BINARY DITHERED OVERSAMPLING ANALOG-TO-DIGITAL CONVERTERS

by

Diego Pienovi

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Electrical & Computer Engineering

Fall 2009

BINARY DITHERED OVERSAMPLING ANALOG-TO-DIGITAL CONVERTERS

by

Diego Pienovi

Approved: _____

Gonzalo R. Arce, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Kenneth E. Barner, Ph.D.
Chair of the Department of Electrical & Computer Engineering

Approved: _____

Michael J. Chajes, Ph.D.
Dean of the College of Engineering

Approved: _____

Debra Hess Norris, M.S.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGEMENTS

I wish to thank all the faculty, staff, friends and collegues from the Electrical and Computer Engineering Deparment at the University of Delaware. In particular, I would like to thank my advisor, Dr. Gonzalo Arce who gave me the opportunity to pursue this research.

I must also thank my parents Serrana and Mario and my family and friends in Uruguay and in the United States. This work is specially dedicated to the memory of my father Mario who has not only given me my life but has also taught me to dream, to think independently, to love my country and to try to live a fair and honest life.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

In today's world, analog-to-digital converters (ADC) play a major role. Our modern society requires and depends on electronic devices that process the analog input data in the digital domain, such as cellphones, audio and video systems, and even domestic appliances. Therefore, the development of fast and accurate ADCs have become a key issue for the industry. In general, a good ADC is the one that achieves high resolution (low quantization noise) with low complexity. One of the most popular techniques to decrease the quantization noise is to digitalize the incoming signal with a sampling frequency many times higher than twice the signal bandwith. This technique is called oversampling. Probably the most popular oversampling converters are the Sigma-Delta (SD) modulators which use a very high sampling frequency and a binary quantizer (for first order SD). In SD, the quantizer is embedded into a feedback loop in such a way that the quantization noise is not only spread over the spectrum (because of oversampling), but it is also shaped to the upper frequencies (this is called noise shaping). The big problem with SD is that the quantization noise spectrum presents undesired harmonics caused by the non-linear nature of the quantizer. To avoid this, the solution is to add an independent signal before the quantization stage called dither. It was proven by Lipshitz and Vanderkooy in [1] that dithering a first order SD modulator is ineffective as it turns the modulator in constant overload. In addition to this problem, because of the feedback loop, higher order SD modulators can be unstable. That being said, the aim of this work is to present a simple oversampling ADC without feedback capable of generating the minimum uncorrelated quantization noise that yields the maximum

possible SNR at the output. For that purpose, this work develops the statistical characteristics of the optimum dither that achieves the mentioned goal for different types of input signals.

Chapter 1

INTRODUCTION

In this chapter, we first introduce the two different categories for analog-to-digital converters (ADCs): the Nyquist Rate converters and the Oversampling converters. Then we focus on the Oversampling ADCs and briefly describe the two main groups in this category: the Oversampled PCM converters and the Sigma-Delta converters. Then, the binary dithered-oversampling A/D converter is presented. The main idea of this converter is to add the optimum dithered signal before the binary quantization stage to maximize the output SNR.

1.1 Nyquist Rate and Oversampling A/D Converters

In A/D conversion technology, there basically exist two different types of converters: Nyquist Rate Converters and Oversampling Converters [2]. The Nyquist Rate Converters use a fairly low sampling frequency (slightly above the Nyquist frequency), but a quite high number of quantization levels in order to achieve a certain SNR. On the other hand, Oversampling Converters can achieve the same SNR but with a lower number of quantization levels and a sampling frequency many times above the Nyquist rate. Therefore, there is a trade-off between sampling frequency and number of quantization levels to achieve a certain SNR.

In Figure 1.2, we see how the in-band quantization noise power is reduced when increasing the sampling frequency.

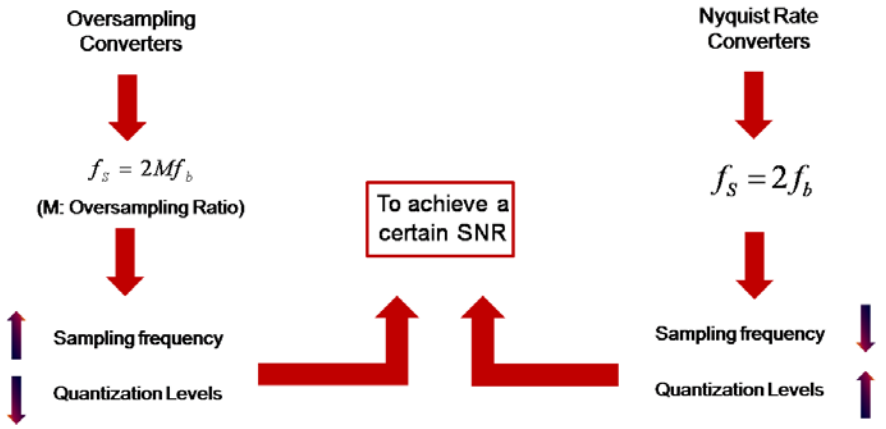

Figure 1.1: Oversampling and Nyquist Rate Converters. There is a trade off between sampling frequency and quantization levels to achieve a certain SNR.

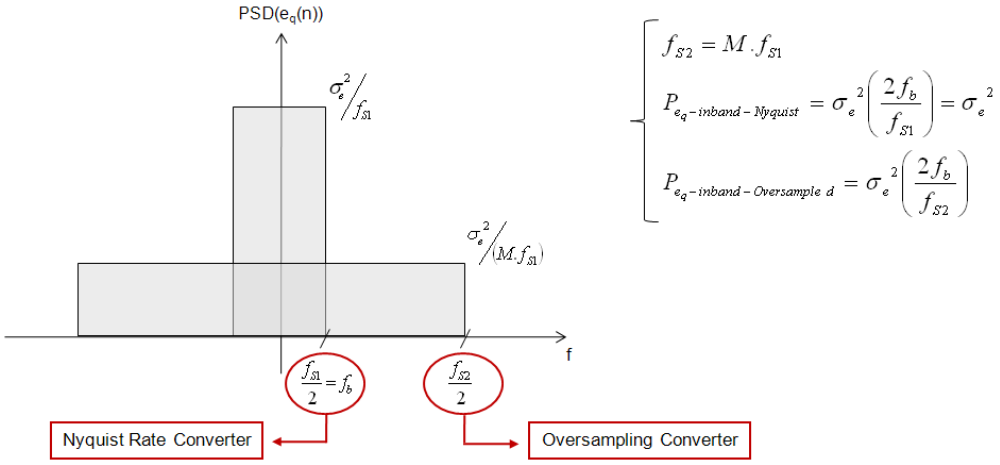

Figure 1.2: Oversampling and quantization noise. The in-band quantization noise is reduced when using oversampling.

If we want to achieve a good performance with few quantization levels, we need to focus on the design of oversampling converters. The two basic configurations for oversampling converters are: the 'Oversampled PCM' and 'Sigma-Delta' [2].

1.1.1 Oversampled PCM

In PCM, the incoming signal is compared to a set of fixed thresholds, and depending on the amplitude of the analog input sample with respect to the thresholds, the output value is selected. These converters simply use a very high sampling frequency (i.e. the sampling frequency is m times higher than the Nyquist limit) in order to spread the quantization noise over the spectrum. In this way, after low-pass filtering, the resulting digitalized signal has a much lower quantization noise (in comparison to the one obtained sampling at the Nyquist frequency). In this case, the conversion is done sample-by-sample, so there is no need to store previous input samples to get the present output. We can say that this is a memoryless A/D conversion. The in-band quantization noise power IQNP in this case is:

$$IQNP = \frac{\sigma_e^2}{m} \tag{1.1}$$

where m is the oversampling ratio and σ_e^2 is the quantization noise power for a linear model of the quantizer [2].

1.1.2 Sigma-Delta Converters

This type of oversampling converter also takes advantage of the high sampling frequency to spread out the quantization noise. After low-pass filtering, the remaining noise is much less than if we were using the Nyquist frequency as the sampling frequency. The interesting thing about Sigma-Delta is that they also perform noise shaping. That means, that the quantization noise is not only spread over the spectrum, but it is also high-pass filtered. In this way, the amount of quantization noise power in the band of interest is significantly reduced. To achieve this,

the system needs feedback and makes use of previous input and output samples to predict the present output. One important disadvantage about Sigma-Delta is that the integrator in the feedback loop may turn the system unstable for certain input signals. In addition to instability, for first-order Sigma-Delta configurations, the output is highly correlated with the input signal, and the spectrum also presents harmonics [3]. One way to solve this problem would be to use dither along with Sigma-Delta [12], but this was proven to be uneffective as the quantizer would be in constant overload [1]. Despite of these issues, Sigma-Delta achieves a much better performance than Oversampled PCM because of the noise shaping property.

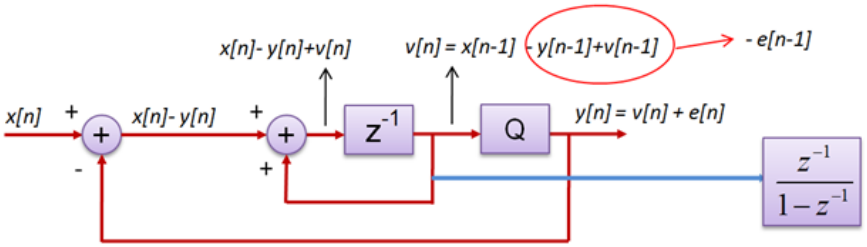

Figure 1.3: First-order single-stage Sigma-Delta. The quantization noise $e[n]$ is shaped by a first order high-pass filter [4].

The in-band quantization noise power in this case is,

$$IQNP = \frac{\sigma_e^2 \pi^2}{3m^3}. \tag{1.2}$$

In Figure 1.4, the in-band quantization noise power is plotted for both cases: Oversampled PCM and First-order Sigma-Delta. It is assumed that $\sigma_e^2 = 1$.

In Figure 1.5, the diagram illustrates how the noise shaping characteristic of Sigma-Delta makes the in-band quantization noise much lower than in the case of PCM. This is the reason why Sigma-Delta performs better in terms of SNR.

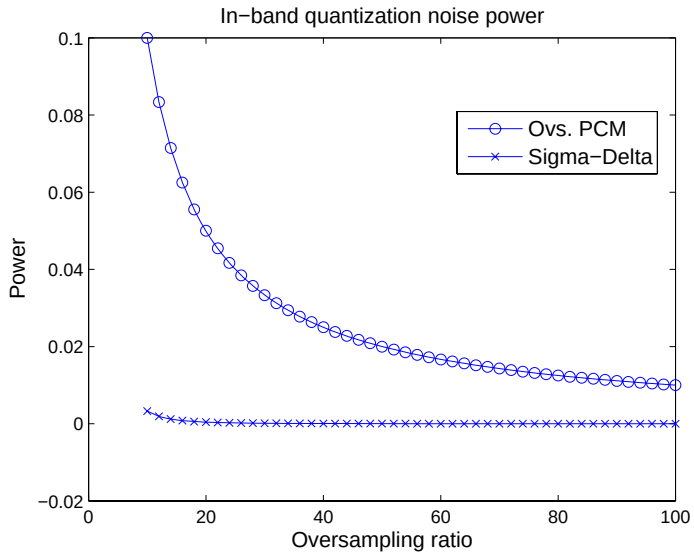

Figure 1.4: In-band quantization noise when $\sigma_e^2 = 1$. Sigma-Delta performs better than oversampling PCM.

1.2 Binary Dithered-Oversampling Analog-to-Digital Conversion

In this section, we are introducing a new approach for analog-to-digital conversion. The main idea of a binary dithered-oversampling converter is to use a dither signal $M[n]$ called 'mask' independent of the analog input $x[n]$ in order to make the quantization error uncorrelated to the input signal. We will focus our attention in the binary quantization problem, as it covers the most general case. For the case when we have more than two possible quantization levels, the problem can be splitted into several binary quantization problems where the output $y[n]$ can take the value a or b with probability p and $(1-p)$ if we know that $x[n]$ lies in (a, b). However, we will not concentrate in this case and it will not be part of the analysis in the following chapters. Moreover, the higher the number of quantization levels considered, the more uncorrelated the quantization error and the input signal will be, so there will be no point in using dither before the quantization stage [2]. Just like any other oversampling converter, the input signal $x[n]$ will be sampled at a rate m times higher than the Nyquist limit in order to spread the quantization noise over a

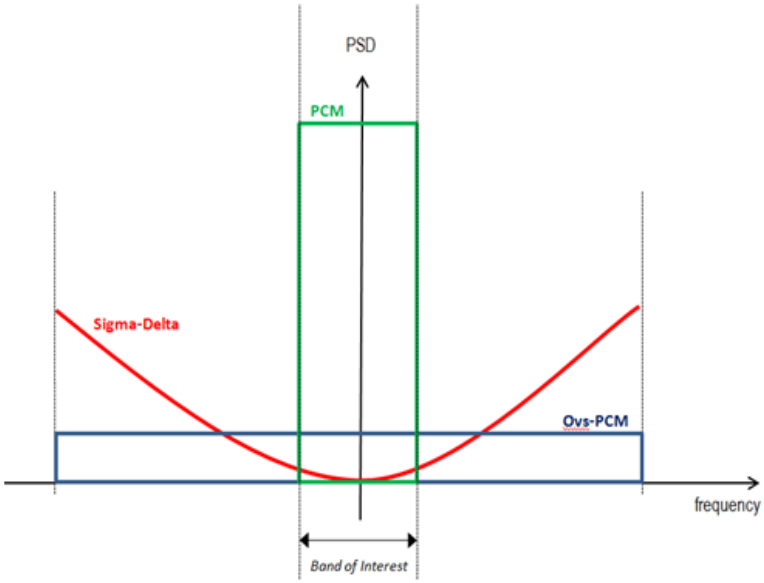

Figure 1.5: Quantization noise spectrum for PCM, Oversampling PCM and Sigma-Delta. The noise shaping property of Sigma-Delta makes this technique to be the one with the lowest in-band quantization noise power.

broad spectrum. Then, decimation has to be performed. The addition of the dither signal $M[n]$ is equivalent to make the quantization thresholds variables (instead of using the fixed ones like PCM). The idea is to control this pseudo-randomness of the quantization levels in order to achieve the best performance possible at the output (i.e. we want to make the in-band quantization noise power as small as possible). In this way, we not only get rid of the harmonics in the quantization noise spectra, but we also improve the output SNR by shaping the quantization noise.

To begin with, let us assume that we have an analog input sample $x[n] \in (-1, 1)$. Its corresponding binary quantized output $y[n]$ will take the value -1 with probability p_0, and the value +1 with probability p_1. Furthermore, the expected

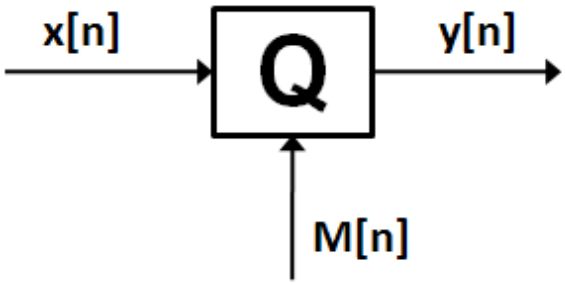

Figure 1.6: Binary Dithered-Oversampling Analog-to-Digital Converter.

value of $y[n]$ must be x[n]. Therefore,

$$\begin{cases} -p_0 + p_1 &= x[n] \\ p_0 + p_1 &= 1. \end{cases} \tag{1.3}$$

Solving for p_0 and p_1 we have,

$$\begin{cases} p_0 &= \frac{1-x[n]}{2} \\ p_1 &= \frac{1+x[n]}{2}. \end{cases} \tag{1.4}$$

The mask signal $M[n] \in (-1, 1)$ is an external signal uncorrelated with the input $x[n]$ with certain statistical properties. In particular it will have a probability density function $f_M(m)$ and an autocorrelation function $r_M(k)$. For every incoming sample $x[n]$, a decision about $y[n]$ being -1 or +1 will be made based on the corresponding value of $M[n]$. If $M[n]$ lies below a certain threshold $t[n]$, we will assign $y[n] = -1$. On the other hand, if $M[n]$ is greater than $t[n]$, $y[n]$ will be +1. Therefore, the probability of $M[n]$ being less than $t[n]$ must be p_0, whereas the probability of $M[n]$ greater than $t[n]$ must be p_1.

$$p(M[n] < t[n]) \;=\; \int_{-1}^{t[n]} f_M(m)\mathrm{d}m \;=\; p_0$$

$$p(M[n] > t[n]) \;=\; \int_{t[n]}^{1} f_M(m)\mathrm{d}m \;=\; p_1. \tag{1.5}$$

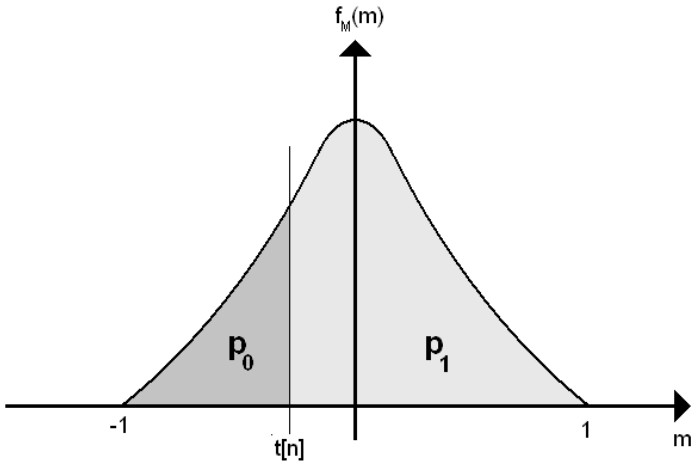

Figure 1.7: Probability density function of $M[n]$. The probability of $M[n]$ being less than $t[n]$ is p_0, whereas the probability of $M[n]$ being greater than $t[n]$ is p_1.

