ANALYSIS OF TWO PROBLEMS IN SIGNAL QUANTIZATION AND A/D CONVERSION

A Thesis
Presented to
The Academic Faculty

by

David Jiménez

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology
August 2008

ANALYSIS OF TWO PROBLEMS IN SIGNAL QUANTIZATION AND A/D CONVERSION

Approved by:

Professor Yang Wang, Advisor
School of Mathematics
Georgia Institute of Technology

Professor Guillermo Goldsztein
School of Mathematics
Georgia Institute of Technology

Professor Christopher Heil
School of Mathematics
Georgia Institute of Technology

Professor Steven McLaughlin
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Doron Lubinsky
School of Mathematics
Georgia Institute of Technology

Date Approved: April 17, 2008

To Dad. I miss you.

ACKNOWLEDGEMENTS

The step from acquiring to generating knowledge is without a doubt one of the most significant challenges I have faced, and without the support, patience and guidance of several people, it would have been unsurpassable. To all of them I owe my deepest gratitude.

First and foremost, I want to express my deepest gratitude to Yang for all that help and support, academic and otherwise, given to me through this years. His guidance, patience, advice and encouragement have played an essencial roll in my first steps as a researcher. I am indebted to him more than he knows.

Many thanks go also to Chris Heil, Doron Lubinsky, Guillermo Goldsztein and Steven McLaughlin for serving on my committee. To them I also owe some gratitude for amazingly interesting times inside a classroom, and at times, also mathematical and colloquial conversations in hallways and offices.

It is delighting to have the oportunity to pay tribute to the entire staff of the School of Mathematics of Georgia Tech. A special mention goes to Genola Turner, Sharon McDowell and Christy Dalton. I also would like to recognize the efficiency and diligence of the IT Support Team, who have successfully guided me on the resolution of all and every technical problem I have confroted here.

There are many mathematicians who have provided me intellectually stimulating insights into areas of Mathematics as diverse as Harmonic Analysis, Number Theory, Algebraic Geometry and Dynamical Systems. My profound gratitude to Howie Weiss, Matt Baker, Sinan Güntürk, Özgür Yılmaz, John Benedetto, Alex Powell and Ingrid Daubechies.

I have found also very rewarding to be part of the community of gradstudents

of the School of Mathematics. Both for mutual feedback on mathematical topics, as well as for endless hours of some of the greatest (and funniest) conversations I have ever had, I thank Luis Hernándes, Guido Kampel, Trevis Litherland, Hua Xu, Eliane Traldi, Csaba Biro, Teena Carroll and Ben Webb. A very special mention should be made of Alex Grigo, who had the misfortune to have his office next to mine for four long years, and has had to endure me more than any other fellow gradstudent.

I would also like to thank the School of Mathematics to trust me the organization of the High School Math Competition. It has been an honor and a pleasure to share that task with Prof. Rena Brakebill, as well as with Nguyen Truong, Mitch Keller, Cindy Phillips and Nicole Larsen.

Nobody gets to gradschool without having been first and undergrad, and I feel compelled to pay a tribute to some of the men and women who inspired me to follow the path of the mathematician. From them, I owe a special gratitude to Pedro Rodríguez, William Alvarado, Santiago Cambronero and Asdrubal Duarte.

Definitively, I wouldn't be here if not by my family. To my mom and all that group of people who endure me while I was growing up, who supported and encouraged me through the years. Infinite thanks to all of you. To my dad, a special mention, he was the person who nurtured my interest for sciences and mathematics since my early years. I will grieve his loss for the rest of my life.

Last but not least, I honestly think there is nobody on earth I owe more gratitude than to Susana, the woman who was brave enough to marry me and remain with me through the years. She has been the one on my side through the happy times, as well as through the tough ones. For loving me as much as she does (at times more than what I deserve), for supporting me at every moment, and just for being there when I need her, I thank her more than words can fairly express.

No acknoledgement list is ever exhaustive. To those people whose names I have overlooked, I apologize for my omission.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF TABLES viii

LIST OF FIGURES ix

SUMMARY x

I A BRIEF INTRODUCTION TO QUANTIZATION 1

1.1 Quantization and A/D Conversion 1

1.2 Quantizers and General Definitions 2

1.3 Pulse Code Modulation (PCM) 3

PART I WHITE NOISE HYPOTHESIS FOR UNIFORM QUANTIZATION ERRORS OF FRAME EXPANSIONS

II BASICS OF FRAME THEORY 5

2.1 Frames 5

2.2 Useful Facts about Finite Frames 7

2.3 Frames and Vector Quantization 8

III THE WHITE NOISE HYPOTHESIS 10

3.1 Historical Background 10

3.2 *A Priori* Error Bounds and MSE under the WNH 11

IV A CLOSER LOOK AT THE WNH 13

4.1 Legitimacy of the WNH 13

4.2 Asymptotic Behavior of Errors: Linear Independence Case 16

4.3 Asymptotic Behavior of Errors: Linear Dependence Case 19

PART II THE ANALYSIS OF BETA-ALPHA ANALOG-TO-DIGITAL ENCODERS

V REPRESENTATIONS OF REAL NUMBERS 25

5.1 Decimal and Binary Representations 25
5.2 f-Expansions for Real Numbers 26
VI THE β-ENCODER 28
6.1 Imperfect Quantizers 28
6.2 β-Expansions 29
6.3 Robustness of the β-Encoder 31
VII $\beta\alpha$-ENCODER 33
7.1 A Non-Canonical β-Expansion 33
7.2 The $\beta\alpha$-Encoder vs. the β-Encoder 36
7.3 Imprecise α-Multiplication 39
VIII ERGODIC PROPERTIES OF THE $\beta\alpha$-ENCODER 41
8.1 Some Invariant Sets for T 41
8.2 Li-Yorke Theorem and Ergodicity of T for $K = 1$ 43
8.3 Ergodicity of T for $K > 1$ 45
APPENDIX A SAMPLING THEOREM 47
APPENDIX B NUMERICAL RESULTS FOR WNH 50
REFERENCES 54

LIST OF TABLES

1 The Harmonic frame in $\mathbb{R}^2$ 51

2 The randomly generated frame in $\mathbb{R}^4$ 52

3 The Harmonic frame in $\mathbb{R}^4$ 52

4 The frame of Example 5.4 in $\mathbb{R}^3$ 53

LIST OF FIGURES

1 Ranges of acceptable outputs $Q_f(x_n)$ for a β-Encoder. 32

2 Ranges of acceptable outputs $Q_f(x_n)$ for a $\beta\alpha$-Encoder. 38

3 Example of T non Ergodic 46

SUMMARY

In this thesis we consider two different problems in quantization theory. During the first part we discuss the so called *Bennett's White Noise Hypothesis*, introduced to study quantization errors of different schemes. Under this hypothesis, one assumes that the reconstruction errors of different channels can be considered as uniform, independent and identically distributed random variables.

We prove that in the case of uniform quantization errors for frame expansions, this hypothesis is in fact false. Nevertheless, we also prove that in the case of fine quantization, the errors of different channels are asymptotically uncorrelated, validating, at least partially, results on the computation of the mean square error of reconstructions that were obtained through the assumption of Bennett's hypothesis.

On the second part of this thesis, we will introduced a new scalar quantization scheme, called a $\beta\alpha$-encoder. We analyse its robustness with respect to the quantizer imperfections. This scheme also induces a challenging dynamical system. We give partial results dealing with the ergodicity of this system.

CHAPTER I

A BRIEF INTRODUCTION TO QUANTIZATION

1.1 Quantization and A/D Conversion

Information technology introduced through the 20^{th} century and whose rapid development continues to this day has allowed mankind the ability to process, store, transmit and retrieve large volumes of data in digital form, this is, finite strings of *digits*, elements of a finite alphabet.

Up to some extend, this represents a limitation: Digital data is in essence discrete, while an important percentage of the information involved in the daily human life comes from sources that are by nature analog. Therefore there is an intrinsic need to transform this analog information into digital data. This is what we call *analog-to-digital* (A/D) *conversion*.

Analog information seldomly requires an exact reproduction, as measurements need to be known up to certain precision, and images as well as sounds have much more detail than that meeting our senses. Thus, as long as the technology available is able to reproduce such information within the appropriate range of accuracy (to be defined according to the application), some of the original information can be sacrified.

As some detail can be ignored from the original data, given the limitation of sensors, whether it's our sensory organs or electronic sensors, we may model our information as a *bandlimited* function f, this is, the support of its Fourier Transform $\hat{f}$ is in $[-\Omega, \Omega]$ for some finite value $\Omega \in \mathbb{R}$. Without loss of generality, we will assume $\Omega = \pi$.

