Telemetry* **SET-8**, 204-209.

[25] Inose H., Yasuda Y. 1963. A unity bit coding method by negative feedback. *Proc. IEEE* **51**, 1524-1535.

[26] Isabelle S. H., Wornell G. W. 1998. *The Digital Signal Processing Handbook, edited by Madisetti V. K., Williams D. B.*, 72.1-72.13. CRC Press, IEEE Press.

[27] Ito R. 1981. Rotation sets are closed. *Math. Proc. Camb. Phil. Soc.* **89**, 107-111.

[28] Iwersen J. E. 1990. Comment on "the structure of the limit cycles in sigma delta modulation". *IEEE Transactions on Communications* **38(8)**, 1117.

[29] Kalman R. E. 1956. Nonlinear aspects of sampled-data control systems. *Mircrowave Research Institute Proc. Symposium on Nonlinear Circuit Analysis* **VI**, 273-313.

[30] Keener J. 1980. Chaotic behavior in piecewise continuous difference equations. *Trans. AMS* **261**, 589-604.

[31] Knuth D. E. 1998. *The Art of Computer Programming, Volume 2, Seminumerical Algorithms, Third Edition*. Addison Wesley Longman.

[32] Langford W. F. 2000. *Course Notes: An Introduction to Deterministic Chaos*.

[33] Li T. Y. and Yorke J. A. 1975. Period three implies chaos. *Am. Math. Monthly* **82**, 985.

[34] Lipshitz S. P., Vanderkooy J. and Wannamaker R. A. 1991. Minimally audible noise shaping. *J. Audio Eng. Soc.* **39(11)**, 836-852.

[35] Lipshitz S. P., Vanderkooy J. and Wannamaker R. A. 1992. Quantization and dither: a theoretical survey. *J. Audio Eng. Soc.* **40(5)**, 355-375.

[36] Lipshitz S. P., Vanderkooy J. 2001. Towards a better understanding of 1-bit sigma-delta modulators, part 2. *AES Convention Paper 5477*, 1-11.

[37] Lipshitz S. P., Vanderkooy J. 2004. Dither myths and facts. *AES Convention Paper 6279*, 1-16.

[38] Lipshitz S. P., Wannamaker R. A., Vanderkooy J. 2004. Dithered noise shapers and recursive digital filters. *J. Audio Eng. Soc.* **52(11)**, 1124-1141.

[39] Løkken I., Vinje A., Sæther T. 2006. Noise power modulation in dithered and undithered high-order sigma-delta modulators. *J. Audio Eng. Soc.* **54(9)**, 841-854.

[40] MacKay R. S. and Tresser C. 1986. Transition to topological chaos for circle maps. *Physica* **19D**, 206-237.

[41] Marchesoni F., Menichella-Saetta E., Pochini M. and Santucci S. 1988. Analog simulation of underdamped stochastic systems driven by colored noise: spectral densities. *Physical Review A* **37(8)**, 3058-3066.

[42] Mickens R. E. 1990. *Difference Equations, Theory and Applications, Second Edition.* New York N. Y.: Van Nostrand Reinhold.

[43] McNamara B. and Wiesenfeld K. 1989. Theory of stochastic resonance. *Physical Review A* **39(9)**, 4854-4869.

[44] Moss F., Pierson D. and O'Gorman D. 1994. Stochastic resonance: tutorial and update. *International Journal of Bifurcation and Chaos* **4(6)**, 1383-1397.

[45] Moss F. and Wiesenfeld K. 1995. The benefits of background noise. *Scientific American* **273(2)**, 66-69.

[46] Newhouse S. E., Palis J. and Takens F. 1983. Bifurcations and stability of families of diffeomorphisms. *Publ. Math. IHES* **57(5)**, 5-71.

[47] Norsworthy S. R. 1997. Quantization errors and dithering in delta-sigma modulators. *Delta-Sigma Data Converters, Theory, Design, and Simulation, edited by Norsworthy S. R., Schreier R., Temes G. C.*, 75-140. Piscataway N.J.: IEEE Press.

[48] Oppenheim A., Schafer A. 1989. *Discrete-Time Signal Processing.* Englewood Cliffs N.J.: Prentice-Hall.

[49] Reefman D., Reiss J. D., Janssen E., Sandler M. B. 2005. Description of limit cycles in sigma-delta modulators. *IEEE Transactions on Circuits and Systems-I: Regular Papers* **52(6)**, 1211-1223.

[50] Reiss J. D. 2001. The analysis of chaotic time series. *Ph.D. Thesis, The Academic Faculty, Georgia Institute of Technology*, 1-219.

[51] Reiss J. D., Sandler M. B. 2001. The benefits of multibit chaotic sigma delta modulation. *Chaos* **11(2)**, 377-383.

[52] Reiss J. D., Sandler M. B. 2003. Dither and noise modulation in sigma delta modulators. *AES Convention Paper 5935*, 1-5.

[53] Risbo L. 1994. Sigma-delta modulators—stability analysis and optimization. *Ph.D. Thesis, Electronics Institute, Technical University of Denmark*, 1-179.

[54] Risbo L. 1995. On the design of tone-free sigma-delta modulators. *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing* **42(1)**, 52-55.

[55] Ruette S. 2003. Chaos for continuous interval maps—a survey of the relationship between the various sorts of chaos. *Preprint, Laboratoire de Mathématiques Topologie et Dynamique, Université Paris-Sud*, 1-123.

[56] Schreier R. 1991. Noise-shaped coding. *Ph.D. Thesis, Dept. of Electrical Engineering, University of Toronto*, 1-98.

[57] Schreier R., Snelgrove W. M. 1991. Sigma-delta modulation is a mapping. *IEEE Proc. ISCAS, Singapore*, 2415-2418.

[58] Schreier R. 1993. Destabilizing limit cycles in delta-sigma modulators with chaos. *IEEE Proc. ISCAS, Chicago*, 1369-1372.

[59] Schuchman L. 1964. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology* **COM-12**, 162-165.

[60] Shannon C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379-423, 623-656.

[61] Vrscay E. R. 2003. *private correspondence.*

[62] Wang H. 1993. A study of sigma-delta modulations as dynamical systems. *Ph.D. Thesis, Graduate School of Arts and Sciences, Columbia University*, 1-152.

[63] Wannamaker R. A., Lipshitz S. P. 1992. Time domain behaviour of dithered quantizers. *AES Convention Preprint, San Francisco*, 1-33.

[64] Wannamaker R. A., Lipshitz S. P. and Vanderkooy J. 2000. Stochastic resonance as dithering. *Physical Review E* **61(1)**, 233-236.

[65] Wannamaker R. A., Lipshitz S. P., Vanderkooy J. and Wright J. N. 2000. A theory of nonsubtractive dither. *IEEE Transactions on Signal Processing* **48(2)**, 499-516.

[66] Whittaker E. T. 1915. On the functions which are represented by the expansions of the interpolation theory. *Proc. R. Soc. Edinburgh* **35**, 181-194.

[67] Wiesenfeld K. and Moss F. 1995. Stochastic resonance and the benefits of noise: from ice ages to crayfish and SQUIDs. *Nature* **373(5)**, 33-36.

[68] Wiesenfeld K., Jaramillo F. 1998. Minireview of stochastic resonance. *Chaos* **8(3)**, 539-548.

[69] Wolf A., Swift J. B., Swinney H. L. and Vastano J. A. 1985. Determining lyapunov exponents from a time series. *Physica* **16D**, 285-317.