In paricular, if $M[n]$ is uniformly distributed, we will have,

$$f_M(m) = \begin{cases} \frac{1}{2} & if \quad -1 \leq m \leq 1 \\ 0 & otherwise \end{cases} \tag{1.6}$$

and

$$F_M(m) = \begin{cases} 0 & if \quad -\infty \leq m \leq -1 \\ \frac{m+1}{2} & if \quad -1 \leq m \leq 1 \\ 1 & if \quad 1 \leq m \leq \infty \end{cases} \tag{1.7}$$

with $F_M(m)$ being the corresponding cumulative distribution function. In this case, recalling equation (1.5),

$$F_M(t[n]) = \frac{t[n]+1}{2} = p_0. \tag{1.8}$$

From equation (1.3) and equation (1.5), it is then very easy to prove that $t[n] = -x[n]$.

$$\begin{aligned}
\int_{t[n]}^{1} f_M(m)\mathrm{d}m - \int_{-1}^{t[n]} f_M(m)\mathrm{d}m &= x[n] \\
\int_{t[n]}^{1} \frac{1}{2}\mathrm{d}m - \int_{-1}^{t[n]} \frac{1}{2}\mathrm{d}m &= x[n] \\
\frac{1-t[n]}{2} - \Big(\frac{t[n]+1}{2}\Big) &= x[n] \\
-t[n] &= x[n]. \qquad (1.9)
\end{aligned}$$

With all these things considered, for uniform masks $M[n]$, the quantization rule is the following: If $M[n] \leq -x[n]$, then $y[n] = -1$. On the other hand, if $M[n] \geq -x[n]$, then $y[n]$ will be $+1$. Therefore, for uniform $M[n]$,

$$y[n] = sgn(x[n] + M[n]). \qquad (1.10)$$

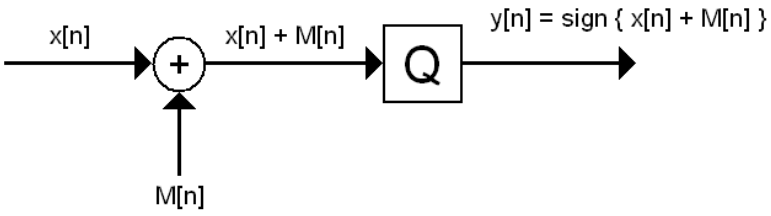

Figure 1.8: Binary Dithered-Oversampling ADC when $M[n]$ is uniformly distributed.

For the case of non-uniform $M[n]$, it is possible to get a similar expression as long as the distribution function is known. In general,

$$y[n] = sgn(M[n] - t[n]). \qquad (1.11)$$

Chapter 2

OPTIMUM DITHER FOR A CONSTANT INPUT

In this chapter we consider the case when we want to quantize an analog DC signal (i.e. $x[n] = \lambda \ \ \forall n$). This case appears to be of interest for oversampling converters as an array of analog samples could be approximated by a constant signal as long as the oversampling ratio is sufficiently large [3]. Our goal will be to find the optimum statistical properties of $M[n]$ to maximize the output SNR.

2.1 Setting up the problem

Let us consider the case where we have an analog DC input $x[n] = \lambda$ sampled at a rate well above the Nyquist frequency and we want to quantize this signal using a binary quantizer and dither. For simplicity, we can assume that $M[n]$ has a uniform distribution. Then, our objective will be to give the optimum joint statistical properties of $M[n]$ such that the output SNR after decimation is maximized (i.e. we want to minimize the in-band quantization noise power (IQNP)). The binary oversampled signal will be in this case $y[n] = sgn(M[n] + \lambda)$.

Our optimization criteria is minimizing the IQNP, therefore, we first need to get an expression for the power spectral density (PSD) of the quantization error $e[n] = y[n] - \lambda$. To do so, we need to start by getting an expression for the autocorrelation of $e[n]$ called $r_e(k)$.

$$r_e(k) = E\Big\{e[n]e[n-k]\Big\}. \tag{2.1}$$

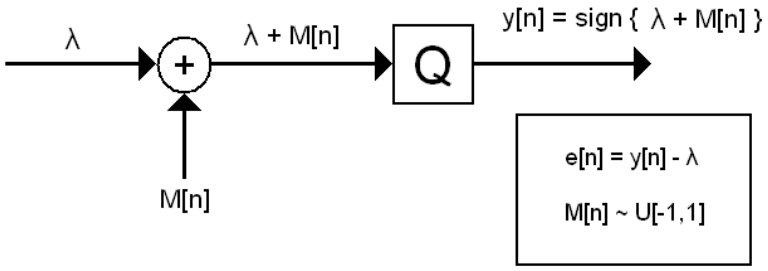

Figure 2.1: Dithered-oversampling binary ADC. The input $x[n]$ is constant (i.e. $x[n] = \lambda$).

Expanding (2.1) we have,

$$\begin{aligned} r_e(k) &= E\Big\{\Big[sgn(M[n]+\lambda)-\lambda\Big]\Big[sgn(M[n-k]+\lambda)-\lambda\Big]\Big\} \\ &= E\Big\{sgn(M[n]+\lambda)sgn(M[n-k]+\lambda)\Big\} - \lambda E\Big\{sgn(M[n]+\lambda)\Big\} \\ &\quad -\lambda E\Big\{sgn(M[n-k]+\lambda)\Big\} + \lambda^2 \\ &= E\Big\{sgn(M[n]+\lambda)sgn(M[n-k]+\lambda)\Big\} \\ &\quad -2\lambda E\Big\{sgn(M[n]+\lambda)\Big\} + \lambda^2. \end{aligned} \tag{2.2}$$

As mentioned above, $M[n]$ is uniform, therefore, its PDF is defined as in (1.6). Let's now work on the term $E\Big\{sgn(M[n]+\lambda)\Big\}$.

$$\begin{aligned} E\Big\{sgn(M[n]+\lambda)\Big\} &= \int_{-\infty}^{+\infty} sgn(m+\lambda) f_M(m)\mathrm{d}m \\ &= \int_{-1}^{+1} sgn(m+\lambda)\frac{1}{2}\mathrm{d}m \\ &= \int_{-\lambda}^{+1}\frac{1}{2}\mathrm{d}m \;-\; \int_{-1}^{-\lambda}\frac{1}{2}\mathrm{d}m \\ &= \frac{1+\lambda}{2} \;-\; \frac{1-\lambda}{2} \\ &= \lambda. \end{aligned} \tag{2.3}$$

Now,

$$\begin{aligned} r_e(k) &= E\Big\{sgn(M[n]+\lambda)sgn(M[n-k]+\lambda)\Big\} - 2\lambda^2 + \lambda^2 \\ &= E\Big\{sgn(M[n]+\lambda)sgn(M[n-k]+\lambda)\Big\} - \lambda^2. \end{aligned} \tag{2.4}$$

For $k = 0$ we get,

$$r_e(0) = 1 - \lambda^2. \tag{2.5}$$

To get the values of $r_e(k)$ for $k > 0$, we need to consider the joint statistics of $M[n]$. Let's define the joint cumulative distribution funcion (CDF) of $M[n]$ and $M[n-k]$ as follows,

$$\begin{aligned} F_{m_0,m_k}(m_0, m_k) &= p(M[n] \leq m_0, M[n-k] \leq m_k) \\ &= \int_{-1}^{m_k}\int_{-1}^{m_0} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \end{aligned} \tag{2.6}$$

with $f_{m_0,m_k}(m_0, m_k)$ being the corresponding joint probability density function (PDF).

$$f_{m_0,m_k}(m_0, m_k) = \frac{\partial^2 F_{m_0,m_k}(m_0, m_k)}{\partial m_0 \partial m_k}. \tag{2.7}$$

We know that $M[n]$ is uniform in $(-1, 1)$. This implies the following conditions for $F_{m_0,m_k}(m_0, m_k)$,

$$F_{m_0,m_k}(1, m_k) = F_M(m_k) = \begin{cases} 0 & if \quad -\infty \leq m_k \leq -1 \\ \frac{m_k+1}{2} & if \quad -1 \leq m_k \leq 1 \\ 1 & if \quad 1 \leq m_k \leq \infty \end{cases} \tag{2.8}$$

$$F_{m_0,m_k}(m_0, 1) = F_M(m_0) = \begin{cases} 0 & if \quad -\infty \leq m_0 \leq -1 \\ \frac{m_0+1}{2} & if \quad -1 \leq m_0 \leq 1 \\ 1 & if \quad 1 \leq m_0 \leq \infty \end{cases} \tag{2.9}$$

and

$$F_{m_0,m_k}(m_0, -1) = F_{m_0,m_k}(-1, m_k) = 0. \tag{2.10}$$

Let's now recall a simple relation between a two dimensional PDF and its corresponding CDF that will be useful for the mathematical derivation of $r_e(k)$. If

$F_{x,y}(x,y)$ is a two dimensional CDF with a corresponding PDF $f_{x,y}(x,y)$, the following relation holds,

$$\int_{y_1}^{y_2}\int_{x_1}^{x_2} f_{x,y}(x,y)\mathrm{d}x\mathrm{d}y = F_{x,y}(x_2,y_2) - F_{x,y}(x_1,y_2) - F_{x,y}(x_2,y_1) + F_{x,y}(x_1,y_1). \tag{2.11}$$

With all these things considered, we can go back to equation (2.4) and work with the term $E\Big\{sgn(M[n]+\lambda)sgn(M[n-k]+\lambda)\Big\} = g(M[n],M[n-k],\lambda)$. Therefore,

$$\begin{aligned} g(M[n],M[n-k],\lambda) &= \int_{-1}^{1}\int_{-1}^{1} sgn(x+\lambda)sgn(y+\lambda)f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-1}^{-\lambda}\int_{-1}^{-\lambda} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \;-\; \int_{-\lambda}^{1}\int_{-1}^{-\lambda} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \\ &\quad - \int_{-1}^{-\lambda}\int_{-\lambda}^{1} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \;+\; \int_{-\lambda}^{1}\int_{-\lambda}^{1} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y. \end{aligned}$$

Then, it is possible to write the four double integrals in terms of $F_{m_0,m_k}(m_0,m_k)$ by making use of the property described in equation (2.11).

$$\begin{aligned} g(M[n],M[n-k],\lambda) &= F_{m_0,m_k}(-\lambda,-\lambda) - F_{m_0,m_k}(-1,-\lambda) - F_{m_0,m_k}(-\lambda,-1) \\ &\quad + F_{m_0,m_k}(-1,-1) - [F_{m_0,m_k}(-\lambda,1) - F_{m_0,m_k}(-1,1) \\ &\quad - F_{m_0,m_k}(-\lambda,-\lambda) + F_{m_0,m_k}(-1,-\lambda)] - [F_{m_0,m_k}(1,-\lambda) \\ &\quad - F_{m_0,m_k}(-\lambda,-\lambda) - F_{m_0,m_k}(1,-1) + F_{m_0,m_k}(-\lambda,-1)] \\ &\quad + F_{m_0,m_k}(1,1) - F_{m_0,m_k}(-\lambda,1) - F_{m_0,m_k}(1,-\lambda) \\ &\quad + F_{m_0,m_k}(-\lambda,-\lambda). \end{aligned} \tag{2.12}$$

From equations (2.8), (2.9), (2.10) and the fact that $F_{m_0,m_k}(1,1) = 1$, we know that

$$F_{m_0,m_k}(x,-1) = F_{m_0,m_k}(-1,y) = 0 \tag{2.13}$$

and

$$F_{m_0,m_k}(-\lambda,1) = F_{m_0,m_k}(1,-\lambda) = \frac{1-\lambda}{2}. \tag{2.14}$$

Therefore,

$$\begin{aligned} g(M[n], M[n-k], \lambda) &= 4F_{m_0,m_k}(-\lambda, -\lambda) + 1 - 4\Big[\frac{1-\lambda}{2}\Big] \\ &= 4F_{m_0,m_k}(-\lambda, -\lambda) + 2\lambda - 1. \end{aligned} \tag{2.15}$$

Now we are ready to substitute in equation (2.4),

$$\begin{aligned} r_e(k) &= E\Big\{sgn(M[n]+\lambda)sgn(M[n-k]+\lambda)\Big\} - \lambda^2 \\ &= 4F_{m_0,m_k}(-\lambda, -\lambda) + 2\lambda - 1 - \lambda^2. \end{aligned} \tag{2.16}$$

Equation (2.16) gives us an expression for the autocorrelation of the error $r_e(k)$ in terms of the statistics of $M[n]$ when the input to the system is a constant $\lambda \in (-1, 1)$. It is a well known result that the power spectral density (PSD) of the error $e[n]$ (that we will call $S_e(w)$) is the Discrete Time Fourier Transform (DTFT) of $r_e(k)$ and can be expressed as follows,

$$S_e(w) = r_e(0) + 2\sum_{k=1}^{+\infty} r_e(k)cos(wk). \tag{2.17}$$

The optimization criteria that we need to consider is the in-band quantization noise power (IQNP), which is in fact the integral of $S_e(w)$ in the band of interest.

$$\begin{aligned} IQNP &= \frac{1}{2\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} S_e(w)\mathrm{d}w \\ &= \frac{1-\lambda^2}{m} + \frac{1}{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \sum_{k=1}^{+\infty} r_e(k)cos(wk)\mathrm{d}w \end{aligned} \tag{2.18}$$

where m is the oversampling ratio. It is reasonable to assume that for a certain integer N, $r_e(k) = 0\ \forall k \geq N$. Therefore, the infinite sum in (2.18) becomes finite and the IQNP can be expressed as follows,

$$\begin{aligned} IQNP &= \frac{1-\lambda^2}{m} + \frac{1}{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \sum_{k=1}^{N} r_e(k)cos(wk)\mathrm{d}w \\ &= \frac{1-\lambda^2}{m} + \sum_{k=1}^{N} \frac{2}{k\pi} sin\Big(\frac{k\pi}{m}\Big) r_e(k). \end{aligned} \tag{2.19}$$

We observe that the IQNP is linear with respect to the coefficients $r_e(k)$. Therefore, given N, we could find an optimal set of $r_e(k)$ with $k = 1, .., N$ that make the IQNP minimum. The set can be found by using linear programming optimization.

When $k \geq N$, we are actually assuming that the samples $M[n]$ and $M[n-k]$ are independent. Therefore, for $k \geq N$,the joint PDF will be the multiplication of two uniform densities in the range $(-1, 1)$.

$$f_{m_0,m_k}(m_0, m_k) = \begin{cases} 1/4 & if \quad -1 \leq m_0 \leq -1 \quad and \quad -1 \leq m_k \leq -1 \\ 0 & otherwise. \end{cases} \tag{2.20}$$

If this happens, the respective CDF for $-1 \leq m_0 \leq -1 \quad and \quad -1 \leq m_k \leq -1$ will be

$$F_{m_0,m_k}(m_0, m_k) = \frac{(1+m_0)(1+m_k)}{4}. \tag{2.21}$$

In this case, substituting in equation (2.16), we get

$$\begin{aligned} r_e(k) &= 4\Big[\frac{(1-\lambda)^2}{4}\Big] + 2\lambda - 1 - \lambda^2 \\ &= (1-\lambda)^2 + 2\lambda - 1 - \lambda^2 \\ &= 0 \end{aligned} \tag{2.22}$$

as expected.

Once we get the optimum $r_e(k)$, by using equation (2.16), we can find the corresponding values of $F_{m_0,m_k}(-\lambda, -\lambda)$ for all k. The set of functions $F_{m_0,m_k}(m_0, m_k)$ satisfying these conditions will be the solution for our optimal mask $M[n]$. That means that the mask given by the joint CDFs $F_{m_0,m_k}(m_0, m_k)$ for all k will be the one that minimizes the in-band quantization noise power (i.e. the one that maximizes the SNR at the output). However, before finding the optimum coefficients, there are some constraints that must be addressed.

Firstly, $F_{m_0,m_k}(m_0, m_k)$ is monotonically non-decreasing and right-continuous in both variables m_0 and m_k. Now, recalling the border conditions given by equations (2.8), (2.9) and (2.10), we end up with an upper and a lower bound for $F_{m_0,m_k}(m_0, m_k)$.

$$0 \leq F_{m_0,m_k}(m_0, m_k) \leq min\Big\{\frac{1+m_0}{2}, \frac{1+m_k}{2}\Big\}. \tag{2.23}$$

Secondly, the PSD of the error $S_e(w)$ has to be positive. $S_e(w)$ is a function of $r_e(k)$, and $r_e(k)$ is a function of $F_{m_0,m_k}(m_0, m_k)$. Therefore, this condition also has impact on $F_{m_0,m_k}(m_0, m_k)$ for all k.

$$\begin{aligned} S_e(w) \geq 0 \quad &\Rightarrow \quad r_e(0) + 2\sum_{k=1}^{N} r_e(k)cos(wk) \geq 0 \\ &\Rightarrow \quad 1 - \lambda^2 + 2\sum_{k=1}^{N}\Big[4F_{m_0,m_k}(-\lambda,-\lambda) + 2\lambda - 1 - \lambda^2\Big]cos(wk) \geq 0. \end{aligned} \tag{2.24}$$

The third point to consider is that the PSD of $M[n]$ (that we will call $S_M(w)$) also needs to be positive. This has impact on the values of the corresponding autocorrelation $r_M(k)$ which is also function of the joint PSD of the mask $f_{m_0,m_k}(m_0, m_k)$ for all k.

$$\begin{aligned} S_M(w) \geq 0 \quad &\Rightarrow \quad r_M(0) + 2\sum_{k=1}^{+\infty} r_M(k)cos(wk) \geq 0 \\ &\Rightarrow \quad 1/3 + 2\sum_{k=1}^{+\infty}\Big[\int_{-1}^{+1}\int_{-1}^{+1} vw f_{m_0,m_k}(v,w)\mathrm{d}v\mathrm{d}w\Big]cos(wk) \geq 0. \end{aligned} \tag{2.25}$$

Lastly, we need to consider that the joint PDFs $f_{m_0,m_k}(m_0, m_k)$ should be such that $\lim_{k\to\infty} r_M(k) = 0$.