The *Sampling Theorem*, also called *Shannon-Nyquist Theorem*, solves, at least up

to some extent, this problem.

Theorem 1.1 (Sampling Theorem) *Let $f : \mathbb{R} \to \mathbb{R}$ be a bandlimited function such that $\hat{f}$ is supported in $[-\pi, \pi]$. Let $\lambda > 1$, and $\varphi \in L^1(\mathbb{R})$ such that $\hat{\varphi}$ is continuous and satisfies*

$$\hat{\varphi}(\xi) = \begin{cases} 1 & \text{if } |\xi| \leq \pi, \text{ and} \\ 0 & \text{if } |\xi| \geq \lambda\pi. \end{cases}$$

Then, the following equality holds in the Cesàro mean for all t.

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f\left(\frac{n}{\lambda}\right) \varphi\left(t - \frac{n}{\lambda}\right).$$

A more detailed discussion of this theorem is given in Appendix A. Assuming that the fixed function φ can be computed for any value, f can be perfectly reconstructed from its *samples* $x_n = f(n\lambda^{-1})$. Therefore, the analog signal f can be expressed in terms of the discrete set $\{x_n\}_{n \in \mathbb{Z}}$. Nevertheless, the samples themselves come from the real numbers, and they are still in nature analog, as their digital representation is almost certainly, infinite and aperiodic. Therefore, there is still the need to transform each of the values x_n to a finite digital expression $\tilde{x}_n$, and certainly, there is a need to represent just finitely many of such values. The process of converting these infinitely many samples to a finite collection of finite strings of digits is called *quantization*.

1.2 Quantizers and General Definitions

Once a *sampling* process has been applied to an analog signal f, a *quantizer* or *quantization scheme* is the process of taking the collection $\{x_n\}_{n \in \mathbb{Z}}$ and *encode* them into $\{\tilde{x}_n\}_{n \in \mathcal{J}}$ (for some discrete set $\mathcal{J}$) at a *low cost* in such a way that a reproduction can be recovered from such collection with as *high quality* as possible, where the cost of the process and the quality of the reproduction are to be defined depending on the specific application.

In a more formal setting, a quantizer can be defined as consisting of a *source space* $\mathcal{X}$ (assumed to be a metric space), a density distribution g over $\mathcal{X}$, a set of *cells*

$\mathcal{S} = \{S_i : i \in \mathcal{I}\}$ for some index set $\mathcal{I}$ (we assume that $\mathcal{S}$ is a partition of $\mathcal{X}$, and $\mathcal{I}$ is a finite or countable set), a set of *output values* or *levels* $\mathcal{C} = \{y_i | i \in \mathcal{I}\}$ together with a *quantizer function* defined by $Q(x) = y_i$ for $x \in S_i$. Unless otherwise stated, $\mathcal{C}$ is assumed to be a set of finite binary strings.

It is intuitively clear that given any source space $\mathcal{X}$ and density distribution g, it should be possible to define a wide range of quantizers. How to choose the most appropriate depends therefore on the inherited concepts of cost and reproduction quality.

Generally, a signal is quantized to be stored or transmitted in digital form, and therefore, the length in bits of the quantized output should be optimized. Hence, it is wise to define the cost, or *bit rate* of the quantizer as

$$R(Q) = \sum_{i \in \mathcal{I}} \ell(y_i) P(S_i), \tag{1.1}$$

where $\ell(y_i)$ is the length in bits of the binary representation of y_i, and $P(S_i)$ the probability of a source imput to belong to S_i.

On the other hand, the quality of the quantizer can be defined as how *accurate* the reconstruction of a source input is. Every $y_i \in \mathcal{C}$ has a unique reconstruction value $x_i \in \mathcal{X}$ associated with it. A useful way to define accuracy is to define a distortion measure $d(x, x_i) = |x - x_i|^2$. It is possible then to quantify the average distortion of the system as

$$D(Q) = \mathcal{E}[d(X, Q(X))] = \sum_{i \in \mathcal{I}} \int_{S_i} d(x, x_i) g(x) dx, \tag{1.2}$$

and thus, a small average distortion translate in a high quality of the quantization scheme and vice versa.

1.3 Pulse Code Modulation (PCM)

One of the first quantization schemes is pulse code modulation or PCM. In this case, $\mathcal{S}$ is a partition of $\mathbb{R}$ into disjoint intervals. For every interval $S_i[a_i, a_{i+1})$ in $\mathcal{S}$, the

quantization rule Q assigns to each x in such cell a preset value (called *levels*) $y_i \in S_i$. The values $\{a_i\}$ are called the *thresholds* of the scheme.

A PCM quantizer is said to be *uniform* if the levels y_i are equispaced, say Δ apart, and the thresholds are midway between adjacent levels. If an infinite number of levels is allowed, then all cells S_i width equal to Δ. If only a finite number of levels is allowed, then all but two of the cells will have width Δ and the two outermost will be semi-infinite. Δ is said to be the *quantization step*.

Throughout this thesis, when we refer to PCM, we refer to the *uniform pulse code modulation* quantization scheme, where $\mathcal{C} = \Delta\mathbb{Z}$, with $\Delta > 0$ to be specified and the quantization rule given by

$$Q_\Delta(t) := \left\lfloor \frac{t}{\Delta} + \frac{1}{2} \right\rfloor \Delta. \tag{1.3}$$

In other words t is replaced by the value in $\mathcal{C}$ closest to t.

PART I

White Noise Hypothesis for Uniform Quantization Errors of Frame Expansions

CHAPTER II

BASICS OF FRAME THEORY

When a signal is processed, it is generally practical to quantize the samples in blocks (either of fixed or variable length) instead of sample by sample. If such blocks are considered to have a fixed length, this is called vector quantization. This is the case we will analyze through the first part of this thesis.

If we assume that samples are consistently grouped in blocks of d scallars, we can consider the *input space* as $\mathbb{R}^d$ instead of $\mathbb{R}$.

Once an *input* or *signal* is given, it is often necessary to make an atomic decomposition of it using a given set of *atoms*, or *dictionary* $\{\mathbf{v}_j\}$. In this approach, a signal $\mathbf{x}$ is represented as a linear combination of $\{\mathbf{v}_j\}$,

$$\mathbf{x} = \sum_j c_j \mathbf{v}_j.$$

In practice $\{\mathbf{v}_j\}$ is a finite set. Furthermore, for the purpose of error correction, recovery from data erasures or robustness, redundancy is built into $\{\mathbf{v}_j\}$, i.e. it has more elements than needed. Instead of a true basis, $\{\mathbf{v}_j\}$ is chosen to be a *frame*. We may without loss of generality assume that this *dictionary* has N elements, with $N \geq d$, and thus, we will denoted by $\{\mathbf{v}_j\}_{j=1}^N$.

2.1 Frames

As it was already discussed, one of the basic processes an input undergoes on most applications of A/D conversion is that of *discretization*, in which the input space is by nature *analog* and its samples have to be described through the use of a finite *dictionary*. Due to the potential presence of noise, it is advisable to implement some sort of redundancy in such process to facilitate better reconstruction of the input later

on.

If the input space is a finite-dimention vector space, intuitively this can be seen as representing the signal as linear combination of the elements of some finite set that generates the complete space, in a way, an *over-complete basis*, a set that spans the complete space, as a basis, nevertheless, the linear independence condition is omited. This is, informally speaking, what a frame is. A more formal definition of a frame is given below.

Definition 2.1 (Frame) *An ordered set* $\{\mathbf{v}_j\}_{j\in\mathcal{I}}$ *of elements of a Hilbert space* H *is called a* frame *if the index set* $\mathcal{I}$ *is finite or countable and there are constants* $A, B > 0$ *such that*

$$A\|\mathbf{x}\|^2 \leq \sum_{j\in\mathcal{I}} |(\mathbf{x}\cdot\mathbf{v}_j)|^2 \leq B\|\mathbf{x}\|^2, \tag{2.1}$$

where $\mathbf{x}\cdot\mathbf{y}$ *denotes the inner product of the vectors* $\mathbf{x}$ *and* $\mathbf{y}$.

The numbers A and B in the definition are called *lower* and *upper frame bounds* respectively. The largest $A > 0$ and smallest $B > 0$ satisfying the frame inequalities on (2.1) for all $\mathbf{x} \in H$ are called the *optimal frame bounds*. Also, if $A = B$ then the frame is said to be *tight*.

