Order Number 9215042

Asymptotic analysis of digital transmission systems for a first order Gauss-Markov process

Shankar, Hari, Ph.D.

University of Hawaii, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

ASYMPTOTIC ANALYSIS OF DIGITAL TRANSMISSION SYSTEMS FOR A FIRST ORDER GAUSS-MARKOV PROCESS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAII IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

ELECTRICAL ENGINEERING

DECEMBER 1991

By

Hari Shankar

Dissertation Committee:

N.T. Gaarder, Chairperson

R. Brown

N. Abramson

S. Lin

E.J. Weldon

Acknowledgements

I wish to thank my advisor, N.T. Gaarder, for his help, guidance and patience during the course of working on this problem. I am also grateful for the luxury of having been able to walk in at any time into his office for a discussion. This work would not have been possible without his help. I also wish to thank all my friends and fellow students, most of all V. Rao Sattiraju for all the discussions and his help in so many other ways.

It is difficult to express fully my gratitude to all of my family. Most of all I thank Amma, Appa, Shiva and Chitra from whom I have learnt so much and for their support through a long and difficult period, Krishna and Usha for their concern and support and especially Lisa for her patience, support and endurance.

Abstract

In this dissertation, we examine the problem of transmitting a first order Gauss-Markov process over a noiseless digital channel for large transmission rates. We address the problem of of minimizing the time averaged mean square error, for a fixed transmission rate of r bits per second, by finding the optimum sampling rate and the optimum number of quantization levels.

The performance of different reconstruction filters is first examined. Optimal nonlinear reconstruction filters are approximated by linear filters. Fine quantization techniques are used to analyze the performance of different quantization schemes. Fine quantization techniques have been traditionally used to evaluate the variance of the quantization error. These techniques have been extended to evaluate expectations of other functions of the quantization error. The results are used to study the performance of mismatched quantizers, i.e. quantizers that are not optimized exactly for their source statistics. The optimum quantization of two different random variables is also studied; approximations for the correlation between the input of one quantizer and the quantization error of the other and for the correlation between their quantization errors are found. These results form a powerful set of tools with which more complex quantization systems can be analyzed.

We study the minimum time averaged error that can be achieved by different quantization systems for a fixed transmission rate and also compare the power spectral densities of their quantization error. PCM, matched and mismatched DPCM and Sigma-Delta modulation are analyzed. The improvement in performance that can be obtained by adding memory to a quantizer is examined. To this end, a modified PCM scheme that contains sufficient memory in the receiver to store the

previous output of the transmitter is examined. We also describe a finite state sliding block quantizer and study its performance as a function of the memory in the transmitter and the receiver.

Table of Contents

List of Figures

Chapter 1

Introduction

In this dissertation we examine the problem of transmitting an analog, continuous time waveform over a noiseless, digital channel. A digital channel is a channel over which one of a finite set of symbols can be transmitted at discrete time instances. The input waveform is thus first converted into a discrete time signal and the amplitude of each sample is then transmitted over the channel. The constraint imposed by the channel is that the number of bits that can be transmitted per second is finite. Since the amplitude of the signal takes a continuum of values, it is not possible to transmit its exact value and hence the original signal cannot be recovered without error. Given a digital channel, we are then faced with the problem of designing a transmission scheme that minimizes the distortion between the input and the estimated signal.

A model for such a digital transmission system is shown in figure 1.1. Most transmission systems consist of the following components. The input signal may be first passed through a preprocessing filter to make it suitable for transmission; typically it is passed through a low pass filter to obtain a bandlimited signal. This is followed by a sampler, which converts the continuous time input into a discrete time signal. In order to transmit the amplitude of the samples over the digital channel, a quantization scheme is used. Any quantization scheme can be decomposed into two parts – a transmitter, which upon observing a sample transmits a symbol over the channel and a receiver which forms an estimate of the input to the transmitter (or of

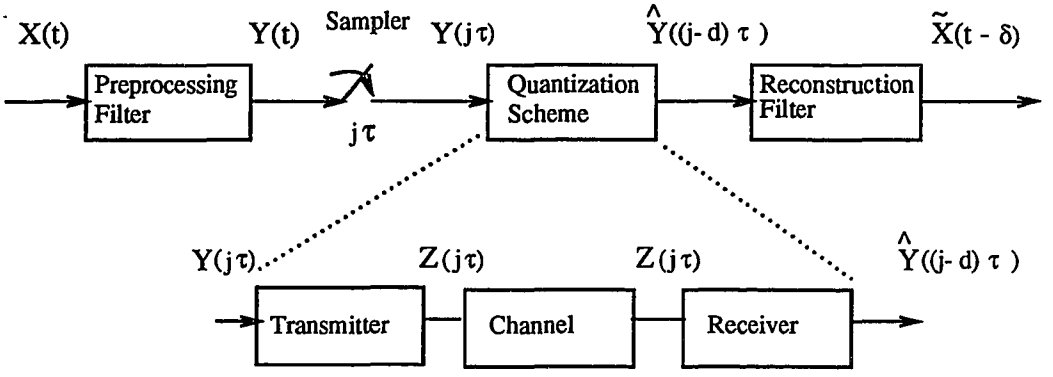

Figure 1.1: A Digital Transmission System.

some previous sample d time instances ago) by observing the output of the channel. Finally we have a reconstruction filter which constructs an estimate of the analog signal from the estimates of the samples. In general there is a delay introduced by the reconstruction filter and the estimated output is a delayed version of the input signal.

1.1 A Mathematical Model

We shall assume that the signal to be transmitted is a zero mean, unit variance, first order Gauss-Markov random process whose autocorrelation function is

$$E\{X(t)X(t+t_0)\} = e^{-|t_0|}. \tag{1.1}$$

The main reason for choosing this model is that it is mathematically tractable and it illustrates the basic problems that are involved in the digital transmission of an analog signal. In this case the sampled process $X(j\tau)$, $j \in \{\ldots, -1, 0, 1, \ldots\}$ is also a first order Markov, Gaussian process.

We assume that the transmission rate over the digital channel is equal to r bits/second. If we sample the input process once every τ seconds and if the output of the transmitter, $Z(j\tau)$, can take n different values, then we find that the transmission rate is

$$r = \frac{\log_2(n)}{\tau} \text{ bits/second.} \tag{1.2}$$

The performance of the overall system will be measured by a time averaged, mean square error which we shall call the smoothed error:

$$\xi_{sm} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} E\left\{(X(t) - \tilde{X}(t))^2\right\}\, dt, \tag{1.3}$$

where $\tilde{X}(t)$ is the output of the reconstruction filter. The average error between the input to the quantization scheme $Y(j\tau)$ and its output $\hat{Y}(j\tau)$ is called the quantization error and will be denoted by ξ_q:

$$\xi_q = E\left\{(Y(j\tau) - \hat{Y}(j\tau))^2\right\}. \tag{1.4}$$

In order to compare the performance of different schemes, the quantization error and the smoothed error are normalized with respect to the input signal power and this ratio is usually expressed in decibels as a signal to noise ratio.

Since the transmission rate is fixed, it is not possible to increase the the sampling rate and the number of quantization levels simultaneously. Given a quantization scheme, a reconstruction filter and a transmission rate of r bits per second, the problem is to find the optimum sampling rate τ_o and the optimum number of values n_o, that the output of the transmitter can take, in order to minimize the smoothed error. This is the basic problem that we investigate in this dissertation.

1.2 Discussion

Most analyses of source coding systems deal only with discrete time sources (exceptions being [6] and [8]). Different transmission systems are compared on the basis

of their quantization error for a fixed transmission rate, where the transmission rate is measured in bits per source symbol. This approach is justified by the following argument (here we quote Jayant and Noll [9]), "minimization of the variance of the reconstruction error in the discrete time domain ensures minimization of the variance of the error between the analog versions as well and these two error signals indeed have the same variances."

The argument is true only if the discrete time signal is obtained from a finite bandwidth, continuous time waveform by sampling above the Nyquist rate and if the continuous time waveform is reconstructed at the receiver by using an ideal low pass filter. The approach then excludes the analysis of infinite bandwidth signals and also neglects the error introduced by non ideal reconstruction filters. More importantly it does not take into consideration the trade off between the number of quantization levels and the sampling rate when the transmission rate is fixed at r bits per second. Hence this approach cannot be used to minimize the smoothed error of a transmission system by optimizing the sampling rate and the number of quantization levels.

Our approach also enables us to compare the performance of different transmission schemes more correctly. The reason being that certain quantization systems like DPCM exploit the correlation between input samples while others like PCM do not. Hence the performance of a DPCM system is far better than that of a PCM system with the same number of quantization levels when the inputs are very highly correlated. But the correlation between the samples can be increased only by increasing the sampling rate which, for a fixed transmission rate of r bits per second, can be accomplished only by decreasing the number of quantization levels. It is then not obvious if we do indeed obtain a large gain with a DPCM system over a PCM system. It is also important to use the time averaged error to compare

the performance of different systems. Else, for any given transmission rate (in bits per second), by sampling the analog input at a sufficiently slow rate, the number of quantization levels can be made arbitrarily large and the quantization error at the sampling instances will then be arbitrarily small. But this would indeed be a very poor transmission system. Hence when comparing two systems like PCM and DPCM, the sampling rate, the number of quantization levels and the smoothed error must be taken into consideration. By considering only the quantization error and disregarding the sampling rate, one cannot compare different transmission schemes correctly.

In order to calculate the smoothed error of a system, we need to analyze the reconstruction filter and the quantization scheme. Optimum reconstruction filters are nonlinear and hard to analyze, so we investigate the performance of some suboptimal linear filters instead. We believe that for large sampling rates their performance will closely match the performance of the optimal reconstruction filter. Since the filters are linear, standard techniques can be used to analyze their performance.

Quantization schemes are much harder to design. Many different kinds of quantization schemes exist and they can be classified according to the type of their structure:

- Scalar quantizers are those that quantize each sample individually, without regard to other samples.

- Vector quantizers are those that group the input samples into blocks or vectors and quantize the vector. Theory shows that by increasing the dimension of the vectors, the performance can be improved.

- Recursive quantizers store information about past inputs by means of a state variable. The output of the quantizer is then a function of the current input

and the state variable. The state variable is updated recursively at each time instant. If the number of values that the state variable can take is finite, then the quantizer is called a finite state quantizer.

- Sliding block quantizers map overlapping input vectors into code words; at each time instance j, the input vector $(X(j\tau), \ldots, X((j-l)\tau))$ is mapped into a code word.

Different schemes are suitable for different types of inputs and some are more complex than others to implement. Scalar quantizers are the simplest to implement but have the largest distortion for a given transmission rate. Their performance can be improved by using vector quantizers but the complexity of the system increases rapidly as the dimension of the vectors increases. Recursive quantizers are suitable for the transmission of signals that vary slowly with time. They are relatively simple to implement but can be sensitive to channel errors. Not much is known about sliding block schemes but, if the block length of the receiver is small, then one advantage is that it would be relatively robust to channel errors.

An exact analysis of most schemes is difficult due to the inherent nonlinear nature of a quantizer; different approximations are used to estimate the performance. Often one has to resort to simulations to test their performance. We mention some approximations that are commonly used.

- A Linear Additive Noise Model: The output of a scalar quantizer is modeled as the sum of the input and an independent error signal. This is an over simplified model of limited application since the error signal is determined exactly by the input and hence is by no means independent of the input (see [9] for details).

- Transform Methods: In the case of uniform, scalar quantizers, the quantization error is a periodic function of the input if the input is sufficiently bounded. It is then possible to express the quantization error as a Fourier series and derive useful formulae for various quantities (see [5] for some applications and results using this method).

- Fine Quantization Techniques: If the density function of the input is in some sense "reasonably smooth," it is possible to derive approximations for the quantization error and other quantities. This is the method that we develop in detail in this dissertation.

The purpose of this dissertation is to develop sufficiently general results that will enable us to study different systems for the transmission of a first order Markov, Gaussian random process.

The material in this dissertation is arranged as follows. For the convenience of the reader, we gather in the next chapter, the main results that have been obtained in this dissertation and discuss their significance. In the third chapter we describe and analyze three different types of reconstruction filters and derive some general expressions for the smoothed error. In the fourth chapter we develop various formulae for analyzing scalar quantizers based on fine quantization techniques. In the fifth chapter we use the results of the previous two chapters to investigate different types of quantization schemes. Finally we conclude this dissertation by mentioning some problems that remain for future research.

And finally a note about notation. Capital letters denote random variables except for the letter 'E' which denotes the expectation operator. For example, $E\{X\}$ is the expectation of the random variable X. Script letters, for example $\mathcal{R}, \mathcal{S}$, are used to denote sets. The value of the density function of the random variable X

at the point x is denoted by $p_X(x)$. For simplicity, a periodically sampled function $x(j\tau)$, $j = \ldots, -1, 0, 1, \ldots$ will be denoted as x_j. The derivative of the function $f(x)$ will be denoted as $f'(x)$. Partial derivatives of multivariate functions will be denoted by explicitly listing the variables with respect to which the derivatives are taken as a superscript; for example $f^{xy}(x, y)$ denotes the partial derivative of the function $f(x, y)$ with respect to the variables x and y.

Chapter 2

Results and Discussion

In this chapter we discuss the results that we have derived in this dissertation and compare the performance of different quantization systems.

The most significant result is that fine quantizer techniques can be used to analyze a wide variety of problems. This is amply demonstrated in chapters three and four. Although fine quantizer results are asymptotic approximations that converge to the exact values only as the number of quantization levels tends to infinity, we find that in general they converge quickly and are accurate even when the number of quantization levels is small; in the case of Gaussian inputs we find that they are applicable in certain cases for quantizers with just four levels.

Fine quantizer approximations were first used to derive approximations for the variance of the quantization error [4, 10, 15]. We have extended this technique to calculate the expectation of a wide class of functions of the quantization error of an optimum quantizer. Another general set of results that we have obtained is in the case when two different inputs are quantized by their optimal quantizers. Under certain general conditions we can calculate approximations for the correlation between the input of one quantizer and the quantization error of the other. We can also approximate the correlation between the quantization errors. These results form a powerful set of tools that can be used to analyze a wide variety of quantization systems in a relatively simple manner.

We have also derived some general approximations for the smoothed error of quantization systems for different types of reconstruction filters. We show that for large transmission rates, the smoothed error is approximately equal to the sum of the quantization error and the error due to the reconstruction filter. Analysis of various systems shows that the error due to the reconstruction filters is larger than the quantization error. It thus seems likely that we stand to gain more by designing better reconstruction filters rather than by concentrating only upon improving the quantization error.

We next discuss the performance of different quantization systems. We examine the quantization error, the smoothed error and the flatness of the power spectral density of the quantization error.

PCM Systems — The quantization error of PCM systems has been examined by numerous authors. An asymptotic analysis of scalar quantizers was first published by Bennet [1]. Optimal quantizers were first examined by Lloyd [10] and Max [11]. In the same paper, Lloyd also derived an asymptotic expression for the variance of the quantization error of optimal scalar quantizers. Zador [15] derived mathematically rigorous arguments for the asymptotic variance of the quantization error of scalar and vector quantizers. The same results were rederived using more intuitive arguments by Gersho [4]. All the above results deal with the quantization error and to our knowledge, no results have been published that examine the smoothed error of PCM systems for a first order Gauss-Markov input.

It is well known that the variance of the quantization error of an optimal quantizer for a unit variance Gaussian input is approximately equal to $2.7/n^2$, where n is the number of quantization levels [10]. To our knowledge, no similar analytical approximation exists for the asymptotic performance of a uniform quantizer in the case of a Gaussian input. We compare the performance of an optimal quantizer and

a uniform quantizer and find that the variance of the quantization error of a uniform quantizer is of the order of $\ln(n)$ times larger than that of the optimal quantizer. For large n the difference can be significant; for example, when $n = 100$, the SNR of a uniform quantizer is smaller than that of an optimal quantizer by about 6.5 dB. We also compare the quantization error of a uniform quantizer with that of an optimal quantizer in the case of a Laplacian input. We find that the decrease in the SNR is much larger in this case than in the case of Gaussian inputs. We attribute this to the fact that a Laplacian density has a larger tail than a Gaussian.

However to design an optimum quantizer, the distribution of the input must be known. We examine the performance of an optimum quantizer when the distribution of the input is not known exactly; namely, we examine the case when the the mean and the variance of the input differ from the values for which the quantizer is designed. We show in the case of Gaussian inputs that it is the ratio of the error in estimating the mean to the standard deviation of the input that is significant in decreasing the SNR. If the error in determining the mean of the input is within one half the standard deviation of the input, then the SNR decreases by at most 3 dB. If there is a 10 percent error in determining the standard deviation of the input, then the SNR decreases by about 0.5 dB while for a 20 percent error the SNR decreases by about 2 dB. For larger errors, the SNR decreases at even a faster rate.

We also examine the smoothed error for a first order Gauss-Markov input. We find that for a transmission rate of r bits per second, the smoothed error decreases asymptotically as $(c\ln(r))/r$ where c is a constant that lies between 0.3 and 0.14 depending upon the reconstruction filter (see section 5.1 for details). We find that it is the error due to the reconstruction filter that is the dominant term in the smoothed error. The smoothed error of a uniform quantizer is larger than that of an optimal quantizer by about 2.3 dB. Unlike the quantization error, the difference

in the smoothed errors between an optimal quantizer and a uniform quantizer is independent of the rate. Hence due to its simplicity, it may be preferable to use a uniform quantizer. When we compare the smoothed error of PCM systems with the distortion rate bound for this input, we find that the SNR of a PCM system is about $\log(r)$ times smaller than the distortion rate bound. Thus the difference between the maximum achievable SNR and the SNR of a PCM system grows infinitely large as the transmission rate increases.

In the case of optimal quantizers we examine the autocorrelation function of the quantization error to determine if its power spectral density is flat. In certain applications like speech, it has been observed that the perceptual quality of the output is better when the power in the quantization error is spread over a wide range of frequencies (i.e., the power spectral density is flat) rather than being concentrated over a narrow range of frequencies [9]. General conditions under which the quantization error is white and independent of the input have been obtained by Sripad and Snyder [14] and Schuchman[12]. The quantization error of a scalar quantizer has been analyzed by Gray [5] when the input is a sinusoid. To our knowledge, the autocorrelation function of the quantization error of a PCM system has not been examined in the case when the input is a first order Gauss-Markov input. We show that even in the case when the correlation coefficient between the inputs equals one, that by adding a dither signal, the quantization error increases by about 1 dB. A more significant result is obtained when we optimize the number of quantization levels and the sampling rate for a given transmission rate to minimize the smoothed error. We find that for this choice of the sampling rate and number of quantization levels, the correlation between the quantization errors at different time instances is almost zero and hence the power spectral density will be flat. This implies that dithering is not needed for PCM systems.

A Modified PCM System — We examine the asymptotic smoothed error of a modified PCM system which consists of a memoryless transmitter and a sliding block receiver of length two. The analysis of such systems has been previously carried out only in the case of a binary quantizer [3]. We derive expressions for the gain in the quantization error of such a system with memory only in the receiver, over that of a PCM system. Results show that the gain in the smoothed error is marginal; for example, at a transmission rate of 1000 bits per second, the gain in the SNR is only 0.06 dB. We believe that modifying a PCM system by adding any amount of memory only in the receiver does not improve the asymptotic smoothed error significantly.

DPCM Systems — There is a large amount of literature that examines the performance of DPCM systems. The analysis of delta modulation systems for first order Gauss-Markov inputs was first carried out by Slepian [13]. He obtained a series solution in Hermite polynomials but the computational procedure did not converge for large sampling rates. This analysis was extended to DPCM systems by Hayashi [6] but it suffered from similar convergence problems. Janardhanan [8] was the first to obtain a computational procedure that converged for large sampling rates for matched and mismatched DPCM systems. The method though is elaborate and can be carried out only numerically. Our analysis, although approximate, is much simpler and yields results that are in good agreement with the more accurate ones of [8].

We have analyzed different versions of DPCM systems. In the case of matched DPCM systems (matched DPCM systems are those where the correlation between successive inputs is known exactly and this value is used in the prediction filter), we show that the quantization error can be approximated by the expression $2.7\ (1\ -\ \rho^2)\ /\ n^2$, where ρ is the correlation between two successive input samples

and n is the number of quantization levels. This is true whether the quantizer is optimized either for its input or for the innovations process corresponding to the sampled process. This result is important because optimizing the quantizer for its input in a DPCM system is a hard problem. On the other hand the standard Lloyd-Max algorithm can be used to optimize the quantizer for the innovations process since the distribution of this process can be easily found. Our results show that settling for this suboptimal quantizer does not cause the quantization error to increase. Another important point to be noted from the approximation for the quantization error is that it is dependent on the number of quantization levels and on the sampling period. This is unlike the case of PCM systems where the quantization error is independent of the sampling rate. Hence the quantization error can be decreased by either increasing the sampling rate or by increasing the number of quantization levels. When we consider the smoothed error, we are better off by increasing the sampling rate since the quantization error and the error due to the reconstruction filters can be simultaneously decreased. This is precisely the reason why DPCM systems perform so much better than PCM systems.

The asymptotic smoothed error of a matched DPCM system is equal to c/r where c is a constant whose value lies between 0.8 and 1.4 depending upon the reconstruction filter used. Hence in this case the smoothed error decreases at the same rate as the distortion rate bound as the transmission rate increases. The SNR of matched DPCM systems is within 1.4 dB of the distortion rate bound for large transmission rates when the input is filtered before sampling and the an ideal lowpass reconstruction filter is used. This is a large improvement over the performance of PCM schemes. In fact the difference between the SNR of a matched DPCM system and a PCM system grows infinitely large as the transmission rate increases, although quite slowly. For example at a transmission rate of 1000 bits per second, the SNR

of a matched DPCM system is about 3 dB larger than the SNR of a PCM system with the same reconstruction filter and about 4 dB larger at a transmission rate of 10,000 bits per second.

We also show that the autocorrelation function of matched DPCM systems decreases geometrically from its maximum value; the ratio term of the geometric series being inversely proportional to the number of quantization levels. Hence we expect the power spectral density to be reasonably flat but PCM systems are better in this respect. The spectral density of the quantization error process can be improved by decreasing the sampling rate and increasing the number of quantization levels but only at the expense of increasing the smoothed error. For example, for a transmission rate of 1000 bits/second, by using twice as many quantization levels as the optimum number of quantization levels (i.e., we use a quantizer with 20 quantization levels instead of 10), the smoothed error of a matched DPCM system with a low pass reconstruction filter increases by 0.5 dB but the correlation between two adjacent quantization errors decreases by 6 dB.

Our analysis of mismatched DPCM systems (these are systems where the exact correlation between successive inputs is not known and an estimate is used in the prediction filter), shows that the quantization error is quadratic in the difference between the correlation between two successive inputs and the estimate of this correlation that is used in the prediction filter. Perfect integrator systems are those that do not use a multiplier in the feedback loop; they perform as well as matched DPCM systems for large transmission rates. This is not surprising since the correlation between successive samples tends to 1 as the transmission rate increases. This considerably decreases the complexity of a DPCM system. However for low transmission rates when the sampling rate is low, the SNR of a perfect integrator

system decreases rapidly and when the correlation between successive samples of the input is less than 0.5, the quantization error is larger than that of a PCM system.

Sigma-Delta Modulation — The Sigma-Delta modulation scheme was first proposed by Inode and Yasuda [7]. A detailed analysis of Sigma-Delta modulation systems has been published by Gray [5], but only for dc and sinusoidal inputs. With our approximations we have been able to obtain an approximate analysis in the case of a first order Gauss-Markov input. Our analysis of Sigma-Delta modulation shows that the quantization error is larger than that of a PCM system. The quantization error is a function of the sampling rate but unlike DPCM it increases as the correlation between the input samples increases. In the worst case when the correlation coefficient between the samples equals one, the variance of the quantization error is roughly 3 dB larger than that of PCM systems.

On the other hand, the asymptotic smoothed error is approximately equal to that of a PCM system. For a given sampling rate, the correlation between quantization errors is smaller than in the case of PCM. Further, the optimum sampling frequency needed to minimize the variance of the quantization error is smaller than in the case of PCM systems. Because of these two factors, the power spectral density of the quantization error is flatter than that of PCM systems. Hence in terms of the smoothed error, a Sigma-Delta modulation system performs worse than a PCM or a DPCM system but the power spectral density of its quantization error is flatter.

A Sliding Block Finite State Quantizer — Finite state quantization systems are difficult to design and analyze and not many results have been published about them except for those in [3]. Nor are there many good examples of finite state quantization systems. We propose a sliding block, finite state quantizer which is a feed forward approximation of a DPCM system. We investigate the amount of memory that is required in the transmitter and the receiver to achieve a performance

close to that of a DPCM system. Our results show that for a transmission rate of r bits/second, to achieve a quantization error that is $(1 + \alpha)$ times that of a DPCM system ($0 < \alpha < 1$), the asymptotic state size of the transmitter increases as $n_o^{r^2} \log(r/\alpha)$ while that of the receiver increases as $n_o^{r \log(r/\alpha)}$. Here n_o is a constant that depends upon the reconstruction filter used but typically, $5 \leq n_o \leq 10$. Thus the state size of the transmitter is larger than that of the receiver. We believe this to be true because it is the transmitter that tracks the state of the receiver and not the other way around.

Chapter 3

Reconstruction Filters

In this chapter we analyze three different types of reconstruction filters and derive expressions for the smoothed error. The optimum reconstruction filters are nonlinear but if the quantization error is small, they can be approximated by linear filters. The advantage in considering linear filters is that they are easier to analyze and standard techniques can be used to determine their performance.

3.1 A Memoryless Reconstruction Filter

We first consider a transmission system which does not have any preprocessing filter. The input to the quantization system at time $j\tau$ is X_j and the output of the quantizer is $\hat{X}_{j-d}$, where $d\tau$ is the decoding delay of the quantization system. A memoryless reconstruction filter then forms an estimate of the input process $X(t)$ in the time interval $\{(j-1/2)\tau, (j+1/2)\tau\}$ using only $\hat{X}_j$, the output of the quantizer at time $j\tau$. Thus the reconstruction filter has a decoding delay of $\tau/2$ seconds and there is an overall decoding delay of $(d+1/2)\tau$ seconds for this system (see figure 3.1).

The optimum estimate of $X(t)$, given that the output of the quantizer $\hat{X}_j$ equals x, is the conditional expectation $E\{X(t)/\hat{X}_j = x\}$. This is difficult to compute and we approximate it with the conditional expectation $E\{X(t)/X_j = x\}$ instead. If the quantization error is small, this approximation should be close to the optimal value. Since the input process $X(t)$ is Gaussian, this conditional expectation equals

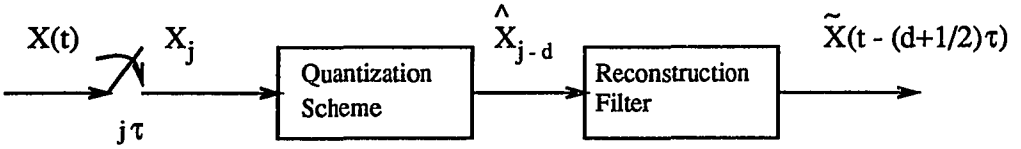

Figure 3.1: A Memoryless Reconstruction Filter.

ρx where ρ is the correlation between $X(t)$ and $X(j\tau)$. Since the autocorrelation function of the input process is equal to $e^{-|t|}$, the output of the reconstruction filter can be expressed as follows:

$$\tilde{X}(t) = e^{-|t-j\tau|}\hat{X}_j, \quad (j-1/2)\tau < t < (j+1/2)\tau. \tag{3.1}$$

We now calculate the smoothed error. Because of stationarity, it is sufficient to average the mean squared quantization error over only one sampling interval[1]. We note that the input at any time t can be expressed as a linear combination of the input at time $t = 0$ and an independent, zero mean, unit variance Gaussian random variable:

$$X(t) = \rho X_0 + \sqrt{1-\rho^2}U. \tag{3.2}$$

[1]We point out that in the mathematical model set up here, the sampled process is not stationary in the strict sense because the origin of the time axis is fixed to coincide exactly with a sample, namely X_0. This can be rectified by making the sampling instances random as follows: we assume that the jth sampling instance is now equal to $(j\tau+T)$, where T is a uniformly distributed random variable over the interval $(0,\tau)$ and independent of the input process. The random processes $X(t), X_j, \hat{X}_j$ and $\tilde{X}(t)$ are now all jointly stationary. The analysis of the system can now be carried out by first conditioning on T. Since this random variable is independent of the input process, the conditioning does not effect the remaining analysis and hence we ignore this point and assume that T equals zero.

The smoothed error can then be expressed as

$$\xi_{sm} = \frac{1}{\tau}\int_{-\tau/2}^{\tau/2} E\left\{\left(\sqrt{1-\rho^2}U + \rho(X_0 - \hat{X}_0)\right)^2\right\} dt. \tag{3.3}$$

The input X_0 and the random variable U are uncorrelated. If the quantization error is small, the estimate $\hat{X}_0$ will be approximately equal to the input X_0 and hence will also be approximately uncorrelated with U. With this approximation the above equation simplifies to

$$\xi_{sm} \approx \frac{1}{\tau}\int_{-\tau/2}^{\tau/2} \left(1 - \rho^2 + \rho^2\xi_q\right) dt. \tag{3.4}$$

The above equation can be easily evaluated and the smoothed error can be expressed as

$$\xi_{sm} \approx \frac{1}{\tau}\left(\tau - (1 - e^{-\tau})(1 + \xi_q)\right). \tag{3.5}$$

When the sampling interval is small, this expression can be approximated by the first few terms of its Taylor series in τ:

$$\xi_{sm} \approx \frac{\tau}{2} + \xi_q. \tag{3.6}$$

The first term is the error due to the reconstruction filter and for large sampling rates it varies linearly with τ. The above equation then states that the smoothed error is approximately the sum of the error due to the reconstruction filter and the quantization error.

3.2 A Reconstruction Filter with a Unit Delay Memory

We next examine a system without any preprocessing filter that uses a reconstruction filter which forms an estimate of $X(t)$ in the time interval $\{(j-1)\tau < t < j\tau\}$ using $\hat{X}_j$ and $\hat{X}_{j-1}$ (see figure 3.2). This reconstruction filter was first analyzed by

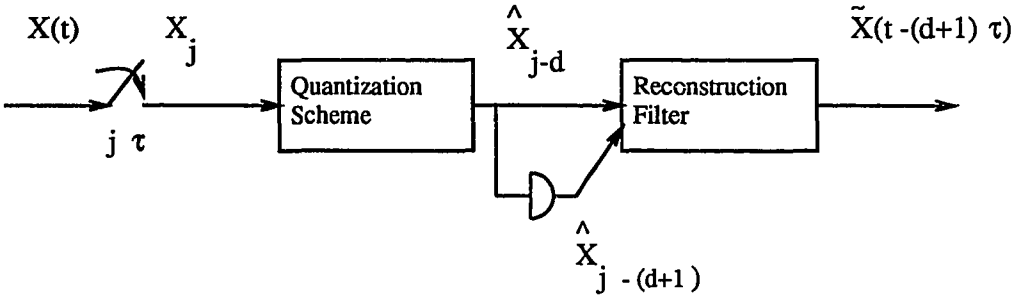

Figure 3.2: A Reconstruction Filter With a Unit Delay Memory.