The problem of finding an optimum mask by finding the optimum set of functions $F_{m_0,m_k}(m_0, m_k)$ under these constraints is not trivial and can easily lead to mistakes. Therefore, we will turn the problem of finding the optimum $F_{m_0,m_k}(m_0, m_k)$

into the problem of finding the optimum mask spectrum (i.e. finding the optimum $S_M(w)$). This means that instead of looking for the set of CDFs, we will define a general $f_{m_0,m_k}(m_0, m_k)$ that will allow us to get any desired spectrum for $M[n]$.

2.2 Model for a general joint PDF for M[n]

In this section we introduce a general joint PDF for $M[n]$. For every k we will have a PDF $f_{m_0,m_k}(m_0, m_k)$ depending on a tunable parameter C_k. By varying C_k for all k it will be possible to get any desired spectrum for $M[n]$.

To begin with, let us recall that $|r_M(k)| \leq r_M(0) = 1/3$. Therefore, $-1/3 \leq r_M(k) \leq 1/3$. Now let's look at the definition of $r_M(k)$.

$$\begin{aligned} r_M(k) &= \int_{-1}^{1}\int_{-1}^{1} m_0 m_k f_{m_0,m_k}(m_0, m_k)\mathrm{d}m_k\mathrm{d}m_0 \\ &= \int_{-1}^{1}\int_{-1}^{1} \frac{m_0 m_k}{2} f_{m_0/m_k}(m_0/m_k)\mathrm{d}m_k\mathrm{d}m_0 \\ &= \int_{-1}^{1} \frac{m_k}{2}\left[\int_{-1}^{1} m_0 f_{m_0/m_k}(m_0/m_k)\mathrm{d}m_0\right]\mathrm{d}m_k \\ &= \int_{-1}^{1} \frac{m_k}{2} E(m_0/m_k)\mathrm{d}m_k. \end{aligned} \tag{2.26}$$

We want the integral in equation (2.26) to be $\frac{C_k}{3}$ with $-1 \leq C_k \leq 1$. Therefore, we will need to find a $f_{m_0/m_k}(m_0/m_k)$ such that $E(m_0/m_k) = C_k m_k$ for $-1 \leq C_k \leq 1$. In this way,

$$\begin{aligned} r_M(k) &= \int_{-1}^{1} \frac{m_k}{2} E(m_0/m_k)\mathrm{d}m_k \\ &= \int_{-1}^{1} \frac{m_k}{2} C_k m_k \mathrm{d}m_k \\ &= \frac{C_k}{3}. \end{aligned} \tag{2.27}$$

It is easy to note that if we define the PDF of m_0 given m_k to be uniform (i.e. $f_{m_0/m_k}(m_0/m_k)$ is uniform with respect to m_0), we get $E(m_0/m_k) = \frac{C_k}{3}$ for $-1 \leq$

$C_k \leq 1$. Therefore,

$$f_{m_0/m_k}(m_0/m_k) = \begin{cases} \Big[2(1 - C_k m_k)\Big]^{-1} & if \quad \begin{cases} 2C_k m_k - 1 \leq m_0 \leq 1 \\ 0 \leq C_k m_k \leq 1 \end{cases} \\ \Big[2(1 + C_k m_k)\Big]^{-1} & if \quad \begin{cases} -1 \leq m_0 \leq 1 + 2C_k m_k \\ -1 \leq C_k m_k \leq 0 \end{cases} \\ 0 & otherwise. \end{cases} \tag{2.28}$$

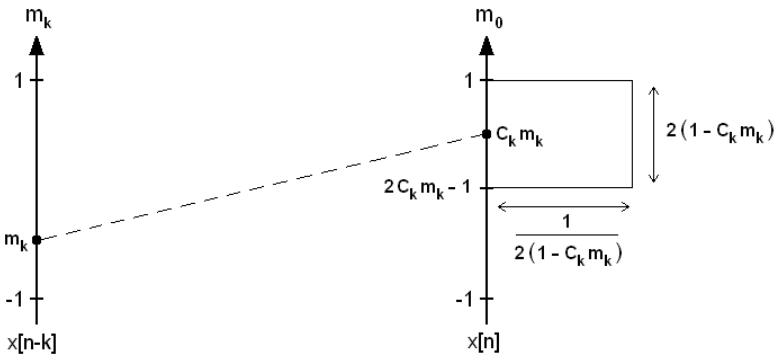

Figure 2.2: Conditional probability density function $f_{m_0/m_k}(m_0/m_k)$. Given m_k, $f_{m_0/m_k}(m_0/m_k)$ is a uniform random variable centered in $C_k m_k$

The joint PDF $f_{m_0,m_k}(m_0, m_k)$ will be the multiplication of $f_{m_0/m_k}(m_0/m_k)$ by the marginal (which is uniform in $(-1, 1)$).

$$f_{m_0,m_k}(m_0, m_k) = \begin{cases} \frac{f_{m_0/m_k}(m_0/m_k)}{2} & if \quad -1 \leq m_0, m_k \leq 1 \\ 0 & otherwise. \end{cases} \tag{2.29}$$

Figures 2.3 and 2.4 show the domain for $f_{m_0,m_k}(m_0, m_k)$. Figures 2.5, 2.6 and 2.7 depict the joint PDF when the parameter C_k is varied. We observe that as C_k approaches zero, $f_{m_0,m_k}(m_0, m_k)$ tends to be uniform.

In this way, we have found a joint PDF for $M[n]$ and $M[n-k]$ that makes the autocorrelation $r_M(k)$ vary between $-1/3$ and $1/3$, so we can pick any desired

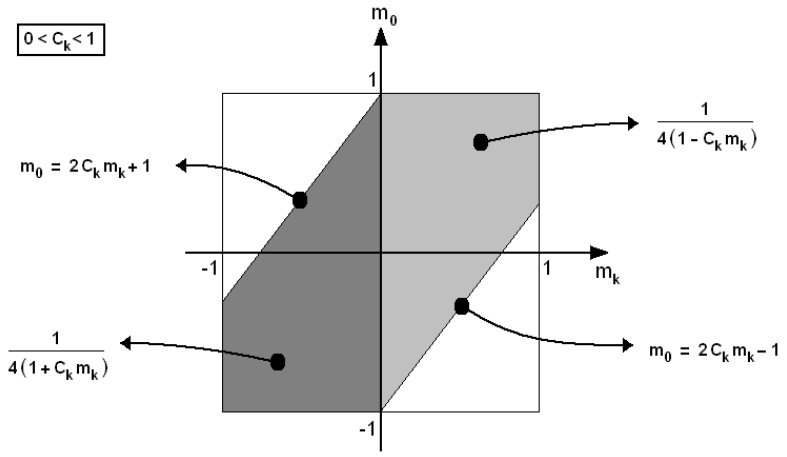

Figure 2.3: Domain for $f_{m_0,m_k}(m_0, m_k)$ when $0 \leq C_k \leq 1$.

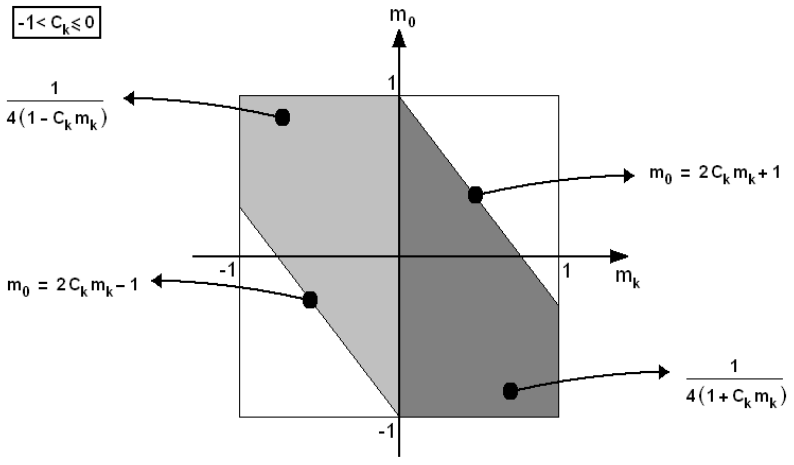

Figure 2.4: Domain for $f_{m_0,m_k}(m_0, m_k)$ when $-1 \leq C_k \leq 0$.

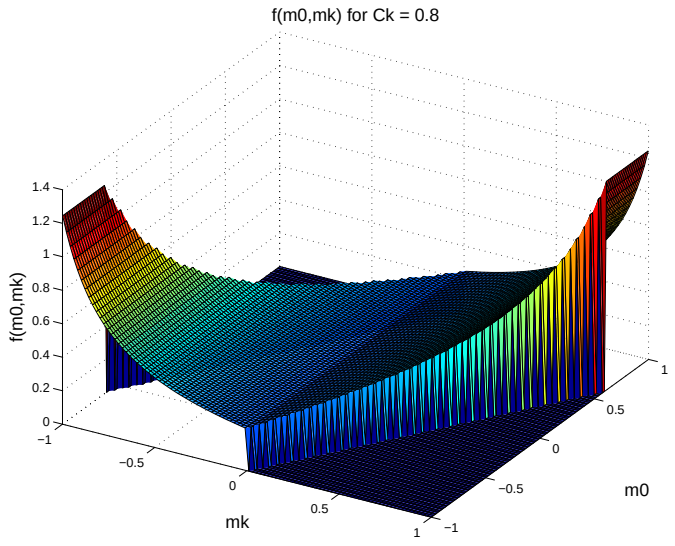

Figure 2.5: Joint PDF $f_{m_0,m_k}(m_0, m_k)$ for $C_k = 0.8$.

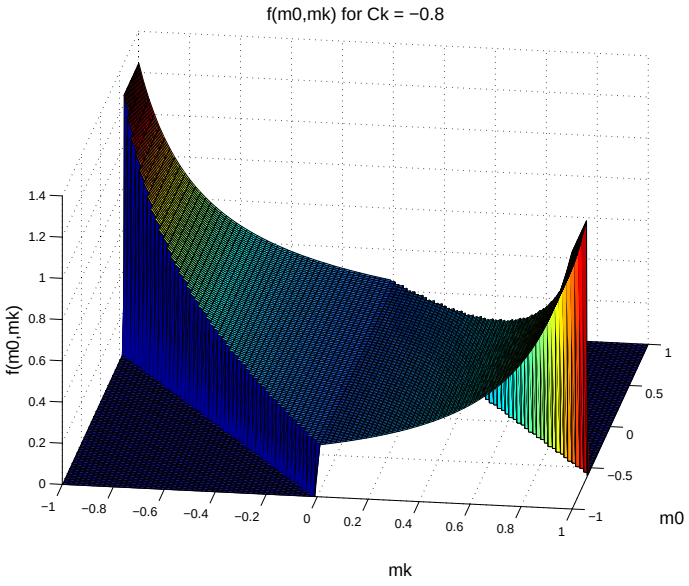

Figure 2.6: Joint PDF $f_{m_0,m_k}(m_0, m_k)$ for $C_k = -0.8$.

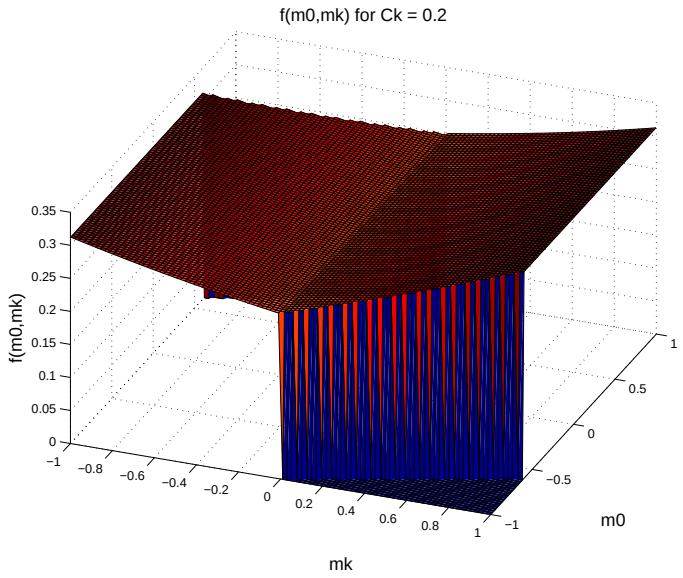

Figure 2.7: Joint PDF $f_{m_0,m_k}(m_0, m_k)$ for $C_k = 0.2$.

spectrum for $M[n]$ by choosing the appropriate value of C_k for all k. The corresponding joint CDF will be described by a large and complex expression defined as follows,

$$
\begin{aligned}
F_{m_0,m_k}(m_0, m_k) &= \int_{-1}^{m_0}\int_{-1}^{m_k} f_{m_0,m_k}(x,y)\mathrm{d}y\mathrm{d}x \\
&= g_1(m_0, m_k)\ if \begin{cases} 1/2 \le C_k \le 1 \\ -1 \le m_k \le 0 \\ -1 \le m_0 \le -2C_k + 1 \end{cases} \\
&\quad g_3(m_0, m_k)\ if \begin{cases} 1/2 \le C_k \le 1 \\ -1 \le m_k \le 0 \\ -2C_k + 1 \le m_0 \le 2C_k m_k + 1 \end{cases} \\
&\quad g_2(m_0, m_k)\ if \begin{cases} 1/2 \le C_k \le 1 \\ -1 \le m_k \le 0 \\ 2C_k m_k + 1 \le m_0 \le 1 \end{cases}
\end{aligned}
$$

$$
g_8(m_0, m_k)\ if\ \begin{cases} 1/2 \le C_k \le 1 \\ 0 \le m_k \le \frac{1-C_k}{C_k} \\ -1 \le m_0 \le 2C_k m_k - 1 \end{cases}
$$

$$
g_6(m_0, m_k)\ if\ \begin{cases} 1/2 \le C_k \le 1 \\ 0 \le m_k \le \frac{1-C_k}{C_k} \\ 2C_k m_k - 1 \le m_0 \le -2C_k + 1 \end{cases}
$$

$$
g_4(m_0, m_k)\ if\ \begin{cases} 1/2 \le C_k \le 1 \\ 0 \le m_k \le \frac{1-C_k}{C_k} \\ -2C_k + 1 \le m_0 \le 1 \end{cases}
$$

$$
g_8(m_0, m_k)\ if\ \begin{cases} 1/2 \le C_k \le 1 \\ \frac{1-C_k}{C_k} \le m_k \le 1 \\ -1 \le m_0 \le -2C_k + 1 \end{cases}
$$

$$
g_7(m_0, m_k)\ if\ \begin{cases} 1/2 \le C_k \le 1 \\ \frac{1-C_k}{C_k} \le m_k \le 1 \\ -2C_k + 1 \le m_0 \le 2C_k m_k - 1 \end{cases}
$$

$$
g_5(m_0, m_k)\ if\ \begin{cases} 1/2 \le C_k \le 1 \\ \frac{1-C_k}{C_k} \le m_k \le 1 \\ 2C_k m_k - 1 \le m_0 \le 1 \end{cases}
$$

$$
g_1(m_0, m_k)\ if\ \begin{cases} 0 \le C_k \le 1/2 \\ -1 \le m_k \le 0 \\ -1 \le m_0 \le -2C_k + 1 \end{cases}
$$

$$
g_3(m_0, m_k)\ if\ \begin{cases} 0 \le C_k \le 1/2 \\ -1 \le m_k \le 0 \\ -2C_k + 1 \le m_0 \le 2C_k m_k + 1 \end{cases}
$$

$$g_2(m_0, m_k)\ if \begin{cases} 0 \le C_k \le 1/2 \\ -1 \le m_k \le 0 \\ 2C_k m_k + 1 \le m_0 \le 1 \end{cases}$$

$$g_8(m_0, m_k)\ if \begin{cases} 0 \le C_k \le 1/2 \\ 0 \le m_k \le 1 \\ -1 \le m_0 \le 2C_k m_k - 1 \end{cases}$$

$$g_6(m_0, m_k)\ if \begin{cases} 0 \le C_k \le 1/2 \\ 0 \le m_k \le 1 \\ 2C_k m_k - 1 \le m_0 \le -2C_k + 1 \end{cases}$$

$$g_4(m_0, m_k)\ if \begin{cases} 0 \le C_k \le 1/2 \\ 0 \le m_k \le 1 \\ -2C_k + 1 \le m_0 \le 1 \end{cases}$$

$$g_{11}(m_0, m_k)\ if \begin{cases} -1 \le C_k \le -1/2 \\ -1 \le m_k \le 0 \\ -1 \le m_0 \le 2C_k m_k - 1 \end{cases}$$

$$g_9(m_0, m_k)\ if \begin{cases} -1 \le C_k \le -1/2 \\ -1 \le m_k \le 0 \\ 2C_k m_k - 1 \le m_0 \le -2C_k - 1 \end{cases}$$

$$g_{10}(m_0, m_k)\ if \begin{cases} -1 \le C_k \le -1/2 \\ -1 \le m_k \le 0 \\ -2C_k - 1 \le m_0 \le 1 \end{cases}$$

$$g_{12}(m_0, m_k)\ if \begin{cases} -1 \le C_k \le -1/2 \\ 0 \le m_k \le \frac{-1-C_k}{C_k} \\ -1 \le m_0 \le -2C_k - 1 \end{cases}$$

$$g_{13}(m_0, m_k)\ if \begin{cases} -1 \leq C_k \leq -1/2 \\ 0 \leq m_k \leq \frac{-1-C_k}{C_k} \\ -2C_k - 1 \leq m_0 \leq 2C_k m_k + 1 \end{cases}$$

$$g_{16}(m_0, m_k)\ if \begin{cases} -1 \leq C_k \leq -1/2 \\ 0 \leq m_k \leq \frac{-1-C_k}{C_k} \\ 2C_k m_k + 1 \leq m_0 \leq 1 \end{cases}$$

$$g_{14}(m_0, m_k)\ if \begin{cases} -1 \leq C_k \leq -1/2 \\ \frac{-1-C_k}{C_k} \leq m_k \leq 1 \\ -1 \leq m_0 \leq 2C_k m_k + 1 \end{cases}$$

$$g_{15}(m_0, m_k)\ if \begin{cases} -1 \leq C_k \leq -1/2 \\ \frac{-1-C_k}{C_k} \leq m_k \leq 1 \\ 2C_k m_k + 1 \leq m_0 \leq -2C_k - 1 \end{cases}$$

$$g_{16}(m_0, m_k)\ if \begin{cases} -1 \leq C_k \leq -1/2 \\ \frac{-1-C_k}{C_k} \leq m_k \leq 1 \\ -2C_k - 1 \leq m_0 \leq 1 \end{cases}$$

$$g_{11}(m_0, m_k)\ if \begin{cases} -1/2 \leq C_k \leq 0 \\ -1 \leq m_k \leq 0 \\ -1 \leq m_0 \leq 2C_k m_k - 1 \end{cases}$$

$$g_{9}(m_0, m_k)\ if \begin{cases} -1/2 \leq C_k \leq 0 \\ -1 \leq m_k \leq 0 \\ 2C_k m_k - 1 \leq m_0 \leq -2C_k - 1 \end{cases}$$

$$g_{10}(m_0, m_k)\ if \begin{cases} -1/2 \leq C_k \leq 0 \\ -1 \leq m_k \leq 0 \\ -2C_k - 1 \leq m_0 \leq 1 \end{cases}$$

$$
\begin{aligned}
g_{12}(m_0, m_k)\ if\ &\begin{cases} -1/2 \le C_k \le 0 \\ 0 \le m_k \le 1 \\ -1 \le m_0 \le -2C_k - 1 \end{cases} \\
g_{13}(m_0, m_k)\ if\ &\begin{cases} -1/2 \le C_k \le 0 \\ 0 \le m_k \le 1 \\ -2C_k - 1 \le m_0 \le 2C_k m_k + 1 \end{cases} \\
g_{16}(m_0, m_k)\ if\ &\begin{cases} -1/2 \le C_k \le 0 \\ 0 \le m_k \le 1 \\ 2C_k m_k + 1 \le m_0 \le 1. \end{cases}
\end{aligned}
\tag{2.30}
$$