It is clear that an orthonormal basis $\{\mathbf{e}_j\}_{j\in\mathcal{J}}$ of a Hilbert space is a frame for such space. One of the nicest properties of such basis is the fact that for every $\mathbf{x} \in H$, the following reconstruction formula is satisfied:

$$\mathbf{x} = \sum_{j\in\mathcal{J}} (\mathbf{x}\cdot\mathbf{e}_j)\mathbf{e}_j.$$

Such property is in general not satisfied for a frame. Nevertheless, for any frame $\{\mathbf{v}_j\}_{j\in\mathcal{I}}$ it is always possible to derive an auxiliar frame $\{\mathbf{u}_j\}_{j\in\mathcal{I}}$ such that for every $\mathbf{x} \in H$,

$$\mathbf{x} = \sum_{j\in\mathcal{I}} (\mathbf{x}\cdot\mathbf{v}_j)\mathbf{u}_j = \sum_{j\in\mathcal{I}} (\mathbf{x}\cdot\mathbf{u}_j)\mathbf{v}_j. \tag{2.2}$$

For a detailed proof of this fact, see [9, §5.6]. If $\{\mathbf{v}_j\}_{j\in\mathcal{I}}$ and $\{\mathbf{u}_j\}_{j\in\mathcal{I}}$ satisfy (2.2), then they are said to be each other's *dual frame*.

2.2 Useful Facts about Finite Frames

All the facts mentioned on the previous section apply to general frames, finite or not. On this section we exploit some of the specific characteristics of finite frames. For this purpose, we consider the Hilbert space H to have a finite dimention d, and without loss of generality we call it $\mathbb{R}^d$. Our frame has N elements and it is denoted by $\{\mathbf{v}_j\}_{j=1}^N$.

For encoding purposes, given the ease of reconstruction of $\mathbf{x}$ introduced by (2.2), it is desirable to find a fast way to compute the data $\{\mathbf{x}\cdot\mathbf{v}_j\}_{j=1}^N$. Note that if we set the matrix $F=[\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_N]$, this is, the $d\times N$ matrix having the vectors $\mathbf{v}_j$ as columns, and set $\mathbf{y}=[\mathbf{x}\cdot\mathbf{v}_1,\mathbf{x}\cdot\mathbf{v}_2,\ldots,\mathbf{x}\cdot\mathbf{v}_N]^T$, then $\mathbf{y}=F^T\mathbf{x}$.

Lemma 2.1 *$\{\mathbf{v}_j\}_{j=1}^N$ is a frame if and only if F has rank d.*

As F has full rank ($F^T\mathbf{x}=\mathbf{0}$ implies $\mathbf{x}=\mathbf{0}$), then FF^T is a positive definite matrix, and therefore, invertible, and all its eigenvalues are positive. Furthermore, FF^T is a symmetric matrix, therefore one can choose an orthonormal basis $\{\mathbf{e}_j\}_{j=1}^d$ for $\mathbb{R}^d$ such that each $\mathbf{e}_j$ is an eigenvector of FF^T. Besides, note that

$$(FF^T)^{-1}F\mathbf{y}=(FF^T)^{-1}FF^T\mathbf{x}=\mathbf{x}. \tag{2.3}$$

In this setting, F is called the *matrix representation of the frame* $\{\mathbf{v}_j\}_{j=1}^N$, and for practical purposes, we should not make any distinction between F and $\{\mathbf{v}_j\}_{j=1}^N$. Let's call $G=(FF^T)^{-1}F$, and denote it as $G=[\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_N]$, where $\mathbf{u}_j$ are the columns of G, then note that

$$\mathbf{x}=GF\mathbf{x}=G\mathbf{y}=\sum_{j=1}^N(\mathbf{x}\cdot\mathbf{v}_j)\mathbf{u}_j, \tag{2.4}$$

an thus, F and G are mutual dual frames. We will call G the *cannonical dual frame* of F.

Call $0<\lambda_{\min}=\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_d=\lambda_{\max}$ the eigenvalues of FF^T. Now, suppose that $\{\mathbf{e}_j\}_{j=1}^d$ is an orthonormal basis of $\mathbb{R}^d$, where each $\mathbf{e}_j$ is the eigenvector

of FF^T associated with λ_j. Note that if $\mathbf{x} = a_1\mathbf{e}_1 + \cdots + a_d\mathbf{e}_d$, then

$$\begin{aligned}
\mathbf{x}^T FF^T\mathbf{x} &= (a_1\mathbf{e}_1 + \cdots + a_d\mathbf{e}_d)\cdot(a_1\lambda_1\mathbf{e}_1 + \cdots + a_d\lambda_d\mathbf{e}_d) \\
&= a_1^2\lambda_1 + a_2^2\lambda_2 + \cdots + a_d^2\lambda_d \\
&\leq \lambda_{\max}(a_1^2 + a_2^2 + \cdots + a_2^2) \\
&= \lambda_{\max}\|\mathbf{x}\|^2.
\end{aligned}$$

The equality is achieved for $\mathbf{x} = \mathbf{e}_d$. Similarly $\lambda_{\min}\|\mathbf{x}\|^2 \leq \mathbf{x}^T FF^T\mathbf{x}$, where the equality is achived for $\mathbf{x} = \mathbf{e}_1$. Finally, note that as $F^T\mathbf{x} = [\mathbf{x}\cdot\mathbf{v}_1, \ldots, \mathbf{x}\cdot\mathbf{v}_N]^T$, then

$$\mathbf{x}^T FF^T\mathbf{x} = (F^T\mathbf{x})^T F^T\mathbf{x} = \|F^T\mathbf{x}\|^2 = \sum_{j=1}^{N} |\mathbf{x}\cdot\mathbf{v}_j|^2.$$

Thus,

$$\lambda_{\min}\|\mathbf{x}\|^2 \leq \sum_{j=1}^{N} |\mathbf{x}\cdot\mathbf{v}_j|^2 \leq \lambda_{\max}\|\mathbf{x}\|^2,$$

and therefore $\lambda_{\min}$ and $\lambda_{\max}$ are the optimal frame bounds for F.

If F is a tight frame, then $\lambda = \lambda_{\min} = \lambda_{\max}$ and $G = \lambda^{-1}F$, and the reconstruction formula become

$$\mathbf{x} = \frac{1}{\lambda}\sum_{j=1}^{N}(\mathbf{x}\cdot\mathbf{v}_j)\mathbf{v}_j. \tag{2.5}$$

2.3 Frames and Vector Quantization

Given a frame $\{\mathbf{v}_j\}_{j=1}^N$ and its canonical dual frame $\{\mathbf{u}_j\}_{j=1}^N$, one would desire to use the coefficients $\{\mathbf{x}\cdot\mathbf{v}_j\}_{j=1}^N$ and (2.4) to obtain a perfect reconstruction of $\mathbf{x}$. Nevertheless, as it has been already discussed, such demand is implausible when using a digital media. Instead, the coefficients are to be quantized. We consider a uniform PCM quantization of each individual coefficient, and thus we use the quantized data $\{Q_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\}_{j=1}^N$, where Q_Δ is defined by (1.3), obtaining an imperfect reconstruction

$$\tilde{\mathbf{x}} = \sum_{j=1}^{N} Q_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\mathbf{u}_j. \tag{2.6}$$

This raises the following question: How good is the reconstruction? This question has been studied in terms of both the worst case error and the mean square error (**MSE**), see e.g. [20]. Note that the error from the reconstruction is

$$\mathbf{x} - \tilde{\mathbf{x}} = \sum_{j=1}^{N} \tau_\Delta \left(\mathbf{x} \cdot \mathbf{v}_j \right) \mathbf{u}_j, \tag{2.7}$$

where $\tau_\Delta(t) := t - Q_\Delta(t) = \left(\left\{ \frac{t}{\Delta} + \frac{1}{2} \right\} - \frac{1}{2} \right) \Delta$, with $\{\cdot\}$ denoting the fractional part.

While an *a priori* error bound is relatively straightforward to obtain, the *mean square error* $\mathbf{MSE} := \mathcal{E}\left(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2 \right)$, assuming certain probability distribution for $\mathbf{x}$, is much harder. To simplify the problem, the so-called *White Noise Hypothesis* (**WNH**), is employed by engineers and mathematicians in this area (see e.g. [3, 2, 20]).

In Chapter 3 we will review the **WNH**, the *a priori* error bound and previous results about the **MSE** obtained under such hypothesis. Later, in Chapter 4 we will give a closer look to the **WNH** and the results obtained through it.