Hayashi [6]. There is now a decoding delay of τ seconds at the receiver. The best estimate of the input equals the conditional mean of $X(t)$ given that the outputs of the quantizer at time $j\tau$ and $(j-1)\tau$ equal x and x' respectively. This is difficult to compute and as in the previous section we approximate it with one that is easier to calculate, namely $E\{X(t)/X_{j-1} = x', X_j = x\}$. When the quantization error is small this estimate will be close to the optimal estimate. Since the process $X(t)$ is Gaussian, this suboptimal estimate is linear and is given by the following expression:

$$\tilde{X}(t) = a(t)x_0 + b(t)x_1 \quad 0 \leq t < \tau, \tag{3.7}$$

where

$$x_0 = q_X(X(0)), \quad x_1 = q_X(X(\tau)) \tag{3.8}$$

and

$$a(t) = \frac{e^{-t} + e^{t-2\tau}}{1 - e^{2\tau}}, \quad b(t) = \frac{e^{-(\tau-t)} - e^{-(\tau+t)}}{1 - e^{-2\tau}}. \tag{3.9}$$

To calculate the smoothed error we proceed in a manner similar to the memoryless case. It can then be shown that

$$\xi_{sm} = \frac{1+\rho^2}{1-\rho^2} - \frac{1}{\tau} + \frac{\frac{\sinh 2\tau}{\tau} - 2}{2\sinh^2\tau}\xi_q + \frac{\tau\cosh\tau - \sinh\tau}{\tau\sinh^2\tau} E\left\{(X_0 - \hat{X}_0)(X_\tau - \hat{X}_\tau)\right\}. \tag{3.10}$$

For sufficiently small τ, this expression can be approximated by the first few terms of its Taylor series expansion in τ and the following approximation can be obtained:

$$\xi_{sm} \approx \frac{\tau}{3} + \frac{2\xi_q}{3} + \frac{1}{3}E\left\{(X_0 - \hat{X}_0)(X_T - \hat{X}_T)\right\}. \tag{3.11}$$

The first term is the error due to the reconstruction filter and we see that it has decreased by approximately 1.8 dB when compared to the memoryless reconstruction filter. Also, the smoothed error now depends not only on the variance of the quantization error but also on the correlation between the quantization errors at two adjacent time intervals. We note that this correlation is less than or equal to the variance of the quantization error and hence the smoothed error is bounded from above by the sum of the error due to the reconstruction filter and the variance of the quantization error. We also note that we would not expect the performance to improve very much by increasing the complexity of the reconstruction filter by adding memory to remember more of the quantizer outputs. This is because the sampled process $\{X_j\}$ is a first order Markov process and hence $\hat{X}_j$ will also be 'approximately' first order Markov. This implies that using more samples to estimate $X(t)$ should not significantly effect the smoothed error.

3.3 A Low Pass Reconstruction Filter

An alternate technique to reconstruct an analog signal from its samples is to first bandlimit the input process by passing it through a lowpass filter of bandwidth $1/(2\tau)$. The output of the filter is then sampled every τ seconds. The sampled

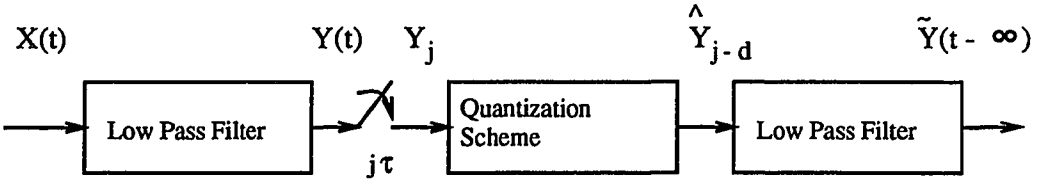

Figure 3.3: A Low Pass Reconstruction Filter.

process is quantized and transmitted. The receiver reconstructs the input process by passing the quantized samples through a lowpass filter of bandwidth $1/(2\tau)$. Bandlimiting the input process before sampling eliminates the aliasing error. The decoding delay for this scheme is infinitely large. We denote the output of the lowpass filter at the transmitter as $Y(t)$, the sampled process as Y_j, the output of the quantizer as $\hat{Y}_j$ and the output of the reconstruction filter as $\tilde{Y}(t)$ (see figure 3.3). The smoothed error can be expressed as follows:

$$\begin{aligned}\xi_{sm} &= \lim_{\theta\to\infty}\frac{1}{\theta}\int_{-\theta/2}^{\theta/2} E\left\{(X(t)-Y(t)^2\right\}+ \\ &\quad 2E\left\{(X(t)-Y(t))(Y(t)-\tilde{Y}(t))\right\}+E\left\{(Y(t)-\tilde{Y}(t))^2\right\}\,dt. \quad (3.12)\end{aligned}$$

We next evaluate each of the expectations within the integral in the above equation. Since the lowpass filter is a linear, time-invariant system, it can be shown using standard techniques that the average mean square error between the input to the lowpass filter $X(t)$ and its output $Y(t)$ is given by

$$E\left\{(X(t)-Y(t))^2\right\}=1-\frac{2}{\pi}\tan^{-1}(\pi/\tau). \quad (3.13)$$

To evaluate the second expectation within the integral in (3.12), we first note that the random process $Y(t)$ is a low pass signal whose power spectral density is zero for frequencies greater than $(1/(4\tau)$. We then note that $(X(t)-Y(t))$ is a high pass process whose power spectral density is zero for frequencies less than $1/(4\tau)$. Since $Y(t)$ is derived from the process $X(t)$ by passing it through a linear, time invariant system, it follows that $Y(t)$ and $(X(t)-Y(t))$ are uncorrelated because their power spectral densities are disjoint. Further since $X(t)$ is a Gaussian process, it follows that the two processes are independent. The random process $\tilde{Y}(t)$ is a function only of $Y(t)$ and hence is also independent of $(X(t)-Y(t))$. Since the mean of $X(t)$ is zero, we then have the result that $(X(t)-Y(t))$ and $(Y(t)-\tilde{Y}(t))$ are uncorrelated (in fact they are independent). The second term is thus equal to zero.

To evaluate the third term, we note that the processes $Y(t)$ and $\tilde{Y}(t)$ are band limited, and hence by the sampling theorem the signals can be represented as a sum of sinc functions:

$$Y(t) = \sum_{j=-\infty}^{\infty} Y_j \,\mathrm{sinc}\left(\frac{t-j\tau}{\tau}\right) \quad \text{and} \tag{3.14}$$

$$\tilde{Y}(t) = \sum_{j=-\infty}^{\infty} \hat{Y}_j \,\mathrm{sinc}\left(\frac{t-j\tau}{\tau}\right). \tag{3.15}$$

We then obtain the following identity:

$$\lim_{\theta\to\infty}\frac{1}{\theta}\int_{-\theta/2}^{\theta/2} E\left\{(Y(t)-\tilde{Y}(t))^2\right\}\,dt = \sum_{j=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} E\left\{(Y_j-\hat{Y}_j)(Y_l-\hat{Y}_l)\right\} \lim_{\theta\to\infty}\frac{1}{\theta}\int_{-\theta/2}^{\theta/2}\mathrm{sinc}\left(\frac{t-j\tau}{\tau}\right)\mathrm{sinc}\left(\frac{t-l\tau}{\tau}\right)\,dt.$$

The functions $\left\{\frac{1}{\sqrt{\tau}}\,\mathrm{sinc}\left(\frac{t-j\tau}{\tau}\right)\right\}$ (here j takes all integer values) form an orthonormal set and hence, in the above summation, the terms when j is not equal to l are equal to zero. Also, the ratio θ/τ is equal to the number of samples in the time interval

θ and hence we can rewrite the above equation as

$$\lim_{\theta\to\infty}\frac{1}{\theta}\int_{-\theta/2}^{\theta/2} E\left\{(Y(t)-\tilde{Y}(t))^2\right\}\,dt = \lim_{m\to\infty}\frac{1}{(2m+1)}\sum_{j=-m}^{m} E\left\{(Y_j-\hat{Y}_j)^2\right\}. \quad (3.16)$$

Since the processes Y_j and $\hat{Y}_j$ are stationary, the mean square error between Y_j and $\hat{Y}_j$ is independent of j. We thus obtain the result that the smoothed error between $Y(t)$ and $\tilde{Y}(t)$ is equal to the error in quantizing the random variable Y_j (note that the variance of this random variable is equal to $(2/\pi)\tan^{-1}(\pi/\tau)$).

The above results can be combined to obtain the following expression for the smoothed error:

$$\xi_{sm} = 1 - \frac{2}{\pi}\tan^{-1}(\pi/\tau) + \frac{2}{\pi}\tan^{-1}(\pi/\tau)\xi_q. \quad (3.17)$$

For small τ, this can be approximated by the first few terms of the Taylor series in τ. The above expression for the smoothed error can then be simplified as follows:

$$\xi_{sm} \approx 0.2\tau + \xi_q. \quad (3.18)$$

3.4 Discussion

Based upon the above results, it is reasonable to make the assumption that the asymptotic smoothed error for any system is of the form

$$\xi_{sm} \approx \alpha\xi_q + \beta\tau, \quad (3.19)$$

where α and β are constants. In general α will depend upon the reconstruction filter and the quantizer while β will depend only upon the reconstruction filter. The term $\beta\tau$ is the error due to the reconstruction filter in reconstructing the estimate of the analog input from the quantized samples. From the above examples we see that β does not vary greatly with the type of reconstruction filter being used. Since it is also independent of the quantizer, increasing the number of quantization levels does not

effect this term. The only way that it can be decreased is by increasing the sampling frequency. The first term is the error arising from the quantization process. If the quantizer is memoryless, this term is independent of the sampling interval τ and can be decreased only by increasing the number of quantization levels. When the transmission rate is fixed, it is not possible to simultaneously increase the sampling rate and the number of quantization levels. This puts a limit on the performance that can be achieved with a memoryless quantizer. But it is possible that by adding memory to the quantizer, the quantization error can be made to depend upon the sampling frequency. Then both the terms contributing to the smoothed error can be decreased simultaneously by increasing the sampling frequency. If this is possible, then it definitely would pay to increase the complexity of the quantizer rather than that of the reconstruction filter.

Chapter 4

Fine Quantizers

In this chapter we examine minimum mean square error fine quantizers. We limit our investigation only to scalar quantizers. Fine quantization techniques have been used to approximate the mean square error that occurs when quantizing an input with a known distribution [4, 10, 15]. In this chapter we extend these techniques to calculate more general expectations. We then use these results to study the effects of some simple mismatches between the input and the quantizer; mainly the case when the mean and the variance of the input differ from the values that the quantizer was designed for. We also study the optimum quantization of two different random variables; we approximate the correlation between the input of the first quantizer and the quantization error of the second and also the correlation between the two quantization errors. While these results are interesting in their own right, their main importance will be seen in the next chapter where they will be used to analyze various quantization schemes.

4.1 Minimum Mean Square Error Quantizers

A scalar quantizer is a function (which we shall denote by $q()$) that maps the real line into a finite set of numbers called representative points. An n level quantizer is a quantizer with n representative points which we shall denote as $q_0, q_1, \ldots, q_{n-1}$. The quantizer partitions the real line into n disjoint regions called quantization regions

which are given by the following equation:

$$\mathcal{S}_i = \{x : q(x) = q_i\} \qquad i = 0, \ldots, n-1. \tag{4.1}$$

In the context of a communications system, this quantizer can be subdivided into a transmitter and a receiver as follows. If the input lies in the quantization region $\mathcal{S}_i$, the transmitter transmits the index i over the channel and the receiver upon observing this value, chooses the representative point q_i as its output. Thus the transmitter is associated with the quantization regions while the receiver is associated with the representative points.

We shall be interested in the case when the input to the quantizer is a continuous random variable X with known density function $p_X(x)$. A minimum mean square error quantizer is one that minimizes the mean square error between its input X and its output $q(X)$. Throughout this dissertation, any such quantizer will be referred to as an optimal quantizer. We denote the quantizer optimized for the random variable X by $q_X()$ and the corresponding quantization error by $\eta_X()$, i.e.,

$$\eta_X(x) = x - q_X(x). \tag{4.2}$$

For ease of notation we will omit the subscript X when referring to a quantizer or the quantization error if it is obvious that the quantizer has been optimized for a particular input and use the notation $q()$ and $\eta()$ instead.

There are two necessary conditions for a quantizer to be optimal (see [10, 11]):

$$q_i = E\{X/X \in \mathcal{S}_i\} \tag{4.3}$$

and

$$\mathcal{S}_i = \left\{x : |x - q_i|^2 = \min_j |x - q_j|^2\right\}. \tag{4.4}$$

These conditions are referred to as the Lloyd-Max conditions for an optimal quantizer. From the first condition we see that the representative point of a region is the

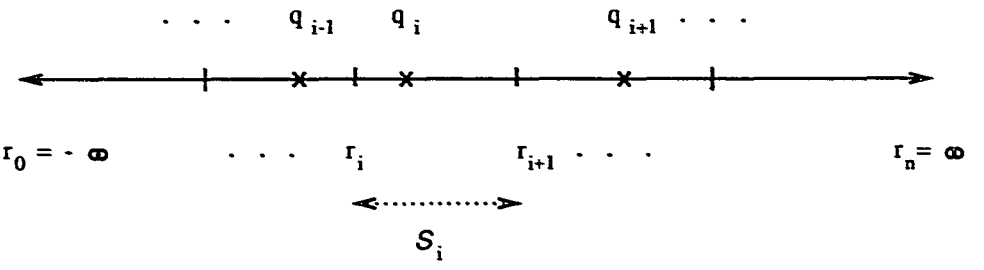

Figure 4.1: A Scalar Quantizer.

centroid of the region. The second condition implies that the optimal quantization regions are intervals whose endpoints lie midway between the representative points. Thus if the representative points are ordered such that $q_0 < q_1 < \ldots < q_{n-1}$, then the quantization regions can be described as follows (also see figure 4.1):

$$\mathcal{S}_i = \{x : r_i < x < r_{i+1}\} \quad \text{i=0,...,n-1.} \tag{4.5}$$

The points r_i, $i = 1, \ldots, n-1$, separate the quantization regions and lie midway between the representative points, i.e.,

$$r_i = (q_{i-1} + q_i)/2 \qquad i = 1, \ldots, n-1, \tag{4.6}$$

and $r_0 = -\infty$ and $r_n = \infty$. The quantization error that occurs when the input lies in either of the two unbounded intervals $\mathcal{S}_0$ or $\mathcal{S}_{n-1}$ is called the overload error and the error that occurs when the input lies in any of the bounded regions is referred to as the granular error.

Other important properties of optimal quantizers follow quite easily from the two conditions given in (4.3) and (4.4) (for proofs, see [9]). We next list some of the properties.

1. The mean of the output equals the mean of the input, i.e.,

$$E\{q_X(X)\} = E\{X\}, \tag{4.7}$$

and hence the mean of the quantization error equals zero.

2. The output and the error of an optimal quantizer are uncorrelated, i.e.,

$$E\{q_X(X)\eta_X(X)\} = 0. \tag{4.8}$$

3. The input and the error are correlated and this correlation equals the mean square error, i.e.,

$$E\{X\eta_X(X)\} = E\{\eta_X^2(X)\}. \tag{4.9}$$

The main difficulty in designing optimal quantizers is that (4.3) and (4.4) cannot be explicitly solved for the representative points and regions, except in some simple special cases. Iterative solutions are used to design the quantizer for a specific input and number of quantization levels. Thus there is no explicit formula for the mean square error of an n level optimal quantizer optimized for an input X with density function $p_X(x)$.

A word is in order about the scaling of quantizers according to the mean and variance of the input. Let $q_X()$ be an optimum quantizer for a unit variance random variable X. Then the optimum quantizer for the random variable $Y = \sigma X + \mu$ is a shifted, scaled version of the optimum quantizer for X. This can be argued as follows. The average mean square error in quantizing Y can be expressed as shown below:

$$E\{(Y - q_Y(Y))^2\} = \sigma^2 E\left\{\left(X - \left(\frac{1}{\sigma}q_Y(\sigma X + \mu) - \frac{\mu}{\sigma}\right)\right)^2\right\}. \tag{4.10}$$

But by the definition of an optimum quantizer, the expectation on the right hand side is minimum when $(\frac{1}{\sigma}q_Y(\sigma X + \mu) - \frac{\mu}{\sigma})$ is the optimum quantizer of X, i.e.,

$$\frac{1}{\sigma}q_Y(\sigma X + \mu) - \frac{\mu}{\sigma} = q_X(X). \tag{4.11}$$

This implies that the optimal quantizer for Y is a shifted, scaled version of the optimal quantizer for X, i.e.,

$$q_Y(Y) = \sigma q_X\left(\frac{Y-\mu}{\sigma}\right) + \mu. \tag{4.12}$$

Thus the random variable Y can be quantized in the following steps. First the mean is subtracted from the input and the difference is divided by the standard deviation σ. The resulting value is quantized by an optimum quantizer for a zero mean, unit variance random variable. Finally, the output of the quantizer is multiplied by the standard deviation σ and added to the mean μ. An alternate interpretation of the above equation is that if the representative points of an optimum quantizer for a zero mean, unit variance input are given by $(q_0, \ldots, q_{n-1})$ and the endpoints of the quantization regions equal $(r_1, \ldots, r_{n-1})$, then the optimum quantizer for an input with mean μ and variance σ^2 has representative points $(\sigma q_0 + \mu, \ldots, \sigma q_{n-1} + \mu)$ and the endpoints of the quantization regions equal $(\sigma r_1 + \mu, \ldots, \sigma r_{n-1} + \mu)$. Further, if the mean square error in quantizing a zero mean, unit variance random variable X is ϵ^2, then the mean square error in quantizing the random variable $Y = \sigma X + \mu$ is $\sigma^2\epsilon^2$.

4.2 Fine Quantizers

We describe a quantizer as fine when the density function of the input is approximately constant within each of the bounded quantization regions and the probability of overload error is negligible. In this section we discuss an approach used to derive approximations for the quantization error of optimal, fine quantizers and state the main results that follow from this approach (for details see [4] or appendix A). We show how these results can be extended to calculate other quantities of interest. We also compare the performance of optimal quantizers with that of uniform quantizers.

4.2.1 Asymptotic Quantization Error of Optimal Quantizers

Fine quantizer arguments have been used to derive approximations for the mean square error of an optimal quantizer. The asymptotic expression for the variance of the quantization error assumes that the number of quantization levels is sufficiently large so that the density function is approximately constant throughout each of the bounded quantization regions and is equal to the value of the function at the midpoint of the interval. We note that this is essentially the first term in the Taylor series expansion of the density function about the midpoint of the interval and thus by using higher order terms in the Taylor series, more accurate approximations can be obtained. Another assumption made in deriving the asymptotic expression is that the probability of overload error is negligible. The main result that has been derived using these approximations is that the mean square error of an optimal quantizer is approximately given by the following equation[1]:

$$E\{\eta_X^2(X)\} = \frac{k_X \sigma_X^2}{n^2} + O(1/n^3), \tag{4.13}$$

where k_X is a constant dependent on the density function of the input:

$$k_X = \frac{1}{12\sigma_X^2}\left(\int_{\infty}^{\infty} p_X^{1/3}(x)dx\right)^3. \tag{4.14}$$

We shall henceforth refer to k_X as the fine quantizer coefficient for the random variable X. For large n, the terms of order $1/n^3$ and higher are negligible and the term in $1/n^2$ is a good approximation to the mean square error. For a Gaussian input the fine quantizer coefficient is approximately 2.72 while for a Laplacian input it is approximately 4.5. A consequence of the assumption that the density function is approximately constant in a region is that the representative point of the region

[1]The notation $x_n = O(f(n))$ means that there exist positive constants c and n_0 such that $|x_n| \leq c|f(n)|$ for all $n > n_0$.

approximately equals the midpoint of the interval. If we define $\bar{r}_i$ to be the midpoint of the quantization region $\mathcal{S}_i$, i.e.,

$$\bar{r}_i = (r_i + r_{i+1})/2 \qquad i = 1, \ldots, n-2, \tag{4.15}$$

then,

$$q_i \approx \bar{r}_i \qquad i = 1, \ldots, n-2. \tag{4.16}$$

We also denote the length of a quantization region $\mathcal{S}_i$ by δ_i, i.e.,

$$\delta_i = r_{i+1} - r_i. \tag{4.17}$$

Another important result of the fine quantizer analysis is that the length of a quantization region (for an optimal quantizer) is proportional to the inverse of the product of the number of quantization levels and the cube root of the density function[2]:

$$\delta_i \approx \frac{(12 k_X \sigma_X^2)^{1/3}}{n p_X^{1/3}(\bar{r}_i)} \qquad i = 1, \ldots, n-2. \tag{4.18}$$

In figure 4.2 the exact mean square error and the error predicted by the asymptotic approximation in (4.13) are plotted for a Gaussian input as a function of the number of quantization levels. We see that for values of n as small as two, the asymptotic approximation is of the same order of magnitude as the exact error. When the number of levels is greater than ten, the asymptotic formula is within twenty percent of the exact error. In figure 4.3 we plot the quantization error for a Laplacian input. As in the Gaussian case we see that the asymptotic formula is a good approximation for all values of n. We thus use the asymptotic formula to approximate the quantization error for inputs with different distributions even for a small number of quantization levels.

[2]We point out that this expression for the length of a quantization interval is obtained by retaining only the term in $1/n^2$ in the approximation for the mean square error. If we wish to calculate the terms of order $1/n^3$, the problem of determining the optimal regions becomes more difficult (see appendix A).

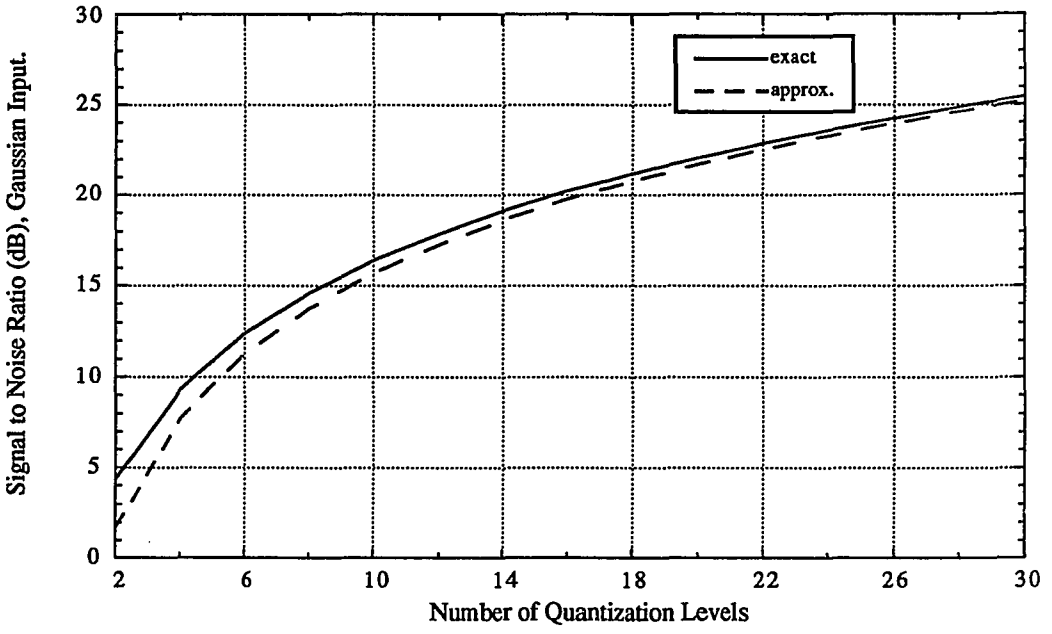

Figure 4.2: SNR of an Optimal Quantizer for a Gaussian Input vs. n.

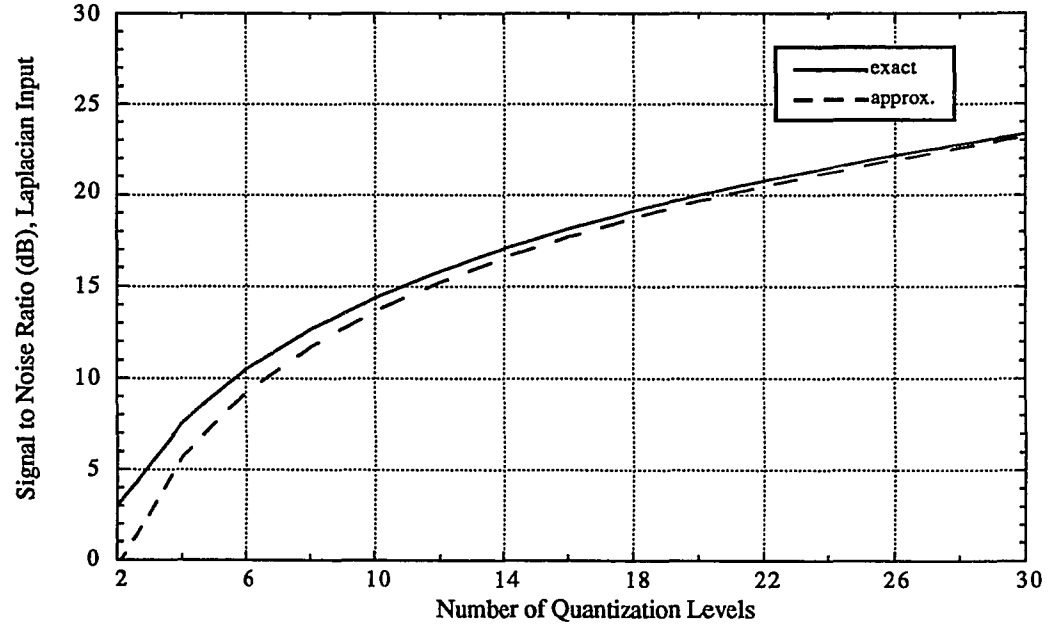

Figure 4.3: SNR of an Optimal Quantizer for a Laplacian Input vs. n.

Some other approximations follow from the properties of optimal quantizers. We state some useful ones below.

1. The correlation between the input and the quantization error can be approximated as follows:

$$E\{X\eta_X(X)\} \approx \frac{k_X\sigma_X^2}{n^2}. \tag{4.19}$$

2. The variance of the output of the quantizer is given by

$$E\{q_X^2(X)\} \approx \sigma_X^2(1 - k_X/n^2). \tag{4.20}$$

   This condition implies that the variance of the output of the quantizer, is smaller than the variance of the input random variable.

3. Given that the input is in the quantization region $\mathcal{S}_i$, the quantization error can be approximated as $p_X(\bar{r}_i)\delta_i^3/12$. Substituting for δ_i from (4.18), the conditional error can be expressed as $12k_X\sigma_X^2/(12n^3)$, which is independent of the region i. Thus we obtain the well known result that each quantization region contributes approximately an equal amount to the quantization error.

4.2.2 Asymptotic Quantization Error of Uniform Quantizers

A uniform scalar quantizer is a scalar quantizer whose quantization regions are all of the same length. Uniform quantizers are easier to implement than optimal quantizers whose quantization regions are in general of varying lengths. Also, unlike optimal quantizers the exact density function of the input need not be known to design a uniform quantizer. In this section we compare the performance of a uniform quantizer to that of an optimal one.

We consider a uniform quantizer with $n+2$ quantization regions, n of which are of a finite of length δ. These n regions extend from r_1 to r_{n+1}. We shall assume that the

density function of the input is approximately constant within each of these bounded quantization regions. The representative points are then approximately equal to the midpoint of the regions. The quantization error can therefore be approximated as

$$\begin{aligned}\xi_q &\approx \sum_{i=1}^{n} \int_{r_i}^{r_{i+1}} (x - \bar{r}_i)^2 p_X(\bar{r}_i)\, dx + \int_{-\infty}^{r_1} (x - q_0)^2 p_X(x)\, dx \\ &\quad + \int_{r_{n+1}}^{\infty} (x - q_{n+1})^2 p_X(x)\, dx. \end{aligned} \tag{4.21}$$

In the above equation, the terms in the summation form the granular error while the remaining two terms are the overload error. We denote the length of the interval from r_1 to r_{n+1} as α, i.e., $r_{n+1} - r_1 = \alpha$ and hence δ equals α/n. For large n, the summation can be approximated by an integral and the quantization error can be approximated as follows:

$$\xi_q \approx \frac{\alpha^2}{12n^2} \int_{r_1}^{r_{n+1}} p_X(x)\, dx + \int_{-\infty}^{r_1} (x - q_0)^2 p_X(x)\, dx + \int_{r_{n+1}}^{\infty} (x - q_{n+1})^2 p_X(x)\, dx. \tag{4.22}$$

The points r_1 and r_{n+1} must be chosen to minimize the quantization error. This choice depends upon the density function of the input.

We first consider the case when X is a zero mean, unit variance, Gaussian random variable. The representative points of the two unbounded intervals are approximately equal to $-\alpha/2$ and $\alpha/2$ respectively. The quantization error can then be expressed as

$$\xi_q \approx \frac{\alpha^2}{12n^2}(1 - 2\phi(\alpha/2)) + \left(-\alpha \frac{e^{-\alpha^2/8}}{\sqrt{2\pi}} + \left(2 + \frac{\alpha^2}{2}\right) \phi(\alpha/2) \right), \tag{4.23}$$

where $\phi()$ is given by

$$\phi(\alpha) = \int_{\alpha}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx. \tag{4.24}$$

This expression for the quantization error can be optimized with respect to α and the minimum quantization error that can be achieved with a uniform quantizer

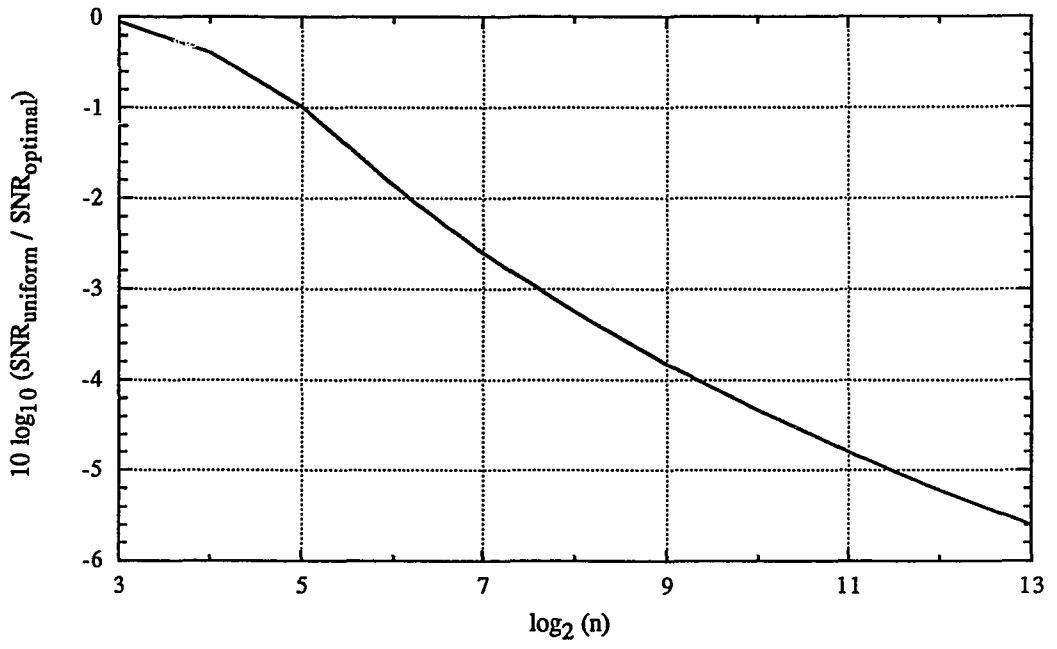

Figure 4.4: $10\log_{10}\left(\frac{SNR_{uniform}}{SNR_{optimum}}\right)$ vs. $\log_2(n)$ for a Gaussian Input.

can be found. We plot the difference between the quantization error of an optimal quantizer and a uniform quantizer in figure 4.4, where the optimum α was found numerically.