The corresponding $g_i(m_0, m_k)$ in equation (2.30) with $i = 1...16$ are defined as follows,

$$
\begin{aligned}
g_1(m_0, m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[(m_0+1)log\Big(\frac{1+C_k m_k}{1-C_k}\Big)\Big] \\
g_2(m_0, m_k) &= \frac{m_k+1}{2} \\
g_3(m_0, m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[m_0 - 1 + 2C_k + (m_0+1)log\Big(\frac{2+2C_k m_k}{m_0+1}\Big)\Big] \\
g_4(m_0, m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[-1 + m_0 + 2C_k(1+m_k) + \\
&\quad (1+m_0)log\Big(\frac{2}{m_0+1}\Big) + (1-m_0)log(1-C_k m_k)\Big] \\
g_5(m_0, m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[(1-m_0)log(1-C_k m_k) - (1-m_0) + \\
&\quad (1+m_0)log\Big(\frac{2}{1+m_0}\Big) + 2C_k(1+m_k)\Big] \\
g_6(m_0, m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[2C_k m_k + (1-m_0)log(1-C_k m_k) - (1+m_0)log(1-C_k)\Big] \\
g_7(m_0, m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[2C_k + m_0\Big(2 + log(4)\Big) + (1-m_0)log(1-m_0) \\
&\quad -(1+m_0)log(1+m_0)\Big] \\
g_8(m_0, m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[(1-m_0)log\Big(\frac{1-m_0}{2}\Big) + (1+m_0)\Big(1 - log(1-C_k)\Big)\Big]
\end{aligned}
$$

$$
\begin{aligned}
g_9(m_0,m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[2C_km_k - 1 - m_0 + (1-m_0)log\Big(\frac{2-2C_km_k}{1-m_0}\Big)\Big] \\
g_{10}(m_0,m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[2C_k + 2C_km_k + (1+2C_k+m_0)log\Big(\frac{1+C_k}{1-C_km_k}\Big) \\
&\quad +(2+2C_k)log\Big(\frac{1-C_km_k}{1+C_k}\Big)\Big] \\
g_{11}(m_0,m_k) &= 0 \\
g_{12}(m_0,m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[(1+m_0)\Big(log(1+C_km_k)-1\Big) + (1-m_0)log\Big(\frac{2}{1-m_0}\Big)\Big] \\
g_{13}(m_0,m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[2C_k + (1+2C_k+m_0)log(1+C_k) + (2+2C_k)log\Big(\frac{1}{1+C_k}\Big) \\
&\quad +(1+m_0)log(1+C_km_k)\Big] \\
g_{14}(m_0,m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[(1+m_0)\Big(log(1+C_km_k)-1\Big) + (1-m_0)log\Big(\frac{2}{1-m_0}\Big)\Big] \\
g_{15}(m_0,m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[(1-m_0)log\Big(\frac{2}{1-m_0}\Big) + (1+m_0)log\Big[\frac{(1+m_0)(1+C_km_k)}{2+2C_km_k}\Big] \\
&\quad +2C_km_k - 2m_0\Big] \\
g_{16}(m_0,m_k) &= \Big[\frac{1}{4C_k}\Big]\Big[1 - m_0 + 2C_km_k + 2C_k + (1+m_0)log\Big[\frac{(1+m_0)(1+C_km_k)}{2+2C_km_k}\Big] \\
&\quad +(1+2C_k+m_0)log(1+C_k) + (2+2C_k)log\Big(\frac{1}{1+C_k}\Big)\Big].
\end{aligned}
\tag{2.31}
$$

For simplicity, denote $f^*_{m_0,m_k}(m_0,m_k)$ and $F^*_{m_0,m_k}(m_0,m_k)$ as the joint PDF and the joint CDF defined in this section (equations (2.29) and (2.30) respectively).

2.3 Optimum coefficients

In the last section we have introduced a fixed mathematical form for the CDF $F_{m_0,m_k}(m_0,m_k)$ that allows us to select any power spectral density $S_M(w)$ for $M[n]$. Therefore, we can now substitute in equation (2.16) to get an expression for the autocorrelation of the error $r_e(k)$.

$$
r_e(k) = 4F^*_{m_0,m_k}(-\lambda,-\lambda) + 2\lambda - 1 - \lambda^2. \tag{2.32}
$$

In this way, any spectrum for $M[n]$ will have a corresponding set of $F^*_{m_0,m_k}(-\lambda,-\lambda)$ for all k (i.e.the autocorrelation $r_M(k)$ will be completely determined by the set C_k for all k). Another important remark, is that for every fixed λ, the value of $F^*_{m_0,m_k}(-\lambda,-\lambda)$ will vary in a certain range, and therefore, the value of $r_e(k)$ will also vary in a certain range. We can call α_λ and β_λ to the lower and upper bound of $r_e(k)$ respectively.

$$\alpha_\lambda \leq r_e(k) \leq \beta_\lambda \tag{2.33}$$

In Figure 2.8 we observe the surface $F^*_{m_0,m_k}(-\lambda,-\lambda)$ when varying C_k between $(-1,1)$ and when we vary λ between $(-1,1)$.

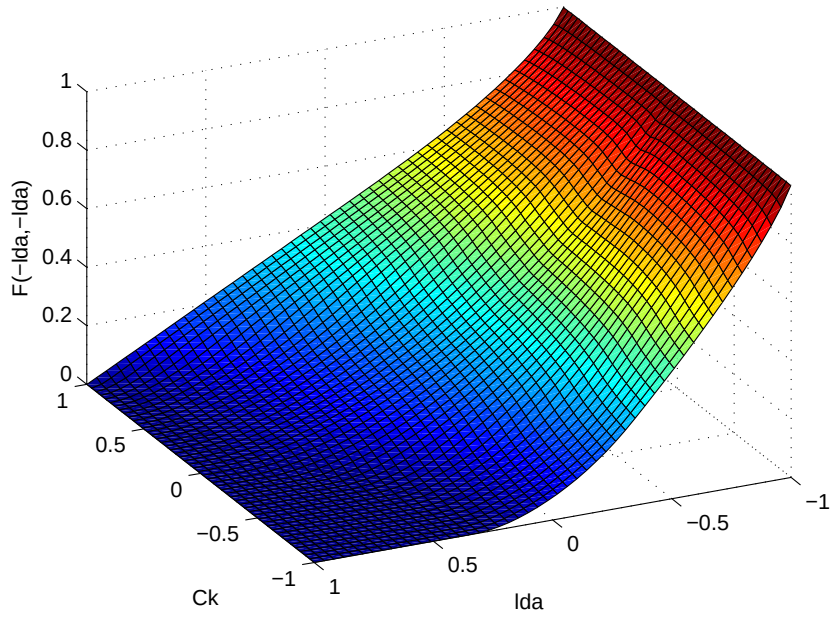

Figure 2.8: Joint CDF $F^*_{m_0,m_k}(-\lambda,-\lambda)$ as a function of λ and C_k.

We can also plot the maximum and minimum possible value of $r_e(k)$ for a given λ between $(-1,1)$. This is shown in Figure 2.9. Unhopefully, there is no analytical expression for the upper and lower bounds for $r_e(k)$ given λ. Figure 2.10

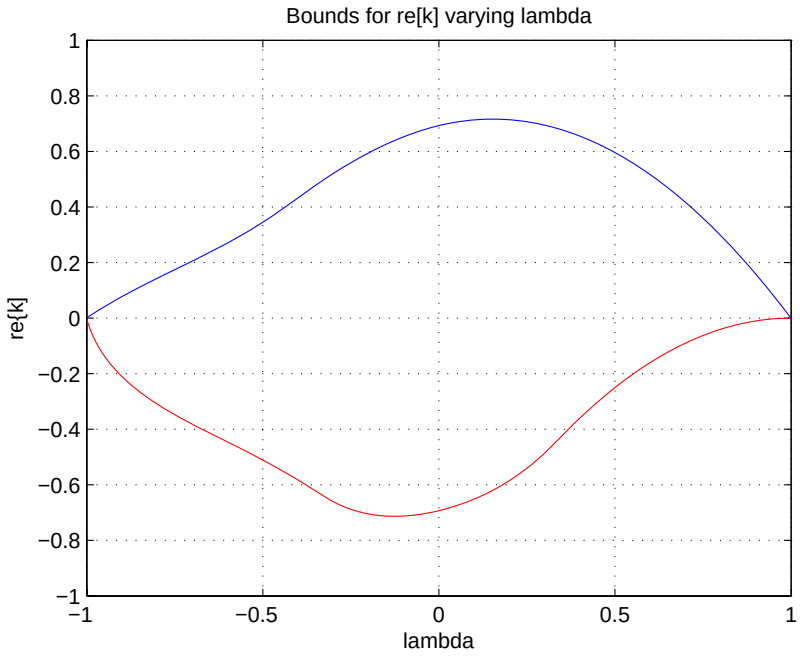

Figure 2.9: Bounds of $r_e(k)$ as a function of λ.

shows how $r_e(k)$ vary with respect to C_k for different values of λ. We observe that when $C_k = 0$, $r_e(k) = 0$ which makes sense with equation (2.22).

Up to this point, given a positive integer N, we want to find the optimum set of coefficients $r_e(k)$ for all $k = 1...N$ that minimizes the in-band quantization noise power (we assume that $r_e(k) = 0$ for $k > N$). We also know that for a fixed λ, the autocorrelation coefficients must lie in a certain range. With these facts considered, we can now perform the optimization. By means of linear programming optimization [5], we can find the set of coefficients $r_e(k)$ that minimize the in-band quantization noise power IQNP from equation (2.19) subject to the constraints in equation (2.33) given λ, N and m (the oversampling ratio).

After the optimization is done, we obtain the following figures. They show the performance of the A/D converter with the optimum mask when varying λ. Figure 2.11 is a plot of the in-band quantization noise (IQNP) as a function of λ for different

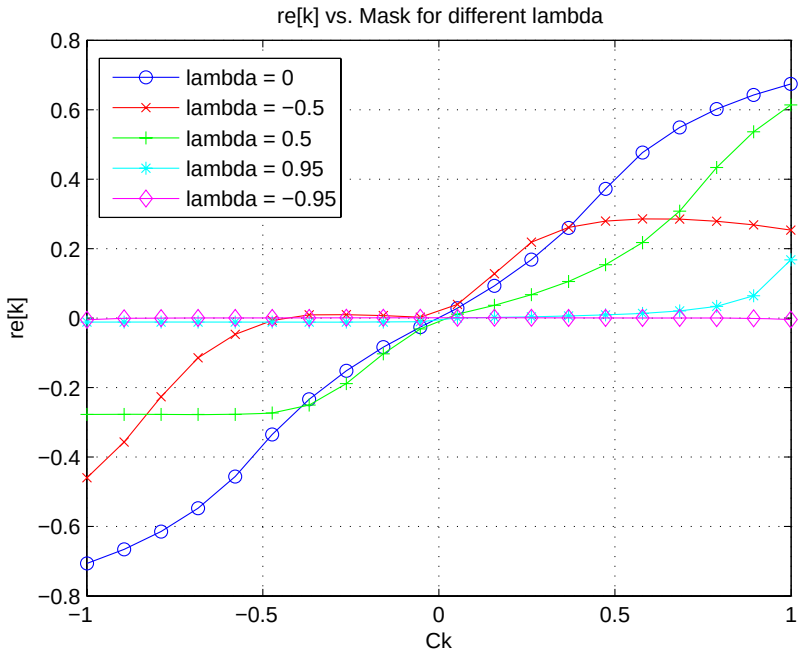

Figure 2.10: Autocorrelation of the error $r_e(k)$ as a function of C_k for different values of λ.

values of N. In the same figure, there are also plots of the IQNP for Sigma-Delta (first, second and third order modulators), so we can compare results. In this case, the oversampling is fixed to $m = 20$. Then, Figure 2.12 is a plot of the optimum coefficient as a function of λ for $N = 1$ (i.e. it is a plot of $r_e(1)$ as a function of λ). In Figures 2.13, 2.14 and 2.15, we can observe the plots of the optimum coefficients as a function of *lambda* for $N = 2$, $N = 3$ and $N = 4$ respectively. For all these plots, the oversampling ratio is also assumed to be $m = 20$. In Figures 2.16 and 2.17, we plot the effective number of bits (ENOB) versus the oversampling ratio for $\lambda = 0.5$ and $\lambda = 0$ respectively.

2.4 Conclusion

In this chapter we have introduced the oversampling A/D converter with dither when dealing with a constant input signal. We have also developed the

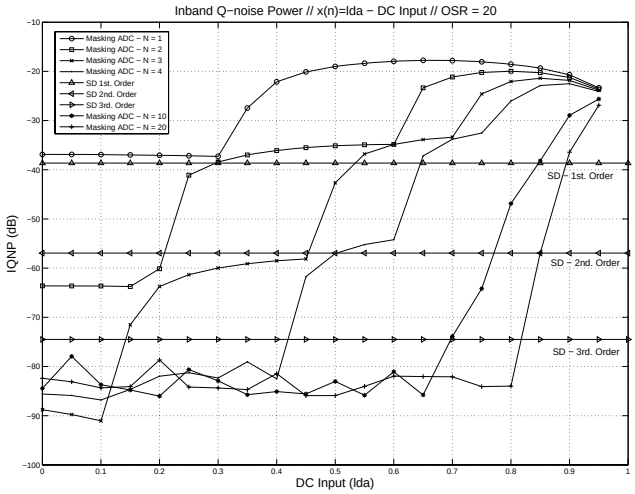

Figure 2.11: In-band quantization noise power as a function of λ for the case when $N = 1, 2, 3, 4, 10$ and 20 when the oversampling ratio is $m = 20$. In the case of Sigma-Delta, the IQNP does not depend on λ.

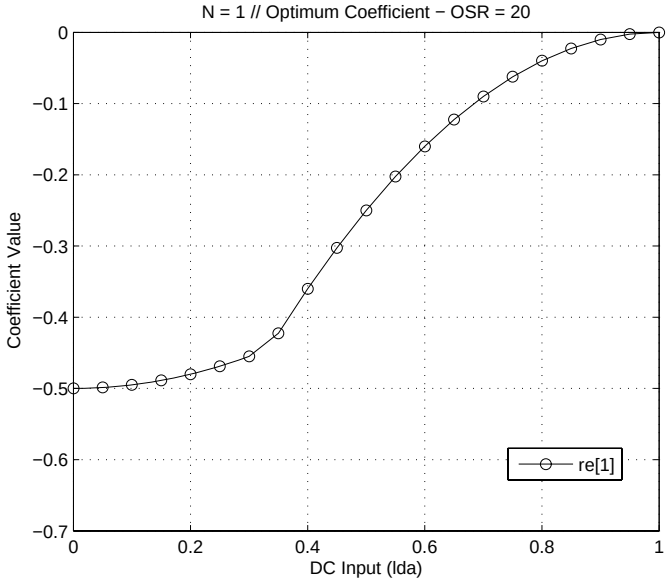

Figure 2.12: Optimum coefficient as a function of λ for $N = 1$. The oversampling ratio is $m = 20$.