CHAPTER III

THE WHITE NOISE HYPOTHESIS

3.1 Historical Background

The **WNH** is often called *Bennett's White Noise Assumption* [3, 2]. Bennett studied quantization error (distortion) in his fundamental paper [4] in the scalar setting.

The **WNH** asserts the following:

- Each $\tau_\Delta(\mathbf{x} \cdot \mathbf{v}_j)$ is uniformly distributed in $[-\Delta/2, \Delta/2)$; hence it has mean 0 and variance $\Delta^2/12$.
- $\{\tau_\Delta(\mathbf{x} \cdot \mathbf{v}_j)\}_{j=1}^N$ are independent random variables.

Bennett demonstrated that under the assumption that the scalar random variable has a smooth density, the quantization error behaves like uniformly distributed "random noise" when Δ is small, resulting in the **MSE** to be approximately $\Delta^2/12$. Bennett also studied quantization errors in the nonuniform quantization setting, which can often be reduced to the uniform setting by the use of companders. The current interest in the **WNH** stems from the study of vector quantization, in which several correlated signals are quantized simultaneously such as in our setting. A vast literature on vector quantization and on vector quantization errors exist, and for an excellent and comprehensive survey on vector quantization see Gray and Neuhoff [22]. A weaker form of the **WNH**, which states that the error components are approximately uncorrelated in the high resolution setting, i.e. when Δ is small, is often found in engineering literatures without rigorous proofs (see [18] and the discussion in [43]).

A rigorous proof of this weaker form of the **WNH** was first given in Viswanathan and Zamir [43]. More precisely, they proved that if two random variables X, Y have

a joint density function then $\frac{1}{\Delta^2}\mathcal{E}\left(\tau_\Delta(X)\tau_\Delta(Y)\right) \longrightarrow 0$ as $\Delta \to 0$. Viswanathan and Zamir also proved similar results in the nonuniform quantization setting, under much stronger assumptions.

3.2 A Priori *Error Bounds and MSE under the WNH*

In this section we derive *a priori* error bounds and a formula for the **MSE** under the **WNH**. These results are not new. We include them for self-containment. We use the following settings throughout this section: Let $\{\mathbf{v}_j\}_{j=1}^N$ be a frame in $\mathbb{R}^d$ with corresponding frame matrix $F = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$. The eigenvalues of FF^T are $0 < \lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d = \lambda_{\max}$. Let $\{\mathbf{u}_j\}_{j=1}^N$ be the canonical dual frame with corresponding matrix $G = (FF^T)^{-1}F$. For any $\mathbf{x} = \sum_{j=1}^N (\mathbf{x}\cdot\mathbf{v}_j)\,\mathbf{u}_j$, using the quantization alphabet $\mathcal{C} = \Delta\mathbb{Z}$ we have the PCM quantized reconstruction

$$\tilde{\mathbf{x}} = \sum_{j=1}^N Q_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\mathbf{u}_j.$$

Proposition 3.1 *For any $\mathbf{x} \in \mathbb{R}^d$ we have*

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \frac{1}{2}\sqrt{\frac{N}{\lambda_{\min}}}\Delta. \tag{3.1}$$

If in addition $\{\mathbf{v}_j\}_{j=1}^N$ is a tight frame with frame constant λ, then

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \frac{1}{2}\sqrt{\frac{N}{\lambda}}\Delta. \tag{3.2}$$

Proof. We have

$$\mathbf{x} - \tilde{\mathbf{x}} = \sum_{j=1}^N \tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\mathbf{u}_j = G\mathbf{y},$$

where $\mathbf{y} = \left[\tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_1\right), \ldots, \tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_N\right)\right]^T$. Thus $\|\mathbf{x}-\tilde{\mathbf{x}}\|^2 = \mathbf{y}^T G^T G\mathbf{y} \leq \rho\left(G^TG\right)\|\mathbf{y}\|^2$ where $\rho(\cdot)$ denotes the spectral radius. Now

$$\rho(G^TG) = \rho(GG^T) = \rho((FF^T)^{-1}) = \frac{1}{\lambda_{\min}}.$$

Observe that $|\tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)| \leq \Delta/2$. Thus $\|\mathbf{y}\|^2 \leq N(\Delta/2)^2$. This yields an *a priori* error bound (3.1). The bound (3.2) is an immediate corollary.

■

Proposition 3.2 *Under the* ***WNH****, the* **MSE** *is*

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)=\frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1}=\frac{\Delta^2}{12}\sum_{j=1}^{N}\|\mathbf{u}_j\|^2. \tag{3.3}$$

In particular, if $\{\mathbf{v}_j\}_{j=1}^N$ *is a tight frame with frame constant* λ*, then*

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)=\frac{d}{12\lambda}\Delta^2. \tag{3.4}$$

Proof. Denote $G^TG=[b_{ij}]_{i,j=1}^N$ and again let $\mathbf{y}=[\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_1),\ldots,\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_N)]^T$. Note that with the **WNH**, $\mathcal{E}(y_iy_j)=\mathcal{E}(\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_i)\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j))=(\Delta^2/12)\delta_{ij}$. Now $\mathbf{x}-\tilde{\mathbf{x}}=G\mathbf{y}$ and hence

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)=\mathcal{E}\left(\mathbf{y}^TG^TG\mathbf{y}\right)=\sum_{i,j=1}^{N}b_{ij}\mathcal{E}\left(y_iy_j\right)=\sum_{i=1}^{N}b_{ii}\frac{\Delta^2}{12}=\frac{\Delta^2}{12}\mathrm{tr}(G^TG).$$

Finally, $\mathrm{tr}(G^TG)=\sum_{j=1}^N\|\mathbf{u}_j\|^2$, and

$$\mathrm{tr}(G^TG)=\mathrm{tr}(GG^T)=\mathrm{tr}((FF^T)^{-1})=\sum_{j=1}^{d}\lambda_j^{-1}.$$

■

Note that using (3.3) the **MSE** for quantization decreases by a factor of 4 if we decrease Δ by a factor of 2. This amounts to an increase in signal to noise ratio of approximately 6dB ($10\log_{10}4\approx6$). This is often referred to as the *6dB-per-bit-rule*.

Remark: The **MSE** formulae (3.3) and (3.4) still hold if the independence of $\{\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\}_{j=1}^N$ in the **WNH** is replaced with the weaker condition of uncorrelation.

CHAPTER IV

A CLOSER LOOK AT THE WNH

4.1 Legitimacy of the WNH

The **WNH** asserts that the error components $\{\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\}_{j=1}^N$ are independent and identically distributed random variables. Intuitively this cannot be true if $N > d$. This is indeed the case in general.

Theorem 4.1 *Let $\mathbf{X} \in \mathbb{R}^d$ be an absolutely continuous random vector. Let $\{\mathbf{v}_j\}_{j=1}^N$ be nonzero vectors in $\mathbb{R}^d$ with $N > d$. Then the random variables $\{\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_j)\}_{j=1}^N$ are not independent.*

Proof. Let F be the frame matrix for the frame $\{\mathbf{v}_j\}$. Then $\dim(\text{range}(F^T)) \le d$, and therefore $\mathcal{L}(\text{range}(F^T)) = 0$ where $\mathcal{L}$ is the Lebesgue measure on $\mathbb{R}^N$. Let $\mathbf{Y} = [Y_1,\dots,Y_N]^T := F^T\mathbf{X}$, and let $\tilde{\mathbf{Y}} = [Q_\Delta(Y_1),\dots,Q_\Delta(Y_N)]^T$ be the quantized $\mathbf{Y}$. Denote $\mathbf{Z} = \mathbf{Y} - \tilde{\mathbf{Y}} = [Z_1,\dots,Z_N]^T$. Note that $Y_j = \mathbf{v}_j\cdot\mathbf{X}$, so each Y_j is absolutely continuous, and therefore so is each Z_j. If $\{Z_j\}$ are independent, then $\mathbf{Z}$ must be absolutely continuous.

Now, Set $\Omega := \text{range}(F^T) + \Delta\mathbb{Z}^N$. Then $\mathcal{L}(\Omega) = 0$ because $\Delta\mathbb{Z}^N$ is a countable set. However, $\mathbf{Z}$ takes values in Ω so $P(\mathbf{Z}\in\Omega) = 1$. This contradicts the absolute continuity of $\mathbf{Z}$.

■

Remark: Actually, for Theorem 4.1 to hold we only need to assume that $\mathbf{X}$ has an absolutely continuous component, i.e. $\mathbf{X} = \mathbf{X}_c + \mathbf{X}_s$ where $\mathbf{X}_c \neq 0$ is absolutely continuous and $\mathbf{X}_s$ is singular. However, the theorem can fail without the absolute continuity condition, even if each component of $\mathbf{X}$ may be absolutely continuous.