To obtain an analytical estimate, we first note that for large α, $\phi(\alpha)$ can be approximated closely as follows:

$$\phi(\alpha) \approx \frac{e^{-\alpha^2/2}}{2}. \tag{4.25}$$

By retaining only the significant terms in α, the quantization error can then be approximated by

$$\xi_q \approx \frac{\alpha^2}{12n^2} + \frac{\alpha^2 e^{-\alpha^2/8}}{4}. \tag{4.26}$$

The first term on the right hand side increases as α increases while the second term decreases for values of α greater than $\sqrt{8}$. Hence for large n, the optimum α at which the quantization error is minimum can be estimated by the point at which

the two terms are equal:

$$\alpha_{opt} \approx \sqrt{8 \ln(3n^2)}. \tag{4.27}$$

The quantization error of a uniform quantizer for a unit variance, Gaussian input can then be approximated as

$$\xi_q \approx \frac{4 \ln(3n^2)}{3n^2}. \tag{4.28}$$

This approximation states that the variance of quantization eror of an n level uniform quantizer is larger than that of an optimal quantizer by a factor of $0.5 \ln(3n^2)$ for a Gaussian input. A comparison with the exact error shows that this approximation is an upper bound.

In a similar fashion it can be shown that for a unit variance Laplacian input the quantization error is given by

$$\xi_q \approx \frac{\alpha^2}{12n^2}(1 - e^{-\alpha/\sqrt{2}}) + \frac{1}{2}e^{-\alpha/\sqrt{2}}. \tag{4.29}$$

This expression must be minimized with respect to α. We plot the difference in the SNR between an optimal quantizer and a uniform quantizer for a Laplacian input in figure 4.5.

We note that the performance of a uniform quantizer relative to an optimal quantizer is worse in the case of a Laplacian input than in the case of a Gaussian input. This is because the tail of a Laplacian distribution decreases more slowly than that of a Gaussian distribution. Hence α is larger in the former case and the width δ of each of the bounded quantization regions is larger than that of a uniform quantizer with the same number of levels for a Gaussian input.

4.2.3 Other Asymptotic Approximations

We now outline a method for calculating expectations of the form $E\{f(X, q_X(X))\}$, where $f()$ is some function of the input X and the output of an optimal quantizer.

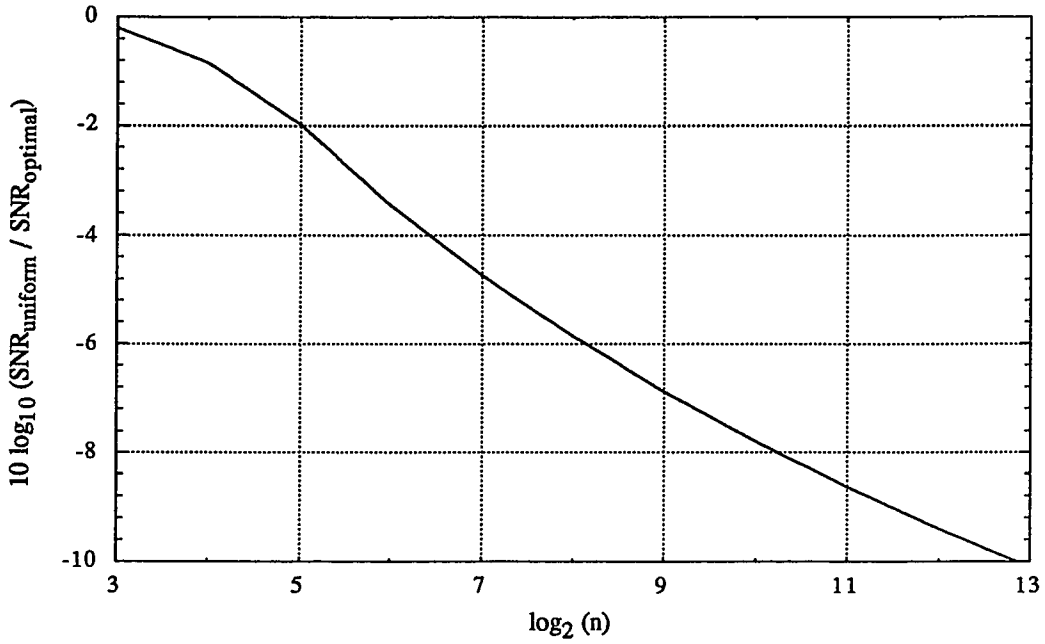

Figure 4.5: $10\log_{10}\left(\frac{SNR_{uniform}}{SNR_{optimum}}\right)$ vs. $\log_2(n)$ for a Laplacian Input.

The approach used will be similar to that used in calculating the fine quantizer approximation for the mean square error. However in order to calculate all the terms of order $1/n^2$, it is necessary to expand the density function as a Taylor series and use the higher order terms as necessary.

We first calculate a better approximation for the representative points. By definition we have,

$$q_i = \frac{\int_{r_i}^{r_{i+1}} x p_X(x)\,dx}{\int_{r_i}^{r_{i+1}} p_X(x)\,dx}.$$

By expanding the density function $p_X(x)$ in a Taylor series about the point $\bar{r}_i$ and assuming that $p_X(\bar{r}_i)$ is not equal to zero we get,

$$q_i = \bar{r}_i + \frac{\int_{r_i}^{r_{i+1}} (x-\bar{r}_i)(p_X(\bar{r}_i) + (x-\bar{r}_i)p'_X(\bar{r}_i) + \ldots)\,dx}{\int_{r_i}^{r_{i+1}} (p_X(\bar{r}_i) + (x-\bar{r}_i)p'_X(\bar{r}_i) + \ldots)\,dx} \tag{4.30}$$

$$= \bar{r}_i + \frac{p'_X(\bar{r}_i)}{12p_X(\bar{r}_i)}\delta_i^2 + O(\delta_i^4) \qquad i = 1,\ldots,n-1. \tag{4.31}$$

For our purposes it is sufficient to calculate all terms up to order δ^3. (Note that for calculating the approximation for the mean square error in (4.13), the first term in the above expression is sufficient).

We now return to the problem of calculating $E\{f(X, q_X(X))\}$. Since $f()$ is a function only of the random variable X, the expectation can be be expressed as follows:

$$E\{f(X, q_X(X))\} = \sum_{i=0}^{n-1} \int_{r_i}^{r_{i+1}} f(x, q_i)\, p_X(x)\, dx. \tag{4.32}$$

We expand the function $f(x, q_i)p_X(x)$, in the ith quantization interval, in a Taylor series in x about the midpoint $\bar{r}_i$ and denote the coefficient of the jth term of this Taylor series by $g_j(\bar{r}_i, q_i)$ to make explicit its dependence on $\bar{r}_i$ and the representative point q_i:

$$f(x, q_i)p_X(x) = \sum_{j=0}^{\infty} g_j(\bar{r}_i, q_i)(x - \bar{r}_i)^j, \quad x \in \mathcal{S}_i. \tag{4.33}$$

By substituting this Taylor series expansion into (4.32) and interchanging the order of summation we obtain

$$E\{f(X, q_X(X))\} = \sum_{j=0}^{\infty} \sum_{i=0}^{n-1} g_j(\bar{r}_i, q_i) \int_{r_i}^{r_{i+1}} (x - \bar{r}_i)^j\, dx. \tag{4.34}$$

The integrals in the above equation can be easily evaluated as follows:

$$\int_{r_i}^{r_{i+1}} (x - \bar{r}_i)^j\, dx = \begin{cases} 0 & \text{j odd} \\ \\ \frac{\delta_i^{j+1}}{2^j(j+1)} & \text{j even.} \end{cases} \tag{4.35}$$

By substituting the above results into (4.34), we obtain the expression given below:

$$E\{f(X, q_X(X))\} = \sum_{j\,even} \sum_{i=0}^{n-1} \left(\frac{g_j(\bar{r}_i, q_i)\delta_i^j}{2^j(j+1)} \right) \delta_i. \tag{4.36}$$

We note that the representative point of a region q_i and the length of a quantization interval δ_i, are functions of the midpoint $\bar{r}_i$ of the interval and the number of quantization regions n (see equations (4.31) and (4.18)). Hence the term within the

parenthesis in the above equation can be expressed as a function of only $\bar{r}_i$ and n. We call this function $h_j(\bar{r}_i, n)$:

$$h_j(\bar{r}_i, n) = \frac{1}{2^j(j+1)} g_j \left(\bar{r}_i, \bar{r}_i + \frac{p'_X(\bar{r}_i)(12k_X\sigma_X^2)^{2/3}}{12n^2 p_X^{5/3}(\bar{r}_i)} \right) \left(\frac{(12k_X\sigma_X^2)^{1/3}}{np_X^{1/3}(\bar{r}_i)} \right)^j. \quad (4.37)$$

The expectation $E\{f(X, q_X(X))\}$ can then be approximated as

$$E\{f(X, q_X(X))\} \approx \sum_{jeven} \sum_{i=0}^{n-1} h_j(\bar{r}_i, n)\delta_i. \quad (4.38)$$

Since the quantizer is fine, the summation over the index i can be approximated by an integral and we obtain the following expression:

$$E\{f(X, q_X(X))\} \approx \sum_{jeven} \int_{-\infty}^{\infty} h_j(x, n)\, dx. \quad (4.39)$$

This formula expresses the required expectation as a function of the number of quantization regions n. We note that the above formula is accurate only in the term in $1/n^2$ since the expression for δ_i in (4.18) is obtained by minimizing only the term in $1/n^2$ of the quantization error (see appendix A for a further discussion regarding this point). In the rest of this dissertation, we shall refer to the above sequence of arguments to approximate an expectation of a function of the output of a fine quantizer as the fine quantizer approximation.

We illustrate the above method by calculating the expectation $E\{X^3\eta_X(X)\}$. By definition we have

$$E\{X^3\eta_X(X)\} = \sum_{i=0}^{n-1} \int_{r_i}^{r_{i+1}} x^3(x - q_i)p_X(x)\, dx. \quad (4.40)$$

The function within the integral can be expanded quite easily in a Taylor series about $\bar{r}_i$. We evaluate the resulting integrals, substitute the approximation for the representative point q_i from (4.31) and arrange the terms in increasing powers of δ_i. We then obtain the following equation:

$$E\{X^3\eta_X(X)\} \approx \sum_{i=0}^{n-1} \frac{\bar{r}_i^2 \delta_i^2 p_X(\bar{r}_i)}{4} \delta_i + O(\delta_i^4). \quad (4.41)$$

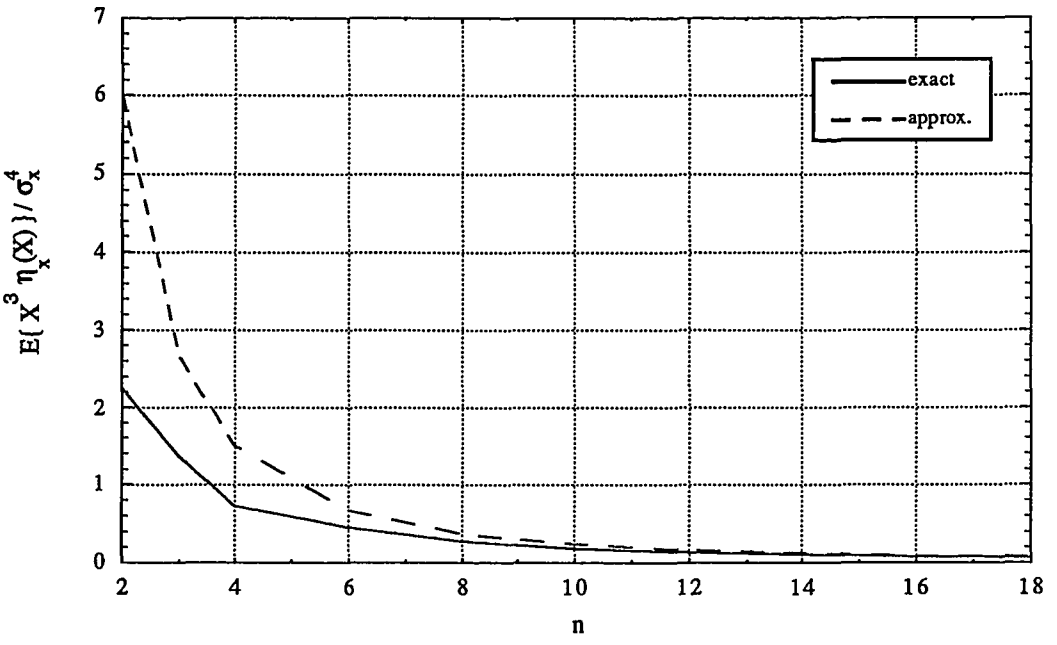

Figure 4.6: $E\{X^3\eta_X(X)\}/\sigma_X^4$ vs. n.

By substituting the approximation for δ_i from (4.18), the above equation can be rewritten as follows:

$$E\{X^3\eta_X(X)\} \approx \sum_{i=0}^{n-1} \left(\frac{\bar{r}_i^2 (12 k_X \sigma_X^2)^{2/3} p_X^{1/3}(\bar{r}_i)}{4n^2} \right) \delta_i + O(\delta_i^4). \tag{4.42}$$

This expression can then be approximated by an integral:

$$E\{X^3\eta_X(X)\} \approx \frac{(12 k_X \sigma_X^2)^{2/3}}{4n^2} \int_{-\infty}^{\infty} x^2 p_X^{1/3}(x)\, dx + O(1/n^4). \tag{4.43}$$

In the case when X is a zero-mean, Gaussian random variable with variance σ_X^2, the integral can be easily evaluated and we get the approximation,

$$E\{X^3\eta_X(X)\} \approx \frac{24}{n^2}\sigma_X^4 + O(1/n^4). \tag{4.44}$$

In figure 4.6 we plot the above asymptotic approximation and the exact value of the expectation. We note that the approximation converges quite rapidly to the exact value.

We state similar approximations for two other expectations that we will need in the next section (here X is a zero mean Gaussian random variable):

$$E\{X^2\eta_X^2(X)\} \approx \frac{3 \times 2.7\sigma_X^4}{n^2} \tag{4.45}$$

$$E\{X^4\eta_X^2(X)\} \approx \frac{27 \times 2.7\sigma_X^6}{n^2}. \tag{4.46}$$

4.3 Fine Quantizer Approximations For Mismatched Quantizers

In this section we consider the problem when the quantizer is not optimized for the input. We examine the case when the mean and the variance of the input differ from the values for which the quantizer was designed. Thus, if the quantizer is designed for a zero-mean random variable X with standard deviation σ_X, we consider the case when the input to the quantizer is the random variable $(1+\epsilon)X+\epsilon'$. The mean of this random variable is equal to ϵ' and the standard deviation is $(1+\epsilon)\sigma_X$. We will show that it is possible to to derive a Taylor series expansion in ϵ and ϵ' for various expectations (like the mean and variance of the quantization error) and that the coefficients of ϵ and ϵ' in this series expansion can be approximated by using fine quantizer approximations.

We first calculate an approximation for the mean of the quantization error. By defining the random variable Y as

$$Y = (1+\epsilon)X + \epsilon', \tag{4.47}$$

we can express the expected value of the quantization error as follows:

$$E\{\eta_X((1+\epsilon)X+\epsilon')\} = \int_{-\infty}^{\infty} \eta_X(y)p_Y(y)\,dy. \tag{4.48}$$

The density function of Y can now be expanded in a Taylor series in the variables

ϵ and ϵ' as shown below:

$$p_Y(y) = \frac{1}{1+\epsilon} p_X\left(\frac{y-\epsilon'}{1+\epsilon}\right)$$

$$\begin{aligned} &\approx p_X(y) - \epsilon(p_X(y) + y p'_X(y)) - \epsilon' p'_X(y) \\ &\quad + \epsilon^2(p_X(y) + 2y p'_X(y) + \frac{y^2}{2} p''_X(y)) \\ &\quad + \frac{\epsilon'^2}{2} p''_X(y) + \epsilon\epsilon'(2p'_X(y) + y p''_X(y)) + \ldots \end{aligned} \tag{4.49}$$

By substituting this expansion into (4.48), we obtain the following expression for the mean of the quantization error:

$$\begin{aligned} E\{\eta_X(Y)\} \approx\ & -\epsilon E\left\{X\eta_X(X)\frac{p'_X(X)}{p_X(X)}\right\} - \epsilon' E\left\{\eta_X(X)\frac{p'_X(X)}{p_X(X)}\right\} \\ & + \epsilon^2 E\left\{\eta_X(X)\left(\frac{2Xp'_X(X)}{p_X(X)} + \frac{X^2 p''_X(X)}{2p_X(X)}\right)\right\} \\ & + \frac{\epsilon'^2}{2} E\left\{\eta_X(X)\frac{p''_X(X)}{p_X(X)}\right\} \\ & + \epsilon\epsilon' E\left\{\eta_X(X)\left(\frac{2p'_X(X)}{p_X(X)} + \frac{Xp''_X(X)}{p_X(X)}\right)\right\}; \end{aligned} \tag{4.50}$$

we have used the result that the mean of the quantization error of an optimal quantizer is zero. The expectations in the above equation can be evaluated using the fine quantizer approximation and thus an approximation for the mean of the quantization error can be derived for any input if its density function is known.

We next consider the case when X is a zero-mean, Gaussian random variable with variance σ_X^2. We note that in this case the density function is even and that $\eta_X(x)$ is an odd function of x. Further the following equations hold:

$$\frac{p'_X(X)}{p_X(X)} = -\frac{X}{\sigma_X^2} \quad \text{and} \quad \frac{p''_X(X)}{p_X(X)} = \left(-1 + \frac{X^2}{\sigma_X^2}\right)\frac{1}{\sigma_X^2}. \tag{4.51}$$

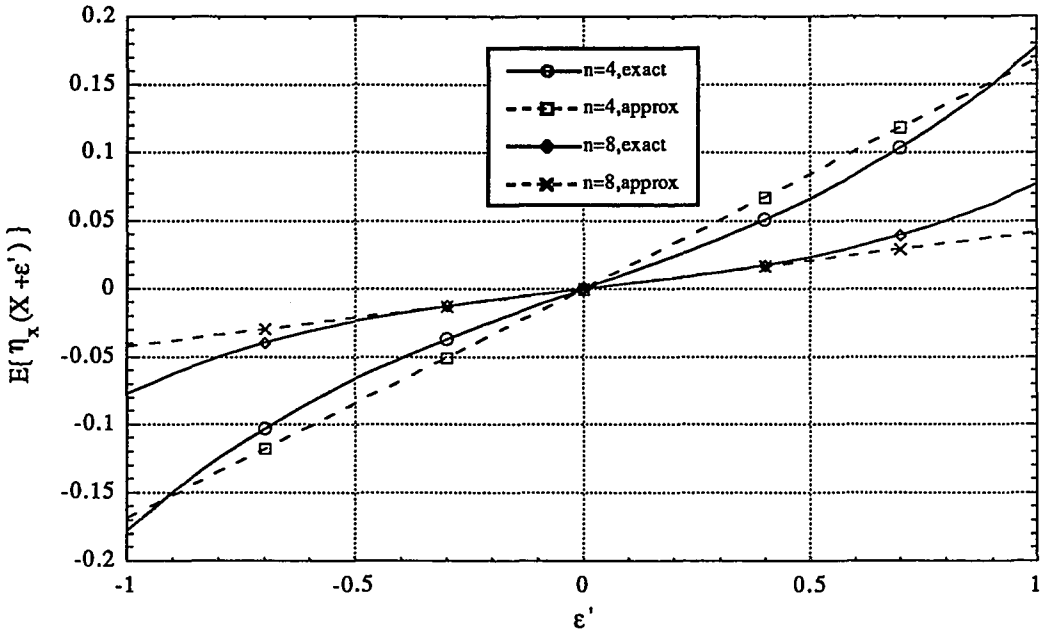

Figure 4.7: $E\{\eta_X(X+\epsilon')\}$ vs. ϵ'.

By substituting the above relations into (4.50) and using the fact that the expectation of an odd function of X is zero, we simplify (4.50) to the following expression:

$$E\{\eta_X((1+\epsilon)X+\epsilon')\} \approx \frac{\epsilon'}{\sigma_X^2}E\{X\eta_X(X)\} + \epsilon\epsilon' E\left\{\left(\frac{X^3}{\sigma_X^4} - \frac{3X}{\sigma_X^2}\right)\eta_X(X)\right\}. \quad (4.52)$$

By using the fine quantizer approximations derived in (4.19) and (4.44), we obtain the following approximation in the case when X is a zero-mean, Gaussian random variable:

$$E\{\eta_X((1+\epsilon)X+\epsilon')\} \approx \epsilon'\left(\frac{2.7+16\epsilon}{n^2}\right). \quad (4.53)$$

We note that the above expression is independent of the variance of X and linear in ϵ' with slope $(2.7+16\epsilon)/n^2$. In figure 4.7 we plot the approximation in (4.53) for the case when $\epsilon = 0$ as a function of ϵ' along with the exact curves which were computed numerically. These curves have been plotted for a 4-level and an 8-level quantizer. We note that the approximation is quite close to the exact curve for

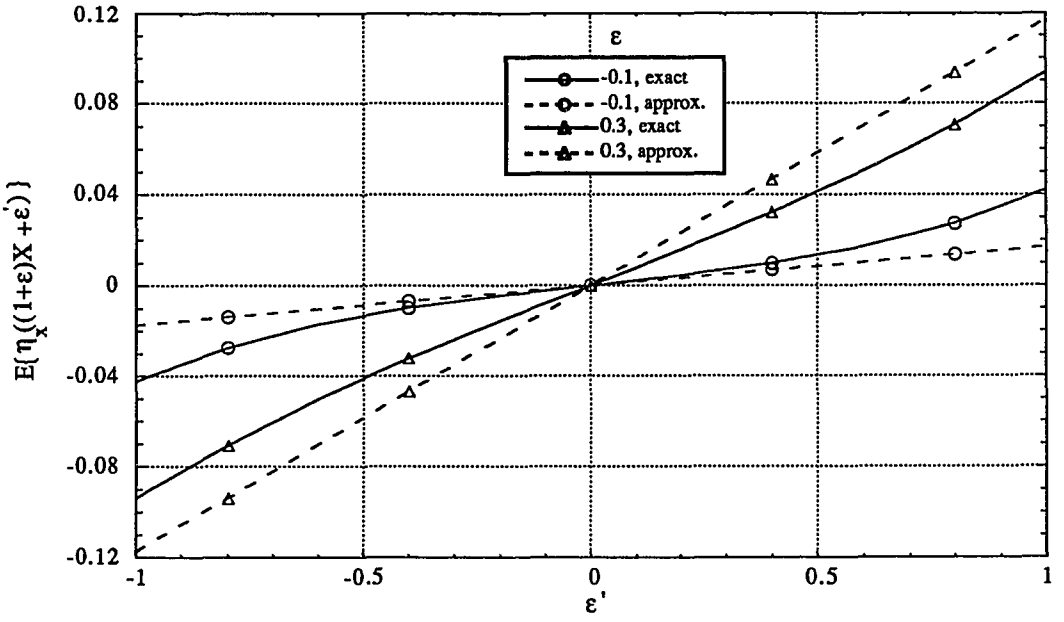

Figure 4.8: $E\{\eta_X((1+\epsilon)X+\epsilon')\}$ vs. ϵ', $n=8$.

values of epsilon as large as 1. When ϵ is not zero (see figure 4.8), the range over which the linear approximation is accurate, decreases. For negative values of ϵ, the variance of the input becomes small and for fixed n, the fine quantizer assumption that the density function is smooth in each quantization region breaks down. In the other case, as ϵ increases, the variance of the input increases and for a fixed n, the assumption that the overload error is negligible breaks down. When ϵ' is zero (i.e., the mean is matched), the mean of the quantization error is zero for all values of ϵ. This is obvious from the symmetry of the input.

In the general case when the input X is not Gaussian, expressions can be derived for the expectations in (4.50). Since the algebra is involved, we do not state these results here but we observe for future reference that the most significant term in n for the expected value of the quantization error is of order $1/n^2$ and can be expressed in the form $k(\epsilon,\epsilon')/n^2$, where $k()$ is some function of ϵ and ϵ'.

Another quantity of interest is the correlation between the input and the quantization error. This correlation will also be useful in analyzing different quantization systems in the next chapter. An approximation for this correlation can be derived in a manner similar to that used to approximate the mean of the quantization error. If we define the random variable Y to be equal to $(1+\epsilon)X+\epsilon'$, then the Taylor series expansion in ϵ and ϵ' for this correlation is as follows:

$$\begin{aligned} E\{Y\eta_X(Y)\} \approx\ & E\{X\eta_X(X)\} - \epsilon E\left\{X\eta_X(X)\left(1+X\frac{p'_X(X)}{p_X(X)}\right)\right\} \\ & - \epsilon' E\left\{X\eta_X(X)\frac{p'_X(X)}{p_X(X)}\right\} \\ & + \epsilon^2 E\left\{X\eta_X(X)\left(1+\frac{2Xp'_X(X)}{p_X(X)}+\frac{X^2p''_X(X)}{2p_X(X)}\right)\right\} \\ & + \epsilon'^2 E\left\{X\eta_X(X)\frac{p''_X(X)}{2p_X(X)}\right\} \\ & + \epsilon\epsilon' E\left\{X\eta_X(X)\left(\frac{2p'_X(X)}{p_X(X)}+\frac{Xp''_X(X)}{p_X(X)}\right)\right\}. \end{aligned} \tag{4.54}$$

Fine quantizer approximations can be used to evaluate the expectations above. When X is a zero-mean, Gaussian random variable, the above equation reduces to the following expression:

$$E\{Y\eta_X(Y)\} \approx \frac{2.7}{n^2}\sigma_X^2 + \epsilon\frac{21\sigma_X^2}{n^2} + \epsilon'^2\frac{11}{n^2}. \tag{4.55}$$

The correlation is linear in ϵ and quadratic in ϵ'. Hence it is more sensitive to small mismatches in the standard deviation than to a small mismatch in the mean.

We will need to evaluate correlations of the form $E\{X\eta_X(X+\epsilon')\}$ later in this dissertation. By using (4.53) and (4.55), we obtain the following relation when X

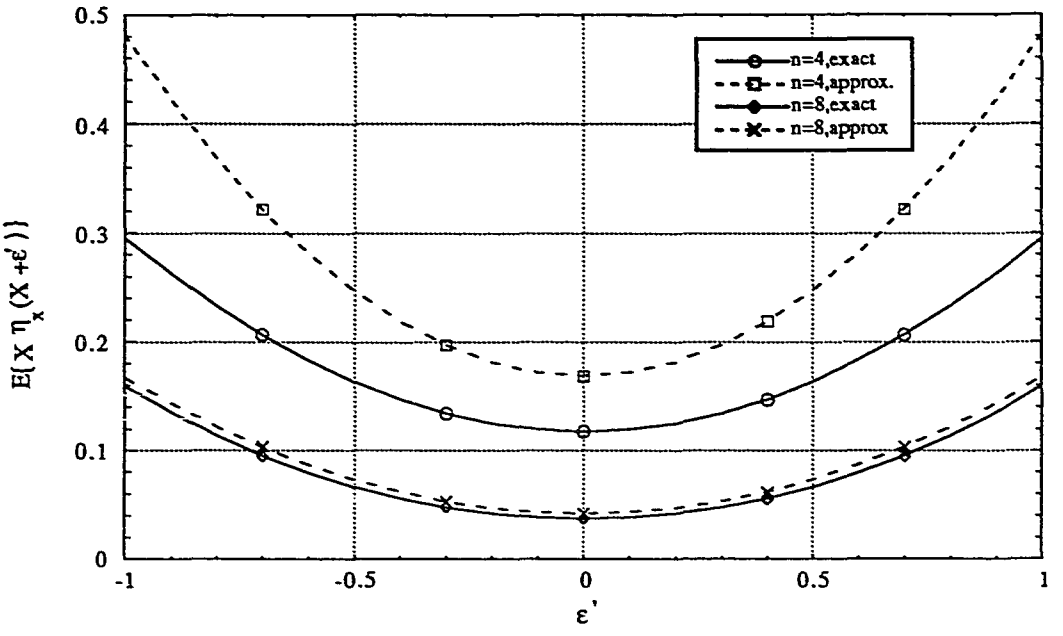

Figure 4.9: $E\{X\eta_X(X+\epsilon')\}$ vs. ϵ'.

is a zero mean Gaussian random variable:

$$E\{X\eta_X(X+\epsilon')\} \approx \frac{2.7\sigma_X^2}{n^2} + \epsilon'^2\frac{8}{n^2}. \tag{4.56}$$

We plot this approximation and the exact curve in figure 4.9. We see that the approximation is valid for values of ϵ' as large as 1, even when the number of quantization regions is as small as 4.