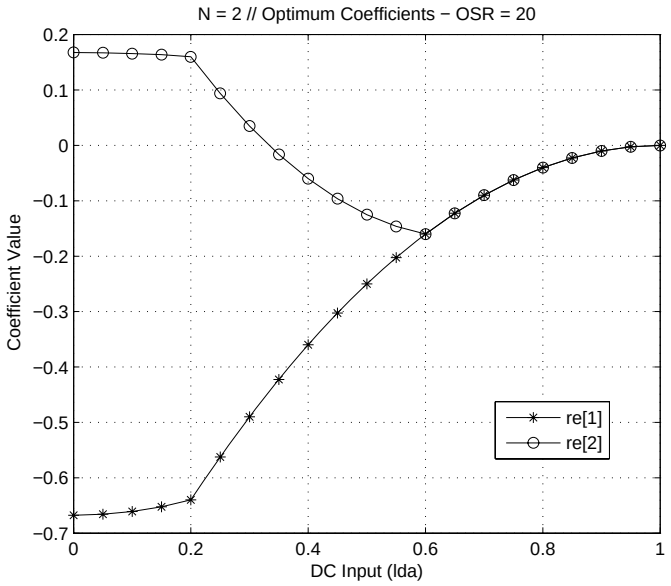

Figure 2.13: Optimum coefficients as a function of λ for $N = 2$. The oversampling ratio is $m = 20$.

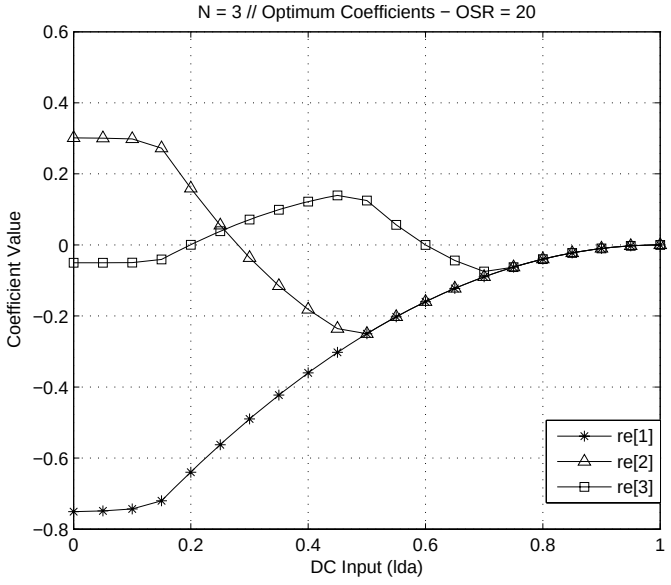

Figure 2.14: Optimum coefficients as a function of λ for $N = 3$. The oversampling ratio is $m = 20$.

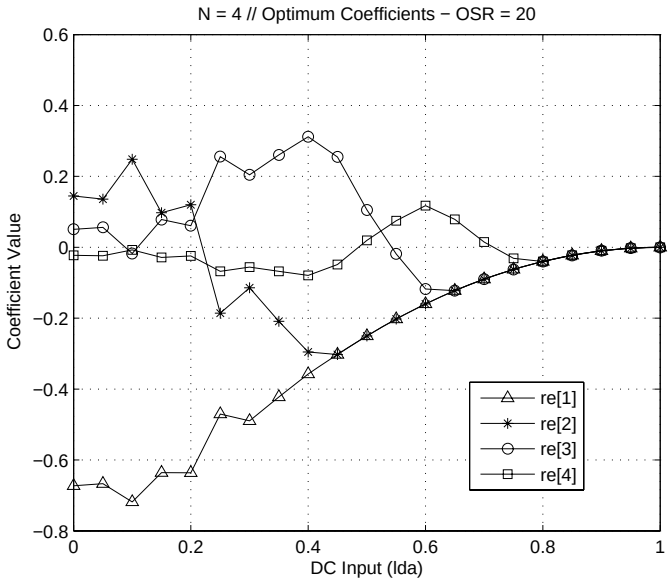

Figure 2.15: Optimum coefficients as a function of λ for $N = 4$. The oversampling ratio is $m = 20$.

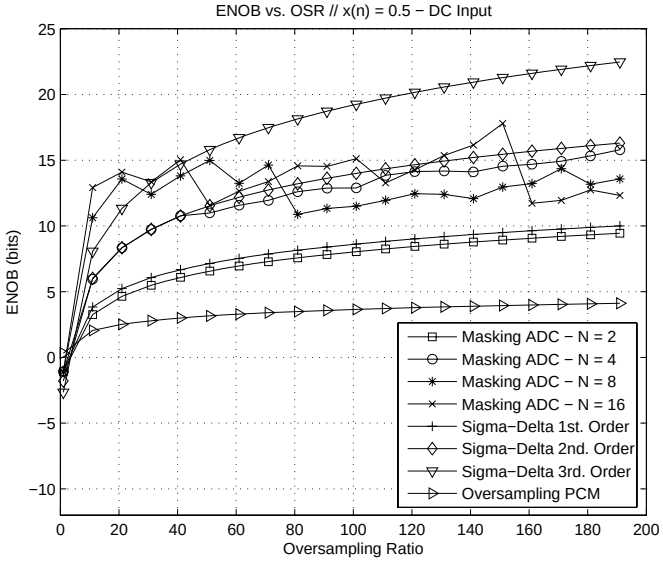

Figure 2.16: Effective number of bits (ENOB) as a function of the oversampling ratio for $\lambda = 0.5$.

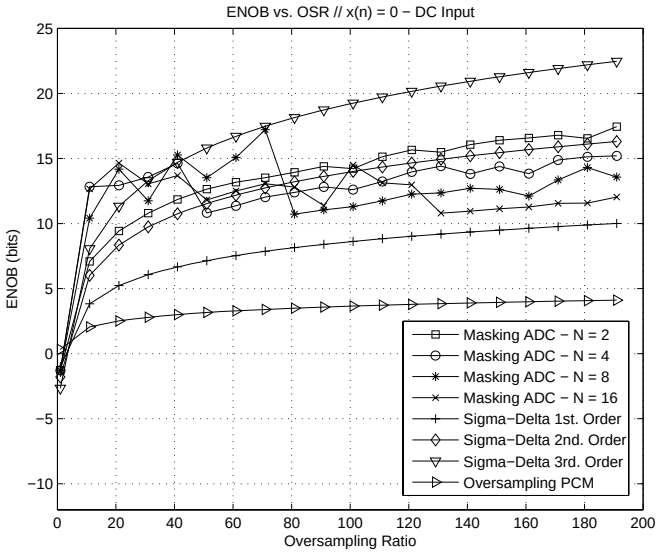

Figure 2.17: Effective number of bits (ENOB) as a function of the oversampling ratio for $\lambda = 0$.

mathematical expression for the autocorrelation of the quantization error $r_e(k)$ as a function of the joint statistics of $M[n]$ (i.e. as a function of the joint CDF of $M[n]$ and $M[n-k]$). Then we have assumed a certain general form for the joint CDF $F_{m_0,m_k}(m_0, m_k)$ such that we can pick any desired power spectral density for $M[n]$ and the problem becomes tractable. After this, we found the optimum N non-zero coefficients $r_e(k)$ that minimizes the IQNP by using linear programming optimization. These coefficients correspond to N values of C_k for $1 \leq k \leq N$. The autocorrelation of the mask $r_M(k)$ is defined as $r_M(k) = C_k/3$, then, once the set of C_k is determined, the power spectral density of the optimum mask $S_M(w)$ is also determined. Therefore, the optimum autocorrelation function and spectrum of $M[n]$ was found.

It is important to remark that in this analysis, the input signal was always known as a constant of value λ. This is why we were able to find a mask $M[n]$ that

perfectly fits the input and we obtained really good results comparable to Sigma-Delta modulation. In Sigma-Delta nothing is assumed about the characteristics of the input signal. However, this technique achieves a very high performance because it uses feedback.

Chapter 3

OPTIMUM DITHER FOR A UNIFORM WHITE NOISE INPUT

In Chapter 2 we have analyzed the case when the input to our dithered-oversampling A/D converter was a constant input. Now, we will study the case when the input is completely random. Just like in the previous chapter, this is an extreme case that has to be addressed in order to complete the analysis of the A/D converter in the limit cases.

3.1 Setting up the problem

In this chapter, the input $x[n]$ to our system will be uniformly distributed white noise. We chose to deal with a uniform distribution as this represents the most general case. As usual, $x[n]$ will be sampled at a rate m times higher than the Nyquist rate. Then, the uniformly distributed dither signal $M[n]$ will be added before the binary quantizer generates the output $y[n]$. The objective will be again to find the optimum statistical characteristics of $M[n]$ that minimize the IQNP at the output (i.e. to maximize the output SNR).

The probability density function of $x[n]$ will be defined as follows,

$$f_X(x) = \begin{cases} \frac{1}{2} & if \quad -1 \leq x \leq 1 \\ 0 & otherwise. \end{cases} \tag{3.1}$$

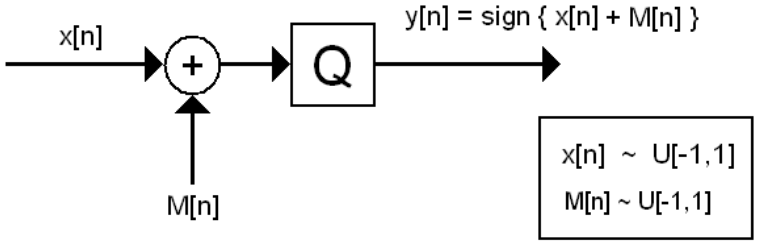

Figure 3.1: Dithered-oversampling binary ADC. The input $x[n]$ is uniformly distributed white noise.

The signal $x[n]$ is also white noise, therefore, the joint PDF of $x[n]$ and $x[n-k]$ will be the multiplication of the marginals.

$$f_{x_0,x_k}(x_0,x_k) = \begin{cases} 1/4 & if \quad -1 \leq x_0 \leq -1 \quad and \quad -1 \leq x_k \leq -1 \\ 0 & otherwise. \end{cases} \tag{3.2}$$

Moreover, as $x[n]$ is uniform in $(-1,1)$, its power $r_X(0) = 1/3$. The corresponding autocorrelation function of the input $r_X(k)$ will be,

$$r_X(k) = \begin{cases} 1/3 & if \quad k = 0 \\ 0 & if \quad k \neq 0. \end{cases} \tag{3.3}$$

The error $e[n]$ will be defined as,

$$e[n] = sgn(M[n] + x[n]) - x[n]. \tag{3.4}$$

Then, the corresponding autocorrelation function $r_e(k)$ will be as follows,

$$\begin{aligned} r_e(k) &= E\Big\{e[n]e[n-k]\Big\} \\ &= E\Big\{\Big[sgn(M[n]+x[n]) - x[n]\Big]\Big[sgn(M[n-k]+x[n-k]) - x[n-k]\Big]\Big\} \\ &= E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} \\ &\quad -E\Big\{sgn(M[n]+x[n])x[n-k]\Big\} - E\Big\{sgn(M[n-k]+x[n-k])x[n]\Big\} \end{aligned}$$

$$
\begin{aligned}
& +E\Big\{x[n]x[n-k]\Big\} \\
= \;& E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} \\
& -2E\Big\{sgn(M[n]+x[n])x[n-k]\Big\}+r_X(k). \qquad (3.5)
\end{aligned}
$$

The mask $M[n]$ is uniform, so its PDF is defined in equation (1.6). Also, the joint PDF of $x[n]$ and $x[n-k]$ is defined in (3.2). Therefore, for $k>0$,

$$
\begin{aligned}
E\Big\{sgn(M[n]+x[n])x[n-k]\Big\} &= \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\Big(\frac{1}{2}\Big)\Big(\frac{1}{4}\Big)sgn(x+y)z\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= 0. \qquad (3.6)
\end{aligned}
$$

For $k>0$, considering equation (3.3), the expression in equation (3.5) becomes,

$$
r_e(k)=E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\}. \qquad (3.7)
$$

When $k=0$,

$$
\begin{aligned}
E\Big\{sgn(M[n]+x[n])x[n]\Big\} &= \int_{-1}^{1}\int_{-1}^{1}\Big(\frac{1}{2}\Big)\Big(\frac{1}{2}\Big)sgn(x+y)x\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \frac{1}{3}. \qquad (3.8)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
r_e(0) &= E\Big\{e[n]e[n]\Big\} \\
&= E\Big\{\Big[sgn(M[n]+x[n])-x[n]\Big]\Big[sgn(M[n]+x[n])-x[n]\Big]\Big\} \\
&= E\Big\{sgn(M[n]+x[n])sgn(M[n]+x[n])\Big\} \\
&\quad -2E\Big\{sgn(M[n]+x[n])x[n]\Big\}+r_X(0) \\
&= 1-2\Big(\frac{1}{3}\Big)+\frac{1}{3} \\
&= \frac{2}{3}. \qquad (3.9)
\end{aligned}
$$

To get an expression for $r_e(k)$ when $k>0$, we need to consider the joint PDF of $M[n]$,$M[n-k]$,$x[n]$ and $x[n-k]$. We know that $M[n]$ and $x[n]$ are independent,

therefore, the joint PDF will be the multiplication of the joint PDF of $M[n]$ and $M[n-k]$ by the joint PDF of $x[n]$ and $x[n-k]$.

$$\begin{aligned} f_{m_0,m_k,x_0,x_k}(m_0,m_k,x_0,x_k) &= f_{m_0,m_k}(m_0,m_k)f_{x_0,x_k}(x_0,x_k) \\ &= \begin{cases} \frac{f_{m_0,m_k}(m_0,m_k)}{4} & if \quad -1 \le m_0,m_k,x_0,x_k \le 1 \\ 0 & otherwise. \end{cases} \end{aligned} \tag{3.10}$$

Then, for $k > 0$,

$$\begin{aligned} r_e(k) &= E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} \\ &= \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} sgn(x+v)sgn(y+w)\left(\frac{1}{4}\right)f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y\mathrm{d}v\mathrm{d}w \\ &= \int_{-1}^{1}\int_{-1}^{1}\left(\frac{1}{4}\right)f_{m_0,m_k}(x,y)\left[\int_{-1}^{1}\int_{-1}^{1} sgn(x+v)sgn(y+w)\mathrm{d}v\mathrm{d}w\right]\mathrm{d}x\mathrm{d}y \\ &= \int_{-1}^{1}\int_{-1}^{1}\left(\frac{1}{4}\right)f_{m_0,m_k}(x,y)\Big[4xy\Big]\mathrm{d}x\mathrm{d}y \\ &= \int_{-1}^{1}\int_{-1}^{1} xy f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-1}^{1}\int_{-1}^{1} m_0 m_k f_{m_0,m_k}(m_0,m_k)\mathrm{d}m_0\mathrm{d}m_k \\ &= E\Big\{M[n]M[n-k]\Big\} \\ &= r_M(k). \end{aligned} \tag{3.11}$$

Therefore, when the analog input $x[n]$ is uniform white noise, the autocorrelation of the error $r_e(k)$ is the same as the autocorrelation of the mask $r_M(k)$ for $k > 0$. In general, for all k,

$$r_e(k) = r_M(k) + \left(\frac{1}{3}\right)\delta(k). \tag{3.12}$$

3.2 Optimum Dither

The expression in equation (3.12) tells us that the PSD of $e[n]$ is the sum of the PSD of $M[n]$ called $S_M(w)$ plus a noise floor. This result is obtained by applying

the DTFT to $r_e(k)$.

$$S_e(w) = S_M(w) + \frac{1}{3}. \tag{3.13}$$

Our goal is to design a mask $M[n]$ to minimize the in-band quantization noise power. Therefore, the optimum $M[n]$ will be the one such that $S_M(w) = 0 \ \forall \ w \in (\frac{-\pi}{m}, \frac{\pi}{m})$ with m being the oversampling ratio. If we were dealing with that optimum $M[n]$, the IQNP would be,

$$\begin{aligned} IQNP &= \frac{1}{2\pi}\int_{-\pi/m}^{\pi/m} S_e(w)\mathrm{d}w \\ &= \frac{1}{2\pi}\int_{-\pi/m}^{\pi/m} S_M(w) + \frac{1}{3}\mathrm{d}w \\ &\simeq \frac{1}{2\pi}\int_{-\pi/m}^{\pi/m} \frac{1}{3}\mathrm{d}w \\ &= \frac{1}{2\pi}\frac{2\pi}{m}\frac{1}{3} \\ &= \frac{1}{3m}. \end{aligned} \tag{3.14}$$

Figure 3.2 shows how the IQNP varies when using the optimum $M[n]$. It also shows the performance for first-order and second-order Sigma Delta. We observe that when we are dealing with uniform white noise as input, the optimum performance is never better than Sigma Delta.

3.3 Conclusion

In this Chapter we have analyzed the performance of the dithered-oversampling A/D converter when the input is uniform white noise and when we use the optimum mask $M[n]$ to maximize the output SNR (i.e. to minimize the IQNP). We have developed the equations for the autocorrelaton of the error $r_e(k)$ and we have seen that it is the same as the autocorrelation of the mask $r_M(k)$ plus a noise floor. If we were using the optimum mask, the performance of the A/D converter will never outperform Sigma-Delta.

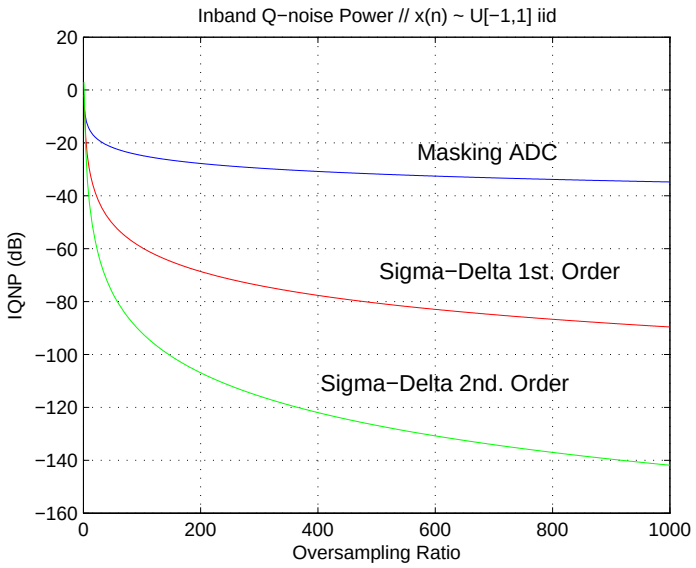

Figure 3.2: In-band quantization noise power for Dithered-oversampling ADC, 1st Order Sigma-Delta and 2nd. order Sigma-Delta for a uniformly distributed white noise input signal.