The simplest example is to take $\mathbf{X} = [X, -X]^T$ where X is any random variable and $\mathbf{v}_1 = [1, 1]^T$ and $\mathbf{v}_2 = [1, -1]^T$.

Even when $N = d$ the **WNH** holds only under rather strict conditions.

Proposition 4.2 *Let $\mathbf{X} = [X_1, \ldots, X_m]^T$ be a random vector in $\mathbb{R}^m$ whose distribution has density function $g(x_1, \ldots, x_m)$.*

1. *The error components $\{\tau_\Delta(X_j)\}_{j=1}^m$ are independent if and only if there exist complex numbers $\{\beta_j(n) : 1 \le j \le m, n \in \mathbb{Z}\}$ such that*

$$\widehat{g}\left(\frac{a_1}{\Delta}, \ldots, \frac{a_m}{\Delta}\right) = \beta_1(a_1) \cdots \beta_m(a_m) \tag{4.1}$$

   for all $[a_1, \ldots, a_m]^T \in \mathbb{Z}^m$.

2. *Let $h_j(t)$ be the marginal density of X_j. Then $\{\tau_\Delta(X_j)\}_{j=1}^m$ are identically distributed if and only if*

$$\sum_{n \in \mathbb{Z}} h_j(t - n\Delta) = H(t) \quad a.e.$$

   for some $H(t)$ independent of j. They are uniformly distributed on $[-\Delta/2, \Delta/2]$ if and only if $H(t) = 1/\Delta$ a.e..

Proof. To prove (1) denote $\mathcal{I}_\Delta = [-\Delta/2, \Delta/2]$. We first observe that $\mathbf{Y} = [\tau_\Delta(X_1), \ldots, \tau_\Delta(X_m)]^T$ has a density

$$G(\mathbf{y}) := \sum_{\mathbf{a} \in \mathbb{Z}^m} g(\mathbf{y} - \Delta\mathbf{a}) \tag{4.2}$$

for $\mathbf{y} \in \mathcal{I}_\Delta^m$. The density $G(\mathbf{y})$ is periodic with period Δ, and it is well known that its Fourier series is given by $G(\mathbf{y}) = \sum_{\mathbf{a} \in \mathbb{Z}^m} c_{\mathbf{a}} e^{2i\pi \frac{\mathbf{a}}{\Delta} \cdot \mathbf{y}}$, where $c_{\mathbf{a}} = \widehat{g}\left(\frac{\mathbf{a}}{\Delta}\right)$. But $\{Y_j\}_{j=1}^m$ are independent if and only if on $\mathcal{I}_\Delta^m$ we have $g(y_1, \ldots, y_m) = g_1(y_1) \cdots g_m(y_m)$. This happens if and only if

$$\widehat{g}\left(\frac{a_1}{\Delta}, \frac{a_2}{\Delta}, \ldots, \frac{a_m}{\Delta}\right) = h_1\left(\frac{a_1}{\Delta}\right) h_2\left(\frac{a_2}{\Delta}\right) \cdots h_m\left(\frac{a_m}{\Delta}\right)$$

for all $\mathbf{a} = [a_1, \ldots, a_m]^T \in \mathbb{Z}^m$, with $h_j(\xi) = \widehat{g}_i(\xi)$. This part of the theorem is proved by setting $\beta_j(n) = h_j(n)$.

The proof of (2) follows directly from the fact that the density of $\tau_\Delta(X_j)$ is $\sum_{n\in\mathbb{Z}} h_j(t - \Delta n)$ for $t \in \mathcal{I}_\Delta$.

■

Proposition 4.2 puts strong constraints on the distribution of $\mathbf{x}$ for the **WNH** to hold. Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with joint density $f(\mathbf{x})$. Let $\{\mathbf{v}_j\}_{j=1}^d$ be linearly independent, and let $\mathbf{Y} = [\mathbf{X}\cdot\mathbf{v}_1, \mathbf{X}\cdot\mathbf{v}_2, \ldots, \mathbf{X}\cdot\mathbf{v}_d]^T$. Then the joint density of $\mathbf{Y}$ is $g(\mathbf{y}) = |\det(F)|^{-1} f\left((F^T)^{-1}\mathbf{y}\right)$ where $F = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d]$. Thus, both the independence and the identical distribution assumptions in the **WNH**, even for $N = d$, will be false unless very exact conditions are met. For instance, if we take $\mathbf{X}$ to be Gaussian and F to be unitary, then the independence property is satisfied only when F diagonalizes the covariance matrix of $\mathbf{X}$.

Corollary 4.3 *Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with joint density $f(\mathbf{x})$ and $\{\mathbf{v}_j\}_{j=1}^d$ be linearly independent vectors in $\mathbb{R}^d$. Let $\mathbf{Y} = F^T\mathbf{X} = [\mathbf{X}\cdot\mathbf{v}_1, \ldots, \mathbf{X}\cdot\mathbf{v}_N]^T$ and $g(\mathbf{y}) = |\det(F)|^{-1} f\left((F^T)^{-1}\mathbf{y}\right)$ where $F = [\mathbf{v}_1, \ldots, \mathbf{v}_d]$.*

1. *$\{\tau_\Delta(Y_j)\}_{j=1}^d$ are independent random variables if and only if there exist complex numbers $\{\beta_j(n) : 1 \le j \le d, n \in \mathbb{Z}\}$ such that*

$$\widehat{g}\left(\frac{a_1}{\Delta}, \ldots, \frac{a_d}{\Delta}\right) = \beta_1(a_1)\cdots\beta_d(a_d) \tag{4.3}$$

   for all $[a_1, \ldots, a_d]^T \in \mathbb{Z}^d$.

2. *Let $h_j(t) = \int_{\mathbb{R}^{d-1}} g(x_1, \ldots, x_{j-1}, t, x_{j+1}, \ldots, x_d)\ dx_1 \cdots\ dx_{j-1}\ dx_{j+1} \ldots\ dx_d$. Then $\{\tau_\Delta(X_j)\}_{j=1}^d$ are identically distributed if and only if $\sum_{n\in\mathbb{Z}} h_j(t - n\Delta) = H(t)$ a.e. for some $H(t)$ independent of j. They are uniformly distributed on $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ if and only if $H(t) = 1/\Delta$ a.e..*

Proof. We only have to observe that $g(\mathbf{y})$ is the density of $\mathbf{Y}$ and that h_j is the marginal density of Y_j. The corollary now follows directly from the theorem.

■

From a practical point of view, with coarse quantization the **MSE** of quantization errors cannot be estimated simply by (3.3). Thus the "6-dB-per-bit" rule may not apply. We shall demonstrate this with numerical results. However, with high resolution quantization the formula (3.3) becomes increasingly accurate. We show this in the next section.

4.2 Asymptotic Behavior of Errors: Linear Independence Case

In many practical applications such as music CD, fine quantizations with 16 bits or more have been adopted. Although the **WNH** is not valid in general, with fine quantizations we prove here that a weaker version of the **WNH** is close to being valid, which yields an asymptotic formula for the PCM quantized **MSE**. Our result here strengthen an asympototic result in [43].

We again consider the same setup as before. Let $\{\mathbf{v}_j\}_{j=1}^N$ be a frame in $\mathbb{R}^d$ with corresponding frame matrix $F = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$. The eigenvalues of FF^T are $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d = \lambda_{\min} > 0$. Let $\{\mathbf{u}_j\}_{j=1}^N$ be the canonical dual frame with corresponding matrix $G = (FF^T)^{-1}F$. For any $\mathbf{x} \in \mathbb{R}^d$ we have $\mathbf{x} = \sum_{j=1}^N (\mathbf{x} \cdot \mathbf{v}_j)\, \mathbf{u}_j$. Using the quantization alphabet $\mathcal{A} = \Delta\mathbb{Z}$ we have the PCM reconstruction (2.6). Note that $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}(\Delta)$ as it depends on Δ. With the **WNH** we obtain the **MSE**

$$\mathbf{MSE} = \mathcal{E}\left(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2\right) = \frac{\Delta^2}{12} \sum_{j=1}^N \lambda_j^{-1}.$$

To study the asymptotic behavior of the error components, we study as $\Delta \to 0^+$ the normalized quantization error

$$\frac{1}{\Delta}(\mathbf{x} - \tilde{\mathbf{x}}) = \sum_{j=1}^N \frac{1}{\Delta} \tau_\Delta\left(\mathbf{x} \cdot \mathbf{v}_j\right) \mathbf{u}_j. \tag{4.4}$$

Theorem 4.4 *Let $\mathbf{X} \in \mathbb{R}^d$ be an absolutely continuous random vector. Let $\{\mathbf{w}_j\}_{j=1}^m$ be a collection of linearly independent vectors in $\mathbb{R}^d$. Then*

$$\left[\frac{1}{\Delta}\tau_\Delta\left(\mathbf{X}\cdot\mathbf{w}_1\right),\ldots,\frac{1}{\Delta}\tau_\Delta\left(\mathbf{X}\cdot\mathbf{w}_m\right)\right]^T$$

converges in distribution as $\Delta \to 0^+$ to a random vector uniformly distributed in $[-1/2,1/2]^m$.