Finally we consider the variance of the quantization error of a mismatched quantizer. It can be shown that this variance can be approximated as follows:

$$E\{\eta_X^2((1+\epsilon)X+\epsilon')\} \approx E\{\eta_X^2(X)\} - \epsilon E\left\{\eta_X^2(X)\left(1+X\frac{p_X'(X)}{p_X(X)}\right)\right\}$$

$$-\epsilon' E\left\{\eta_X^2(X)\frac{p_X'(X)}{p_X(X)}\right\}$$

$$+\epsilon^2 E\left\{\eta_X^2(X)\left(1+\frac{2Xp_X'(X)}{p_X(X)}+\frac{X^2p_X''(X)}{2p_X(X)}\right)\right\}$$

$$+\frac{\epsilon'^2}{2}E\left\{\eta_X^2(X)\frac{p_X''(X)}{p_X(X)}\right\}$$

$$+\epsilon\epsilon' E\left\{\eta_X^2(X)\left(\frac{2p_X'(X)}{p_X(X)}+\frac{Xp_X''(X)}{p_X(X)}\right\}\right). \qquad (4.57)$$

In the case when X is a zero-mean, Gaussian random variable, the expectations in the above expression can be evaluated using fine quantizer approximations and we obtain the approximation

$$E\{\eta_X^2((1+\epsilon)X+\epsilon')\} \approx \frac{2.7\sigma_X^2}{n^2}\left(1+2\epsilon+7\epsilon^2+\epsilon'^2/\sigma_X^2\right). \qquad (4.58)$$

From the above equation we see that the variance of the quantization error is quadratic in ϵ' and hence any mismatch in the mean leads to its increase. On the other hand for sufficiently small ϵ, it is approximately linear in the mismatch of the variance and hence the noise power decreases if the variance of the input decreases which is to be expected. We plot these approximations in figures 4.10 and 4.11 and note that the approximations are quite reasonable even when the number of quantization levels is as small as 4.

We next use the above approximation to examine the signal to noise ratio of a mismatched quantizer. The variance of the input Y is equal to $(1+\epsilon)^2$ and using the above approximation for the variance of the quantization noise, we obtain the following expression for the SNR of a mismatched quantizer:

$$\begin{aligned} 10\log_{10}\left(\frac{\sigma_Y^2}{E\{\eta_X^2(Y)\}}\right) &= 10\log_{10}\left(\frac{1}{2.7/n^2}\right)+10\log_{10}\left(\frac{1}{1+\left(\frac{\epsilon'}{\sigma_X}\right)^2}\right) \\ &\quad +10\log_{10}\left(1-10\epsilon^2\right). \end{aligned} \qquad (4.59)$$

The first term in the right hand side of the above equation is the SNR of a matched

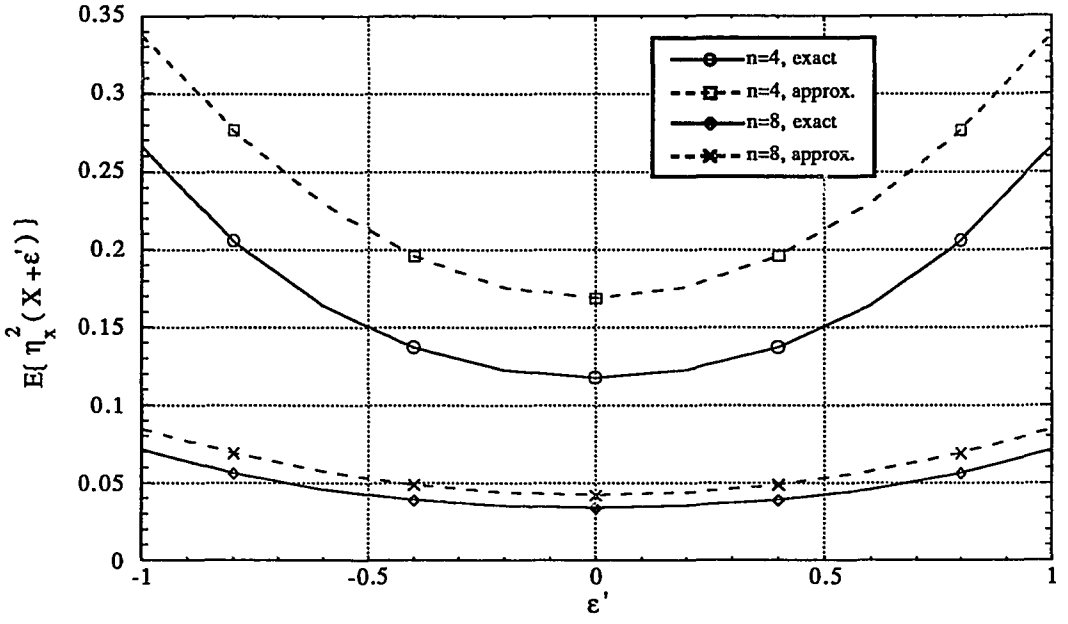

Figure 4.10: $E\{\eta_X^2(X+\epsilon')\}$ vs. ϵ'.

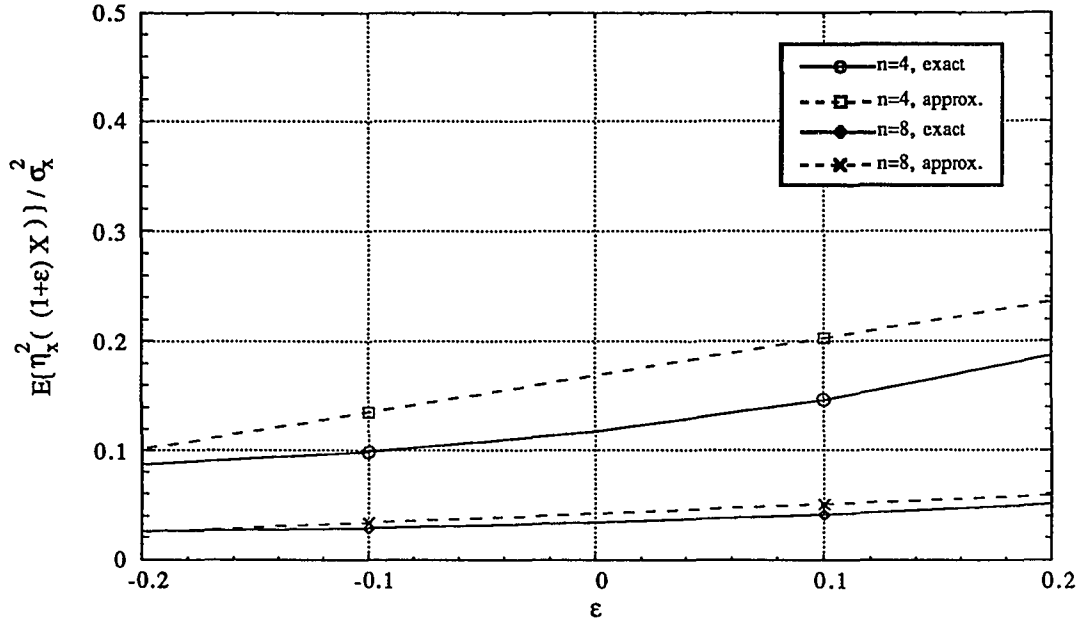

Figure 4.11: $E\{\eta_X^2((1+\epsilon)X)\}/\sigma_X^2$ vs. ϵ.

quantizer (in dB). The second term is the error due to the mismatch in the mean from which we see that it is the mismatch in the mean relative to the variance of the input that determines the decrease in the SNR. The third term is the error due to the mismatch in the variance. We note that any mismatch in the mean or variance causes the signal to noise ratio to decrease. When the variance is matched, i.e., $\epsilon = 0$, we see that for $\epsilon' = 0.5\sigma_X$, the SNR decreases by 1 dB and for $\epsilon' = \sigma_X$, the SNR decreases by 3 dB. On the other hand when the mean is matched, i.e., $\epsilon' = 0$, we see that for $\epsilon = 0.1$ (which corresponds to a 10 percent error in estimating the standard deviation of the input or a 21 percent error in estimating the variance), the SNR decreases by 0.46 dB while for $\epsilon = 0.2$ (which corresponds to a 20 percent error in estimating the standard deviation or a 44 percent error in estimating the variance), the SNR decreases by 2.2 dB.

4.4 Fine Quantizer Approximations For Two Optimal Quantizers

We next consider the problem where two different random variables X and Y are quantized by their optimal quantizers, $q_X()$ and $q_Y()$ respectively. We shall use the results derived here to analyze various quantization schemes in the next chapter. We calculate the correlation between the input of the first quantizer, say X, and the quantization error of the second quantizer, $\eta_Y(Y)$. We also calculate the correlation between the quantization errors of the two systems. The difficulty that arises when working with two different random variables is that although the marginal densities of the random variables may be smooth (and thus the quantizers $q_X()$ and $q_Y()$ are individually fine), the joint density of the two random variables may not be smooth. Hence when the quantizers are viewed jointly as a two dimensional quantizer, the fine quantizer assumptions may not be valid. Our approach is to express one of

the random variables as a linear combination of the other and an uncorrelated random variable. Then we assume that the joint density of these new uncorrelated random variables is sufficiently smooth and hence fine quantizer approximations can be used. Note that although two random variables are uncorrelated, their joint density function need not be smooth as we assume[3]. Thus there are situations where the results of this section will not hold.

4.4.1 An Approximation for $E\{\eta_X(X)Y\}$

In this section we calculate the correlation $E\{\eta_X(X)Y\}$. The simplest case in which this correlation can be calculated is when the two random variables are independent. Since the mean of the quantization error of an optimal quantizer is zero, the correlation in this case is zero. In the general case when the random variables are not independent, we assume that X and Y are zero-mean, unit-variance, random variables. It is then always possible to express Y as a sum of X and an uncorrelated random variable U as follows:

$$Y = \gamma X + \sqrt{1-\gamma^2}U, \tag{4.60}$$

where γ is the correlation coefficient between X and Y, i.e.,

$$\gamma = E\{XY\} \tag{4.61}$$

and

$$E\{XU\} = 0. \tag{4.62}$$

The correlation between Y and the quantization error in X can then be expressed as follows:

$$E\{\eta_X(X)Y\} = \gamma E\{\eta_X(X)X\} + \sqrt{1-\gamma^2}E\{\eta_X(X)U\}. \tag{4.63}$$

[3]Let X be a zero-mean, unit-variance Gaussian random variable and let Y equal X^2. Then X and Y are uncorrelated but their joint density is impulsive and definitely not smooth.

The first expectation on the right side of the above equation can be approximated by using the fine quantizer approximation in (4.19). In order to evaluate the second expectation we assume that since X and U are uncorrelated, the conditional density of X given U is smooth. Hence it can be approximated by the first few terms of its Taylor series expansion in x about the point $\bar{r}_i$ within each of the bounded quantization intervals $\mathcal{S}_i$. With this assumption, the expectation in the second term can be evaluated as follows:

$$E\{\eta_X(X)U\} \approx \int_{-\infty}^{\infty} u p_U(u) \sum_{i=0}^{n-1} \int_{r_i}^{r_{i+1}} \left(x - \bar{r}_i - \frac{p'_X(\bar{r}_i)\delta_i^2}{12 p_X(\bar{r}_i)} \right)$$

$$\left(p_{X/U}(\bar{r}_i/u) + (x - \bar{r}_i) p^x_{X/U}(\bar{r}_i/u) \right) dx\, du, \qquad (4.64)$$

where we have used the approximation in (4.31) for the representative point q_i. By integrating with respect to x and applying the usual fine quantizer approximations, the above expression can be approximated as follows:

$$E\{\eta_X(X)U\} \approx \frac{(12k_X)^{2/3}}{n_X^2} E\left\{ \frac{U\, p^x_{U/X}(U/X)}{p_X^{2/3}(X)\, p_{U/X}(U/X)} \right\}. \qquad (4.65)$$

We note that if X and U are independent, then the above term is equal to zero as expected. By using (4.63), we can now obtain an approximation for the correlation between the random variable Y and the quantization error of X when X and Y are unit variance random variables:

$$E\{\eta_X(X)Y\} \approx \gamma \frac{k_X}{n_X^2} + \frac{(12k_X)^{2/3}}{n_X^2} E\left\{ \frac{U\, p^x_{U/X}(U/X)}{p_X^{2/3}(X)\, p_{U/X}(U/X)} \right\}, \qquad (4.66)$$

where U is defined in (4.60).

If X and Y are not unit variance random variables but still have zero mean, then by defining the unit variance random variables R and S as

$$R = X/\sigma_X \quad \text{and} \quad S = Y/\sigma_Y, \qquad (4.67)$$

we can express the the correlation between Y and the quantization error in X as follows:

$$E\{Y\eta_X(X)\} = \sigma_X\sigma_Y E\{\eta_R(R)S\}. \tag{4.68}$$

The expectation on the right hand side of the above equation can be evaluated using (4.66).

From the above results we see that if X and Y are jointly Gaussian, then

$$E\{Y\eta_X(X)\} \approx \sigma_X\sigma_Y \frac{k_X}{n_X^2}\gamma. \tag{4.69}$$

Hence the correlation between the random variable Y and the error in quantizing X is linear in the correlation between X and Y. Finally we note that if $Y = \eta_Z(Z)$, i.e., Y equals the error in quantizing the random variable Z, then the expression in (4.66) cannot be calculated unless the joint density of X and $\eta_Z(Z)$ is known. Such joint densities are hard to find in general and in this case we would like to find an expression for the correlation between the quantization errors in terms of the joint density of the inputs. We address this problem in the next section.

4.4.2 Approximations for $E\{\eta_X(X)\eta_Y(Y)\}$

In this section we calculate the correlation between the quantization errors $\eta_X(X)$ and $\eta_Y(Y)$. A bound on the magnitude of this correlation is obtained quite simply by observing that

$$|E\{\eta_X(X)\eta_Y(Y)\}| \leq \sqrt{E\{\eta_X^2(X)\eta_Y^2(Y)\}}. \tag{4.70}$$

Using the fine quantizer approximation for the mean square error, we obtain the following inequality:

$$|E\{\eta_X(X)\eta_Y(Y)\}| \leq \frac{\sqrt{k_X k_Y \sigma_X^2 \sigma_Y^2}}{n_X n_Y}, \tag{4.71}$$

where n_X and n_Y are the number of quantization levels in quantizers $q_X()$ and $q_Y()$ and k_X and k_Y are the fine quantizer coefficients for X and Y respectively. Thus when both quantizers have an equal number of quantization levels, the maximum value that the correlation between the quantization errors can take is of the order $1/n^2$.

The simplest case when the correlation can be calculated is when X and Y are independent. Then,

$$E\{\eta_X(X)\eta_Y(Y)\} = E\{\eta_X(X)\}E\{\eta_Y(Y)\} \tag{4.72}$$

and since the mean of the quantization error of an optimal quantizer is zero, the quantization errors are uncorrelated.

We next consider the general case when X and Y are not independent. Our approach is to consider two different cases; one is when the correlation between the inputs is very small and the other is when the inputs are very highly correlated. We use these asymptotes to approximate the correlation between the quantization errors as a function of the correlation between the inputs. For convenience we shall assume that the inputs have zero mean and unit variance and express Y as a linear combination of X and an uncorrelated random variable as before (see (4.60)).

Weakly Correlated Inputs

We first examine the case when γ, the correlation between the random variables X and Y, is small. By conditioning on X, the correlation between the quantization errors can be expressed as follows:

$$E\{\eta_X(X)\eta_Y(Y)\} = E\left\{\eta_X(X)f(X)\right\}, \tag{4.73}$$

where

$$f(x) = E\left\{\eta_Y\left(\sqrt{1-\gamma^2}U + \gamma x\right)/X = x\right\}. \tag{4.74}$$

We next make the assumption that the optimum quantizer for the random variables U and Y are the same; i.e., $q_U() \approx q_Y()$. We note from (4.60) that this is true when γ equals zero and hence when the correlation between the inputs X and Y is small, this is a reasonable assumption to make. With this approximation, $f(x)$ is the expected value of the quantization error of a mismatched quantizer. We assume that the conditional density $p_{U/X}(u/x)$ is relatively smooth and that the number of quantization regions n_Y is sufficiently large so that the fine quantizer approximations can be used to evaluate this conditional expectation.

The conditional expectation can then be expressed as a function of γ, X and n_Y, where it follows from the discussion on mismatched quantizers that the most significant term in n_Y is of the form $k(\gamma, x)/n_Y^2$. The exact form of $k(\gamma, x)$ will depend upon the conditional density function $p_{U/X}(u/x)$ but can be calculated from (4.50). Thus we find that the correlation between the quantization errors of two weakly correlated random variables is

$$E\{\eta_X(X)\eta_Y(Y)\} \approx \frac{1}{n_Y^2} E\{\eta_X(X)k(\gamma, X)\}. \tag{4.75}$$

This remaining expectation can be evaluated by again using the fine quantizer approximations. In the case when X and Y are jointly Gaussian, the conditional density $p_{U/X}(u/x)$ is Gaussian and does not depend upon X. The result of (4.53) is thus directly applicable here and we obtain the following expression for $k(\gamma, X)$:

$$k(\gamma, X) = 2.7\gamma X - 8\gamma^3 X. \tag{4.76}$$

By retaining only the terms in γ, one can easily show that the correlation between the quantization errors is approximately

$$E\{\eta_X(X)\eta_Y(Y)\} \approx \frac{7.2\gamma}{n_X^2 n_Y^2}. \tag{4.77}$$

This expression is true when X and Y are unit variance, weakly correlated, Gaussian random variables. When X and Y are not unit variance random variables, but are still weakly correlated, the correlation between their quantization errors is

$$E\{\eta_X(X)\eta_Y(Y)\} \approx \frac{7.2\gamma\sigma_X\sigma_Y}{n_X^2 n_Y^2}. \tag{4.78}$$

Thus in the case when $q_X()$ and $q_Y()$ are identical quantizers and the inputs are jointly Gaussian and not highly correlated, the correlation between the quantization errors is of the order of $1/n^4$.

Strongly Correlated Inputs

We next investigate the case when X and Y are strongly correlated, i.e.,

$$\gamma = 1 - \epsilon \tag{4.79}$$

where ϵ is small. For convenience, we assume that X and Y are unit variance random variables. We shall restrict ourselves to the case when the marginal density of X and Y are identical; this implies that the optimal quantizers $q_X()$ and $q_Y()$ are also identical, i.e., $q_X() = q_Y() = q()$. Although the quantizers are individually fine, the random variables X and Y are now highly correlated and hence when $q_X()$ and $q_Y()$ are viewed jointly as a two dimensional quantizer, fine quantizer approximations are no longer valid for the vector input (X, Y).

We begin by examining the expectation $E\{(\gamma\eta_X(X) - \eta_Y(Y))^2\}$. We note that the correlation between the quantization errors can be expressed quite simply in terms of this expectation:

$$E\{\eta_Y(Y)\eta_X(X)\} = \frac{1}{2\gamma}\left(\gamma^2 E\{\eta_X^2(X)\} + E\{\eta_Y^2(Y)\} - E\left\{(\gamma\eta_X(X) - \eta_Y(Y))^2\right\}\right). \tag{4.80}$$

It turns out that the algebra in evaluating this expectation is less involved than the algebra in calculating the correlation between the quantization errors directly. We first express Y as a linear combination of X and an uncorrelated random variable U as in (4.60). We then have the relation

$$E\left\{(\gamma\eta_X(X)-\eta_Y(Y))^2\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n}\int_{r_i}^{r_{i+1}}\int_{l_j}^{l_{j+1}}\left(\sqrt{1-\gamma^2}u-q_j+\gamma q_i\right)^2$$

$$\times p_{X,U}(x,u)\,du\,dx, \qquad (4.81)$$

where

$$l_j = \frac{r_j-\gamma x}{\sqrt{1-\gamma^2}}. \qquad (4.82)$$

In order to evaluate the integrals we note that when the number of quantization levels is fixed and the correlation between the inputs is sufficiently large (i.e., ϵ is sufficiently small), most of the probability mass of the random variable U will lie within the intervals $(-\delta_i/\sqrt{2\epsilon}, \delta_i/\sqrt{2\epsilon})$. This assumption implies that if the input X lies in the quantization region $\mathcal{S}_i$, then the input Y lies in either the same quantization region or in one of the two adjacent quantization regions $\mathcal{S}_{i-1}$ or $\mathcal{S}_{i+1}$. This is to be expected when X and Y are strongly correlated. Then for each i, we can approximate the integral by integrating from $-\delta_i/\sqrt{2\epsilon}$ to $\delta_i/\sqrt{2\epsilon}$ along the U-axis. We next make the assumption that the number of quantization levels is large and the marginal density of the inputs is sufficiently smooth so that it is approximately constant within each of the quantization intervals. Hence the representative points q_i are approximately equal to the midpoints $\bar{r}_i$ of the regions[4]. We also make the assumption that the conditional density of X given U is sufficiently smooth, so that it is also approximately constant in the intervals (r_i, r_{i+1}). The integration can then

[4]We shall retain only terms of order $1/n$ in the final expression and hence this is a sufficient approximation for the representative points and the more exact approximation of (4.31) is not needed.

be conveniently carried out by interchanging the order of integration and breaking up the region of integration between into five disjoint parts as shown in figure 4.12. By retaining only the terms in $\sqrt{\epsilon}$ and ϵ, we obtain the following approximation:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{r_i}^{r_{i+1}}\int_{l_j}^{l_{j+1}}\left(\sqrt{1-\gamma^2}u-q_j+\gamma q_i\right)^2 p_{X,U}(x,u)\,du\,dx \approx$$

$$\sum_{i=1}^{n}\int_{\sqrt{\frac{\epsilon}{2}}r_{i+1}}^{\frac{\delta_i}{\sqrt{2\epsilon}}}\left(\int_{r_i}^{\alpha_{i+1}}(\sqrt{2\epsilon}u-\epsilon\bar{r}_i)^2 p_{X/U}(\bar{r}_i/u)\,dx\right.$$

$$\left.+\int_{\alpha_{i+1}}^{r_{i+1}}\left(\sqrt{2\epsilon}u-\epsilon\bar{r}_i-\delta_i\right)^2 p_{X/U}(\bar{r}_i/u)\,dx\right)p_U(u)\,du$$

$$+\int_{\sqrt{\frac{\epsilon}{2}}r_i}^{\sqrt{\frac{\epsilon}{2}}r_{i+1}}\int_{r_i}^{r_{i+1}}(\sqrt{2\epsilon}u-\epsilon\bar{r}_i)^2 p_{X/U}(x/u)\,dx\,p_U(u)du$$

$$+\int_{\frac{-\delta_i}{\sqrt{2\epsilon}}}^{\sqrt{\frac{\epsilon}{2}}r_i}\left(\int_{r_i}^{\alpha_i}(\sqrt{2\epsilon}u-\epsilon\bar{r}_i+\delta_i)^2 p_{X/U}(\bar{r}_i/u)\,dx\right.$$

$$\left.+\int_{\alpha_i}^{r_{i+1}}(\sqrt{2\epsilon}u-\epsilon\bar{r}_i)^2 p_{X/U}(\bar{r}_i/u)\,dx\right)p_u(u)\,du, \tag{4.83}$$

where

$$\alpha_i=\frac{r_i-\sqrt{2\epsilon}u}{1-\epsilon}. \tag{4.84}$$

These integrals can be evaluated using the usual fine quantizer approximations and we then obtain the following approximation (for details see appendix B):

$$E\{(\eta_Y(Y)-\gamma\eta_X(X))^2\} \approx \frac{\sqrt{2}(12k_X)^{1/3}}{n}E\left\{\frac{|U|}{p_X^{1/3}(X)}\right\}\sqrt{\epsilon}$$

$$-\left(2+\frac{(12k_X)^{1/3}}{n}E\left\{\frac{X\mathrm{sgn}(U)}{p_X^{1/3}(X)}\right\}\right)\epsilon$$

$$+\left(\frac{2\sqrt{2}(12k_X)^{1/3}}{n}E\left\{\frac{|U|}{p_X^{1/3}(X)}\right\}\right)\epsilon\sqrt{\epsilon}. \tag{4.85}$$

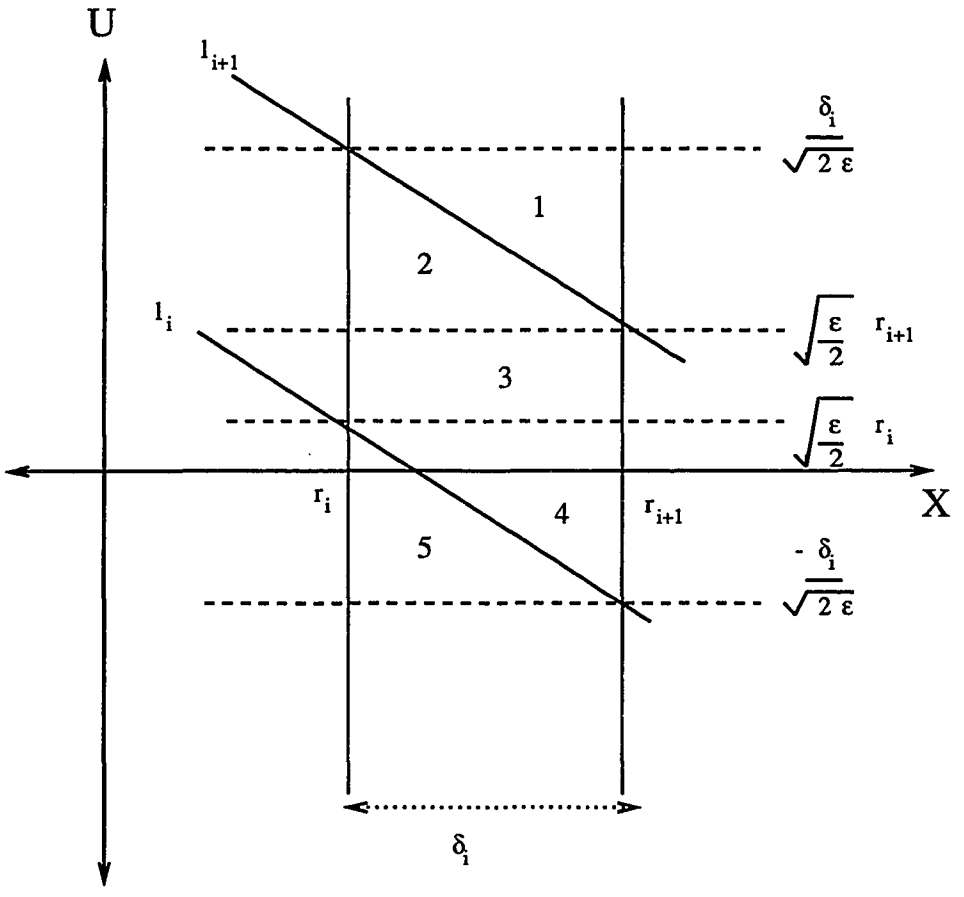

Figure 4.12: Regions of Integration, $E\{\eta_X(X)\eta_Y(Y)\}$.

By substituting this result in (4.80), we can obtain an approximation for the correlation between the quantization errors of two highly correlated, unit variance inputs with identical marginal distributions, when quantized by identical quantizers.

In the case when X and Y are jointly Gaussian, the random variable U is also Gaussian and independent of X. The expectations in the above equation can be evaluated and by using (4.80) we obtain the following approximation for the correlation between the quantization errors of two highly correlated, Gaussian random variables:

$$E\{\eta_X(X)\eta_Y(Y)\} \approx \left(\frac{2.7}{n^2} - \frac{3}{n}\sqrt{1-\gamma} + (1-\gamma) - \frac{9}{n}(1-\gamma)^{3/2}\right)\sigma_X\sigma_Y. \qquad (4.86)$$

We note that the above approximation is a Taylor series expansion in $\sqrt{1-\gamma}$ about the point zero. By choosing a sufficient number of terms of the expansion, we can obtain good approximations for the correlation between the quantization errors of highly correlated random variables.

In figures 4.13, 4.14, 4.15 and 4.16 we plot the the above approximation and the exact correlation between the quantization errors as a function of the correlation between the input random variables for different quantizers. We see from the graphs that this approximation is a reasonably good estimate of the exact curve for large γ. The asymptote at $\gamma = 0$ is also a good approximate of the exact quantization error. As γ decreases from its maximum value of one, the correlation between the quantization errors decreases very rapidly and is almost equal to zero. If we approximate the correlation between the quantization errors for large values of γ by the first two terms of the Taylor expansion in (4.86), then we see that the correlation is equal to zero when $\epsilon \geq 0.81/n^2$. Thus a reasonable approximation is that the quantization errors are uncorrelated when the correlation coefficient between the

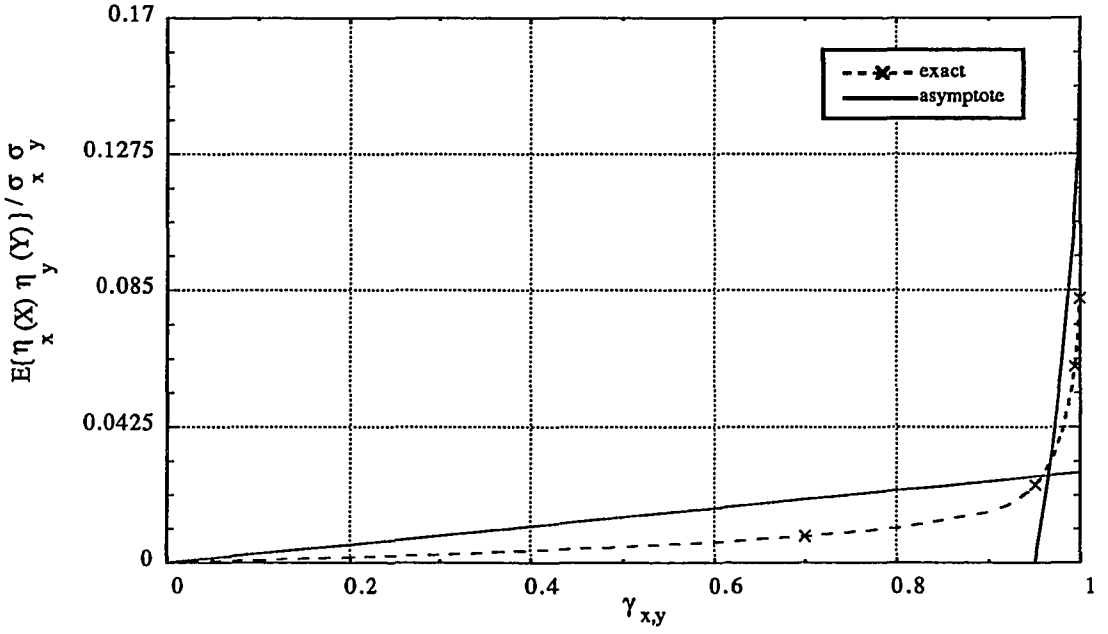

Figure 4.13: $E\{\eta_X(X)\eta_Y(Y)\}$ vs. γ, $n = 4$.