Chapter 4

OPTIMUM DITHER FOR A GENERAL INPUT WITH KNOWN AUTOCORRELATION FUNCTION

To complete the analysis of the dithered-oversampling A/D converter, this chapter describes the case when the analog input is any unformly distributed signal in $(-1, 1)$ with known autocorrelation function.

4.1 Setting up the problem

In this chapter we will consider the case when we have an analog input $x[n]$ and we know its autocorrelation function (i.e. we know its power spectral density). As usual, it is assumed that the input signal is sampled at a rate m times greater than the Nyquist frequency and that a binary quantizer is used to obtain the output $y[n]$. Again, our objective will be to find the optimum $M[n]$ that minimizes the in-band quantization noise power (i.e. maximizes the output SNR). We will assume that the input $x[n]$ is equally likely to take any value between $(-1, 1)$ (i.e. $x[n]$ is uniform in $(-1, 1)$). The PDF of $x[n]$ is then defined in equation (3.1). The error $e[n]$ between the output and the input is defined as

$$\begin{aligned} e[n] &= y[n] - x[n] \\ &= sgn(M[n] + x[n]) - x[n]. \end{aligned} \tag{4.1}$$

Now, we need to consider the autocorrelation function of $e[n]$.

$$r_e(k) = E\Big\{e[n]e[n-k]\Big\}$$

$$
\begin{aligned}
&= E\Big\{\Big[sgn(M[n]+x[n])-x[n]\Big]\Big[sgn(M[n-k]+x[n-k])-x[n-k]\Big]\Big\} \\
&= E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} \\
&\quad -E\Big\{sgn(M[n]+x[n])x[n-k]\Big\}-E\Big\{sgn(M[n-k]+x[n-k])x[n]\Big\} \\
&\quad +E\Big\{x[n]x[n-k]\Big\} \\
&= E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} \\
&\quad -2E\Big\{sgn(M[n]+x[n])x[n-k]\Big\}+r_X(k). \qquad (4.2)
\end{aligned}
$$

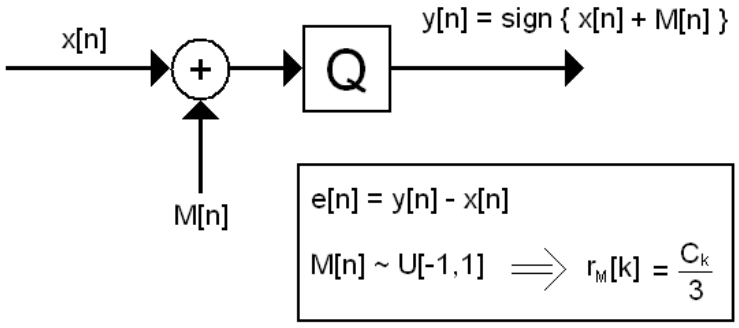

Figure 4.1: Dithered-oversampling binary ADC. The input $x[n]$ is uniformly distributed with known autocorrelation function.

To figure out the term $E\Big\{sgn(M[n]+x[n])x[n-k]\Big\}$, we have to consider the joint PDF of $M[n]$, $x[n]$ and $x[n-k]$. The input signal and the mask are independent, so the joint PDF will be the multiplication of the joint PDF of $x[n]$ and $x[n-k]$ by the PDF of $M[n]$ defined in equation (1.6).

$$
\begin{aligned}
f_{m_0,x_0,x_k}(m_0,x_0,x_k) &= f_M(m_0)f_{x_0,x_k}(x_0,x_k) \\
&= \begin{cases} \frac{f_{x_0,x_k}(x_0,x_k)}{2} & if \quad -1 \le m_0,x_0,x_k \le 1 \\ 0 & otherwise \end{cases}
\end{aligned} \qquad (4.3)
$$

with $f_{x_0,x_k}(x_0, x_k)$ being the joint PDF of $x[n]$ and $x[n-k]$.

$$\begin{aligned} E\Big\{sgn(M[n]+x[n])x[n-k]\Big\} &= \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} sgn(x+v)w\Big(\frac{1}{2}\Big)f_{x_0,x_k}(v,w)\mathrm{d}x\mathrm{d}v\mathrm{d}w \\ &= \int_{-1}^{1}\frac{w}{2}\int_{-1}^{1}\Big[\int_{-1}^{1} sgn(x+v)f_{x_0,x_k}(v,w)\mathrm{d}x\Big]\mathrm{d}v\mathrm{d}w. \end{aligned}$$

The integral into straight brackets can be splitted as follows,

$$\begin{aligned} \int_{-1}^{1} sgn(x+v)f_{x_0,x_k}(v,w)\mathrm{d}x &= \int_{-v}^{1} f_{x_0,x_k}(v,w)\mathrm{d}x - \int_{-1}^{-v} f_{x_0,x_k}(v,w)\mathrm{d}x \\ &= 2vf_{x_0,x_k}(v,w). \end{aligned} \tag{4.4}$$

Then,

$$\begin{aligned} E\Big\{sgn(M[n]+x[n])x[n-k]\Big\} &= \int_{-1}^{1}\frac{w}{2}\int_{-1}^{1}\Big[2vf_{x_0,x_k}(v,w)\Big]\mathrm{d}v\mathrm{d}w \\ &= \int_{-1}^{1}\int_{-1}^{1} vwf_{m_0,m_k}(v,w)\mathrm{d}v\mathrm{d}w \\ &= E\Big\{x[n]x[n-k]\Big\} \\ &= r_X(k). \end{aligned} \tag{4.5}$$

Therefore, going back to equation (4.2),

$$\begin{aligned} r_e(k) &= E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} - 2r_X(k) + r_X(k) \\ &= E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} - r_X(k). \end{aligned} \tag{4.6}$$

To complete the expression for $r_e(k)$, we now need to work with the expectation term in equation (4.6). To do so, we need to consider the joint PDF of $M[n]$, $M[n-k]$, $x[n]$ and $x[n-k]$. As the signals $x[n]$ and $M[n]$ are independent, this joint PDF will be the multiplication of the joint PDF of $M[n]$ and $M[n-k]$ by the joint PDF of $x[n]$ and $x[n-k]$.

$$f_{m_0,m_k,x_0,x_k}(m_0, m_k, x_0, x_k) = f_{m_0,m_k}(m_0, m_k)f_{x_0,x_k}(x_0, x_k) \tag{4.7}$$

where $f_{x_0,x_k}(x_0, x_k)$ is the joint PDF of $x[n]$ and $x[n-k]$ and $f_{m_0,m_k}(m_0, m_k)$ is the joint PDF of $M[n]$ and $M[n-k]$. Now, developing the expectation term,

$$\begin{aligned} E\Big\{sgn(M[n]+x[n])sgn(M[n-k]+x[n-k])\Big\} &= \\ \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} sgn(x+v)sgn(y+w)f_{x_0,x_k}(v,w)f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y\mathrm{d}v\mathrm{d}w &= \\ \int_{-1}^{1}\int_{-1}^{1} f_{x_0,x_k}(v,w)\Big[\int_{-1}^{1}\int_{-1}^{1} sgn(x+v)sgn(y+w)f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y\Big]\mathrm{d}v\mathrm{d}w. \end{aligned} \tag{4.8}$$

Considering the property from equation (2.11), we can express the double integral in straight brackets as a function of the cumulative distribution function of $M[n]$ (i.e. as a function of $F_{m_0,m_k}(m_0, m_k)$). This integral will be a function of v and w, so we can call it $h(v, w)$.

$$h(v,w) = \int_{-1}^{1}\int_{-1}^{1} sgn(x+v)sgn(y+w)f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y. \tag{4.9}$$

Therefore,

$$\begin{aligned} h(v,w) &= \int_{-w}^{1}\int_{-v}^{1} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y + \int_{-1}^{-w}\int_{-1}^{-v} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \; - \\ &\quad \int_{-1}^{-w}\int_{-v}^{1} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y - \int_{-w}^{1}\int_{-1}^{-v} f_{m_0,m_k}(x,y)\mathrm{d}x\mathrm{d}y \\ &= F_{m_0,m_k}(1,1) - F_{m_0,m_k}(-v,1) - F_{m_0,m_k}(1,-w) + F_{m_0,m_k}(-v,-w) + \\ &\quad F_{m_0,m_k}(-v,-w) - F_{m_0,m_k}(-1,-w) - F_{m_0,m_k}(-v,-1) + \\ &\quad F_{m_0,m_k}(-1,-1) - \Big[F_{m_0,m_k}(1,-w) - F_{m_0,m_k}(1,-1) - \\ &\quad F_{m_0,m_k}(-v,-w) + F_{m_0,m_k}(-v,-1)\Big] - \Big[F_{m_0,m_k}(-v,1) - \\ &\quad F_{m_0,m_k}(-v,-w) - F_{m_0,m_k}(-1,1) + F_{m_0,m_k}(-1,-w)\Big]. \end{aligned} \tag{4.10}$$

From equations (1.7), (2.8), (2.9), (2.10) and the fact that $F_{m_0,m_k}(1,1) = 1$, we know that

$$\begin{cases} F_{m_0,m_k}(-1,w) = F_{m_0,m_k}(v,-1) = 0 \\ F_{m_0,m_k}(1,w) = F_M(w) = (1+w)/2 \\ F_{m_0,m_k}(v,1) = F_M(v) = (1+v)/2. \end{cases} \tag{4.11}$$

Then, $h(v,w)$ can be simplified,

$$\begin{aligned} h(v,w) &= F_{m_0,m_k}(1,1) - F_{m_0,m_k}(-v,1) - F_{m_0,m_k}(1,-w) + F_{m_0,m_k}(-v,-w) \\ &\quad + F_{m_0,m_k}(-v,-w) - F_{m_0,m_k}(1,-w) + F_{m_0,m_k}(-v,-w) \\ &\quad - F_{m_0,m_k}(-v,1) + F_{m_0,m_k}(-v,-w) \\ &= 1 + 4F_{m_0,m_k}(-v,-w) - 2F_{m_0,m_k}(-v,1) - 2F_{m_0,m_k}(1,-w) \\ &= 1 + 4F_{m_0,m_k}(-v,-w) - (1-v) - (1-w) \\ &= v + w - 1 + 4F_{m_0,m_k}(-v,-w). \end{aligned} \tag{4.12}$$

Now we are ready to substitute in equation (4.8),

$$\begin{aligned} E\Big\{ sgn(M[n] + x[n])sgn(M[n-k] + x[n-k]) \Big\} &= \\ \int_{-1}^{1}\int_{-1}^{1} f_{x_0,x_k}(v,w)\Big[h(v,w)\Big]\mathrm{d}v\mathrm{d}w &= \\ \int_{-1}^{1}\int_{-1}^{1} f_{x_0,x_k}(v,w)\Big[v + w - 1 + 4F_{m_0,m_k}(-v,-w)\Big]\mathrm{d}v\mathrm{d}w &= \\ \int_{-1}^{1}\int_{-1}^{1} v f_{x_0,x_k}(v,w)\mathrm{d}v\mathrm{d}w + \int_{-1}^{1}\int_{-1}^{1} w f_{x_0,x_k}(v,w)\mathrm{d}v\mathrm{d}w - & \\ \int_{-1}^{1}\int_{-1}^{1} f_{x_0,x_k}(v,w)\mathrm{d}v\mathrm{d}w + 4\int_{-1}^{1}\int_{-1}^{1} f_{x_0,x_k}(v,w)F_{m_0,m_k}(-v,-w)\mathrm{d}v\mathrm{d}w &= \\ E(x_0) + E(x_k) - 1 + 4E_{x_0,x_k}\Big[F_{m_0,m_k}(-x_0,-x_k)\Big]. \end{aligned} \tag{4.13}$$

We know that $x[n]$ is uniform in $(-1,1)$, so $E(x_0) = E(x_k) = 0$. Then,

$$E\Big\{ sgn(M[n] + x[n])sgn(M[n-k] + x[n-k]) \Big\} = 4E_{x_0,x_k}\Big[F_{m_0,m_k}(-x_0,-x_k)\Big] - 1. \tag{4.14}$$

Therefore, we can now substitute in equation (4.6) to get an expression for $r_e(k)$ in terms of the joint CDF of the mask $M[n]$ and the joint statistics of the input.

$$r_e(k) = 4E_{x_0,x_k}\Big[F_{m_0,m_k}(-x_0,-x_k)\Big] - 1 - r_X(k). \tag{4.15}$$

The expression in equation (4.15) is valid for $k > 0$. For the case when $k = 0$, recalling equation (4.2), we have,

$$\begin{aligned} r_e(0) &= E\Big\{sgn(M[n]+x[n])sgn(M[n]+x[n])\Big\} \\ &\quad -2E\Big\{sgn(M[n]+x[n])x[n]\Big\} + r_X(0) \\ &= 1 - 2E\Big\{sgn(M[n]+x[n])x[n]\Big\} + \frac{2}{3}. \end{aligned} \tag{4.16}$$

The signals $M[n]$ and $x[n]$ are both uniformly distributed and independent, therefore, the term $E\Big\{sgn(M[n]+x[n])x[n]\Big\}$ was already calculated in equation (3.8). Then,

$$\begin{aligned} r_e(0) &= 1 - 2\Big(\frac{1}{3}\Big) + \frac{1}{3} \\ &= \frac{2}{3}. \end{aligned} \tag{4.17}$$

Now that we have an expression for $r_e(k)$, we can take its DTFT and find the power spectral density $S_e(w)$. Then, we could find the optimum function $r_e(k)$ that makes $S_e(w)$ minimum in the band of interest (i.e. we would be minimizing the quantization noise power for $w \in (-\pi/m, \pi/m)$ where m is the oversampling ratio).

$$S_e(w) = r_e(0) + 2\sum_{k=1}^{+\infty} r_e(k)cos(wk). \tag{4.18}$$

The in-band quantization noise power will be the integral of $S_e(w)$ in the band of interest.

$$\begin{aligned} IQNP &= \frac{1}{2\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} S_e(w)\mathrm{d}w \\ &= \frac{r_e(0)}{m} + \frac{1}{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \sum_{k=1}^{+\infty} r_e(k)cos(wk)\mathrm{d}w \\ &= \frac{r_e(0)}{m} + \sum_{k=1}^{+\infty} \frac{2}{k\pi} sin\Big(\frac{k\pi}{m}\Big) r_e(k). \end{aligned} \tag{4.19}$$

4.2 Optimum Dither

To calculate the expectation term in equation (4.15), we would need the joint PDF of $x[n]$, namely $f_{x_0,x_k}(x_0, x_k)$. To simplify the analysis, as $x[n]$ is uniform in $(-1, 1)$, we can use the general joint PDF defined in section 2.2 to describe $f_{x_0,x_k}(x_0, x_k)$. This will allow us to select any desired autocorrelation function (i.e. any desired power spectral density) for $x[n]$. Therefore, let's define $f_{x_0/x_k}(x_0/x_k)$.

$$f_{x_0/x_k}(x_0/x_k) = \begin{cases} \Big[2(1 - B_k x_k)\Big]^{-1} & if \quad \begin{cases} 2B_k x_k - 1 \leq x_0 \leq 1 \\ 0 \leq B_k x_k \leq 1 \end{cases} \\ \Big[2(1 + B_k x_k)\Big]^{-1} & if \quad \begin{cases} -1 \leq x_0 \leq 1 + 2B_k x_k \\ -1 \leq B_k x_k \leq 0 \end{cases} \\ 0 & otherwise. \end{cases} \tag{4.20}$$

Now, the joint PDF of $x[n]$ and $x[n-k]$ will be the multiplication of $f_{x_0/x_k}(x_0/x_k)$ by the uniform density function in $(-1, 1)$.

$$f_{x_0,x_k}(x_0, x_k) = \begin{cases} \frac{f_{x_0/x_k}(x_0/x_k)}{2} & if \quad -1 \leq x_0, x_k \leq 1 \\ 0 & otherwise. \end{cases} \tag{4.21}$$

To sum up, the joint PDF $f_{x_0,x_k}(x_0, x_k)$ and the cumulative PDF $F_{x_0,x_k}(x_0, x_k)$ of $x[n]$ will be:

$$\begin{cases} f_{x_0,x_k}(x_0, x_k) &= f^*_{x_0,x_k}(x_0, x_k) \\ F_{x_0,x_k}(x_0, x_k) &= F^*_{x_0,x_k}(x_0, x_k) \end{cases} \tag{4.22}$$

where we substitute C_k by B_k in equation (2.30) for the definition of $F_{x_0,x_k}(x_0, x_k)$. In this way, the autocorrelation of $x[n]$ will be,

$$r_X(k) = \frac{B_k}{3} \qquad \forall\ k > 0 \tag{4.23}$$

with B_k varying between $(-1, 1)$. Therefore, by selecting the proper coefficients B_k, it is possible to get any desired spectrum for $x[n]$. For example, if $B_k = 0$ for all $k > 0$, $x[n]$ would be uniformly distributed white noise.