Proof. Denote $Y_j = \mathbf{X}\cdot\mathbf{w}_j$. Since $\{\mathbf{w}_j\}$ are linearly independent, $\mathbf{Y} = [Y_1,\ldots,Y_m]^T$ is absolutely continuous with some joint density $f(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^m$. As a consequence of (4.2) one has that the distribution of $\mathbf{Z} = [Z_1,\ldots,Z_m]^T$, where $Z_j = \frac{1}{\Delta}\tau_\Delta(Y_j) = \left\{\frac{Y_j}{\Delta}+\frac{1}{2}\right\}-\frac{1}{2}$, is

$$f_\Delta(\mathbf{x}) := \Delta^m \sum_{\mathbf{a}\in\mathbb{Z}^m} f(\Delta\mathbf{x}-\Delta\mathbf{a}). \tag{4.5}$$

for $\mathbf{x}\in[-1/2,1/2]^m$. Again denote $\mathcal{I}_1 := [-1/2,1/2]$. It is easy to see that $\|f_\Delta\|_{L^1(\mathcal{I}_1^m)} \le \|f\|_{L^1(\mathbb{R}^m)}$, for

$$\begin{aligned}
\|f_\Delta\|_{L^1(\mathcal{I}_1^m)} &= \int_{\mathcal{I}_1^m} |f_\Delta(\mathbf{x})|\,d\mathbf{x} \\
&\le \sum_{\mathbf{a}\in\mathbb{Z}^m}\int_{\mathcal{I}_1^m} \Delta^m\,|f(\Delta\mathbf{x}-\Delta\mathbf{a})|\;d\mathbf{x} \\
&= \sum_{\mathbf{a}\in\mathbb{Z}^m}\int_{\Delta{\mathcal{I}_1}^m+\Delta\mathbf{a}} |f\left(\mathbf{y}\right)|\;d\mathbf{y} \\
&= \int_{\mathbb{R}^m} |f\left(\mathbf{y}\right)|\;d\mathbf{y} \\
&= \|f\|_{L^1(\mathbb{R}^m)}.
\end{aligned}$$

Now, if $\Omega = [a_1,b_1]\times\cdots\times[a_m,b_m]$ and $f(\mathbf{x}) = \mathbf{1}_\Omega(\mathbf{x})$, then for $\mathbf{x}\in\mathcal{I}_1^m$ observe that $f_\Delta(\mathbf{x}) = \Delta^m K_\Delta$ where $K_\Delta(\mathbf{x}) = \#\{\mathbf{a}\in\mathbb{Z}^m : \Delta\mathbf{x}+\Delta\mathbf{a}\in\Omega\}$. Obviously, $K_\Delta(\mathbf{x}) = s/\Delta^m + O(\Delta^{-m+1})$ where $s = \mathcal{L}(\Omega)$ is the Lebesgue measure of Ω. Then $f_\Delta \to s\mathbf{1}_{\mathcal{I}_1^m}$ in $L^1(\mathcal{I}_1^m)$ as $\Delta\to 0^+$.

Coming back to the case when $f(\mathbf{x})$ is the density of $\mathbf{Y}$. For any $\varepsilon>0$ it is possible to choose a $g(\mathbf{x})\in L^1(\mathbb{R}^m)$ such that $\|f-g\|_{L^1} < \frac{\varepsilon}{3}$, and furthermore,

$g(\mathbf{x}) = \sum_{j=1}^{N} c_j \mathbf{1}_{E_j}(\mathbf{x})$ is a simple function where $c_j \in \mathbb{R}$ and each E_j is a product of finite intervals. Observe that $\int_{\mathbb{R}^m} g = \sum_{j=1}^{N} c_j \mathcal{L}(E_j)$. Since $(\mathbf{1}_{E_j})_\Delta \to \mathcal{L}(E_j)\mathbf{1}_{\mathcal{I}_1^m}$ in L^1 we have $g_\Delta \to \left(\int_{\mathbb{R}^m} g\right)\mathbf{1}_{\mathcal{I}_1^m}$ as $\Delta \to 0$. Hence there exists a $\delta > 0$ such that $\left\|g_\Delta - \left(\int_{\mathbb{R}^m} g\right)\mathbf{1}_{\mathcal{I}_1^m}\right\|_{L^1} < \varepsilon/3$ whenever $\Delta < \delta$. Now, for $\Delta < \delta$,

$$\begin{aligned}
\left\|f_\Delta - \mathbf{1}_{\mathcal{I}_1^m}\right\|_{L^1(\mathcal{I}_1^m)} &= \|f_\Delta - g_\Delta\|_{L^1(\mathcal{I}_1^m)} + \left\|g_\Delta - \left(\int_{\mathbb{R}^m} g\right)\mathbf{1}_{\mathcal{I}_1^m}\right\|_{L^1(\mathcal{I}_1^m)} \\
&\quad + \left|1 - \left(\int_{\mathbb{R}^m} g\right)\right| \|\mathbf{1}_{\mathcal{I}_1^m}\|_{L^1(\mathcal{I}_1^m)} \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \left|1 - \left(\int_{\mathbb{R}^m} g\right)\right| \\
&= \frac{2\varepsilon}{3} + \left|\left(\int_{\mathbb{R}^m} g\right) - \left(\int_{\mathbb{R}^m} g\right)\right| \\
&< \varepsilon.
\end{aligned}$$

■

Remark: We in fact proved a slightly stronger result, namely the densities converge in L^1. Applying the above theorem to the **MSE**, if $\{\mathbf{v}_j\}_{j=1}^N$ are pairwise linearly independent then the error components $\{\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_j)\}_{j=1}^N$ become asymptotically pairwise independent and each uniformly distributed in $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$.

Corollary 4.5 *Let $\mathbf{X} \in \mathbb{R}^d$ be an absolutely continuous random vector. If $\{\mathbf{v}_j\}_{j=1}^N$ are pairwise linearly independent, then as $\Delta \to 0^+$ we have*

$$\mathcal{E}\left(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2\right) = \frac{\Delta^2}{12}\sum_{j=1}^{d} \lambda_j^{-1} + o(\Delta^2) = \frac{\Delta^2}{12}\sum_{j=1}^{N} \|\mathbf{u}_j\|^2 + o(\Delta^2). \tag{4.6}$$

Proof. Denote by F the frame matrix associated with $\{\mathbf{v}_j\}_{j=1}^N$, $H = (FF^T)^{-1}$, $Y_j = \mathbf{X}\cdot\mathbf{v}_j$, $Z_j = \left\{\frac{Y_j}{\Delta} + \frac{1}{2}\right\} - \frac{1}{2}$, and $\mathbf{Z} = [Z_1, \ldots, Z_m]^T$. By Theorem 4.4, $\mathcal{E}(Z_i) \to 0$

and $\mathcal{E}(Z_i Z_j) \to \frac{1}{12}\delta_{ij}$ as $\Delta \to 0^+$. It follows from the proof of Proposition 3.2 that

$$\begin{aligned}
\frac{1}{\Delta^2}\mathcal{E}(\|\mathbf{X}-\tilde{\mathbf{X}}\|^2) &= \mathcal{E}(\mathbf{Z}^T H \mathbf{Z}) \\
&= \mathcal{E}\left(\sum_{i,j=1}^{N} Z_i Z_j h_{ij}\right) \\
&= \sum_{i,j=1}^{N} h_{ij}\mathcal{E}(Z_i Z_j) \\
&= \frac{1}{12}\sum_{i=1}^{N} h_{ii} + o(1) \\
&= \frac{1}{12}\sum_{j=1}^{d} \lambda_j^{-1} + o(1),
\end{aligned}$$

and hence

$$\mathcal{E}\left(\|\mathbf{X}-\tilde{\mathbf{X}}\|^2\right) = \frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1} + o(\Delta^2) = \frac{\Delta^2}{12}\sum_{j=1}^{N}\|\mathbf{u}_j\|^2 + o(\Delta^2).$$