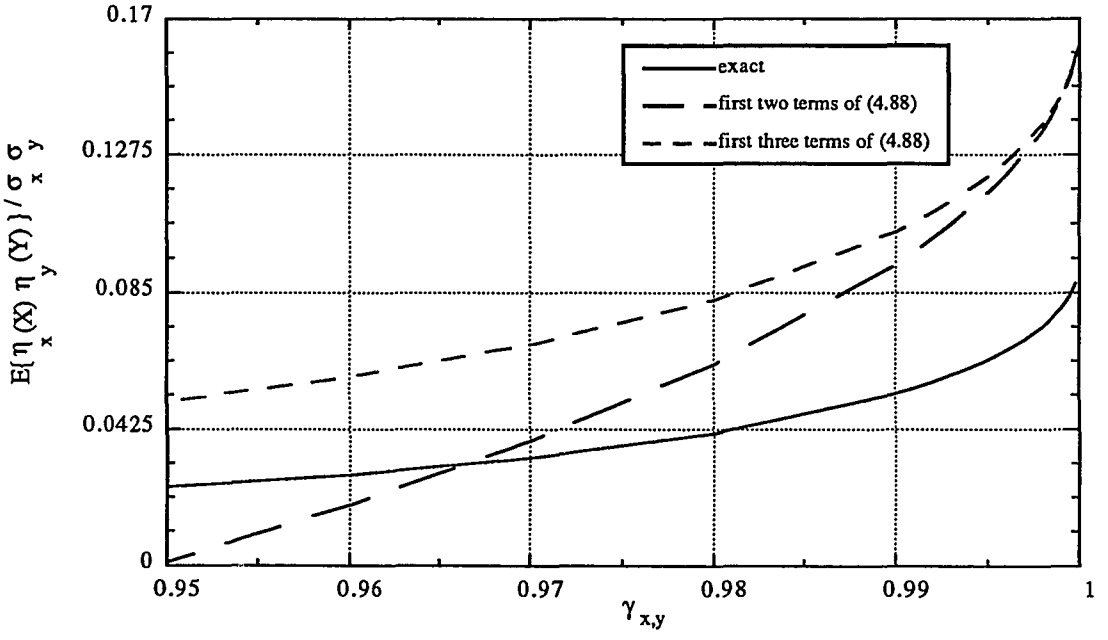

Figure 4.14: $E\{\eta_X(X)\eta_Y(Y)\}$ vs. γ, $0.95 < \gamma < 1$, $n = 4$.

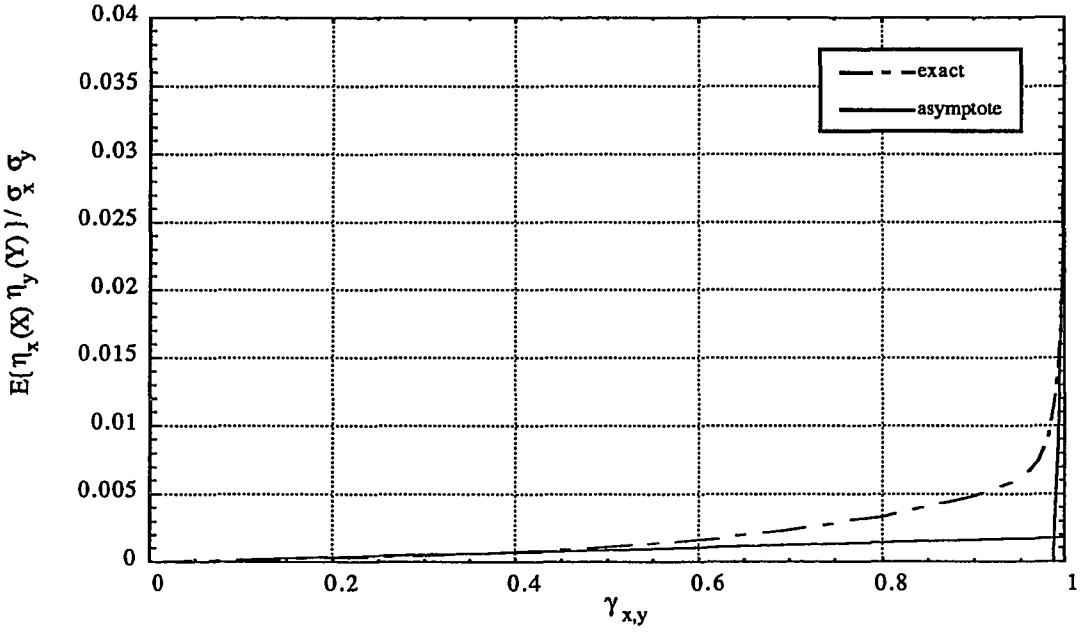

Figure 4.15: $E\{\eta_X(X)\eta_Y(Y)\}$ vs. γ, $n = 8$.

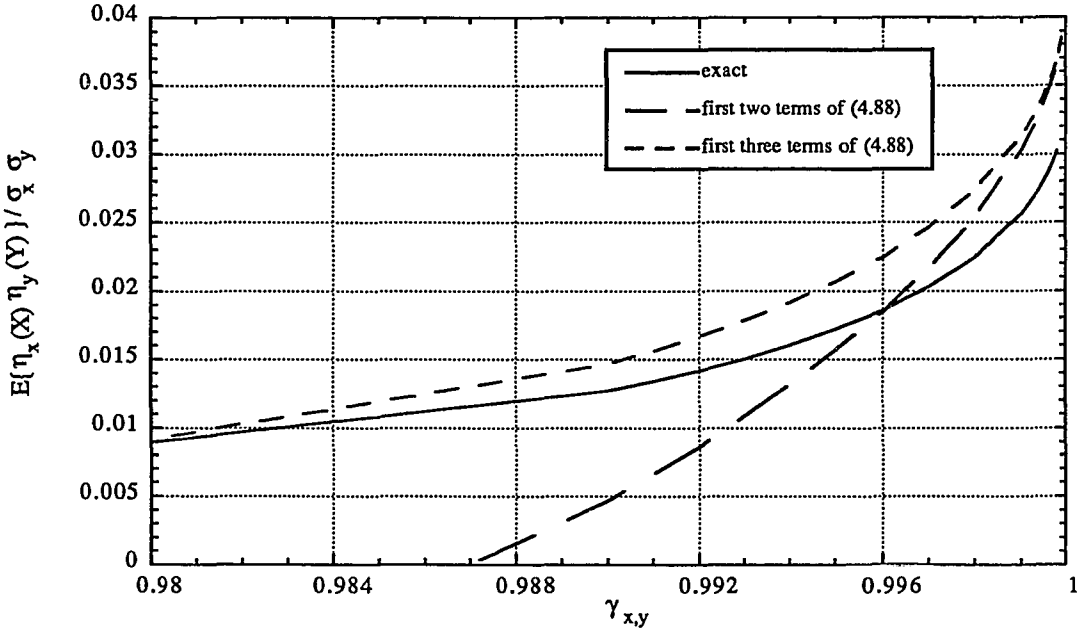

Figure 4.16: $E\{\eta_X(X)\eta_Y(Y)\}$ vs. γ, $0.98 < \gamma < 1$, $n = 8$.

inputs is less than $1 - 0.81/n^2$. On the basis of these results, we next examine the effect of adding a dither signal to the input.

4.5 Dithering

In certain applications, as in the quantization of speech signals, the quality of the quantized signal depends not only on the mean square error but also on the shape of the power spectral density of the quantization error process. It is believed that it is desirable for the quantization error to have a flat power spectral density, i.e. the quantization errors at different time instances should be uncorrelated (see [9]). If the input is a sequence of independent random variables, then it is obvious that the quantization errors at two different time instances are also independent and the power spectral density will be white. When the inputs are correlated, a sequence of independent, identically distributed random variables, known as the dither signal, is first added to the input and the resulting process is quantized. The dither signal is also independent of the input signal. It is obvious that there will be a degradation in the mean square error but one gains a flatter power spectral density for the quantization error. In this section we investigate the effect of adding a particular dither signal to a sequence of correlated Gaussian random variables.

Let the input $\{X_j\}$ be a sequence of unit variance, correlated, Gaussian random variables and let $\{N_j\}$ be a sequence of independent, identically distributed, unit variance, Gaussian random variables which is also independent of the input sequence $\{X_j\}$. The input sequence is scaled by $(1 - \epsilon)$ and the sequence $\{N_j\}$ by $\sqrt{2\epsilon - \epsilon^2}$. This scaling ensures that the input to the quantizer has unit variance. The two are added to form a unit variance signal which is the input to the quantizer:

$$U_j = (1 - \epsilon)X_j + \sqrt{2\epsilon - \epsilon^2}N_j. \tag{4.87}$$

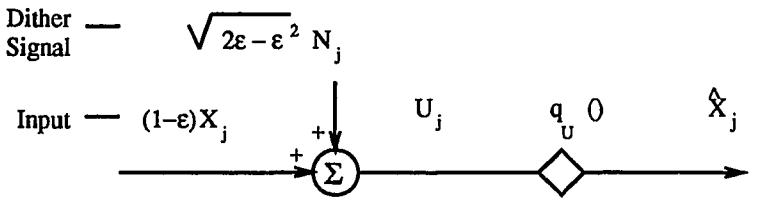

Figure 4.17: A Dithered Quantization System.

We first calculate the mean square error for a dithered quantization system. We note that the random variable U_j is Gaussian and hence the optimal quantizer for X_j and U_j are the same. The mean square error can then be expressed as follows:

$$\begin{aligned} E\{(X_j - \hat{X}_j)^2\} &= E\{\eta_U^2(U_j)\} + 2\epsilon + 2\epsilon E\left\{X_j\,\eta_X\left((1-\epsilon)X_j + \sqrt{2\epsilon-\epsilon^2}N_j\right)\right\} \\ &\quad - 2\sqrt{2\epsilon-\epsilon^2}E\left\{N_j\,\eta_X\left((1-\epsilon)X_j + \sqrt{2\epsilon-\epsilon^2}N_j\right)\right\}. \end{aligned} \tag{4.88}$$

The expectations in the above equations can be approximated for small dither by first conditioning on N_j and then using the mismatched quantizer approximations in (4.53) and (4.55). Note that X_j and N_j are independent as required. We then obtain the approximation

$$E\{(X_j - \hat{X}_j)^2\} \approx \frac{2.7}{n^2} + 2\epsilon\left(1 - \frac{2.7}{n^2}\right). \tag{4.89}$$

The results of the previous section show that the correlation between the quantization errors of two Gaussian random variables is approximately equal to zero, if the correlation coefficient between the inputs is less than $1 - 0.8/n^2$ (we assume that the first two terms of (4.86) are a sufficiently good approximation in this case).

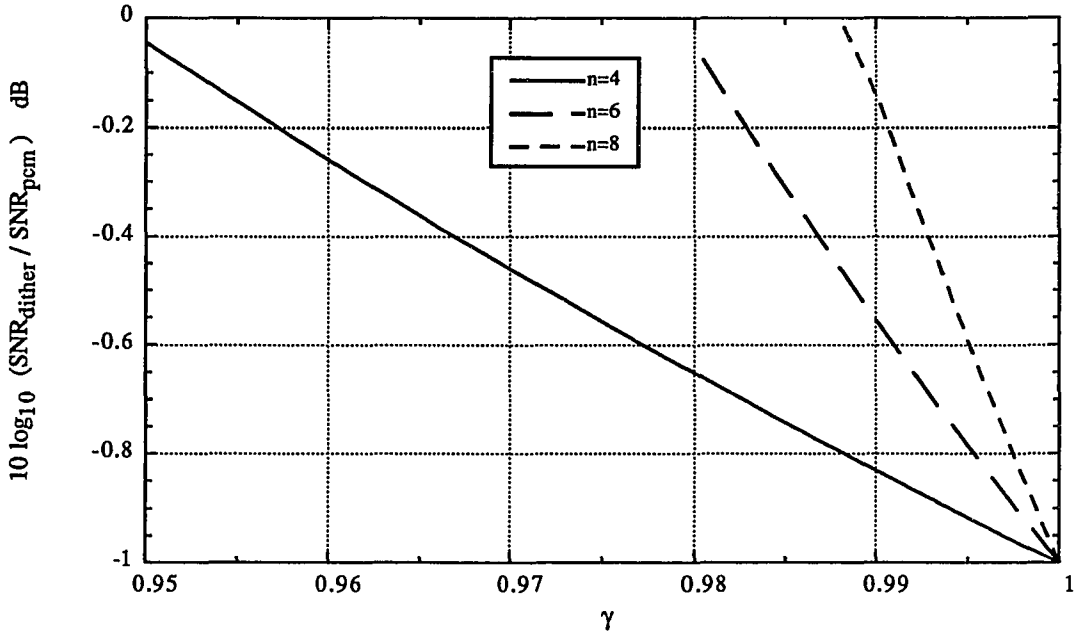

Figure 4.18: $10\log_{10}\left(\frac{SNR_{dither}}{SNR_{pcm}}\right)$ vs. γ.

Thus if we ensure for any input signal that the correlation coefficient between U_j and U_{j+k} is less than $1 - 0.8/n^2$ for all $k \neq 0$, then the quantization errors will be approximately uncorrelated. The correlation between U_j and U_{j+k} is equal to $(1-\epsilon)^2 E\{X_j X_{j+k}\}$ and we obtain the following condition that must be satisfied if the quantization errors are to be uncorrelated:

$$\epsilon \approx 1 - \left(1 - \frac{0.4}{n^2}\right)\frac{1}{\sqrt{\gamma}}, \tag{4.90}$$

where

$$\gamma = \max_{k\neq 0} E\{X_j X_{j+k}\} \geq 1 - \frac{0.8}{n^2}. \tag{4.91}$$

With this value of ϵ, the quantization error equals

$$E\{(X_j - \hat{X}_j)^2\} \approx 2\left(1 - \frac{1}{\sqrt{\gamma}}\right) + \frac{1}{n^2}\left(\frac{6.1}{\sqrt{\gamma}} - 2.7\right). \tag{4.92}$$

In the worst case when the inputs are constant, i.e., $X_j = X_{j+k}$ for all j and k and $\gamma = 1$, the quantization error is about 1 dB larger than that of a PCM system without dither. In figure 4.18, we plot the difference in the signal to noise ratios of a dithered system and an undithered PCM system as a function of γ for different quantizers. It is obvious that as the number of quantization levels increases, the degradation in the signal to noise ratio due to the addition of a dither signal decreases. In certain applications this degradation incurred due to the decrease in the signal to noise ratio is more than offset by the improvement in the spectrum of the error process.

Chapter 5

Analysis Of Some Digital Transmission Systems

In this chapter we investigate different methods of transmitting a first order, Gauss-Markov process over a noiseless, digital channel. We are mainly interested in the minimum smoothed error that can be achieved by a particular scheme for a given transmission rate. In order to minimize the smoothed error with the constraint that the transmission rate is fixed, we are led to investigate the trade off between the number of quantization levels and the sampling period. Depending upon the type of reconstruction filter used, the smoothed error can be expressed in terms of the variance and other statistics of the quantization error. For the schemes under consideration, it is possible to derive difference equations for these statistics but the equations are nonlinear and exact solutions are difficult to obtain. We use the fine quantizer approximations developed in the previous chapter to obtain approximate solutions instead. In the design of the quantizer itself, the criterion that one usually tries to minimize is the variance (average power) of the quantization error. This is equivalent to minimizing the area under the power spectral density of the quantization error. In certain applications like speech, the shape of the power spectral density of the quantization error also plays an important role in determining the quality of the reproduced output (see [9]). Often a flat power spectral density is preferred because the energy is then distributed over a wide range of frequencies. We shall also investigate this aspect of different quantization schemes.

5.1 PCM

In PCM systems the quantization scheme used is a scalar, memoryless quantizer and each sample is quantized into one of n levels. In this section we study the performance of optimal and uniform quantizers.

5.1.1 Optimum Quantizers

The quantization error of an optimum quantizer can be approximated with the fine quantizer approximation given in (4.13). Using (3.19), the smoothed error of this system for a unit variance, Gaussian input can be approximated as follows:

$$\xi_{sm} \approx \alpha \frac{2.7}{n^2} + \beta\tau. \tag{5.1}$$

Given the constraint that the transmission rate is r bits/second, we are interested in designing the system to minimize the mean square error, i.e., we wish to find the optimum sampling rate τ_o and the optimum number of quantization levels n_o. But the number of quantization levels is equal to $2^{r\tau}$ and by substituting this expression for n into the above expression, we obtain the following equation relating the smoothed error to the transmission rate and the sampling period:

$$\xi_{sm} \approx \alpha \frac{2.7}{2^{2r\tau}} + \beta\tau. \tag{5.2}$$

We can now minimize the smoothed error with respect to the sampling period τ to obtain an expression for the optimum sampling period τ_o:

$$\tau_o \approx \frac{\ln(r)}{1.4r} + \frac{\ln(3.7\alpha/\beta)}{1.4r} \text{ seconds.} \tag{5.3}$$

The optimum number of quantization levels n_o, is then given by

$$n_o \approx \sqrt{\frac{3.7\alpha r}{\beta}}\ . \tag{5.4}$$

Thus as the transmission rate increases, the number of quantization levels increases, which implies that the complexity of the quantizer increases. Also, as r tends to infinity, we have the result

$$\lim_{r\to\infty} \frac{\tau_o n_o^2}{\ln\left(\frac{\beta n_o^2}{3.7\alpha}\right)} \approx \frac{2.6\alpha}{\beta}. \tag{5.5}$$

This implies that the sampling period decreases at a rate faster than that at which the number of quantization levels increases. The minimum smoothed error that can be achieved for a transmission rate of r bits/second is given by

$$\xi_{sm} \approx \frac{\beta \ln r}{1.4r} + \frac{1}{r}\left(\frac{\beta\left(3.78 + 3.7\ln\left(3.7\alpha/\beta\right)\right)}{5.18}\right). \tag{5.6}$$

For large rates the first term dominates and thus the smoothed error decreases asymptotically as $O(\log(r)/r)$. This term arises due to the reconstruction filter. Hence in the case of PCM, it pays to improve the performance of the reconstruction filter.

We next examine the autocorrelation function of the quantization error to determine if the power spectral density of the quantization error is constant. It follows from the results in section 4.4.2 (see equations 4.78 and 4.86) that the correlation between the quantization errors at two different time instances can be approximated as follows:

$$E\{\eta_X(X_j)\eta_X(X_{j+l})\} \approx \begin{cases} \frac{2.7}{n^2} - \frac{3\sqrt{1-\rho}}{n} & \text{if} \quad \rho > 1 - \frac{0.81}{n^2} \\ \frac{7.2\rho}{n^4} & \text{otherwise.} \end{cases} \tag{5.7}$$

Here ρ is the correlation between X_j and X_{j+l} and is equal to $e^{-|l\tau|}$. For large transmission rates, the number of quantization levels is large and hence we can approximate the term of order $1/n^4$ by zero. The above equation can then be rewritten to obtain a condition when the the correlation between quantization errors

at two different time instances is negligible:

$$E\{\eta_X(X_j)\eta_X(X_{j+l})\} \approx 0 \quad \text{if} \quad |l| > \frac{0.28\beta}{\alpha \ln(3.7r\alpha/\beta)}. \tag{5.8}$$

By substituting the values of α and β for different reconstruction filters, we see that for small vaules of r, two adjacent quantization errors are approximately uncorellated.

We now calculate the smoothed error for different reconstruction filters. For a memoryless reconstruction filter, α is equal to 1 and β is equal to 0.5 (see equation 3.6). Hence the optimum number of number of quantization levels, the optimum sampling rate and the smoothed error can be approximated as follows:

$$n_o \approx \sqrt{7.5r}, \tag{5.9}$$

$$\tau_o \approx \frac{\ln 7.5r}{1.4r}, \tag{5.10}$$

$$\xi_{sm} \approx \frac{1.1}{r} + \frac{\ln r}{2.8r}. \tag{5.11}$$

It can also be seen that for these values of α and β, the quantization errors are approximately uncorrelated since the condition in (5.8) is satisfied even for $l = 1$ for rates as low as 5 bits/second. Thus the power spectral density of the quantization error is approximately white. This is a desirable property and it implies that dithering is not necessary in this case.

If a reconstruction filter with a unit delay memory is used, we would expect the optimum sampling rate to be less than or equal to the optimum sampling rate for the memoryless reconstruction filter (for after all we now have a better reconstruction filter). The correlation between the quantization errors at two adjacent time instances will then be negligible. It then follows from (3.11) that α is equal to 2/3 and β is equal to 1/3. The optimum number of quantization levels, the optimum

sampling period and the smoothed error for this system are given by

$$n_o \approx \sqrt{7.5r}, \tag{5.12}$$

$$\tau_o \approx \frac{\ln(7.5r)}{1.4r}, \tag{5.13}$$

$$\xi_{sm} \approx \frac{0.72}{r} + \frac{\ln(r)}{4.2r}. \tag{5.14}$$

Note that the number of quantization levels and the sampling rate are the same as in the case of the memoryless reconstruction filter. The smoothed error though is smaller and for large rates the improvement in the SNR is approximately 1.8 dB.

When a low pass filter is used as a reconstruction filter, the optimum number of quantization levels, the optimum sampling period and the smoothed error equal

$$n_o \approx \sqrt{18.7r}, \tag{5.15}$$

$$\tau_o \approx \frac{\ln(18.7r)}{1.4r}, \tag{5.16}$$

$$\xi_{sm} \approx \frac{0.56}{r} + \frac{\ln(r)}{6.9r}. \tag{5.17}$$

For the same transmission rate, the optimum number of quantization levels is about 1.6 times larger than that of the other reconstruction filters which implies a more complex quantizer. For large rates, the SNR is about 2.2 dB larger than that of a PCM system with a reconstruction filter with a unit delay memory and about 3.9 dB larger than the smoothed error of a PCM system with a memoryless reconstruction filter.

By comparing the smoothed errors we can see that in terms of the smoothed error and the power spectral density of the quantization error, a low pass reconstruction filter is the best, followed by the reconstruction filter with a unit delay and the smoothed error is largest for the memoryless reconstruction filter. Not surprisingly, to obtain better performance, a more complex system is needed. Not only does the complexity of the reconstruction filter increase, but the quantizer becomes more

complex since, for a given transmission rate, the number of quantization levels is larger. But for large rates, the smoothed error in all three cases is of the same order of magnitude, namely $O(\log(r)/r)$. Also, the power spectral density of the quantization error for all three systems is approximately white. A final point to note is that the sampling period decreases at a rate faster than that at which the number of quantization levels increases.

We next compare these results with the distortion-rate bound. We note that the power spectral density of this first order Markov, Gaussian random process equals $2/(1 + (2\pi f)^2)$. It can then be shown that for large rates the minimum possible distortion achievable can be approximated by (see [2], problem 4.23)

$$\xi_{sm,minimum} \approx \frac{0.58}{r}. \tag{5.18}$$

Thus for large rates we see that there is a big gap between the performance of a PCM scheme and the rate distortion bound. For example, at $r = 1000$ bits/second, the PCM scheme with prefiltering and a lowpass reconstruction filter is 4.3 dB away from the distortion-rate bound while at $r = 10,000$ bits/second, it is 5.1 dB away. As the transmission rate tends to infinity, the difference between the distortion rate bound and the smoothed error of a PCM system also becomes infinitely large.

5.1.2 Uniform Quantizers

In this section we analyze the performance of a uniform quantizer in a PCM system. We first note that for the same number of quantization levels, the quantization error of a uniform quantizer is greater that that of an optimum quantizer. Hence for a given transmission rate, we will expect the optimum number of quantization levels for a uniform quantizer to be larger than that for an optimum quantizer with the same reconstruction filter. Since the quantization errors at different time instances

are almost uncorrelated for an optimal quantizer, we will expect the same to be true for uniform quantizers. Thus the smoothed error can be approximated as follows (see equation (4.28)):

$$\xi_{sm} \approx \frac{4\ln(3n^2)}{3n^2} + \beta\tau, \tag{5.19}$$

where β depends upon the type of reconstruction filter being used.

For a given transmission rate, this expression can be minimized with respect to the sampling period τ as in the previous section. In this case we end up with a transcendental equation which can only be solved numerically. Instead we can find an upper bound to the performance by assuming that the optimal sampling period in this case is approximately equal to the optimal sampling period for a PCM system with an optimal quantizer. The smoothed error is then approximately equal to

$$\xi_{sm} \approx \frac{1.2\beta\ln(r/\beta)}{r}. \tag{5.20}$$

Comparing this approximation with that for the smoothed error of an optimal quantizer (5.6), we see that for the same reconstruction filter we lose at most 2.3 dB in the smoothed error by using a uniform quantizer instead of an optimal one.

5.2 A Modified PCM Scheme

In this section we consider a modified PCM quantization scheme in which we add a unit delay memory in the receiver. The output of the receiver is thus a function of the current and the previous output of the transmitter; hence the receiver is a sliding block machine of length two. The transmitter itself is a memoryless device. In spite of the apparent simplicity of this system, we have been unable to find any method of optimizing the transmitter-receiver pair. This is unlike the case of memoryless quantizers where the Lloyd-Max conditions give us an iterative algorithm by which locally optimal quantizers can be found (see [10]).

To get around this problem, we assume that the optimal transmitter in this case is similar to the transmitter of a memoryless quantizer. Hence the quantization regions are disjoint segments as shown in figure 4.1. We believe that such a transmitter will also be optimal in this case. Once the transmitter is fixed, the optimal representative points (i.e., the receiver) can be calculated. The quantization error of such systems is smaller than that of memoryless systems (see [3] for an analysis of such a system with a binary quantizer). In this section we investigate if with this addition of memory in the receiver, the asymptotic smoothed error is better than that of PCM systems for large transmission rates.

The quantization error of this modified PCM scheme is given by the expectation

$$\xi_q = E\left\{(X_k - q(X_k, X_{k-1}))^2\right\}. \tag{5.21}$$

For convenience we shall denote the event that X_k lies in the quantization region $\mathcal{S}_i$ and X_{k-1} lies in the quantization region $\mathcal{S}_j$, as $\mathcal{R}_{i,j}$. Given the event $\mathcal{R}_{i,j}$, the output of the receiver is denoted by $q_{i,j}$. The quantization error can then be expressed as follows:

$$\xi_q = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} E\left\{(X_k - q_{i,j})^2/\mathcal{R}_{i,j}\right\} Pr\{\mathcal{R}_{i,j}\} \tag{5.22}$$

By adding and subtracting the midpoint of the ith quantization interval, the above expression can be expressed as follows:

$$\begin{aligned}\xi_q \;=\; & \sum_{i=0}^{n-1} E\left\{(X_k - \bar{r}_i)^2\right\} Pr\{X_k \in \mathcal{S}_i\} \\ & + \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} Pr\{\mathcal{R}_{i,j}\}\left(\bar{r}_i - q_{i,j} - E\left\{X_k - \bar{r}_i/\mathcal{R}_{i,j}\right\}\right)^2 \\ & - \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} Pr\{\mathcal{R}_{i,j}\}\left(E\left\{X_k - \bar{r}_i/\mathcal{R}_{i,j}\right\}\right)^2 \end{aligned} \tag{5.23}$$

We note that only the second term on the right hand side of the above equation is a function of $q_{i,j}$ and that this term is nonnegative. Hence in order to minimize the

quantization error, the representative points should be chosen as follows:

$$q_{i,j} = \bar{r}_i - E\{X_k - \bar{r}_i / \mathcal{R}_{i,j}\}. \tag{5.24}$$

We also note that the first term in the right hand side of (5.23) is the quantization error of a PCM system.

The quantization error of the modified PCM system can then be expressed as follows:

$$\xi_q \approx \xi_{q,pcm} - \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \frac{\left(\int_{r_i}^{r_{i+1}}\int_{r_j}^{r_{j+1}} (x-\bar{r}_i) p_{X_k,X_{k-1}}(x,y)\,dx\,dy\right)^2}{\int_{r_i}^{r_{i+1}}\int_{r_j}^{r_{j+1}} p_{X_k,X_{k-1}}(x,y)\,dx\,dy} \tag{5.25}$$

The summation on the right hand side is the gain that we obtain over a PCM system by adding a unit delay memory in the receiver. Each term in the summation is positive and thus the quantization error of this system is indeed smaller than that of PCM.

To evaluate this gain we expand the joint density function $p_{X_k,X_{k-1}}(x,y)$ in a Taylor series in x and y about the point $(\bar{r}_i, \bar{r}_j)$. For ease of notation, we drop the subscripts and denote $p_{X_k,X_{k-1}}(x,y)$ by $p(x,y)$. With some algebra we obtain the following expression for the quantization error:

$$\begin{aligned}\xi_q \approx \xi_{q,pcm} - \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \Bigg(& \frac{(p^x(\bar{r}_i,\bar{r}_j))^2\delta_i^4}{144p(\bar{r}_i,\bar{r}_j)} + \frac{p^x(\bar{r}_i,\bar{r}_j)p^{xxx}(\bar{r}_i,\bar{r}_j)\delta_i^6}{2880p(\bar{r}_i,\bar{r}_j)} \\ & + \frac{2p^x(\bar{r}_i,\bar{r}_j)p^{xyy}(\bar{r}_i,\bar{r}_j)\delta_i^4\delta_j^2}{3456p(\bar{r}_i,\bar{r}_j)} - \frac{(p^x(\bar{r}_i,\bar{r}_j))^2 p^{xx}(\bar{r}_i,\bar{r}_j)\delta_i^6}{3456(p(\bar{r}_i,\bar{r}_j))^2} \\ & - \frac{(p^x(\bar{r}_i,\bar{r}_j))^2 p^{yy}(\bar{r}_i,\bar{r}_j)\delta_i^4\delta_j^2}{3456(p(\bar{r}_i,\bar{r}_j))^2} \Bigg) \delta_i\delta_j + O(\delta^{10}).\end{aligned} \tag{5.26}$$

This expression can be simplified with the usual fine quantizer approximations. We have been unable to solve the problem for an optimal quantizer (see appendix A for a discussion of the difficulties involved). Instead we assume that the quantization

regions are all of the same length δ (as in a uniform quantizer) and lie in the interval from $-\alpha(n)/2$ to $\alpha(n)/2$. Note that α is a function of the number of quantization levels and it must be chosen to minimize the quantization error. The above equation can then be simplified and the quantization error can be approximated as

$$\xi_q \approx \xi_{q,pcm} - \frac{\alpha^4(n)}{144n^4} E\left\{\frac{(p^x(x,y))^2}{p^2(x,y)}\right\} - \frac{\alpha^6(n)}{2880n^6} E\left\{\frac{p^x(x,y)p^{xxx}(x,y)}{p^2(x,y)}\right\}$$

$$-\frac{\alpha^6(n)}{3456n^6}\left(E\left\{\frac{2p^x(x,y)p^{xyy}(x,y)}{p^2(x,y)} - \frac{(p^x(x,y))^2 p^{xx}(x,y)}{p^3(x,y)}\right.\right.$$

$$\left.\left. - \frac{(p^x(x,y))^2 p^{yy}(x,y)}{p^3(x,y)}\right\}\right). \tag{5.27}$$

The expectations can be evaluated and by expanding the resulting expression in a Taylor series in τ, we obtain the following approximation:

$$\xi_q \approx \frac{\alpha^2(n)}{12n^2} - \frac{\alpha^4(n)}{288n^4\tau} + \frac{\alpha^6(n)}{3456n^6\tau^2}. \tag{5.28}$$

For a given transmission rate of r bits/second, the optimum number of quantization levels n_o, the optimum sampling period τ_o and the optimum $\alpha(n)$ can be found so as to minimize the smoothed error. We perform this optimization numerically and find that the gain in the smoothed error is marginal. The smoothed error for a system with a low pass reconstruction filter is smaller than that of a PCM system with a memoryless, uniform quantizer by 0.02 dB at a transmission rate of 1000 bits/seconds and by 0.06 dB at 10,000 bits/second. We believe that a similar result will hold even if an optimum transmitter were designed.