Now that we know the joint PDF of $x[n]$, to finally obtain the expression for $r_e(k)$ defined in equation (4.15), we need to assume a joint PDF for $M[n]$ and $M[n-k]$. As we have done before, we will select the joint PDF $f_{m_0,m_k}(m_0, m_k)$ to be the one defined in section 2.2. Then,

$$\begin{cases} f_{m_0,m_k}(m_0, m_k) &= f^*_{m_0,m_k}(m_0, m_k) \\ F_{m_0,m_k}(m_0, m_k) &= F^*_{m_0,m_k}(m_0, m_k). \end{cases} \tag{4.24}$$

The corresponding autocorrelation function for $M[n]$ will be,

$$r_M(k) = \frac{C_k}{3} \qquad \forall\ k > 0 \tag{4.25}$$

with C_k varying between $(-1, 1)$. In this way, equation (4.15) becomes,

$$\begin{aligned} r_e(k) &= 4E_{x_0,x_k}\Big[F_{m_0,m_k}(-x_0, -x_k)\Big] - 1 - r_X(k) \\ &= 4E_{x_0,x_k}\Big[F_{m_0,m_k}(-x_0, -x_k)\Big] - 1 - \frac{B_k}{3}. \end{aligned} \tag{4.26}$$

For every input $x[n]$ with any autocorrelation function $r_X(k)$ and for every possible $M[n]$ with any autocorrelation function $r_M(k)$ we will have a value of $r_e(k)$ given by the equation (4.26). Then, if we vary B_k between $(-1, 1)$ and C_k between $(-1, 1)$, we can get all possible values for $r_e(k)$. To get a closed form analytical expression for equation (4.26), we would need to consider the analytical form of $F^*_{m_0,m_k}(m_0, m_k)$ given by equation (2.30). It is indeed very hard to handle such long expression, so we decided to run a numerical simulation to get the value of $r_e(k)$ given B_k and C_k. Figure 4.2 shows the function $r_e(k)$ as a function of C_k and B_k both varying between $(-1, 1)$. It is interesting to note that for any fixed B_k, the function $r_e(k)$ is a line with slope $1/3$ through the origin as seen in Figure 4.3.

This means, that no matter the value of B_k (i.e. no matter the value of the autocorrelation $r_X(k)$), the autocorrelation of the error $r_e(k)$ is equal to $\frac{C_k}{3}$. Then, for $k > 0$,

$$\begin{aligned} r_e(k) &= \frac{C_k}{3} \\ &= r_M(k). \end{aligned} \tag{4.27}$$

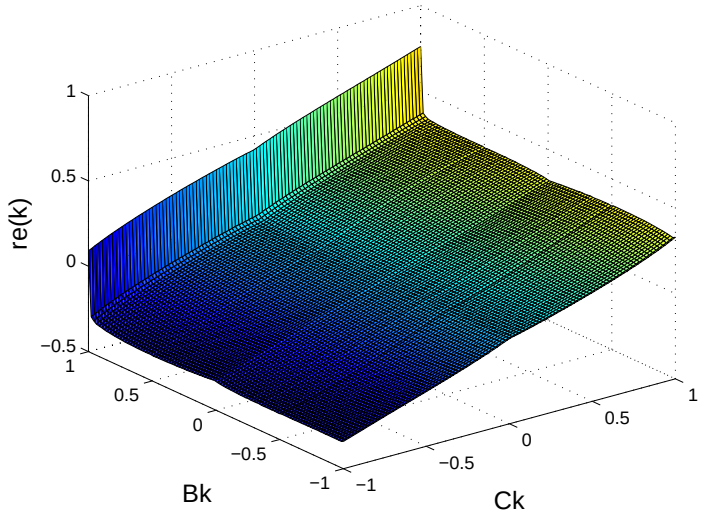

Figure 4.2: Autocorrelation function of the error $e[n]$ as a function of C_k and B_k.

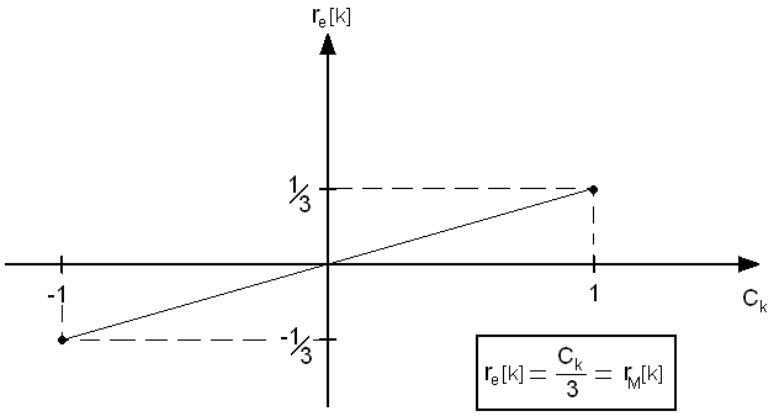

Figure 4.3: Autocorrelation function of the error $e[n]$ as a function of C_k for all B_k.

From equation (4.17), we also know that $r_e(0) = 2/3$. Therefore,

$$r_e(k) = r_M(k) + \left(\frac{1}{3}\right)\delta(k) \tag{4.28}$$

for all k. Just like in the case from Chapter 3 when the input was uniformly distributed white noise, when taking the DTFT in equation (4.28), the power spectral density of the error $S_e(w)$ will be the sum of the power spectral density of the mask $S_M(w)$ plus a noise floor.

$$S_e(w) = S_M(w) + \frac{1}{3}. \tag{4.29}$$

Then, the optimum dither or optimum mask $M[n]$ will be the one that yields zero power in the band of interest (i.e. $S_M(w) = 0 \quad \forall w \in (-\pi/m, \pi/m)$). In this case, like in the case of equation (3.14), the minimum IQNP achievable will be,

$$IQNP = \frac{1}{3m} \tag{4.30}$$

where m is the oversampling ratio.

This result means that when dealing with a uniformly distributed input $x[n]$, no matter what is its power spectral density (i.e. regardless its autocorrelation function $r_X(k)$), the minimum in-band quantization noise power achievable is $\frac{1}{3m}$. This lower bound is achievable when our dither $M[n]$ has zero power in the band of interest. This result is the same as the one obtained in Chapter 3, and tells us that we cannot find any dither $M[n]$ to get a better performance than any Sigma-Delta modulator even when the PSD of the input is known.

4.3 Conclusion

In this chapter we have studied the case when the input to our dithered-oversampling A/D converter is a uniformly distributed signal $x[n]$ with known autocorrelation function. This means that $x[n]$ is uniform, but it has a certain known power spectral density. We have assummed a general joint PDF for $x[n]$ and $x[n-k]$ and a general joint PDF for $M[n]$ and $M[n-k]$ that allow us to describe any possible

power spectral density for $x[n]$ and $M[n]$ and to make the problem tractable. We have also developed a mathematical expression for the autocorrelation function of the error $r_e(k)$ that depends on the joint PDF of the input and on the joint CDF of $M[n]$. Then, we have plotted the value of $r_e(k)$ for all possible input joint PDFs and all possible mask joint CDFs. Finally, we have observed that $r_e(k)$ is equal to $r_M(k)$ regardless of the value of $r_X(k)$. This result is the same as the one obtained when dealing with uniformly distributed white noise in Chapter 3 and tells us that the lowest IQNP possible is $\frac{1}{3m}$ regardless of the input spectrum $S_X(w)$.

Another important remark is that the extreme cases analyzed in previous chapters fit in this general framework. For example, let's see what happens with $r_e(k)$ when we are dealing with a constant input (i.e. $x[n] = \lambda \quad \forall\, n$). In this case, $r_X(k) = \lambda^2$ and $f_{x_0,x_k}(x_0, x_k) = \delta(x_0 - \lambda, x_k - \lambda)$. Therefore, recalling equation (4.15), and considering the fact that $E(x_0) = \lambda$ and $E(x_k) = \lambda$ we have,

$$\begin{aligned}
r_e(k) &= 4E_{x_0,x_k}\Big[F_{m_0,m_k}(-x_0,-x_k)\Big] - 1 - r_X(k) + E(x_0) + E(x_k) \\
&= 4\int_{-1}^{1}\int_{-1}^{1} f_{x_0,x_k}(x_0,x_k)F_{m_0,m_k}(-x_0,-x_k)\mathrm{d}x_0\mathrm{d}x_k - 1 - \lambda^2 + 2\lambda \\
&= 4\int_{-1}^{1}\int_{-1}^{1} \delta(x_0-\lambda,x_k-\lambda)F_{m_0,m_k}(-x_0,-x_k)\mathrm{d}x_0\mathrm{d}x_k - 1 - \lambda^2 + 2\lambda \\
&= 4\int_{-1}^{1}\int_{-1}^{1} \delta(x_0-\lambda,x_k-\lambda)F_{m_0,m_k}(-\lambda,-\lambda)\mathrm{d}x_0\mathrm{d}x_k - 1 - \lambda^2 + 2\lambda \\
&= 4F_{m_0,m_k}(-\lambda,-\lambda)\int_{-1}^{1}\int_{-1}^{1} \delta(x_0-\lambda,x_k-\lambda)\mathrm{d}x_0\mathrm{d}x_k - 1 - \lambda^2 + 2\lambda \\
&= 4F_{m_0,m_k}(-\lambda,-\lambda) - 1 - \lambda^2 + 2\lambda. \qquad (4.31)
\end{aligned}$$

As expected, equation (4.31) is exactly the same as the equation (2.16) when we treated the $x[n] = \lambda$ case.

For the case when $x[n]$ is uniformly distributed white noise, we have already seen that the autocorrelation function of the error $r_e(k)$ is exactly the one described in equation (4.28).

Finally, after analyzing this last general case for our dithered-oversampling A/D converter, we can say that adding a dither signal $M[n]$ before the binary quantizer, helps us to make the quantization noise uncorrelated with the input and this is why we end up with a noise floor term $\left(\frac{1}{3}\right)$ in the equation for $S_e(w)$ [6]. This is a desired effect, as if we were not adding dither, we would end up with undesired harmonics in the spectrum of the output $y[n]$ [6]. However, the negative aspect is that the fact of adding the dither signal $M[n]$ will not help us to decrease the in-band quantization noise power if we don't know the precise characteristics of the input signal that we are dealing with. For instance, when we are dealing with a known DC input (i.e. $x[n] = \lambda$), we can find an optimum dither to decrease the IQNP and get a performance even better than Sigma Delta modulation as observed in Figures 2.11, 2.16 and 2.17. On the other hand, however, if we know that in general the input is unformly distributed, and even if we kwow its power spectral density $S_X(w)$, the optimum mask $M[n]$ will not help us to decrease the IQNP.

Chapter 5

SIMULATIONS AND POSSIBLE APPLICATIONS

5.1 Simulations

5.1.1 Blue Mask M[n]

In Chapter 4 we have concluded that for any input signal $x[n]$ with any autocorrelation function $r_X(k)$ (i.e. with any power spectral densiy), the optimum dither $M[n]$ is the one with zero power in the band of interest. This means that $S_M(w) = 0$ for $|w| \leq -\pi/m$. Therefore, all the signal power lies in the upper portion of the spectrum (i.e. $M[n]$ is a high-pass signal). Any type of noise with this caracteristic is called 'blue noise' [7]. Then, the optimum mask $M[n]$ will be called 'blue mask'.

There are some algorithms already developed for generating 'blue masks' like Direct Binary Search (DBS) or Void-and-Cluster (VAC) [8] [9]. These algorithms generate two dimensional blue-noise patterns that are used for digital halftoning in image processing. Just like in the two dimensional case, we can use these algorithms to create one-dimensional blue noise masks $M[n]$. In the following simulations we will use a 256-levels mask generated with Void-and-Cluster. In Figure 5.1 we observe the power spectral density of this mask.

5.1.2 Sinusoidal Input

In this section we observe and compare the performance of the Dithered-oversampling ADC versus Sigma-Delta modulation when the analog input is a full

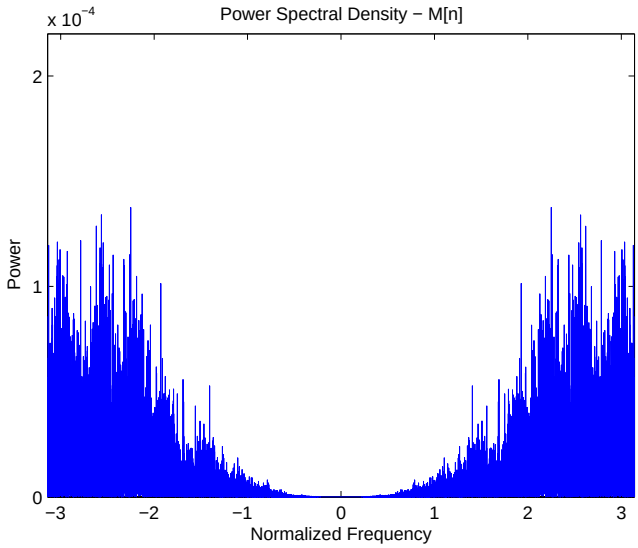

Figure 5.1: Power Spectral Density of M[n] generated with Void-and-Cluster algorithm.

scale sinusoid. In Figure 5.2, the power spectral density of the binary output is plotted for both cases where the normalized frequency of the sine is $\frac{\pi}{200}$. Then, Figure 5.3 shows the output SNR as a function of the oversampling ratio for Dithered-oversampling ADC and Sigma-Delta. The SNR is calculated as the quotient between the signal power and the quantization noise power in the band of interest. In this case, the band of interest is assumed to be twice the frequency of the input signal.

From Figure 5.2 we can see that the quantization noise spectrum in Sigma-Delta presents several harmonics, whereas in the Dithered-oversampling ADC case it is more uniformly randomized. It is also noted in Figure 5.3 that Sigma-Delta performs better than Dithered-oversampling ADC. This is because it is not possible to shape the fixed quantization noise floor that is present when we do not use feedback.

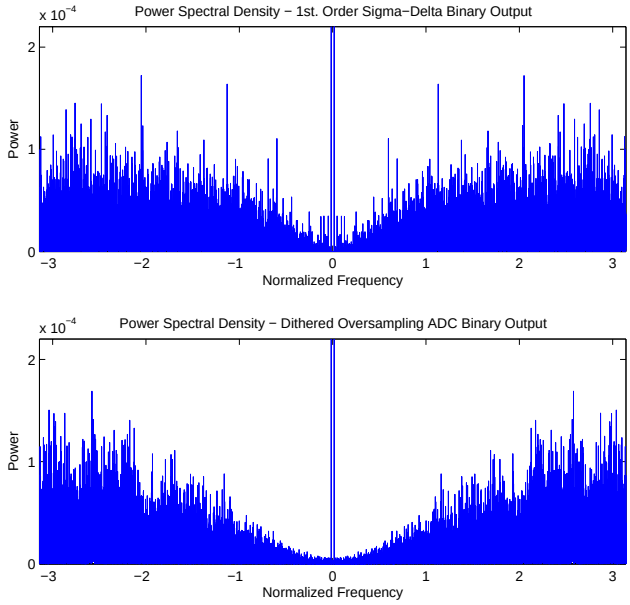

Figure 5.2: Power Spectral Density of the binary output. The top figure shows the Sigma-Delta case, whereas the Dithered-oversampling case is shown in the bottom. The input to the system is a full scale sine with normalized frequency equal to $\frac{\pi}{200}$.

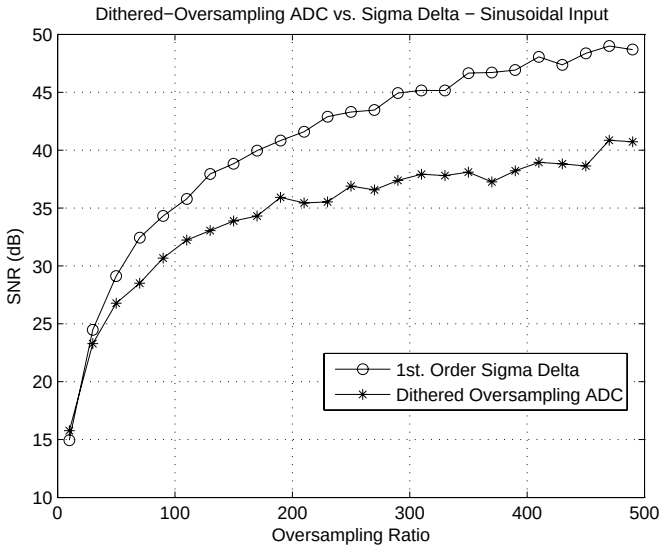

Figure 5.3: SNR versus Oversampling ratio for Sigma-Delta and Dithered-oversampling ADC when using full scale sinusoidal inputs.

5.1.3 DC Input

When we deal with constant analog inputs, the conclusions are similar to the ones drawn in the sinusoidal input case. Sigma-Delta performs better than Dithered-oversampling ADC because of the feedback loop but its spectrum presents several undesired harmonics. For all the simulations in this section, we have used a DC input of amplitude 0.6. In Figure 5.4 we observe the quantization noise spectrum for both cases, whereas in Figure 5.5, the SNR versus oversampling ratio is plotted.

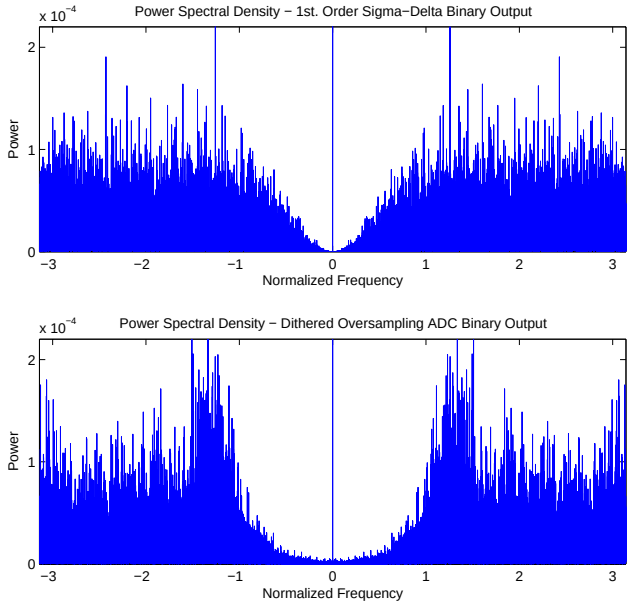

Figure 5.4: Power Spectral Density of the binary output. The top figure shows the Sigma-Delta case, whereas the Dithered-oversampling case is shown in the bottom. The input to the system is a DC input of amplitude 0.6.