■

4.3 Asymptotic Behavior of Errors: Linear Dependence Case

In this section we consider the case in which some vectors in the frame may be parallel. This can happen, for example, if the frame contains redundant elements. Mathematically it would be interesting to understand how the **MSE** behaves as $\Delta \to 0^+$. We return to previous calculations and note that

$$\mathcal{E}(\|\mathbf{X}-\tilde{\mathbf{X}}\|^2) = \sum_{i,j=1}^{N} h_{ij}\mathcal{E}\left(\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_i)\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_j)\right).$$

Our main result in this section is:

Theorem 4.6 *Let X be an absolutely continuous real random variable. Let $\alpha \in$*

$\mathbb{R} \setminus \{0\}$. *Then*

$$\lim_{\Delta \to 0^+} \frac{1}{\Delta^2} \mathcal{E}\left(\tau_\Delta(X)\tau_\Delta(\alpha X)\right) = \begin{cases} 0, & \alpha \notin \mathbb{Q}, \\ \dfrac{1}{12pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is even}, \\ -\dfrac{1}{24pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is odd}, \end{cases} \tag{4.7}$$

where p, q are coprime integers.

Proof. Denote $g(x) := \{x + \frac{1}{2}\} - \frac{1}{2}$. Let $\phi(x) \geq 0$ be an even C^∞ function such that $\mathrm{supp}(g) \subseteq [-1, 1]$ and $\int_{\mathbb{R}} \phi = 1$. Let $g_n(x) = g * \phi_n$ where $\phi_n(x) = n\phi(nx)$. It is standard to check that

(a) $|g_n(x)| \leq 1/2$;

(b) $\mathrm{supp}(g(x) - g_n(x)) \subseteq [\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}] + \mathbb{Z}$;

(c) $g_n(x) \in C^\infty$, and is $\mathbb{Z}$-periodic;

(d) $\int_{\mathbb{R}} g_n(x)\, dx = 0$.

$g_n(x)$ represents a small perturbation of $g(x)$ that "smoothes out" the discontinuities of $g(x)$. Now, set

$$\begin{aligned} E(\Delta) &:= \mathcal{E}\left(\frac{1}{\Delta^2}\tau_\Delta(X)\tau_\Delta(\alpha X)\right) \\ &= \mathcal{E}\left(g\left(\frac{X}{\Delta}\right) g\left(\frac{\alpha X}{\Delta}\right)\right) \\ &= \int_{\mathbb{R}} g\left(\frac{x}{\Delta}\right) g\left(\frac{\alpha x}{\Delta}\right) f(x)\, dx, \end{aligned}$$

and

$$E_n(\Delta) := \int_{\mathbb{R}} g_n\left(\frac{x}{\Delta}\right) g_n\left(\frac{\alpha x}{\Delta}\right) f(x)\, dx.$$

Claim: $E_n(\Delta) \to E(\Delta)$ *as* $n \to \infty$ *uniformly for all* $\Delta > 0$.

Proof of the Claim. Let f be the density of $\mathbf{X}$. For any $\varepsilon > 0$,

$$\begin{aligned} |E_n(\Delta) - E(\Delta)| &= \left| \int_{\mathbb{R}} \left[g_n\left(\frac{x}{\Delta}\right) g_n\left(\frac{\alpha x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) g\left(\frac{\alpha x}{\Delta}\right) \right] f(x)\, dx \right| \\ &\leq \frac{1}{2} \int_{\mathbb{R}} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) \right| f(x)\, dx \\ &\quad + \frac{1}{2} \int_{\mathbb{R}} \left| g_n\left(\frac{\alpha x}{\Delta}\right) - g\left(\frac{\alpha x}{\Delta}\right) \right| f(x)\, dx. \end{aligned}$$

Now there exists an $M > 0$ such that $\int_{[-M,M]^c} f(x)\, dx < \frac{\varepsilon}{2}$. So

$$\int_{\mathbb{R}} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) \right| f(x)\, dx \leq \int_{-M}^{M} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) \right| f(x)\, dx + \frac{\varepsilon}{2}.$$

Furthermore, let $A_n(\Delta, M) := \operatorname{supp}(g_n(x/\Delta) - g(x/\Delta)) \cap [-M, M]$. Then we have

$$A_n(\Delta, M) \subseteq \Delta\Big([\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}] + \mathbb{Z} \Big) \cap [-M, M].$$

Hence $\mathcal{L}(A_n(\Delta, M)) \leq \frac{2M}{\Delta} \cdot \frac{2\Delta}{n} = \frac{4M}{n}$, and thus

$$\int_{-M}^{M} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) \right| f(x)\, dx \leq \int\limits_{A_n(\Delta,M)} f(x)\, dx < \frac{\varepsilon}{2}$$

by choosing n sufficiently large (independent of Δ), which yields

$$\int_{\mathbb{R}} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) \right| f(x)\, dx < \varepsilon.$$

Similarly we have

$$\int_{\mathbb{R}} \left| g_n\left(\frac{\alpha x}{\Delta}\right) - g\left(\frac{\alpha x}{\Delta}\right) \right| f(x)\, dx < \varepsilon$$

for sufficiently large n, proving the Claim. □

Now consider the Fourier Series of $g_n(t)$,

$$g_n(t) = \sum_{k \in \mathbb{Z}} c_k^{(n)} e^{2\pi i k t}.$$

It is well known that the Fourier series converges to $g_n(t)$ uniformly for all t, see e.g. [45]. Furthermore, since $g_n(t)$ is C^∞ we have $|c_k^{(n)}| = o\left((|k|+1)^{-L}\right)$ for all $L > 0$, giving absolute convergence of the Fourier series. Thus

$$\begin{aligned} E_n(\Delta) &= \lim_{K \to \infty} \int_{\mathbb{R}} \Big(\sum_{|k| \leq K} c_k^{(n)} e^{2\pi i k t \Delta^{-1}} \Big) \Big(\sum_{|k| \leq K} c_k^{(n)} e^{2\pi i k \alpha t \Delta^{-1}} \Big) f(t)\, dt \\ &= \lim_{K \to \infty} \sum_{|k|, |\ell| \leq K} c_k^{(n)} c_\ell^{(n)} \widehat{f}\Big(-\frac{k + \alpha \ell}{\Delta} \Big). \end{aligned}$$

Observe that $|\widehat{f}(\xi)| \le \|f\|_{L^1} = 1$, and $|c_k^{(n)}| = o\big((|k|+1)^{-L}\big)$ for any $L > 0$. So the series converges absolutely and uniformly in Δ. Thus

$$E_n(\Delta) = \sum_{k,\ell\in\mathbb{Z}} c_k^{(n)} c_\ell^{(n)} \widehat{f}\Big(-\frac{k+\alpha\ell}{\Delta}\Big). \tag{4.8}$$

For any $n > 0$ we have

$$\lim_{\Delta\to 0^+} E_n(\Delta) = \sum_{k,\ell\in\mathbb{Z}} c_k^{(n)} c_\ell^{(n)} \lim_{\Delta\to 0^+} \widehat{f}\left(-\frac{k+\alpha\ell}{\Delta}\right)$$

because the series converges absolutely and uniformly. Suppose $\alpha \notin \mathbb{Q}$. Then $k+\alpha\ell \neq 0$ if either $k \neq 0$ or $\ell \neq 0$. Thus $\big|-\frac{k+\alpha\ell}{\Delta}\big| \to \infty$, and hence $\lim_{\Delta\to 0^+} \widehat{f}\big(-\frac{k+\alpha\ell}{\Delta}\big) = 0$ as $f \in L^1(\mathbb{R})$. Note also that $c_0^{(n)} = \int_{\mathbb{R}} g_n = 0$. It follows that

$$\lim_{\Delta\to 0^+} E_n(\Delta) = 0.$$

But $E_n(\Delta) \to E(\Delta)$ as $n \to \infty$ uniformly in Δ, which yields $E(\Delta) \to 0$ as $\Delta \to 0^+$.