From the above results, we conjecture that adding a unit delay memory in the quantizer does not significantly improve the asymptotic smoothed error. Further, since the input to the quantizer is a first order Markov process, it seems unlikely that using more samples will improve the quantization error significantly. Thus

we conjecture that the addition of any amount of memory in the receiver end of a memoryless quantizer will not significantly improve the performance of the system.

5.3 DPCM

DPCM is a recursive quantization scheme that takes into account the correlation between input samples. The transmitter uses the outputs of the receiver to predict the next input (note that the receiver output can be duplicated at the transmitter since we assume the channel to be noiseless). The difference between the input and its predicted value (this signal is called the error signal) is quantized and transmitted. The receiver adds the output of the transmitter to the predicted value of the input to construct its output. For a given transmission rate, the main problem then is to find the optimal sampling rate, the quantizer and the prediction filter. We shall examine DPCM systems which use simple linear prediction schemes where only the previous output of the receiver is used to estimate the current input.

5.3.1 Optimized DPCM Systems

Figure 5.1 shows a DPCM scheme in which the estimate of the current input is calculated by multiplying the previous output of the receiver with the correlation between two successive input samples. We shall assume that the quantizer is optimized for its input $\tilde{N}_j$ (this is the reason that we call the system an optimized DPCM system). The error signal $\tilde{N}_j$ is a function of the current input and the past output of the quantizer and hence its statistics depend upon the quantizer. At the same time, the design of the quantizer also depends upon the statistics of its input. Due to this circular dependency, it is difficult, if not impossible, to optimize the quantizer for its input, but we shall proceed with this assumption.

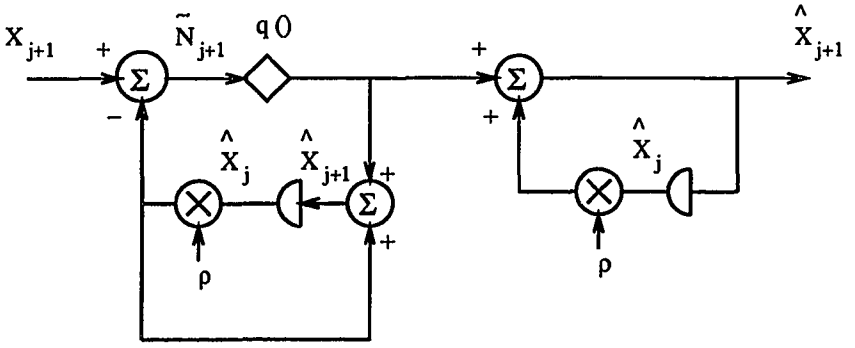

Figure 5.1: A DPCM System.

We first note that the sampled process $\{X_j\}$ can be modeled as the output of a linear, time-invariant system driven by a sequence of independent, zero-mean, unit-variance, normal random variables $\{N_j\}$ (known as the innovations process). The following equation describes this system:

$$X_{j+1} = \rho X_j + \sqrt{1-\rho^2} N_{j+1}. \tag{5.29}$$

Here ρ is the correlation between two successive input samples, i.e.,

$$\rho = E\{X_{j+1} X_j\}. \tag{5.30}$$

By cascading this system with the DPCM system, we can view the innovations process as the input. The innovations are independent and this simplifies the analysis.

The DPCM scheme is described by the following set of equations:

$$\hat{X}_{j+1} = q_{\tilde{N}}(\tilde{N}_{j+1}) + \rho \hat{X}_j \quad j \geq 0, \tag{5.31}$$

$$\tilde{N}_{j+1} = X_{j+1} - \rho \hat{X}_j \quad j \geq 0, \tag{5.32}$$

$$\hat{X}_0 = 0. \tag{5.33}$$

The last equation describes the initial condition of the system. We note that the transmitter and the receiver have the same initial state and since we assume the channel to be noiseless, the state of the transmitter and receiver are equal for all j.

We next derive some important properties of this system. By adding (5.31) and (5.32), we obtain the relation

$$X_{j+1} - \hat{X}_{j+1} = \tilde{N}_{j+1} - q_{\tilde{N}}(\tilde{N}_{j+1}). \tag{5.34}$$

This is the well known result (see [9]) that the difference between the input X_{j+1} and the output of the receiver $\hat{X}_{j+1}$ is equal to the error in quantizing the error signal $\tilde{N}_{j+1}$. Another important relation can be derived from (5.29) and (5.31):

$$\tilde{N}_{j+1} = \rho\, \eta_{\tilde{N}}(\tilde{N}_j) + \sqrt{1-\rho^2} N_{j+1}. \tag{5.35}$$

This is a nonlinear, first order difference equation in the variable $\tilde{N}_j$. An important result that follows from it is that $\tilde{N}_j$ is independent of the random variables N_{j+l}, $l \geq 1$.

Variance of The Quantization Error

The quantization error can be estimated using the above equations. By squaring (5.35) and taking expectations, we obtain the following expression for the variance of the error signal:

$$\sigma^2_{\tilde{N}_{j+1}} = \rho^2 E\{\eta^2_{\tilde{N}}(\tilde{N}_j)\} + (1-\rho^2). \tag{5.36}$$

As j tends to infinity, we assume that the system reaches a steady state. We denote this steady state variance by $\sigma^2_{\tilde{N}}$. An approximation for the variance can then be obtained from the above equation by substituting the fine quantizer approximation of (4.13):

$$\sigma^2_{\tilde{N}} \approx \rho^2 \sigma^2_{\tilde{N}} \frac{k_{\tilde{N}}}{n^2} + (1-\rho^2). \tag{5.37}$$

We can solve this equation for the variance of $\tilde{N}_j$. By keeping terms only of the order of $1/n^2$, we obtain the approximation

$$\sigma^2_{\tilde{N}} \approx (1-\rho^2)\left(1+\frac{\rho^2 k_{\tilde{N}}}{n^2}\right). \tag{5.38}$$

The variance of the error in quantizing $\tilde{N}_j$ can then be approximated as shown below:

$$E\{\eta^2_{\tilde{N}}(\tilde{N}_j)\} \approx \frac{(1-\rho^2)k_{\tilde{N}}}{n^2}. \tag{5.39}$$

In order to determine $k_{\tilde{N}}$ we need to know the distribution of $\tilde{N}_j$. This distribution is not known and instead we make the approximation that $\tilde{N}_j$ is approximately Gaussian which is reasonable especially if the error in quantizing $\tilde{N}_j$ is small (see equation 5.35). Then $k_{\tilde{N}}$ is approximately equal to 2.7 and the variance of the difference between X_j and $\hat{X}_j$ (which by (5.34) equals the error in quantizing $\tilde{N}_j$) can be approximated as

$$E\{(X_j-\hat{X}_j)^2\} \approx \frac{2.7(1-\rho^2)}{n^2}. \tag{5.40}$$

An important point to be noted from this equation is that the quantization error is a function of the sampling rate and the number of quantization levels. Hence the quantization error can be made small by either increasing the number of quantization levels or by increasing the sampling rate.

Autocorrelation Function of the Quantization Error

We next calculate the autocorrelation function of the quantization error. An approximation can be obtained by assuming that $\tilde{N}_j$ and $\tilde{N}_{j+l}$ are jointly Gaussian and using the results of section 4.4.2 (note that this assumption is stronger than assuming that the variables are only marginally Gaussian). We need to first derive an expression for the correlation between the inputs $\tilde{N}_j$ and $\tilde{N}_{j+l}$ (for convenience

we assume that l is positive). By substituting for $\tilde{N}_{j+l}$ from (5.35) we obtain the relation

$$E\{\tilde{N}_j\tilde{N}_{j+l}\} = E\{\tilde{N}_j\sqrt{1-\rho^2}N_{j+l}\} + \rho E\{\tilde{N}_j\eta_{\tilde{N}}(\tilde{N}_{j+l-1})\}. \tag{5.41}$$

As noted earlier, $\tilde{N}_j$ is independent of N_{j+l} for $l > 0$ and hence their correlation equals zero. The correlation between $\tilde{N}_j$ and $\eta_{\tilde{N}}(\tilde{N}_{j+l-1})$ can be approximated using the result in (4.69). We then obtain the approximation

$$E\{\tilde{N}_j\tilde{N}_{j+l}\} \approx \frac{\rho k_{\tilde{N}}}{n^2} E\{\tilde{N}_j\tilde{N}_{j+l-1}\} \tag{5.42}$$

This is a linear, first order difference equation whose initial condition is given by (5.38). By solving this equation, we obtain the result that the correlation between $\tilde{N}_j$ and $\tilde{N}_{j+l}$ can be approximated as follows for all l:

$$E\{\tilde{N}_j\tilde{N}_{j+l}\} \approx \left(\frac{\rho k_{\tilde{N}}}{n^2}\right)^{|l|} \sigma^2_{\tilde{N}}. \tag{5.43}$$

Hence it follows that the correlation coefficient between $\tilde{N}_j$ and $\tilde{N}_{j+l}$ is given by

$$\gamma_{\tilde{N}_j,\tilde{N}_{j+l}} \approx \left(\frac{\rho k_{\tilde{N}}}{n^2}\right)^{|l|}. \tag{5.44}$$

Since $k_{\tilde{N}}$ is approximately equal to 2.7, the correlation coefficient between the variables $\tilde{N}_j$ and $\tilde{N}_{j+l}$ is smaller than 1 when l is not equal to 0.

We can now apply the results in section 4.4.2 (see equation 4.78), to calculate the correlation between the quantization error at two different time instances:

$$\begin{aligned} E\{\eta_{\tilde{N}}(\tilde{N}_j)\eta_{\tilde{N}}(\tilde{N}_{j+l})\} \;\approx\; & \frac{2.7(1-\rho^2)}{n^2}\psi(l) \\ & + \frac{7.2(1-\rho^2)}{n^4}\left(\frac{2.7\rho}{n^2}\right)^{|l|}(1-\psi(l)). \end{aligned} \tag{5.45}$$

Here $\psi(l)$ is the Kronecker delta function which is defined as follows:

$$\psi(l) = \begin{cases} 1 & l = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.46}$$

From the above equations we see that the autocorrelation function is the sum of two terms. The first term is a Kronecker delta function whose magnitude is of the order of $1/n^2$ and the second term is an exponentially decreasing series. From this approximation we see that the autocorrelation function decreases rapidly as l increases. For example, when $n = 5$, the correlation between adjacent quantization errors is about 19 dB smaller than the variance of the quantization error. Hence the power spectral density of the quantization error process is quite flat.

An alternate expression can be derived for the correlation between the quantization errors with one further assumption; namely, N_j and $\tilde{N}_{j+l}$ are jointly Gaussian for all l. By using (5.35) we can obtain the following expression for the correlation between the errors in quantizing $\tilde{N}_j$ and $\tilde{N}_{j+l}$:

$$E\{\eta_{\tilde{N}}(\tilde{N}_j)\eta_{\tilde{N}}(\tilde{N}_{j+l})\} = \frac{1}{\rho^2}\left(E\{\tilde{N}_{j+1}\tilde{N}_{j+l+1}\} - \rho E\{\sqrt{1-\rho^2}N_{j+1}\eta_{\tilde{N}}(\tilde{N}_{j+l})\}\right). \tag{5.47}$$

The first expectation on the right hand side has already been evaluated. We next calculate the correlation between the innovation at time $j+1$ and the error in quantizing the error signal at time $j+l$ where l is positive.

We make the assumption that that $\tilde{N}_{j+l}$ and N_{j+1} are jointly Gaussian. We then obtain the following approximation using (4.69):

$$E\{\sqrt{1-\rho^2}N_{j+1}\eta_{\tilde{N}}(\tilde{N}_{j+l})\} \approx \frac{k_{\tilde{N}}}{n^2}E\{\sqrt{1-\rho^2}N_{j+1}\tilde{N}_{j+l}\}. \tag{5.48}$$

By substituting for $\tilde{N}_{j+l}$ from (5.35) and noting that N_{j+1} and N_{j+l} are independent when $l > 1$, we obtain the difference equation

$$E\{\sqrt{1-\rho^2}N_{j+1}\eta_{\tilde{N}}(\tilde{N}_{j+l})\} \approx \frac{\rho k_{\tilde{N}}}{n^2}E\{\sqrt{1-\rho^2}N_{j+1}\eta_{\tilde{N}}(\tilde{N}_{j+l-1}\}), \quad l > 1. \tag{5.49}$$

The initial condition for the above difference equation can be obtained by evaluating

the correlation between $\sqrt{1-\rho^2}N_{j+1}$ and $\eta_{\tilde{N}}(\tilde{N}_{j+1})$ in a similar manner:

$$E\{\sqrt{1-\rho^2}N_{j+1}\eta_{\tilde{N}}(\tilde{N}_{j+1})\} \approx (1-\rho^2)\frac{k_{\tilde{N}}}{n^2}. \tag{5.50}$$

By solving the above equation we obtain the result

$$E\{\sqrt{1-\rho^2}N_{j+1}\eta_{\tilde{N}}(\tilde{N}_{j+l})\} \approx \frac{(1-\rho^2)}{\rho}\left(\frac{\rho k_{\tilde{N}}}{n^2}\right)^{l-1} \quad l \geq 1. \tag{5.51}$$

We can now calculate the autocorrelation function of the error process. Substituting the results from (5.43) and (5.51) into (5.47), we obtain the expression

$$E\{\eta_{\tilde{N}}(\tilde{N}_j)\eta_{\tilde{N}}(\tilde{N}_{j+l})\} \approx \frac{(1-\rho^2)k_{\tilde{N}}}{n^2}\left(\frac{\rho k_{\tilde{N}}}{n^2}\right)^{|l|}, \quad l = \ldots,-1,0,1,\ldots. \tag{5.52}$$

This result states that the autocorrelation function decreases exponentially with l. Even for small n, the correlation between two adjacent quantization errors is small. For example, when $n = 5$, the correlation between two adjacent quantization errors is approximately 9.7 dB smaller than the variance of the quantization error. Further, the power spectral density can be made flatter by increasing the number of quantization levels and decreasing the sampling rate but, only at the expense of increasing the quantization error. At a transmission rate of 1000 bits per second, by doubling the number of quantization levels from 5 to 10, the smoothed error of a DPCM system with a unit delay reconstruction filter increases by about 0.5 dB and the correlation between two adjacent quantization errors decreases by about 6 dB.

The first approximation (5.45) predicts the power spectral density to be flatter than the approximation in (5.52). But in either case the power spectral density of the quantization error process is almost white. We have been unable to resolve the question as to which of the two solutions is a better approximation to the autocorrelation function of the quantization error. An analysis of a similar system in the next section using different methods leads us to believe that the result in

(5.52) is more accurate. The difference in the two solutions also leads us to question the validity of the assumption that $\tilde{N}_j$ and $\tilde{N}_{j+l}$ are jointly Gaussian.

Variance of the Output

Other statistics like the variance of the output can also be estimated. From (5.31) we obtain the relation

$$E\{\hat{X}_{j+1}^2\} = E\{q_{\tilde{N}}^2(\tilde{N}_{j+1})\} + \rho^2 E\{\hat{X}_j^2\} + 2\rho E\{\hat{X}_j q_{\tilde{N}}(\tilde{N}_{j+1})\}. \tag{5.53}$$

By assuming steady state and by making the approximation that $\hat{X}_j$ and $\tilde{N}_{j+1}$ are jointly Gaussian and by using the fine quantizer approximations in (4.20) and (4.69), one finds that the above equation can be simplified as follows:

$$(1-\rho^2)E\{\hat{X}_{j+1}^2\} \approx \left(1 - \frac{k_{\tilde{N}}}{n^2}\right) E\{\tilde{N}_{j+1}^2\} + 2\rho \left(1 - \frac{k_{\tilde{N}}}{n^2}\right) E\{\hat{X}_j \tilde{N}_{j+1}\}. \tag{5.54}$$

In order to calculate the correlation on the right hand side of the above equation, we use (5.31) and (5.35) to obtain the nonlinear equation below:

$$E\{\hat{X}_j \tilde{N}_{j+1}\} = E\{(q_{\tilde{N}}(\tilde{N}_j) + \rho \hat{X}_{j-1})(\rho \eta_{\tilde{N}}(\tilde{N}_j) + \sqrt{1-\rho^2} N_{j+1})\}. \tag{5.55}$$

We note that N_{j+1} is independent of $\tilde{N}_j$ and $\hat{X}_{j-1}$ and that the output of an optimum quantizer is uncorrelated with the quantization error. By using the fine quantizer approximation in (4.69), the above equation then simplies to

$$E\{\hat{X}_j \tilde{N}_{j+1}\} \approx \frac{\rho^2 k_{\tilde{N}}}{n^2} E\{\hat{X}_{j-1} \tilde{N}_j\}. \tag{5.56}$$

By assuming steady state, we then obtain the result that the output $\hat{X}_j$ and $\tilde{N}_{j+1}$ are uncorrelated. By substituting this result into (5.54) and using the approximation for the variance of $\tilde{N}_j$ from (5.38), we obtain an approximation for the variance of the output:

$$E\{\hat{X}_j^2\} \approx 1 - \frac{(1-\rho^2)k_{\tilde{N}}}{n^2}. \tag{5.57}$$

This result implies that the variance of the estimate is smaller that that of the input X_j.

Optimization of the Smoothed Error

We next calculate the optimum number of quantization levels and the optimum sampling period to minimize the smoothed error. For large sampling rates, the correlation between adjacent samples is approximately equal to $1 - \tau$. It then follows that the smoothed error can be approximated by the equation below (see equation 3.19):

$$\xi_{sm} \approx \alpha \frac{5.4\tau}{n^2} + \beta\tau. \tag{5.58}$$

We now find the optimum number of quantization levels and the optimum sampling period in order to minimize the smoothed error. For a transmission rate of r bits per second and a sampling period of τ, the number of quantization levels equals $2^{nr\tau}$. By minimizing the smoothed error with respect to τ, it is easy to show that the following condition must be satisfied by the optimum sampling interval τ_o:

$$\beta 2^{2r\tau_o} - \alpha 10.8(\ln 2) r\tau_o \ln(2) - \alpha 5.4 = 0. \tag{5.59}$$

We rewrite the above equation in terms of the optimum number of quantization levels:

$$\beta n_o^2 - \alpha 10.8(\ln 2)\ln(n_o) + \alpha 5.4 = 0. \tag{5.60}$$

This equation is independent of the transmission rate r and hence the optimum number of quantization levels must be independent of the transmission rate.

For a memoryless reconstruction filter, $\alpha = 1$ and $\beta = 0.5$. The closest integer solution to (5.60) is then equal to 5. By substituting these values in (5.58), we find that the optimum number of quantization levels, the optimum sampling period and

the smoothed error equal

$$n_o \approx 5, \tag{5.61}$$

$$\tau_o \approx \frac{2.3}{r}, \tag{5.62}$$

$$\xi_{sm} \approx \frac{1.6}{r}. \tag{5.63}$$

Thus for large rates, a relatively coarse quantizer is sufficient but the sampling rate is high which makes the input samples highly correlated. For large rates we now have an improvement over PCM since the smoothed error is now of the order of $1/r$. Thus the smoothed error decreases at the same rate as the distortion-rate function (see (5.18)) and for large rates the SNR of this scheme is 4.4 dB smaller than the distortion-rate bound.

We next calculate the smoothed error for the reconstruction filter with a unit delay memory. For this reconstruction filter, β is equal to $1/3$. Since the quantization errors are approximately uncorrelated, α equals $2/3$. The optimum number of quantization levels, the optimum sampling period and the smoothed error equal

$$n_o \approx 5, \tag{5.64}$$

$$\tau_o \approx \frac{2.3}{r}, \tag{5.65}$$

$$\xi_{sm} \approx \frac{1}{r}. \tag{5.66}$$

The quantizer and the sampling rate are the same as that for a memoryless reconstruction filter but the the SNR is larger by 2 dB. When compared to the distortion-rate bound, the SNR of this scheme is 2.4 dB smaller.

For a low pass reconstruction filter, $\alpha = 1$ and $\beta = 0.2$. It then follows that for this system, the optimum number of quantization levels, the optimum sampling period and the smoothed error equal

$$n_o \approx 10, \tag{5.67}$$

$$\tau_o \approx \frac{3.3}{r}, \tag{5.68}$$

$$\xi_{sm} \approx \frac{0.8}{r}. \tag{5.69}$$

The smoothed error is smaller than that of the other two systems. The SNR of this system is larger than that of a system with a memoryless reconstruction filter by 3 dB and larger than that of a system with a unit memory reconstruction filter by 1 dB. It is smaller than the distortion-rate bound by 1.4 dB. Also, since the quantizer is finer in this case, the power spectral density of the quantization error will also be flatter.

Thus the low pass reconstruction filter performs the best but for large rates the smoothed error is of the same order of magnitude for all the three systems. Also, the quantizers are coarse and the number of levels does not increase with the transmission rate as in the case of PCM systems. This is an advantage in terms of the complexity of the quantizer in addition to the fact that the smoothed error is also smaller. We also see that the spectrum of the the quantization error can be made flatter by increasing the number of quantization levels but at the expense of increasing the smoothed error. The performance of these systems is close to the rate distortion bound and it is unlikely that any system of comparable complexity will outperform a matched DPCM system.

5.3.2 Unoptimized DPCM Systems

An assumption made in the previous section is that it is possible to optimize the quantizer within the feedback loop of a DPCM system. We pointed out that this optimization is difficult and it is more realistic to use a suboptimal quantizer. Since the error signal is in some sense similar to the innovations process, a good quantizer

will be one optimized for the random variable $\sqrt{1-\rho^2}N_j$. We denote this quantizer by $q_N()$. Since the quantizer is not optimized for its input, we call this system an unoptimized DPCM system. We now analyze this unoptimized DPCM system.

It can be easily verified that equations 5.31 through 5.35 hold for this system if $q_{\tilde{N}}()$ is replaced by $q_N()$. Thus the property that difference between the random variables X_j and $\hat{X}_j$ equals the error in quantizing the error signal $\tilde{N}_j$ holds. Further we have the relation

$$\tilde{N}_j = \rho\eta_N(\tilde{N}_{j-1}) + \sqrt{1-\rho^2}N_j. \tag{5.70}$$

As in the case of optimized DPCM systems, this nonlinear difference equation in $\tilde{N}_j$ is the key to analyzing this system.

Variance of the Quantization Error

By using the above equation, we express the variance of the quantization error as follows:

$$E\{\eta_N^2(\tilde{N}_j)\} = E\{\eta_N^2(\sqrt{1-\rho^2}N_j + \rho\eta_N(\tilde{N}_{j-1}))\}. \tag{5.71}$$

Since the random variables $\tilde{N}_j$ and $\sqrt{1-\rho^2}N_j$ are similar, the variance of $\tilde{N}_j$ will be approximately equal to $(1-\rho^2)$. Thus the probability that $|\rho\eta_N(\tilde{N}_j)|$ exceeds $(1-\rho^2)$ will be negligible if the quantization error is small. If we make this assumption, then we are justified in using the results derived in section 4.3 for mismatched quantizers.

To evaluate the expectation on the right hand side of the above equation, we first condition on $\tilde{N}_{j-1}$. Since N_j and $\tilde{N}_{j-1}$ are independent random variables, the conditional expectation is nothing but the variance of the error in quantizing a random variable with a quantizer whose mean is not matched to that of the input. By using the result in (4.58), we then obtain the result

$$E\{\eta_N^2(\tilde{N}_j)\} \approx \frac{2.7}{n^2}\left((1-\rho^2) + \rho^2 E\{\eta_N^2(\tilde{N}_{j-1})\}\right). \tag{5.72}$$

If we assume that the process $\{\tilde{N}_j\}$ is stationary, the variance of the quantization error can be solved for and we obtain the following result:

$$E\{\eta_N^2(\tilde{N}_j)\} \approx \frac{2.7(1-\rho^2)}{n^2}. \tag{5.73}$$

This result is the same as that obtained for a matched quantizer in the previous section. Hence we conclude that the performance, as measured in terms of the variance of the quantization error, does not degrade by using a mismatched quantizer. This simplifies the design of a DPCM system. Further, the results derived in the previous section for the smoothed error hold good in this case too. Thus for large rates a 5 level quantizer is optimal when using a memoryless reconstruction filter or a reconstruction filter with a unit delay memory. If we prefilter the input process before sampling and use the same low pass filter as the reconstruction filter, then a 10 level quantizer is optimal. For a transmission rate of r bits/second, the smoothed error varies approximately as $1/r$ for these systems. A suboptimal DPCM system with a reconstruction filter with unit delay was analyzed by Janardhanan [8] using elaborate numerical techniques. There is less than a 5 percent error between his exact results and our approximations.

Autocorrelation Function of the Quantization Error

We next examine the autocorrelation function of the quantization error. Using (5.70) we can express the correlation between quantization errors l time instances apart as follows:

$$E\{\eta_N(\tilde{N}_{j+l})\eta_N(\tilde{N}_j)\} = E\left\{\eta_N\left(\sqrt{1-\rho^2}N_{j+l} + \rho\eta_N(\tilde{N}_{j+l-1})\right)\eta_N(\tilde{N}_j)\right\}. \tag{5.74}$$

Since N_{j+l} is independent of $\tilde{N}_j$ and $\tilde{N}_{j+l}$, we can use the result (4.53) for mismatched quantizers to convert the above nonlinear equation into a linear difference equation:

$$E\{\eta_N(\tilde{N}_{j+l})\eta_N(\tilde{N}_j)\} \approx \frac{2.7\rho}{n^2} E\{\eta_N(\tilde{N}_{j+l-1})\eta_N(\tilde{N}_j)\}. \tag{5.75}$$

The initial condition is provided by (5.73). Solving this difference equation, we get an approximation for the autocorrelation function of the quantization error:

$$E\{\eta_N(\tilde{N}_{j+l})\eta_N(\tilde{N}_j)\} \approx \frac{2.7(1-\rho^2)}{n^2} \left(\frac{2.7\rho}{n^2}\right)^{|l|}. \tag{5.76}$$

This result is the same as that obtained in (5.52) for an optimized DPCM system. Since we expect the performance of an optimized and an unoptimized DPCM system to be similar, we are more apt to believe the result in (5.52) than the result (5.45) since (5.76) has been derived without making any assumptions about the joint density of the random variables $\tilde{N}_{j+l}$ and $\tilde{N}_j$. As noted earlier (see the discussion following (5.52)), the above result implies that the power spectral density of the quantization error is approximately white.

5.3.3 Mismatched DPCM Systems

In this section we examine the the prediction filter used in DPCM systems. In the previous sections, the input sample was predicted by multiplying the previous output of the receiver by ρ, which is the correlation between two adjacent input samples. We now choose a prediction filter whose output is the product of the output of the receiver and some other constant ρ' (when ρ is not equal to ρ', the system is called a mismatched DPCM system). We first derive an approximation for the mean square error and find the optimum ρ'. We also examine the effect on the mean square error when ρ' is not equal to the optimal value. Finally we study the case when ρ' equals 1 (such predictors are called perfect integrators). Perfect integrator systems are

practically important because the system is simplified by eliminating the multiplier in the feedback loop. Also, we usually do not know the exact value of the correlation between input samples. But by sampling at a very high rate, we can ensure that the correlation between the samples is close to 1 which would justify the use of a perfect integrator. To simplify the analysis we assume that the quantizer is optimized for its input. From the results of the previous section, it is reasonable to assume that the choice of the quantizer does not substantially effect the mean square error. Hence an analysis of this scheme should still provide a reasonable estimate of the behavior of practical DPCM schemes.

The mismatched DPCM system is described by the following equations:

$$\hat{X}_{j+1} = \rho' \hat{X}_j + q_{\tilde{N}}(\tilde{N}_{j+1}) \quad j \geq 0 \tag{5.77}$$

$$\tilde{N}_{j+1} = X_{j+1} - \rho' \hat{X}_j \quad j \geq 0 \tag{5.78}$$

$$\hat{X}_0 = 0. \tag{5.79}$$

By adding (5.77) and (5.78), we see that the difference between the input X_{j+1} and the estimate $\hat{X}_{j+1}$ equals the error in quantizing $\tilde{N}_{j+1}$. The following equation can also be easily derived from (5.29) and (5.77):

$$\tilde{N}_{j+1} = \rho' \eta_{\tilde{N}}(\tilde{N}_j) + (\rho - \rho') X_j + \sqrt{1-\rho^2} N_{j+1}. \tag{5.80}$$

This is the basic difference equation that describes the system.

Variance of the Quantization Error

By squaring the above equation and taking expectations, we obtain the following nonlinear equation for the variance of the error signal $\tilde{N}_j$:

$$E\{\tilde{N}_{j+1}^2\} = \rho'^2 E\{\eta_{\tilde{N}}^2(\tilde{N}_j)\} + (\rho - \rho')^2 + (1-\rho^2) + 2\rho'(\rho - \rho') E\{\eta_{\tilde{N}}(\tilde{N}_j) X_j\}. \tag{5.81}$$

To simplify the above expression we assume that the system reaches a steady state. We denote this steady state variance of $\tilde{N}_j$ by $\sigma^2_{\tilde{N}}$. The first expectation on the right hand side of the above equation can be estimated using the fine quantizer approximation in (4.13). To evaluate the second expectation, we assume that the input X_j and $\tilde{N}_j$ are jointly Gaussian. Then, using (4.69), we obtain the following equation:

$$\sigma^2_{\tilde{N}} \approx \frac{\rho'^2 k_{\tilde{N}}}{n^2}\sigma^2_{\tilde{N}} + (\rho - \rho')^2 + (1 - \rho^2) + \frac{2k_{\tilde{N}}}{n^2}\rho'(\rho - \rho')E\{\tilde{N}_j X_j\}. \tag{5.82}$$

We next calculate the correlation between X_j and $\tilde{N}_j$. From (5.29) and (5.80) we see that

$$E\{X_j\tilde{N}_j\} = \rho\rho' E\{X_{j-1}\eta_{\tilde{N}}(\tilde{N}_{j-1})\} + \rho(\rho - \rho') + (1 - \rho^2). \tag{5.83}$$

This equation can once again be simplified by assuming steady state and by using the fine quantizer approximation from (4.69). The correlation between $\tilde{N}_j$ and X_j can then be approximated as follows:

$$E\{\tilde{N}_j X_j\} \approx ((1 - \rho^2) + \rho(\rho - \rho'))\left(1 + \frac{\rho\rho' k_{\tilde{N}}}{n^2}\right). \tag{5.84}$$

By substituting the above expression into (5.82), we obtain the following approximation for the variance of $\tilde{N}_j$:

$$\begin{aligned}\sigma^2_{\tilde{N}} \approx\ & \left((\rho - \rho')^2 + (1 - \rho^2)\right)\left(1 + \frac{\rho'^2 k_{\tilde{N}}}{n^2}\right) \\ & + \frac{2k_{\tilde{N}}}{n^2}\rho'(\rho - \rho')((1 - \rho^2) + \rho(\rho - \rho'))\left(1 + \frac{\rho\rho' k_{\tilde{N}}}{n^2}\right)^2.\end{aligned} \tag{5.85}$$

The variance of the quantization error can then be approximated by

$$E\left\{\eta^2_{\tilde{N}}(\tilde{N}_j)\right\} \approx \frac{(1 - \rho^2)k_{\tilde{N}}}{n^2} + \frac{(\rho - \rho')^2 k_{\tilde{N}}}{n^2}. \tag{5.86}$$

Thus the optimum value of ρ' is equal to ρ. When ρ' is not equal to this optimum value, the quantization error increases quadratically with the difference between ρ' and ρ. We also see from the above equation that the square error between ρ and ρ' must be much smaller than $(1-\rho^2)$, or else the second term on the right hand side will dominate and the performance of the system will degrade significantly.