5.2 Possible VLSI implementation

For a VLSI implementation of the dithered-oversampling A/D converter, we will need to deal with the optimum blue mask. Up to this point we have developed the theoretical analysis for uniformly distributed masks, but the implementation of a uniform $M[n]$ with certain power spectral density it is a quite challenging task.

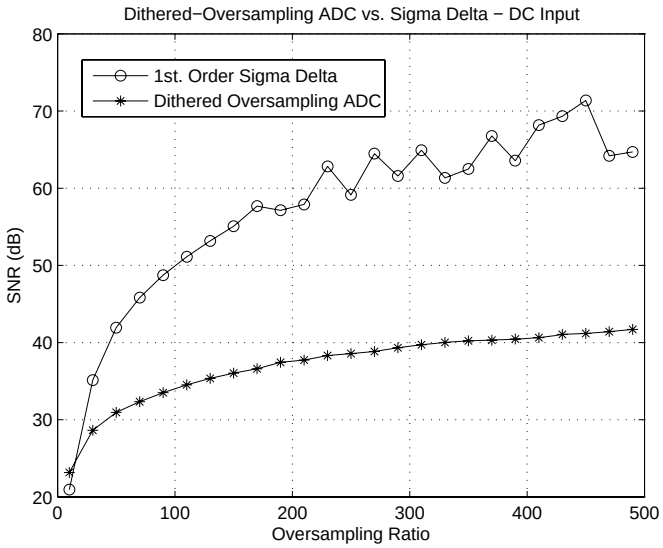

Figure 5.5: SNR versus Oversampling ratio for Sigma-Delta and Dithered-oversampling ADC when using DC inputs.

5.2.1 Non-uniform Mask

In practice, the easiest way to generate any blue noise signal is by high pass filtering white noise. If we do so, we would get the desired spectrum, but the probability density function will definitely not be uniform. One solution for dealing with uniformly distributed blue masks is to generate the mask values in advance and store them in memory. However, this is not a very good solution as it requires a lot of resources and high power consumption which is not desirable [10]. Therefore, one way of solving this problem, is to decide not to use uniform masks. In Chapter 1 we have studied that we can actually use any $M[n]$ with any distribution as long as equation (1.3) holds. For uniform $M[n]$ the comparison threshold is $t[n] = -x[n]$ and therefore $y[n] = sgn(x[n] + M[n])$. On the other hand, if $M[n]$ is not uniform, $y[n] = sgn(M[n] - t[n])$ as stated in equation (1.11).

Now, we will assumme that we have a white noise generator source $v[n]$ which is normal distributed with zero-mean and a certain variance σ_v^2. Then, we will high-pass filter $v[n]$ to obtain our mask $M[n]$. Because of the Central Limit Theorem,

the output $M[n]$ will also be Gaussian distributed with zero-mean and variance σ_M^2.

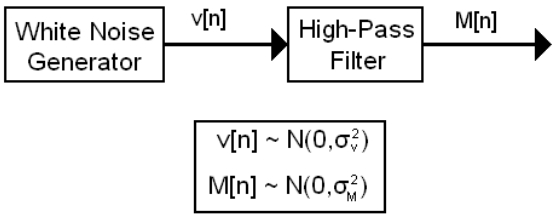

Figure 5.6: M[n] obtained by filtering white noise. The distribution of M[n] will not be uniform, but its spectrum will be blue.

Just like in equation (1.9), we can now find the threshold $t[n]$.

$$\int_{t[n]}^{+\infty} f_M(m)\mathrm{d}m - \int_{-\infty}^{t[n]} f_M(m)\mathrm{d}m = x[n]$$

$$\int_{t[n]}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_M^2}} \exp\left(\frac{-m^2}{2\sigma_M^2}\right)\mathrm{d}m - \int_{-\infty}^{t[n]} \frac{1}{\sqrt{2\pi\sigma_M^2}} \exp\left(\frac{-m^2}{2\sigma_M^2}\right)\mathrm{d}m = x[n]$$

$$Q\Big(\frac{t[n]}{\sigma_M}\Big) + \Big[Q\Big(\frac{t[n]}{\sigma_M}\Big) - 1\Big] = x[n]$$

$$2Q\Big(\frac{t[n]}{\sigma_M}\Big) - 1 = x[n]. \qquad (5.1)$$

To solve for $t[n]$, we need to consider the inverse Q-function $Q^{-1}(x)$.

$$t[n] = \sigma_M Q^{-1}\Big(\frac{x[n]+1}{2}\Big) \qquad (5.2)$$

We don't want to implement such complex function like the inverse Q-function in VLSI, so we will make a first order approximation of equation (5.2). Therefore, given the value of σ_M^2, we can find the coefficients $a_{\sigma_M^2}$ and $b_{\sigma_M^2}$.

$$t[n] \simeq a_{\sigma_M^2} x[n] + b_{\sigma_M^2}. \qquad (5.3)$$

The inverse Q-function is an odd function through the origin, so in general, $b_{\sigma_M^2} = 0$. Then,

$$t[n] \simeq a_{\sigma_M^2} x[n]. \qquad (5.4)$$

Consequently, the binary output $y[n]$ will be,

$$y[n] = sgn(M[n] - a_{\sigma_M^2} x[n]). \tag{5.5}$$

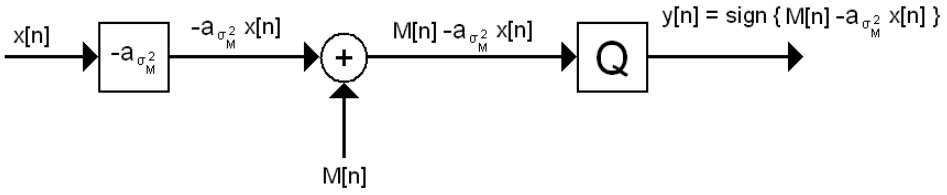

Figure 5.7: Block diagram of the Dithered-oversampling ADC when dealing with a non-uniform M[n].

In this way, we have shown that it would be possible to implement the dithered-oversampling A/D converter with an optimum blue mask. To simplify the implementation, we need to deal with a normal distributed $M[n]$ instead of the uniform one. Therefore, the $M[n]$ will be easily generated by filtering a white noise signal coming out from a white noise generator. In this case, instead of simply add the mask $M[n]$ to $x[n]$ before quantizing (uniform mask case), we first need to multiply $x[n]$ by a constant $a_{\sigma_M^2}$ and then perform quantization.

Figure 5.8 shows the plot of the real value of $t[n]$ from equation (5.2) and its corresponding linear approximation from equation (5.4) for $\sigma_M^2 = 1$. The approximation is done under the least-squares sense.

The table in Figure 5.9 shows the approximation coefficients $a_{\sigma_M^2}$ for different values of σ_M^2.

Once the dither issue is resolved, the oversampling, binary quantization and decimation stages are performed just like in oversampling PCM or Sigma-Delta.

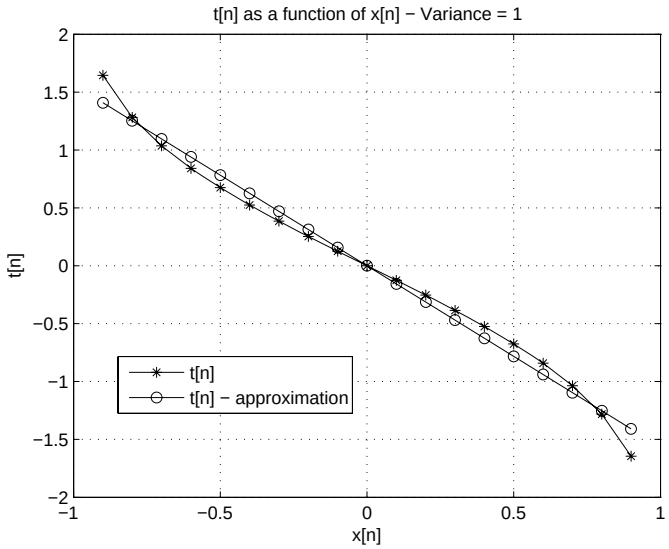

Figure 5.8: t[n] as a function of x[n] for the case of a normal distributed M[n] and its corresponding linear approximation.

σ_M^2	$a_{\sigma_M^2}$
0.5	-1.11
1.0	-1.56
1.5	-1.92

Figure 5.9: Coefficient value $a_{\sigma_M^2}$ for different values of σ_M^2.

5.3 Optical Applications

The rapid development of optical devices in areas such as telecommunications, sensors and imaging, has encouraged the development of optical A/D converters. Optical A/D converters offer better performance than conventional electronic A/D converters because optical signals do not interact with electronic noise and radiation, and are thus immune to electromagnetic interference. All optical A/D converters promise to eliminate the complexity and speed limitations of electrical-to-optical and optical-to-electrical conversions in photonic networks. Recent approaches for optical ADCs basically split the incoming signal energy into several channels, and then compare the energy of each channel to a certain threshold. In this way, with N splitters, we have N+1 possible outputs. For example, if we have three splitters, four possible states are possible, and therefore, a 2-bits ADC has been built [11]. This configuration for A/D conversion is equivalent to PCM.

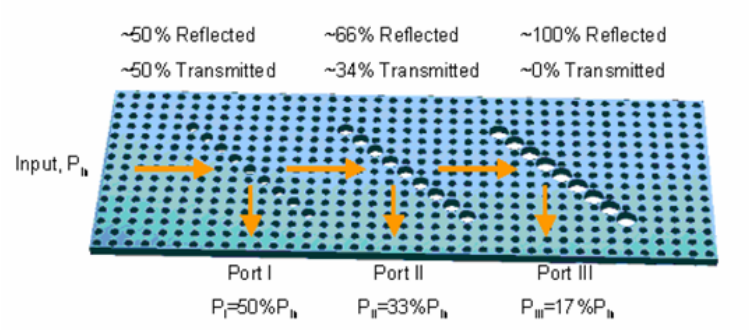

Figure 5.10: Two-bit optical A/D converter consisting of three beam splitting structures in a self-guiding photonic crystal.

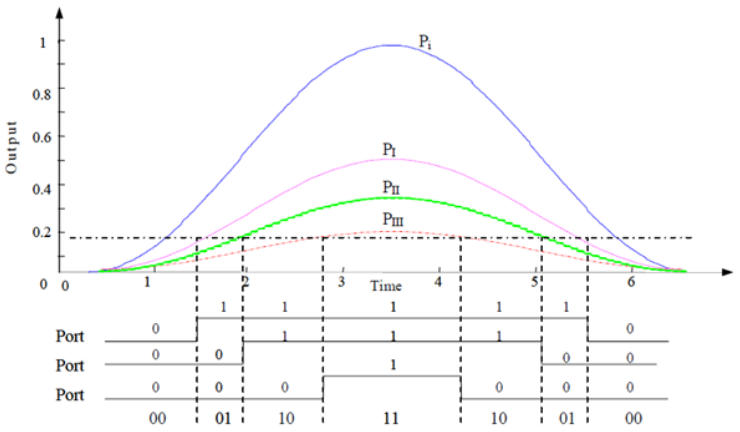

Figure 5.11: Concept of two-bit optical A/D converter.

5.3.1 PCM and Dithered Optical A/D conversion

For optical A/D conversion, Sigma-Delta seems to be impossible or at least very difficult to implement. As mentioned before, to perform a Sigma-Delta conversion we need to store previous samples (not a memoryless conversion). Is it possible to 'store' light or 'integrate' light samples? Until today, the answer seems to be negative. Therefore, memoryless converters seem to be the only valid option when dealing with the design of an optical converter. Oversampled PCM will work in this case, but when dealing with a binary quantizer (or few levels quantizer), the addition of optimum dither $M[n]$ will definitely help to linearize the quantization noise without adding undesired power in the band of interest.

The benefits of using blue noise dither $M[n]$ before quantizing are clearly shown in the simulation presented in Figure 5.12. In this plot, a full scale sinusoid with oversampling ratio $m = 400$ is quantized with a two bits quantizer. The dashed line corresponds to pure PCM and the solid line to the dithered converter (using the optimum blue $M[n]$). In the former, the quantization error power is basically

composed by undesired spikes in the spectrum. On the other hand, when using dither (blue $M[n]$), the quantization noise is 'blue' (meaning that most of the power is pushed to the upper portion of the spectrum) with a very low noise floor in the band of interest.

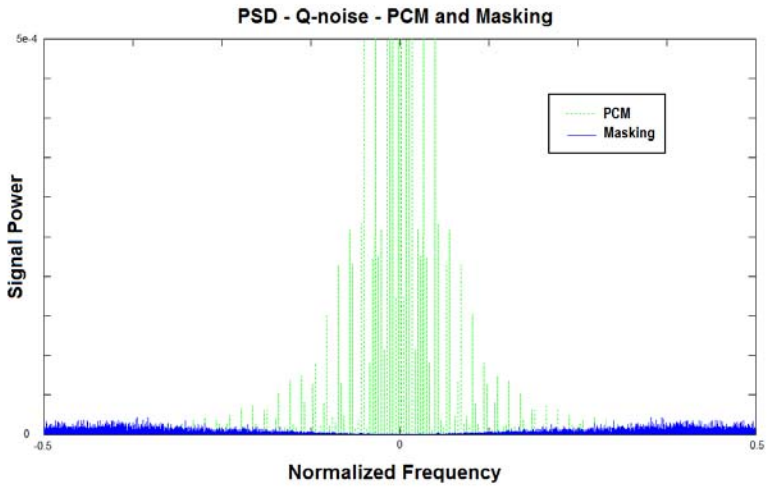

Figure 5.12: Quantization noise spectra for PCM and Dithered ADC. For PCM, the spectrum presents strong spurious tones. On the other hand, when using dither M[n], most of the power is pushed up to the high frequencies resulting in a very low in-band noise floor. Besides, it presents no harmonics.

5.3.2 Band-pass dihtered ADC

One important observation is that up to this point, we are always dealing with low pass signals, so when we decimate, we just care about the quantization noise in the lower portion of the spectrum. Therefore, the optimum mask will be 'blue' (i.e. the optimum is the one with zero power in the lower frequencies). However, in certain applications, we might be dealing with band pass signals where the band of

interest is not necessarily in the low frequencies. In this case, the optimum mask will not be 'blue', but it will be the one with zero power in that particular frequency band of interest. This might be the case of some very high frequency optical applications. If the optimization criteria is changed from low pass to band pass signals, it is also possible to develop a similar analysis as the one through the previous chapters.

5.4 Conclusion

Throughout this work we have analyzed the problem of the binary quantization of an oversampling analog-amplitude signal when using dither $M[n]$. We have studied the cases when the input signal is constant, when it is white noise and when its autocorrelation (or equivalently its spectrum) is known. At the end, we have concluded that the optimum dither $M[n]$ is the one with zero power in the band of interest (i.e. in the low frequencies). As we are dealing with low pass signals, the optimum $M[n]$ is the so called 'blue' mask. That being said, we can now conclude that the use of $M[n]$ makes sense when we are dealing with a binary quantizer or with a quantizer with very few quantization levels, as it is in this case when the quantization error will be correlated with the input. Otherwise, the more quantization levels we have, the more uncorrelated the quantization error to the input will be, and therefore, the linear model of the quantizer becomes acceptable [2]. If we are using a binary quantizer, or if we have few quantization levels, dither is required to 'linearize' the quantizer. However, the addition of any type of noise before the quantization stage will add an extra noise floor level that will decrease the output SNR performance. To avoid this effect, optimum dither will be required. In the case of an oversampling converter where the band of interest is the lower portion of the spectrum, the optimum dither $M[n]$ will be 'blue', meaning that all its power is concentrated in the high frequencies, and no power remains in the lower portion of the spectrum.

BIBLIOGRAPHY

[1] S. P. Lipshitz and J. Vanderkooy, "Why 1-Bit Sigma-Delta Conversion is Unsuitable for High-Quality Applications", presented at the 110th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 2001 May 12-15.

[2] P. M. Aziz, H. V. Sorensen and J. Van Der Spiegel, "An Overview of Sigma-Delta Converters", *IEEE Signal Processing Magazine*, pp. 61-84, January 1996.

[3] R. M. Gray, "Spectral Analysis of Quantization Noise in a Single-Loop Sigma-Delta Modulator with DC Input", *IEEE Transactions on Communications*, Vol. 37, No. 6, pp. 588-599, June 1989.

[4] G. I. Bourdopoulos, A. Pnevmatikakis, V. Anastassopoulos and T. L. Deliyannis, "Delta-Sigma Modulators: Modeling, Design and Applications", Imperial College Press (2003).

[5] D. A. Pierre, "Optimization Theory with Applications", Dover Publications Inc., New York (1986).

[6] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and Dithering: A Theoretical Survey", *J. Audio Eng. Soc.*, Vol. 40, pp. 355-375, May 1992.

[7] R. A. Ulichney, "Dithering wih Blue Noise", *Proceedings of the IEEE*, Vol. 76, No. 1, pp. 56-61, January 1988.

[8] R. A. Ulichney, "The Void-and-Cluster Method for Dither Array Generation", *Proc. SPIE*, Vol. 1913, pp. 332-343.

[9] M. Analoui and J. P. Allebach, "Model-based halftoning by Direct Binary Search", *Proc. SPIE/IS&T Symposium on Electronic Imaging Science and Technology*, pp. 96-108, San Jose, CA, 1992 February 9-14.

[10] B. Le, T. W. Rondeau, J. H. Reed and C. W. Bostian, "Analog-to-Digital Converters", *IEEE Signal Processing Magazine*, pp. 69-77, November 2005.

[11] B. Miao, C. Chen, A. Sharkway, S. Shi and D. W. Prather, "Two bit optical analog-to-digital converter based on photonic crystals", *Optics Express*, Vol. 14, No. 17, pp. 7966-7973, 2006 August 21.

[12] W. Chou and R. M. Gray, "Dithering and Its Effects on Sigma-Delta and Multistage Sigma-Delta Modulation", *IEEE Transactions on Information Theory*, Vol. 37, No. 3, pp. 500-513, May 1991.