Perfect Integrator Systems

When ρ' is equal to 1 (i.e. a perfect integrator), the quantization error is given by

$$\xi_q \approx 2(1-\rho)\frac{k_{\tilde{N}}}{n^2}. \tag{5.87}$$

The error thus increases linearly with the difference between 1 and ρ and when ρ is less than 0.5, the quantization error of this system is larger than that of PCM. If we assume that $\tilde{N}_j$ is Gaussian, then by comparing the above result with (5.40), we see that for the same input $\{X_j\}$, the quantization error of a matched DPCM system is $(1+\rho)/2$ times smaller than that of a DPCM system with a perfect integrator. Thus for large sampling rates, i.e. as $\rho \to 1$, the performance of the two systems is almost the same.

From the above expression for the quantization error, we see that the asymptotic smoothed error for a perfect integrator system can be expressed as:

$$\xi_{sm} \approx \alpha\frac{5.4\tau}{n^2} + \beta\tau, \tag{5.88}$$

where we have assumed that $k_{\tilde{N}} \approx 2.7$. This is the same expression as the smoothed error for a matched DPCM system and hence the same results hold for perfect integrator systems for large transmission rates.

5.3.4 Comparison With Other Methods of Analysis

In this section we compare the results derived for the mean square error in the previous sections with those obtained by a different approach (see [9]). By squaring both sides of (5.34) and taking expectations we obtain

$$E\{(X_{j+1} - \hat{X}_{j+1})^2\} \approx \sigma^2_{\tilde{N}} \frac{k_{\tilde{N}}}{n^2}. \tag{5.89}$$

In order to estimate the variance of $\tilde{N}_j$, we approximate $\hat{X}_j$ by X_j and rewrite (5.32) as follows:

$$\tilde{N}_{j+1} = X_{j+1} - \rho X_j. \tag{5.90}$$

This approximation is good when the error between the input X_j and the estimate $\hat{X}_j$ is small. The variance of $\tilde{N}_j$ is then approximately equal to $(1 - \rho^2)$. The mean square error can thus be approximated as follows:

$$E\{(X_{j+1} - \hat{X}_{j+1})^2\} \approx (1 - \rho^2) \frac{k_{\tilde{N}}}{n^2}. \tag{5.91}$$

This result is the same as the result in 5.39. In a similar fashion, the results in (5.86) and (5.87) can be derived. This is an assurance that our method of solving these types of nonlinear difference equations is valid.

It does not seem possible though, to generalize the above method to calculate other quantities like the autocorrelation function of the quantization error or the variance of the estimate $\hat{X}_j$. Further, it is crucial that (5.34) should hold. This particular relation is true only for systems where the state of the transmitter and receiver are the same (such systems are called tracking systems). But it may be possible to obtain difference equations for nontracking systems and thus derive approximations for the mean square error and other statistics, which would not be possible using this method.

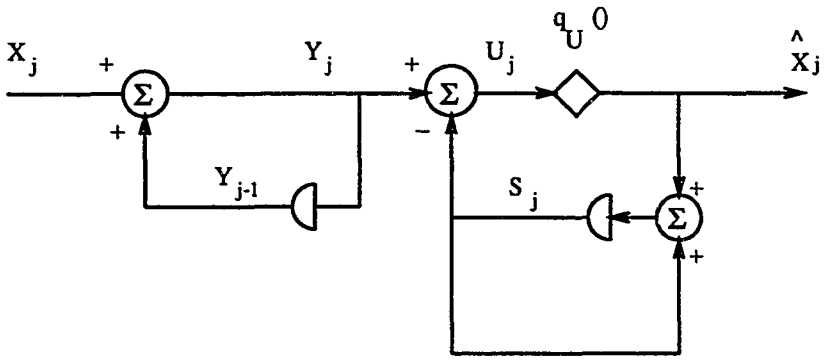

Figure 5.2: A Sigma-Delta Modulator.

5.4 Sigma-Delta Modulation

Sigma-Delta modulation is a recursive quantization scheme which was first proposed to overcome certain limitations of Delta modulation systems like the inability to transmit dc signals and that the inputs need to be highly correlated. The sampled signal is first integrated and then quantized with a DPCM transmitter. By integrating the input, the correlation between the samples is increased. Further, the receiver is simplified since there is no need for an integrator. Figure 5.2 shows a schematic of a Sigma-Delta modulation scheme which uses a DPCM transmitter with a perfect integrator. It is customary to combine the integrator at the input with the integrator in the feed back loop of the DPCM system and have instead one integrator before the quantizer. The receiver also usually passes its input through a lowpass filter (see [9]). We shall ignore the effect of this filter at the receiver in our calculations.

The Sigma-Delta modulation scheme is described by the following equations:

$$Y_j = Y_{j-1} + X_j \tag{5.92}$$

$$U_j = Y_j - S_j \tag{5.93}$$

$$S_{j+1} = S_j + q_U(U_j) \tag{5.94}$$

$$\hat{X}_j = q_U(U_j). \tag{5.95}$$

By manipulating (5.92), (5.93) and (5.94), we can derive the following difference equation for the random variable U_j:

$$U_j = X_j + \eta_U(U_{j-1}). \tag{5.96}$$

Thus the input to the quantizer is the signal X_j plus the quantization error from the previous time instance. This is similar to the addition of a dither signal to the input. The process $\{X_j\}$ can be considered as the output of a linear, time invariant system driven by a sequence of independent, identically distributed Gaussian random variables (see equation (5.29)). From the above equation we then see that U_j is a function only of the variables $N_j, N_{j-1}, \ldots$ and hence is independent of $N_{j+1}, N_{j+2}, \ldots$.

Variance of the Quantization Error

In order to calculate the mean squared error between the input X_j and its estimate, we first calculate the correlation $E\{X_j U_{j-l}\}$ for $l \geq 0$. By using (5.29) and (5.96) we get

$$E\{X_j U_{j-l}\} = E\{(\rho X_{j-1} + \sqrt{1-\rho^2} N_j)(X_{j-l} + \eta_U(U_{j-l-1}))\}. \tag{5.97}$$

By making the assumption that X_j and U_{j-l} are jointly Gaussian, we can use the fine quantizer approximation (4.69) to obtain the relation

$$E\{X_j U_{j-l}\} \approx \rho^l \left(1 + \frac{\rho k_U}{n^2}\right) \quad l \geq 0. \tag{5.98}$$

We next compute the variance of U_j. By squaring (5.96) and using the approximations (5.98) and (4.13), we obtain the result

$$\sigma_U^2 \approx 1 + \frac{k_U}{n^2}(1 + 2\rho) \tag{5.99}$$

and hence the mean square error in quantizing U_j equals

$$E\{\eta_U^2(U_j)\} \approx \frac{k_U}{n^2}. \tag{5.100}$$

We are now in a position to calculate the mean square error between X_j and its estimate $\hat{X}_j$. By substituting for $\hat{X}_j$ from (5.95) and using the approximations derived above, we obtain the following result:

$$E\{(X_j - \hat{X}_j)^2\} \approx \frac{(1+\rho)k_U}{n^2}. \tag{5.101}$$

With the assumption that U_j is Gaussian, we have the approximation that $k_U \approx 2.7$ and hence the mean square error of this system is larger than that of a PCM system. For large sampling rates, $\rho \to 1$ and the quantization error is 3 dB larger than that of a PCM system.

Autocorrelation Function of the Quantization Error

We next examine the autocorrelation function of the quantization error. To calculate the correlation between U_j and U_{j-l}, we use (5.96) and obtain the following equation:

$$E\{U_j U_{j-l}\} = E\{X_j U_{j-l}\} + E\{\eta_U(U_{j-1})U_{j-l}\}. \tag{5.102}$$

The first expectation has already been evaluated in (5.98)). By assuming that U_{j-1} and U_{j-l} are jointly Gaussian, we can use the fine quantizer approximation in (4.69) to evaluate the second expectation on the right hand side of the above equation. We then obtain the linear difference equation shown below

$$E\{U_j U_{j-l}\} \approx \rho^l \left(1 + \frac{\rho k_U}{n^2}\right) + \frac{k_U}{n^2} E\{U_{j-1}U_{j-l}\}. \tag{5.103}$$

The initial condition is given by (5.99) and solving this difference equation we obtain

$$E\{U_j U_{j-l}\} \approx \rho^{l-1}\left(\rho + \frac{k_U}{n^2}(1+\rho^2)\right). \tag{5.104}$$

The correlation coefficient γ, between U_j and U_{j-l} is then given by

$$\gamma \approx \rho^l\left(1 - \frac{\rho^{l+1}k_U}{n^2}\right). \tag{5.105}$$

Hence the correlation between U_j and U_{j-l} is smaller than the correlation between X_j and X_{j-l}. If we assume that U_j and U_{j-l} are jointly Gaussian, we then see that the spectrum is better than that of a PCM system.

Smoothed Error

With the assumption that the correlation between quantization errors at different time instances is negligible, the smoothed error of this system can be approximated as follows

$$\xi_{sm} \approx \frac{5.4}{n^2} + \beta r, \tag{5.106}$$

where β is a constant that depends upon the reconstruction filter that is used (see (3.19)). For a given transmission rate, the optimum number of quantization levels and the optimum sampling period can be found as in the case of PCM systems:

$$\begin{aligned} r_{opt} &\approx \frac{1}{1.4r}\ln\left(\frac{7.5r}{\beta}\right) \\ n_{opt} &\approx \sqrt{\frac{7.5r}{\beta}}. \end{aligned} \tag{5.107}$$

Retaining only the significant terms in r, the smoothed error is given by

$$\xi_{sm} \approx \frac{\beta \ln(r)}{1.4r}. \tag{5.108}$$

From the above results we see that the optimum number of quantization levels for a Sigma-Delta modulation system is larger than that for a PCM system for the

same transmission rate and the sampling rate is smaller. The asymptotic smoothed error is the same for both systems; it is the error due to the reconstruction error that dominates in both cases. The spectra of the quantization error is flatter in the case of Sigma-Delta modulation than in the case of PCM systems. But as we have seen in our analysis of PCM systems, if we choose the optimal number of quantization levels and sampling rate for a given transmission rate, then the quantization process is already reasonably white and thus there is not much to gain by adding a dither signal.

5.5 A Finite State Sliding Block Quantizer

From the results of the previous sections we know that PCM is a memoryless quantization system with an asymptotic smoothed error that decreases as $O(\log(r)/r)$ while DPCM is a recursive, infinite state system whose asymptotic smoothed error decreases as $O(1/r)$. Thus by adding memory to a PCM system, we would expect that it is possible to construct a family of quantization systems whose performance approaches that of DPCM as the state size increases. We are also lead to the more general question – what is the best finite state quantizer for a given state size and transmission rate? This problem has been defined more formally in [3] but is a difficult problem that has not been answered. In this section we provide an example of a finite state quantizer, which in the limit as its state size becomes infinitely large, is equivalent to a DPCM system. We examine the number of states needed in the transmitter and in the receiver in order to achieve a performance close to that of DPCM.

The system that we shall describe is a sliding block quantization system. A sliding block quantizer is one that at each time instance j, maps the input block $(X_j, \ldots, X_{j-l+1})$ (where l is some positive integer called the block length) to

its output. Thus a sliding block transmitter encodes overlapping vectors into their channel alphabet and a sliding block receiver reconstructs the input by decoding overlapping channel vectors. It is possible to have systems where both the transmitter and the receiver are sliding block or to have systems where only one of them is a sliding block machine. Although it has been shown that it is possible to achieve performances arbitrarily close to the rate distortion bound with time invariant sliding block coders and decoders by making the block length arbitrarily large, no algorithms exist for designing such codes. We would like to point out that although the input is a sequence of continuous random variables, the system that we shall propose is a finite state quantizer. The transmitter does not explicitly store the past inputs $X_{j-1}, \ldots, X_{j-l+1}$, but only a finite valued function of these variables. Thus the system that we will describe is a finite state, sliding block quantizer.

In order to illustrate some of the difficulties involved in designing sliding block quantizers, we first design a sliding block (although not finite state) transmitter by modifying DPCM in the following simple manner. We replace the estimate $\hat{X}_j$ (see figure 5.1) by the input X_j and thus obtain a sliding block transmitter of block length two (if this works, it can be further modified by storing a quantized version of X_j instead, to obtain a finite state transmitter). Then at time j, the transmitter quantizes the innovation $\sqrt{1-\rho^2}N_j$ (see equation 5.29). If we use a DPCM receiver with this transmitter, then this system is described by the following equations:

$$\hat{X}_j = q(X_j - \rho X_{j-1}) + \rho \hat{X}_{j-1}, \quad j > 1, \tag{5.109}$$

with the initial condition of the receiver given by

$$\hat{X}_0 = 0. \tag{5.110}$$

We now show that such a simple approach does not work by calculating the mean square error. The mean square error of this system can be expressed in terms

of the innovations as follows:

$$E\{(X_j - \hat{X}_j)^2\} = E\left\{\left(\sum_{i=0}^{j} \rho^i \sqrt{1-\rho^2} N_i - \rho^i q\left(\sqrt{1-\rho^2} N_i\right)\right)^2\right\}. \qquad (5.111)$$

The innovations are a sequence of iid random variables and the above expectation can be easily evaluated by expanding the square since the expectation of the cross terms is equal to zero. The mean square error between X_j and $\hat{X}_j$ is thus the sum of the mean square errors in quantizing the innovations from time 0 to time j. For large j we thus obtain the result that

$$E\{(X_j - \hat{X}_j)^2\} = E\{(N_j - q(N_j))^2\}. \qquad (5.112)$$

This is the same quantization error that we would obtain by using a PCM system.

We note that the innovations are obtained by passing the input sequence $\{X_j\}$ through a linear filter. The receiver in this case is also a linear filter which is the inverse of the filter at the transmitter. We believe that any such system where the input is passed through a linear filter before quantization and where the receiver is a linear filter will not lead to any appreciable improvement in performance over a PCM system. Our belief is based on our attempts to design such filters (using gradient descent techniques), none of which met with any success. It thus seems likely that nonlinear functions are needed to map the input block to its output in order to achieve a good performance.

The reason that the system described in (5.109) and (5.110) fails is that the quantization errors accumulate with time. In DPCM this does not happen because the state of the transmitter and the receiver are equal at each time instant. Such machines are called tracking systems and it has been conjectured that tracking systems are optimal (see [3]). In order to have a tracking system with a sliding block transmitter, one would require the block length of the transmitter to be infinitely

long. By storing all the past inputs, the transmitter could then calculate the state of the receiver at any time instant. But an infinite block length is impractical. We next discuss how a finite state, sliding block transmitter that approximately tracks the state of the receiver can be obtained from a DPCM system.

5.5.1 A Finite State Sliding Block Transmitter

The transmitter that we describe in this section is a time-invariant, feed-forward approximation of a time varying DPCM system, in which a different quantizer is used at each time instant to quantize the error signal. The following set of equations describe the sliding block transmitter:

$$\tilde{X}_{j+1,0} = q_0(X_{j+1}), \tag{5.113}$$

$$\tilde{X}_{j+1,s} = q_s(\tilde{N}_{j+1,s}) + \rho\tilde{X}_{j,s-1}, \quad s = 1,\ldots,l-1, \tag{5.114}$$

$$\tilde{N}_{j+1,s} = X_{j+1} - \rho\tilde{X}_{j,s-1}, \quad s = 1,\ldots,l. \tag{5.115}$$

This structure is obtained by explicitly expressing the output of a time varying DPCM system at any time instance j, as a function of the current and the past $j-1$ inputs. We then obtain an 'unfolded' DPCM structure. The structure grows infinitely large as j tends to infinity, but the size can be kept finite by truncating the structure (see figure 5.3).

There are $(l+1)$ levels (which are labeled from 0 through l and are indexed by the variable s) in the transmitter described by the above equations. The input to the transmitter at time $(j+1)$ is X_{j+1} and the output is $q_l(\tilde{N}_{j+1,l})$. The quantizer at level s is labeled $q_s()$ and the number of quantization levels is n_s. The variables $\tilde{X}_{j,s}$, $s = 0,\ldots,l$ are each stored at level s and take one of a finite set of values. For example, since $\tilde{X}_{j,0}$ is the output of quantizer $q_0()$, it can take n_0 different values. Since $\tilde{X}_{j+1,1}$ is a linear combination of $\tilde{X}_{j,0}$ and the output of quantizer $q_1()$, it can take $n_0 n_1$

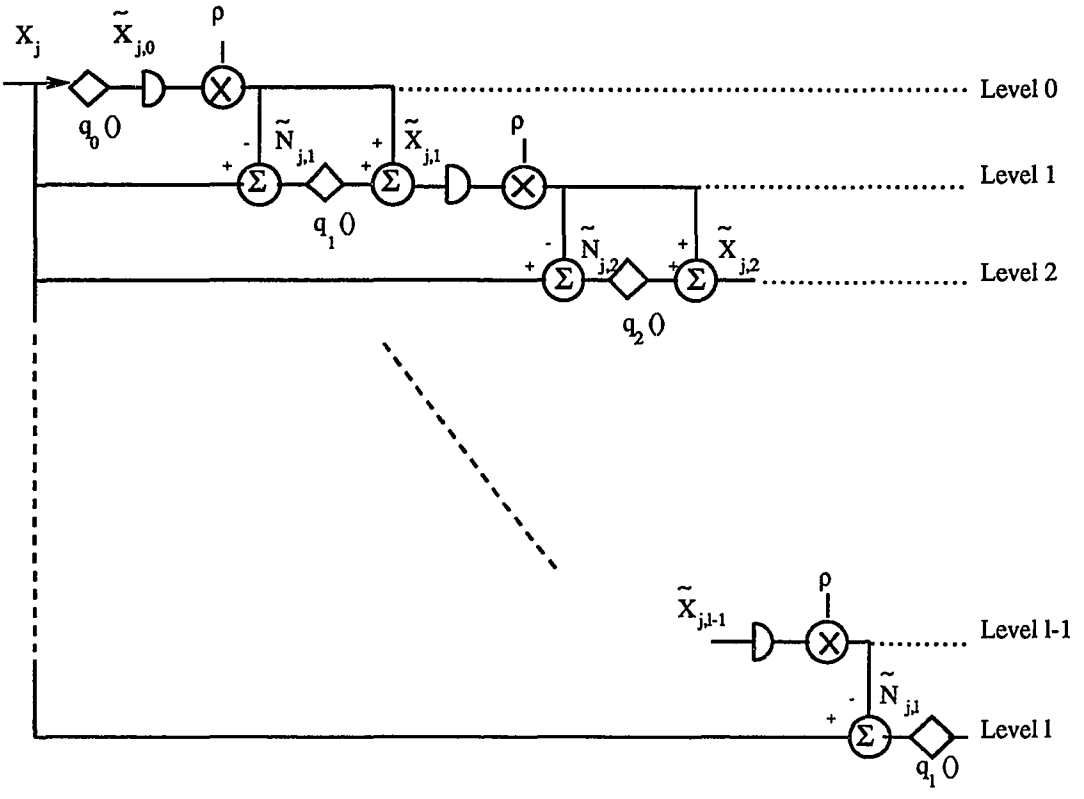

Figure 5.3: A Finite State Sliding Block Transmitter.

distinct values. In general $\tilde{X}_{j,s}$ can take $n_0 n_1 \ldots n_{s-1}$ distinct values. These variables can be viewed as successive approximations to the output of an infinite state DPCM system. They also represent the state of the transmitter since the knowledge of the input and these variables at any time instant is sufficient to calculate the output and the value of the state variables at the next time instant. The number of states is equal to the number of distinct values that the state vector $(\tilde{X}_{j,0}, \tilde{X}_{j,1}, \ldots, \tilde{X}_{j,l-1})$ can take which equals $n_0 n_1 \ldots n_{l-1}$. Hence if the number of levels in the transmitter is finite, it is a finite state transmitter. It can also be easily verified that a transmitter with $l+1$ levels is a sliding block machine whose block length is also equal to $l+1$. Finally, we shall see that it is more convenient to work with the variables $\tilde{N}_{j+1,s}$, $s = 1, \ldots, l$, instead of the state variables $\tilde{X}_{j+1,s}$, $s = 0, \ldots, l$.

We next derive a set of difference equations in the variables $\tilde{N}_{j+1,s}$, $s = 1, \ldots, l$. By using (5.29), (5.114) and (5.115), we obtain the following relation:

$$\tilde{N}_{j+1,s} = \rho \eta_{s-1}(\tilde{N}_{j,s-1}) + \sqrt{1-\rho^2} N_{j+1}, \quad s = 2, \ldots, l+1. \tag{5.116}$$

By using (5.29), (5.113) and (5.115), we obtain the relation

$$\tilde{N}_{j+1,1} = \rho \eta_0(X_j) + \sqrt{1-\rho^2} N_{j+1}. \tag{5.117}$$

By defining the variable $\tilde{N}_{j,0}$ as equal to X_j, the above equations can be expressed as

$$\tilde{N}_{j+1,s} = \rho \eta_{s-1}(\tilde{N}_{j,s-1}) + \sqrt{1-\rho^2} N_{j+1}, \quad s = 1, \ldots, l. \tag{5.118}$$

This is a system of first order nonlinear difference equations with initial condition

$$\tilde{N}_{j+1,0} = \rho \tilde{N}_{j,0} + \sqrt{1-\rho^2} N_{j+1}. \tag{5.119}$$

From (5.118) and (5.119) it can be seen that the variables $\tilde{N}_{j+1,s}$, $s = 0, \ldots, l$ are a function only of the variables $N_{j+1}, N_j, \ldots$ and hence are independent of N_{j+k} for all $k > 1$.

We next analyze the transmitter and calculate some quantities of interest. We shall assume that as j tends to infinity, the transmitter reaches a steady state and that an equilibrium distribution for the state variables exists. In our analysis we shall assume that the quantizers $q_0(), \ldots, q_l()$ have been optimized for their inputs (there is no feedback in the structure and unlike the case of DPCM, the distribution of the inputs to the quantizers can be calculated and hence the quantizers can be optimized). It then follows that the variables $\tilde{N}_{j+1,s}$ and $\tilde{X}_{j+1,s}$ are all zero mean random variables.

Variance of $\tilde{N}_{j,s}$

We first calculate the variance of the random variables $\tilde{N}_{j+1,s}$. Since N_{j+1} and $\tilde{N}_{j,s}$ are independent, by squaring equation 5.118 we obtain the following relation:

$$E\{\tilde{N}^2_{j+1,s}\} = \rho^2 E\{\eta^2_{s-1}(\tilde{N}_{j,s-1})\} + (1-\rho^2), \qquad s = 1, \ldots, l. \tag{5.120}$$

We denote the steady state variance of the variables $\tilde{N}_{j,s}$, $s = 0, \ldots, l$ by $\sigma^2_{\tilde{N}_s}$. By substituting the fine quantizer approximation for the variance of the quantization error from equation 4.13, we obtain the linear difference equation

$$\sigma^2_{\tilde{N}_s} \approx \frac{\rho^2 k_{\tilde{N}_{s-1}}}{n^2_{s-1}} \sigma^2_{\tilde{N}_{s-1}} + (1-\rho^2), \qquad s = 1, \ldots, l, \tag{5.121}$$

whose initial condition can be obtained from (5.119):

$$\sigma^2_{\tilde{N}_0} = 1. \tag{5.122}$$

By solving this difference equation we obtain the following expression for the variance of the random variables $\tilde{N}_{j,s}$, $s = 1, \ldots, l$:

$$\sigma^2_{\tilde{N}_s} \approx (1-\rho^2) + (1-\rho^2) \sum_{j=1}^{s-1} \left(\prod_{i=j}^{s-1} \frac{\rho^2 k_{\tilde{N}_i}}{n_i^2} \right) + \prod_{i=0}^{s-1} \frac{\rho^2 k_{\tilde{N}_i}}{n_i^2}. \tag{5.123}$$

To simplify the above expression, we consider the case when the number of quantization levels in each level is the same (to say n). We make the assumption that the fine quantization coefficients $k_{\tilde{N}_{l-1}}, \ldots, k_{\tilde{N}_0}$ are all equal (to say k). The variance of the random variable $\tilde{N}_{j,s}$ can then be approximated as

$$\sigma^2_{\tilde{N}_s} \approx (1-\rho^2)\left(1+\frac{\rho^2 k}{n^2}\right)+\left(\frac{\rho^2 k}{n^2}\right)^s. \tag{5.124}$$

The first term is the steady state solution and the second term is a transient term. If we assume that the variables $\tilde{N}_{j,s}$ are Gaussian, then the fine quantizer coefficient k is equal to 2.7 and the transient term decreases as s increases. This transient term can be neglected when it is smaller than the steady state solution. By defining the variable l^* as

$$l^* = \frac{\log\left((1-\rho^2)\left(1+\rho^2 k/n^2\right)\right)}{\log\left(\rho^2 k/n^2\right)}, \tag{5.125}$$

we see that for $s > l^*$, the variance of $\tilde{N}_{j,s}$ is approximately constant. We denote this variance by $\sigma^2_{\tilde{N}}$ and note that it can be approximated as

$$\sigma^2_{\tilde{N}} \approx (1-\rho^2)\left(1+\frac{2.7\rho^2}{n^2}\right). \tag{5.126}$$

We note that l^* is typically quite small; for $\rho = 0.9$ and $n = 4$, $l^* = 1$ while for $\rho = 0.999$ and $n = 4$, $l^* = 4$. As n tends to infinity or as ρ tends to 1, we see that l^* increases, although very slowly. These results were found to agree well with simulation results.

Approximations for $E\{\eta_l(\tilde{N}_{j,l})\,\eta_s(\tilde{N}_{j',s})\}$

To simplify the algebra, we shall henceforth assume that the number of quantization levels is the same for all $q_s()$, $s = 0, \ldots, l-1$. We first examine the correlation between the quantization errors of the variables $\tilde{N}_{j,l}$ and $\tilde{N}_{j',s}$, $j \neq j'$, where for

convenience we assume that $j' < j$. By substituting from (5.118) we obtain the relation

$$E\left\{\tilde{N}_{j,l}\,\tilde{N}_{j',s}\right\} = \rho E\left\{\sqrt{1-\rho^2}N_{j'}\,\eta_{s-1}(\tilde{N}_{j-1,s})\right\} + \rho^2 E\left\{\eta_{l-1}(\tilde{N}_{j-1,l-1})\,\eta_{s-1}(\tilde{N}_{j'-1,s-1})\right\}. \quad (5.127)$$

Both the terms on the right hand side are at most of order $1/n^2$ and hence the correlation between the inputs is small. Thus we make the approximation that the correlation between the quantization errors of the variables $\tilde{N}_{j,l}$ and $\tilde{N}_{j',s}$ is negligible when $j \neq j'$.

We next examine the case when $j = j'$ and $s = l - 1$. For convenience we denote the correlation coefficient between $\tilde{N}_{j,s+1}$ and $\tilde{N}_{j,s}$ by γ_{s+1} and the correlation between their quantization errors by θ_{s+1}. From (5.118) and using the fact that for $s > l^*$, the variance of $\tilde{N}_{j,s}$ is approximately constant, we obtain the following relation for the correlation coefficient γ_{s+1}:

$$\gamma_{s+1} = \frac{1-\rho^2+\rho^2\theta_s}{\sigma^2_{\tilde{N}}}, \quad s > l^*. \quad (5.128)$$

We can obtain a lower bound on the correlation coefficient γ_{s^*+1} by assuming that θ_{l^*} is negligible. Then

$$\gamma_{l^*+1} \approx 1 - \frac{\rho^2 k}{n^2}. \quad (5.129)$$

We can also relate the correlation between the quantization errors of $\tilde{N}_{j,s+1}$ and $\tilde{N}_{j,s}$ to the correlation coefficient between them by assuming that $\tilde{N}_{j,s+1}$ and $\tilde{N}_{j,s}$ are jointly Gaussian and that their correlation coefficient is sufficiently large to use the approximation in (4.86). By substituting for γ_s from (5.128) we then obtain the following difference equation in θ_s:

$$\theta_{s+1} \approx \frac{2.7\sigma^2_{\tilde{N}}}{n^2} - \frac{3\rho\sigma_{\tilde{N}}}{n}\sqrt{\frac{2.7\sigma^2_{\tilde{N}}}{n^2} - \theta_s} + \rho^2\left(\frac{2.7\sigma^2_{\tilde{N}}}{n^2} - \theta_s\right), \quad s > l^* + 1. \quad (5.130)$$

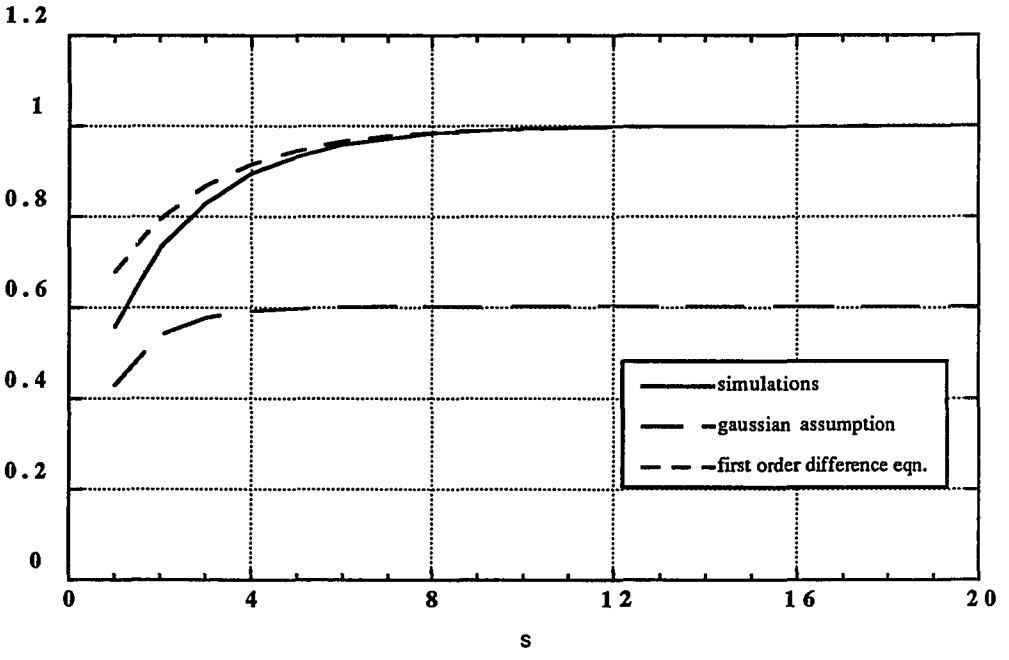

Figure 5.4: $\theta_s/(2.7\sigma^2_{\tilde{N}}/n^2)$ vs. s, $n = 4$, $\rho = 0.4$.

We can obtain an initial condition for this equation by assuming that the correlation between the quantization errors is negligible when $s = l^*$.

By substitution we can verify that the above difference equation has two steady state solutions:

$$\lim_{s\to\infty} \theta_s = \frac{2.7\sigma^2_{\tilde{N}}}{n^2} \quad \text{and} \quad \lim_{s\to\infty} \theta_s = \frac{2.7\sigma^2_{\tilde{N}}}{n^2}\left(1 - \frac{9\rho^2}{2.7(1+\rho^2)^2}\right). \tag{5.131}$$

For the given initial condition, the steady state solution equals the second solution. This is much smaller than the variance of the quantization error and is intuitively incorrect. Simulations also show that this solution underestimates the actual value of the correlation (see figures 5.4, 5.5, 5.6 and 5.7). This leads us to believe that our assumption that $\tilde{N}_{j,s+1}$ and $\tilde{N}_{j,s}$ are jointly Gaussian is incorrect.

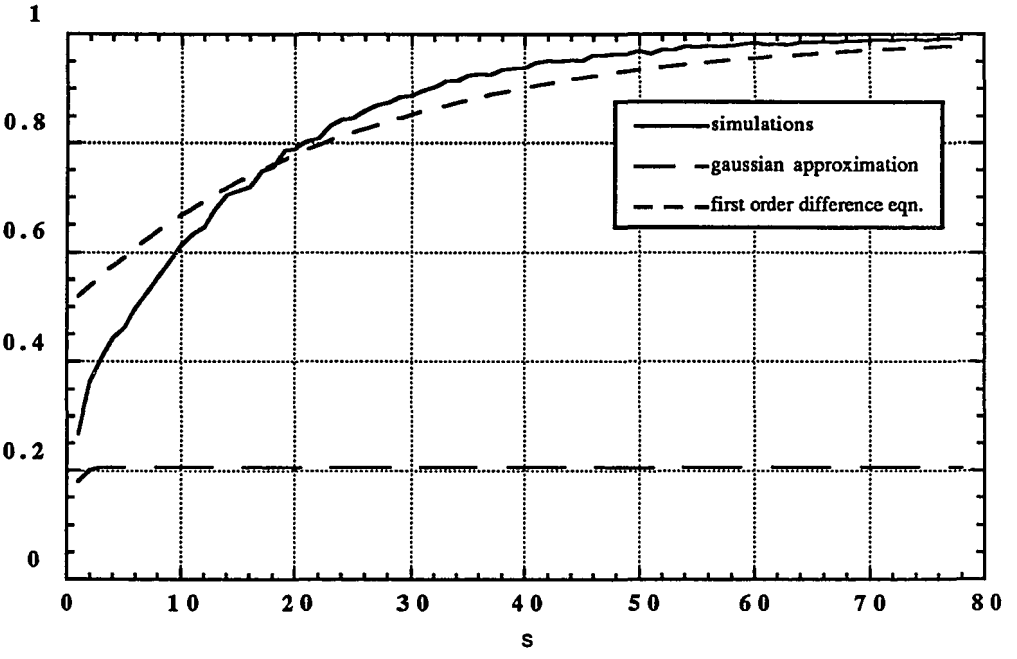

Figure 5.5: $\theta_s/(2.7\sigma^2_{\tilde{N}}/n^2)$ vs. s, $n = 4$, $\rho = 0.8$.

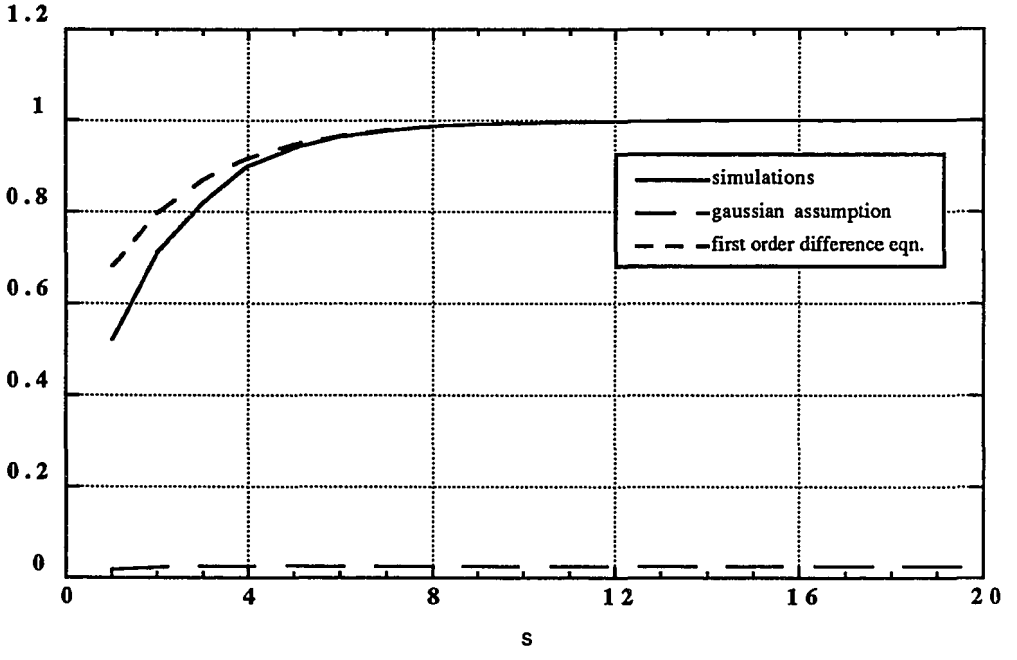

Figure 5.6: $\theta_s/(2.7\sigma^2_{\tilde{N}}/n^2)$ vs. s, $n = 8$, $\rho = 0.4$.

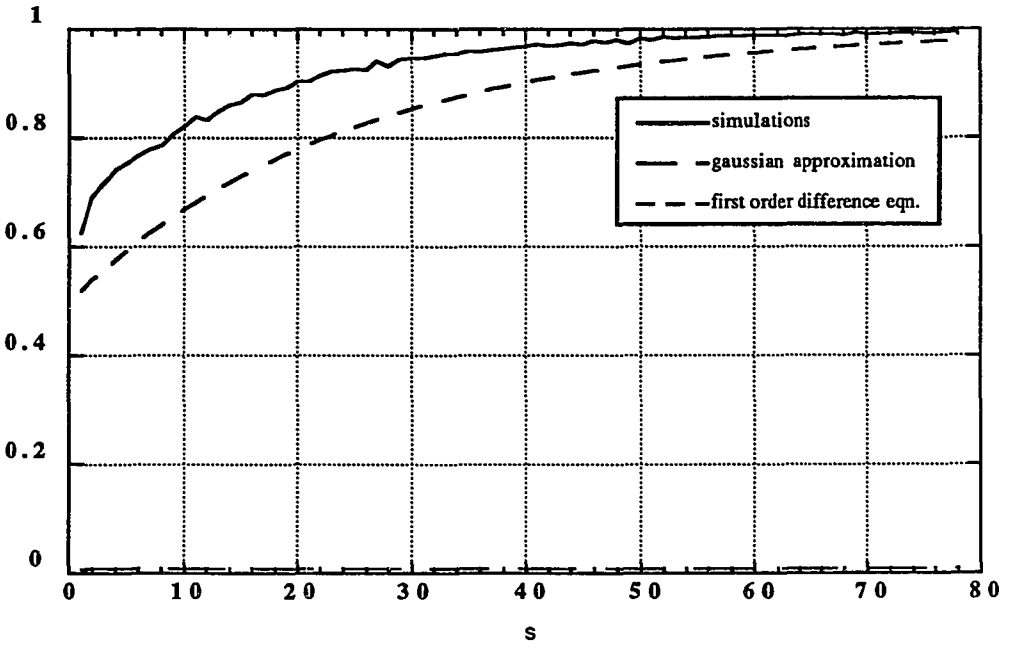

Figure 5.7: $\theta_s/(2.7\sigma^2_{\tilde{N}}/n^2)$ vs. s, $n = 8$, $\rho = 0.8$.

However on the basis of simulations we hypothesize that the correlation between the quantization errors satisfies the following difference equation:

$$\theta_{s+1} = \theta_s + (1-\rho)^2 \left(\frac{2.7\sigma^2_{\tilde{N}}}{n^2} - \theta_s \right), \quad s > l^*. \tag{5.132}$$

Unlike (5.130), this is a first order linear difference equation. Also by substituting for θ_s from (5.128), we can obtain a relation between θ_{s+1} and γ_{s+1}:

$$\theta_{s+1} = (\gamma_{s+1} - 1)\frac{(2-\rho)\sigma^2_{\tilde{N}}}{\rho} + \frac{2.7}{n^2}\left((2-\rho)\sigma^2_{\tilde{N}} + (1-\rho^2)(1-\rho)^2\right) \tag{5.133}$$

for all γ_{s+1} such that θ_{s+1} is nonnegative. This equation implies that the correlation between the quantization error is linear in the correlation between the inputs. For $\gamma_{s+1} = 1$, the correlation between the quantization errors equals $2.7(1-\rho^2)/n^2$ and it decreases linearly with γ_{s+1}. This is unlike the case of jointly Gaussian

random variables, where the correlation between the quantization errors decreases as $\sqrt{1-\gamma_{s+1}}$.

We also make the approximation that $\theta_{l^*} = 0$. With this initial condition the difference equation in θ_s can be solved easily:

$$\theta_s = \begin{cases} \frac{2.7(1-\rho^2)\left(1-(2\rho-\rho^2)^{s-l^*}\right)}{n^2}, & s > l^* \\ 0, & s \leq l^*. \end{cases} \tag{5.134}$$

These solutions have been plotted in the above figures and we see that they agree well with the simulation resuts. We note that as ρ increases, θ_s converges more slowly to its limiting value. We also note that this limiting value equals the variance of the quantization error as expected.

Approximations for $E\left\{X_{j-m}\left(\eta_l(\tilde{N}_{j',l}) - \rho\eta_{l-1}(\tilde{N}_{j'-1,l-1})\right)\right\}$, $j' \geq j-m$

We first consider the correlation $E\{X_{j-m}\eta_l(\tilde{N}_{j',l})\}$, $j' \geq j-m$. By assuming that X_{j-m} and $\tilde{N}_{j',l}$ are jointly Gaussian, we obtain the approximation

$$E\{X_{j-m}\eta_l(\tilde{N}_{j',l})\} \approx \frac{2.7}{n^2}E\{X_{j-m}\tilde{N}_{j',l}\}. \tag{5.135}$$

By substituting for $\tilde{N}_{j',l}$ from (5.118), we can then show that

$$E\left\{X_{j-m}\left(\eta_l(\tilde{N}_{j',l}) - \rho\eta_{l-1}(\tilde{N}_{j'-1,l-1})\right)\right\} \approx \frac{2.7}{n^2}E\{X_{j-m}\sqrt{1-\rho^2}N_{j'}\}. \tag{5.136}$$

Hence the above correlation is zero for $j' > j-m$ and equals $(1-\rho^2)2.7/n^2$ when $j' = j-m$. These results agree with the simulation results.

5.5.2 The Receiver

The receiver of a DPCM system is a linear, recursive filter. It is thus equivalent to an infinite length sliding block system and a finite state approximation is obtained

quite simply by limiting the block length. This receiver is described by the following equation:

$$\hat{X}_j = \sum_{i=0}^{m} q_l(\tilde{N}_{j-i,l}). \tag{5.137}$$

The block size of the receiver is m, i.e., it stores the past m inputs. Thus the state size of the receiver is equal to n_l^m, where n_l is the number of quantization levels of the quantizer q_l in the transmitter.

5.5.3 Mean Square Error

In this section we calculate the quantization error $E\{(X_{j+1} - \hat{X}_{j+1})^2\}$. By using equations (5.29) and (5.137) we can express the quantization error as follows:

$$\xi_q = E\{(\rho^{m+1} X_{l-m} + \sum_{i=0}^{m} \rho^i (\sqrt{1-\rho^2} N_{j-i} - q_l(\tilde{N}_{j-i,l})))^2\}. \tag{5.138}$$

We next use (5.118) to rewrite the above expression in terms of the quantization errors at level $l-1$ and l in the transmitter as follows:

$$\xi_q = E\left\{\left(\rho^{m+1} X_{j-m} + \sum_{i=0}^{m} \rho^i (\eta_l(\tilde{N}_{j-i,l}) - \rho\eta_{l-1}(\tilde{N}_{j-i-1,l-1}))\right)^2\right\}. \tag{5.139}$$

To evaluate this expectation, we expand the square and take the expectation of each term. Since the random processes $\{X_j\}$ and $\{\tilde{N}_{j,s}\}$ are stationary, the above expression can be simplified as follows:

$$\begin{aligned}
\xi_q \;=\; & E\{\rho^{2(m+1)} X_{j-m}^2\} + \frac{1-\rho^{2(m+1)}}{1-\rho^2} E\{(\eta_l(\tilde{N}_{j,l}) - \rho\eta_{l-1}(\tilde{N}_{j-1,l-1}))^2\} + \\
& 2\sum_{i=0}^{m}\sum_{k=1}^{m-i} \rho^{2i+k} E\{(\eta_l(\tilde{N}_{j,l}) - \rho\eta_{l-1}(\tilde{N}_{j-1,l-1}))\,(\eta_l(\tilde{N}_{j-k,l}) - \rho\eta_{l-1}(\tilde{N}_{j-k-1,l-1}))\} \\
& + 2\rho^{m+1}\sum_{i=0}^{m} \rho^i E\{X_{j-m}(\eta_l(\tilde{N}_{j-i,l}) - \rho\eta_{l-1}(\tilde{N}_{j-i-1,l-1}))\}.
\end{aligned} \tag{5.140}$$

We next examine each term in the above expression. Since X_j is a unit variance random variable, the first term in the above equation is $\rho^{2(m+1)}$. To evaluate the

expectation $E\{(\eta_l(\tilde{N}_{j,l}) - \rho\eta_{l-1}(\tilde{N}_{j-1,l-1}))^2\}$, we recall from section 4.3 that the quantization errors at two different time intervals are almost uncorrelated We thus obtain the approximation

$$E\{(\eta_l(\tilde{N}_{j,l}) - \rho\eta_{l-1}(\tilde{N}_{j-1,l-1}))^2\} \approx E\{\eta_l^2(\tilde{N}_{j,l})\} + \rho^2 E\{\eta_{l-1}^2(\tilde{N}_{j-1,l-1})\}. \quad (5.141)$$

To evaluate the expectations in the third term, we expand the product and evaluate the expectation of each resulting term. Since the quantization errors at different time instances are uncorrelated, all the terms can be neglected except for the case when $k = 1$. We then obtain the following approximation:

$$2\sum_{i=0}^{m}\sum_{k=1}^{m-i}\rho^{2i+k}E\{(\eta_l(\tilde{N}_{j,l}) - \rho\eta_{l-1}(\tilde{N}_{j-1,l-1}))(\eta_l(\tilde{N}_{j-k,l}) - \rho\eta_{l-1}(\tilde{N}_{j-k-1,l-1}))\} \approx \frac{2\rho^3(1-\rho^{2(m+1)})}{1-\rho^2}E\{\eta_l(\tilde{N}_{j-1,l})\eta_{l-1}(\tilde{N}_{j-1,l-1})\}.$$

The last term in (5.140) has already been evaluated earlier (5.136).

By combining these results, we can approximate the quantization error as

$$\xi_q \approx \frac{1-\rho^{2(m+1)}}{1-\rho^2}\left((1+\rho^2)\frac{2.7(1-\rho^2)}{n^2} - 2\rho^2 E\{\eta_l(\tilde{N}_{j,l})\,\eta_{l-1}(\tilde{N}_{j,l-1})\}\right) + \rho^{2(m+1)}\left(1 + \frac{5.4(1-\rho^2)}{n^2\rho}\right), \quad (5.142)$$

where the correlation in the above equation can be evaluated using (5.134).

5.5.4 Memory Requirements

We now examine the memory requirements in the transmitter and the receiver. We first note that as the number of structural levels, l, in the transmitter becomes infinitely large, the correlation between the quantization errors of $\tilde{N}_{j,l}$ and $\tilde{N}_{j,l-1}$ equals $2.7(1-\rho^2)/n^2$ (see equation (5.134)). Thus as the block size of the receiver, m, also becomes infinitely large, the quantization error of the system equals that of

a DPCM system:

$$\lim_{l,m\to\infty} \xi_q = \frac{2.7(1-\rho^2)}{n^2}. \tag{5.143}$$

To study the memory requirements in the receiver, we set the number of structural levels in the transmitter equal to infinity. The quantization error can then be approximated as

$$\xi_q \approx \frac{2.7(1-\rho^2)}{n^2} + \rho^{2(m+1)}. \tag{5.144}$$

We note that the first term on the right hand side of the above equation is equal to the quantization error of a DPCM system. We denote the ratio of the difference between the errors of the sliding block and the DPCM system to the quantization error of a DPCM system as α:

$$\alpha = \frac{\xi_q - \xi_{q,dpcm}}{\xi_{q,dpcm}}. \tag{5.145}$$

We can then express the block length of the receiver in terms of n,ρ and α as follows:

$$m = \frac{1}{2\tau} \ln\left(\frac{n^2}{5.4\alpha\tau}\right). \tag{5.146}$$

For large transmission rates and large receiver block length, we know that the performance of this system approaches that of a DPCM system. Hence the optimal number of quantization levels becomes independent of the rate and equals a constant n_o and the optimal sampling rate equals β/r where β is a constant that depends upon the reconstruction filter (see section 5.3.1). Hence the block length m can be expressed as

$$m \approx \frac{r}{2\beta} \ln\left(\frac{n_o^2 r}{5.4\alpha\beta}\right). \tag{5.147}$$

Typical values of n_o and β are 5 and 2.0 respectively. Hence for a given value of α, the block length of the receiver increases asymptotically as $O(r\ln(r))$. Thus for an infinite state size transmitter, the state size of the receiver increases asymptotically as $n_o^{r\ln(r)}$.

We next examine the memory requirements in the transmitter. We assume that the block length of the receiver is infinite. With the same assumptions used in analyzing the memory requirements of the receiver, we can approximate the number of structural levels in the transmitter as

$$l = \frac{r^2}{\beta^2} \ln\left(\frac{r}{\alpha\beta}\right). \tag{5.148}$$

The state size of the transmitter then increases as $n_o^{r^2 \ln(r)}$.

Thus we see that more memory is required in the transmitter than in the receiver. Intuitively we can see that the it is the transmitter that makes the effort to track the state of the receiver and not the other way around and hence we would expect that the state size of the transmitter is larger than that of the receiver.

Chapter 6

Future Work

To conclude this dissertation we mention some directions in which this work can be extended and some problems that remain to be investigated.

Extensions of the problem that we have examined would be to investigate the performance of digital transmission systems when the input is a second or higher order Markov process or when the input is not Gaussian. An important property of Gaussian random variables is that if two jointly Gaussian random variables are uncorrelated then they are also independent. We have used this property extensively in solving the difference equations that describe a quantization system. It remains to be seen if this property is crucial or if our methods of analysis can be applied to non Gaussian inputs. The difficulty with second and higher order Markov processes is the fact that the corresponding sampled process is not Markov. This may not be a major difficulty, since in our analysis we do not make use of the Markov property directly; it is more important that difference equations, similar to those obtained in this dissertation, can still be derived. It may then be possible to find approximations to the solutions of these difference equations using fine quantization techniques, even though the sampled process is not Markov.

A problem that remains to investigated is to determine the joint distribution of the state variables of a DPCM system at two different time instances. Our results indicate that they are not jointly Gaussian although their marginal distributions are approximately Gaussian. This can be further investigated by deriving the

appropriate Chapman-Kolmogorov equations and solving than numerically as in [6, 8]. This would also help in determining the exact behavior of the sliding block, finite state system investigated in chapter 5 where we have had to resort to simulations.

As we have pointed out earlier, the fine quantization approximations that we have developed are valid in evaluating terms up to the order of $1/n^2$, but higher order terms cannot be evaluated. The problem lies in determining the lengths of the quantization regions more accurately. In appendix A we have obtained the required equations but have been unable to find any simple solutions. This is another problem that needs to be examined.

Appendix A

Asymptotic Mean Square Error of Optimum Quantizers

In this appendix we outline the steps in deriving the fine quantizer approximation for the mean square error of a scalar quantizer. We also demonstrate the difficulty in extending the approach if we retain terms of order $1/n^4$ and higher. Our approach follows that of [4].

We assume that the density function $p_X(x)$ of the input X can be approximated by the first few terms of its Taylor series in x within each of the bounded quantization regions and that the overload error can be neglected. The quantization error can then be approximated as follows:

$$\xi_{pw} \approx \sum_{i=1}^{n-2} \int_{r_i}^{r_{i+1}} (x - q_i)^2 \left(p_X(\bar{r}_i) + (x - \bar{r}_i) p'_X(\bar{r}_i) + \frac{1}{2}(x - \bar{r}_i)^2 p''_X(\bar{r}_i) \right) dx, \quad \text{(A.1)}$$

where the representative point q_i can be approximated as in (4.31):

$$q_i \approx \bar{r}_i + \frac{p'(\bar{r}_i)\delta_i^2}{12p(\bar{r}_i)} + \left(\frac{p'''(\bar{r}_i)}{480 p_X(\bar{r}_i)} - \frac{p'(\bar{r}_i)p''(\bar{r}_i)}{288 p^2(\bar{r}_i)} \right) \delta_i^4 \quad \text{(A.2)}$$

By evaluating the integrals in the above equation and retaining the terms up to δ_i^5, we obtain the following approximation:

$$\xi_{pw} \approx \sum_{i=1}^{n-2} \left(\frac{\delta_i^2 p_X(\bar{r}_i)}{12} + \frac{\delta_i^4 p''(\bar{r}_i)}{160} - \frac{\delta_i^4 p'^2_X(\bar{r}_i)}{72 p_X(\bar{r}_i)} \right) \delta_i. \quad \text{(A.3)}$$

In order to simplify the above equation, we assume that for each n there exists a function $\lambda_n()$ such that the length of the quantization region S_i is given by

$$\delta_i \approx \frac{1}{n\lambda(\bar{r}_i)}. \quad \text{(A.4)}$$

The intuitive idea behind this assumption is that the length of a quantization region will vary inversely with the number of quantization regions. Further the length of an interval centered at $\bar{r}_i$ will be a function of $p_X(\bar{r}_i)$. If the density function at $\bar{r}_i$ is small, then the length of the interval at this point will be large and vice versa. The function $\lambda_n()$ is introduced to take this factor into account. We also assume that as n tends to infinity, the function $\lambda_n()$ tends to a limiting function $\lambda()$.

We first note that by rearranging the terms in equation A.4 and summing over the regions, we obtain the relation

$$\sum_{i=0}^{n-1} \lambda(\bar{r}_i)\delta_i = 1. \tag{A.5}$$

In the limit as n tends to infinity, we obtain the relation

$$\int_{-\infty}^{\infty} \lambda(x)\, dx = 1. \tag{A.6}$$

Also in the limit as the number of quantization intervals tends to infinity, the expression for the quantization error in (A.3) can be approximated as follows:

$$\xi_{pw} \approx \int_{-\infty}^{\infty} \left(\frac{p_X(x)}{12n^2\lambda^2(x)} + \frac{1}{n^4\lambda^4(x)} \left(\frac{p''(x)}{160} - \frac{p_X'^2(x)}{72p_X(x)} \right) \right) dx. \tag{A.7}$$

Thus the function $\lambda()$ has to chosen so as to minimize this expression for the quantization error while satisfying the constraint in (A.6). This is a problem in the calculus of variations and it can be shown that $\lambda()$ must satisfy the following equation:

$$\frac{\lambda^2(x)p_X(x)}{4n^2} + \frac{1}{n^4} \left(\frac{p''(x)}{40} - \frac{p_X'^2(x)}{18p_X(x)} \right) = \mu\lambda^5(x), \tag{A.8}$$

where μ is a constant that must be chosen to satisfy the constraint in (A.6). We do not know of any way to solve this equation for $\lambda(x)$. But we note that if the term of order $1/n^4$ is neglected, then it is easy to solve for $\lambda()$:

$$\lambda(x) = \frac{p_X^{1/3}(x)}{\int_{-\infty}^{\infty} p_X^{1/3}(x)\, dx}. \tag{A.9}$$

This leads to the result in (4.18). This expression minimizes only the term in $1/n^2$. This can be easily seen by substituting the above expression into (A.7) and evaluating the terms in $1/n^4$.

Appendix B

Derivation of $E\{\eta_X(X)\eta_Y(Y)\}$ for Highly Correlated Inputs

In this appendix we derive the result in (4.85). We first note that (4.83) can be written as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{r_i}^{r_{i+1}}\int_{l_j}^{l_{j+1}}\left(\sqrt{1-\gamma^2}u-q_j+\gamma q_i\right)^2 p_{X,U}(x,u)\,du\,dx \approx$$

$$\sum_{i=1}^{n}\int_{-\frac{\delta_i}{\sqrt{2\epsilon}}}^{\frac{\delta_i}{\sqrt{2\epsilon}}}\int_{r_i}^{r_{i+1}}\left(\sqrt{2\epsilon}u-\epsilon\bar{r}_i\right)^2 p_{X,U}(\bar{r}_i,u)\,dx\,du$$

$$\int_{\sqrt{\frac{\epsilon}{2}}r_{i+1}}^{\frac{\delta_i}{\sqrt{2\epsilon}}}\int_{\alpha_{i+1}}^{r_{i+1}}\left(-2(\sqrt{2\epsilon}u-\epsilon\bar{r}_i)\delta i+\delta_i^2\right)p_{X,U}(\bar{r}_i,u)\,dx\,du$$

$$\int_{-\frac{\delta_i}{\sqrt{2\epsilon}}}^{\sqrt{\frac{\epsilon}{2}}r_i}\int_{r_i}^{\alpha_i}\left(2(\sqrt{2\epsilon}u-\epsilon\bar{r}_i)\delta i+\delta_i^2\right)p_{X,U}(\bar{r}_i,u)\,dx\,du \qquad \text{(B.1)}$$

Evaluating the integrals with respect to x is trivial since the integrand is not a function of x. We also note that as the correlation between the input random variables tends to one, the ratio $\delta_i/\sqrt{\epsilon}$, $i=1,\ldots,n-1$ tends to infinity when the number of quantization levels is fixed. We can thus substitute $+\infty$ and $-\infty$ for $\delta_i/\sqrt{\epsilon}$ and $-\delta_i/\sqrt{\epsilon}$ in the limits of integration in the above equation. If we assume that the joint density of X and U is small for large values of X, then we can approximate the quantity $r_i\sqrt{\epsilon/2}$ $i=1,\ldots,n$ by 0. By substituting these

approximations into the above equation, we obtain the following approximation:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{r_i}^{r_{i+1}}\int_{l_j}^{l_{j+1}}\left(\sqrt{1-\gamma^2}u-q_j+\gamma q_i\right)^2 p_{X,U}(x,u)\,du\,dx \approx$$

$$\sum_{i=1}^{n}\int_{-\infty}^{\infty}(\sqrt{2\epsilon}u-\epsilon\bar{r}_i)^2\delta_i p_{X,U}(\bar{r}_i,u)\,du +$$

$$\int_0^{\infty}\left(-2\sqrt{2\epsilon}\delta_i u+2\epsilon\bar{r}_i\delta_i+\delta_i^2\right)(r_{i+1}-\alpha_{i+1})p_{X,U}(\bar{r}_i,u)\,du +$$

$$\int_{-\infty}^{0}\left(2\sqrt{2\epsilon}\delta_i u-2\epsilon\bar{r}_i\delta_i+\delta_i^2\right)(\alpha_i-r_i)p_{X,U}(\bar{r}_i,u)\,du \qquad \text{(B.2)}$$

We substitute for α_i from (4.84) and simplify the integrand by retaining only the terms up to the order of δ_i^2 and $\epsilon^{3/2}$. This expression can be further simplified with the usual fine quantizer approximations, i.e., we substitute for δ_i from (4.18) and approximate the summation over i by an integral. We then obtain (4.85).

Bibliography

[1] W. R. Bennet, "Spectra of Quantized Signals," Bell System Technical Journal, pp. 446-472, July 1948.

[2] T. Berger, "Rate Distortion Theory," Prentice-Hall, 1971.

[3] N. T. Gaarder and D. Slepian, "On Optimal Finite-State Digital Transmission Systems," IEEE Trans. on Information Theory, vol. IT-28, pp. 167-186, March 1982.

[4] A. Gersho, "Asymptotically Optimal Block Quantization," IEEE Trans. on Information Theory, vol. IT-25, pp. 373-380, July 1979.

[5] R. M. Gray, "Quantization Noise Spectra," IEEE Trans. on Information Theory, vol. IT-36, pp. 1220-1224, November 1990.

[6] A. Hayashi, "An Exact Analysis of Some Digital Transmission Systems with Gaussian Inputs," Ph.D. Dissertation, University of Hawaii, August 1976.

[7] H. Inose and Y. Yasuda, "A Unity Bit Coding Method by Negative Feedback," Proc. IEEE, vol. 51, November 1963, pp. 1524-1535.

[8] E. Janardhanan, " An Analysis of Leaky Integrator Differential PCM Systems with Stationary First Order Gauss Markov Input," Ph.D. Dissertation, University of Hawaii, May 1979.

[9] N. S. Jayant and P. Noll, "Digital Coding of Waveforms," Prentice-Hall, 1984.

[10] S. P. Lloyd, "Least Squares Quantization in PCM," IEEE Trans. on Information Theory, vol. IT-28, pp. 129-137, March 1982.

[11] J. Max, "Quantization for Minimum Distortion," IRE Trans. on Information Theory, pp. 7-12, March 1960.

[12] L. Schuchman, "Dither Signals and their Effect on Quantization Noise," IEEE Trans. on Commn. Tech., vol COM-12, pp. 162-165, December 1964.

[13] D. Slepian, "On Delta Modulation," Bell System Tech. J., pp. 2101-2137, December 1972.

[14] A. B. Sripad and D. L. Snyder, "A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White," IEEE Trans. on Acoustics, Speech and Signal Processing, pp. 442-449, October 1977.

[15] P. Zador, "Asymptotic Quantization of Continuous Random Variables," unpublished memorandum, Bell Laboratories, 1966.