Next, suppose $\alpha = \frac{p}{q}$ where $p, q \in \mathbb{Z}$, $(p,q) = 1$. We observe that $k + \alpha\ell = 0$ if and only if $k = pm$ and $\ell = -qm$ for some $m \in \mathbb{Z}$. In such a case

$$\widehat{f}\Big(-\frac{k+\alpha\ell}{\Delta}\Big) = \widehat{f}(0) = \int_{\mathbb{R}} f = 1.$$

It follows that

$$\lim_{\Delta\to 0^+} E_n(\Delta) = \sum_{m\in\mathbb{Z}} c_{pm}^{(n)} c_{-qm}^{(n)} \widehat{f}(0) = \sum_{m\in\mathbb{Z}} c_{pm}^{(n)} c_{-qm}^{(n)} = \sum_{m\in\mathbb{Z}} c_{pm}^{(n)} \overline{c_{qm}^{(n)}}.$$

For $r \in \mathbb{Z}, r \neq 0$ set

$$G_r^{(n)}(x) := \sum_{m\in\mathbb{Z}} c_{rm}^{(n)} e^{2\pi i m x}.$$

By Parseval we have

$$\lim_{\Delta\to 0} E_n(\Delta) = \left\langle G_q^{(n)}, G_p^{(n)} \right\rangle_{L^2([0,1])}.$$

It is easy to check that

$$G_r^{(n)} = \frac{1}{|r|} \sum_{j=0}^{|r|-1} g_n\Big(\frac{x+j}{r}\Big).$$

Hence $G_r^{(n)}$ converges in $L^2([0,1])$ to $G_r(x) := \frac{1}{|r|}\sum_{j=0}^{|r|-1} g\big(\frac{x+j}{r}\big)$, which has Fourier series $G_r(x) = \sum_{m\in\mathbb{Z}} c_{rm} e^{2\pi i m x}$ with $c_0 = 0$ and $c_k = \frac{(-1)^{k-1}}{2\pi i k}$ for $k \neq 0$. This yields

$$\lim_{n\to\infty}\lim_{\Delta\to 0^+} E_n(\Delta) = \lim_{n\to\infty}\left\langle G_q^{(n)}, G_p^{(n)}\right\rangle = \langle G_q, G_p\rangle = \sum_{m\in\mathbb{Z}} c_{qm}\overline{c_{pm}}.$$

Finally

$$\begin{aligned}\sum_{m\in\mathbb{Z}} c_{qm}\overline{c_{pm}} &= \sum_{m\in\mathbb{Z}\setminus\{0\}} \Big(\frac{(-1)^{qm-1}}{2\pi i m q}\Big)\overline{\Big(\frac{(-1)^{pm-1}}{2\pi i m p}\Big)} \\ &= \frac{1}{2pq\pi^2}\sum_{m=1}^{\infty}\frac{(-1)^{(p+q)m}}{m^2}.\end{aligned}$$

Note that if $p+q$ is even then $\sum_{m=1}^{\infty}\frac{(-1)^{(p+q)m}}{m^2} = \sum_{m=1}^{\infty}\frac{1}{m^2} = \frac{\pi^2}{6}$. On the other hand, if $p+q$ is odd then $\sum_{m=1}^{\infty}\frac{(-1)^{(p+q)m}}{m^2} = \sum_{m=1}^{\infty}\frac{(-1)^m}{m^2} = -\frac{\pi^2}{12}$. The theorem follows.

■

Corollary 4.7 *Let* $\mathbf{X}$ *be an absolutely continuous random vector in* $\mathbb{R}^d$, $\mathbf{w} \neq 0$, $\mathbf{w} \in \mathbb{R}^d$ *and* $\alpha \in \mathbb{R}\setminus\{0\}$. *Then*

$$\lim_{\Delta\to 0^+}\frac{1}{\Delta^2}\mathcal{E}\left(\tau_\Delta(\mathbf{w}\cdot\mathbf{X})\tau_\Delta(\alpha\mathbf{w}\cdot\mathbf{X})\right) = \begin{cases} 0, & \alpha \neq \mathbb{Q}, \\ \dfrac{1}{12pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is even}, \\ -\dfrac{1}{24pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is odd}, \end{cases} \tag{4.9}$$

where p, q *are coprime integers.*

Proof. We only need to note that $\mathbf{w}\cdot\mathbf{X}$ is an absolutely continuous random variable. The corollary follows immediately from Theorem 4.4.

■

We can now characterize completely the asymptotic behavior of the MSE in all cases. For any two vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ define $r(\mathbf{w}_1, \mathbf{w}_2)$ by

$$r(\mathbf{w}_1,\mathbf{w}_2) = \begin{cases} \dfrac{1}{pq}\mathbf{w}_1\cdot\mathbf{w}_2, & \mathbf{w}_1 = \dfrac{p}{q}\mathbf{w}_2, \text{ and } p+q \text{ is even}, \\ -\dfrac{1}{2pq}\mathbf{w}_1\cdot\mathbf{w}_2, & \mathbf{w}_1 = \dfrac{p}{q}\mathbf{w}_2, \text{ and } p+q \text{ is odd}, \\ 0, & \text{otherwise}, \end{cases}$$

where p, q are coprime integers.

Corollary 4.8 *Let* $\mathbf{X} \in \mathbb{R}^d$ *be an absolutely continuous random vector. Then as* $\Delta \longrightarrow 0^+$ *the* **MSE** *satisfies*

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right) = \frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1} + \frac{\Delta^2}{6}\sum_{1\le i<j\le N} r(\mathbf{u}_i,\mathbf{u}_j) + o(\Delta^2), \tag{4.10}$$

Proof. In the proof of (4.5) we showed that

$$\lim_{\Delta\to 0^+}\frac{1}{\Delta^2}\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right) = \sum_{i,j} h_{ij}\mathcal{E}\left(Z_iZ_j\right)$$

with the notations there. The result is immediate from Theorem 4.7. ■

For fixed quantization step $\Delta > 0$ we shall denote

$$\mathbf{MSE}_{ideal} = \frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1} + \frac{\Delta^2}{6}\sum_{1\le i<j\le N} r(\mathbf{u}_i,\mathbf{u}_j), \tag{4.11}$$

and call it the *ideal* ***MSE***. If $\{\mathbf{v}_j\}_{j=1}^N$ are pairwise linearly independent, then the $\mathbf{MSE}_{ideal}$ is simply $\frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1}$, the **MSE** under the **WNH**.

We should point out that even though the **WNH** is not true asympototically if some vectors in a frame are parallel, the contribution from the second part of (4.11) is often small enough that the **MSE** under the **WNH** is close enough to the ideal **MSE**. In Appendix we shall show some numerical data, comparing the actual **MSE** with the ideal **MSE**.

PART II

The Analysis of Beta-Alpha Analog-to-Digital Encoders

CHAPTER V

REPRESENTATIONS OF REAL NUMBERS

With the development of writing systems, as well as the growth in the complexity of the trade and engineering experienced by mankind around the 4^{th} millennium BC, the need of numeral systems was as self evident as it is today. The historical development of such concepts is far from the scope of this work. Nevertheless we give a brief introduction to some of the most widely used systems today.

5.1 Decimal and Binary Representations

The decimal representation is without a doubt the most widely used numeral system in every day life around the world. The decimal system is a positional notation numeral system. This means, the symbols 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9 represent respectively the first ten non-negative integer numbers, and are called *digits*, and thus on any given representation of a number, a position is related to the next by the common ratio of 10, that is called the *base* or *radix*. As the reader should be more than familiar, the string 1981 represents $1 \cdot 10^3 + 9 \cdot 10^2 + 8 \cdot 10^1 + 1 \cdot 10^0$ and the string 37.125 represents $3 \cdot 10^1 + 7 \cdot 10^0 + 1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 5 \cdot 10^{-3}$.

The election of 10 as the radix of the system is not arbitrary, as the fingers of our hands were the first counting machines available, but any positive integer other than 1 can be used for such purpose. The first, and in some sense, the "mathematically most natural" choice for such radix would be the number 2, or *binary*. In this case, each number is represented by a finite or infinite string of zeros and ones. In this case, the number that was previously represented as 37.125 in decimal notation, would be represented in binary by 10011.001.

The historical success of positional systems comes from the fact that they ease

the symbolic computation of the basic arithmetic computations, although, in modern Mathematics, there are other systems used. Most of them can be considered as particular cases of f-expansions for real numbers. These would be introduced in the next section.

5.2 f-Expansions for Real Numbers

A mathematically interesting way to express real numbers is the use of continued fractions. For example, the number obtained by dividing 59 by 26 can be represented in decimal notation by $2.26923076923\cdots = 2.2\overline{692307}$. On the other hand, note that

$$\frac{59}{26} = 2 + \frac{1}{3 + \frac{1}{1+\frac{1}{2+\frac{1}{2}}}}$$

and thus such number could be represented also by $[2; 3, 1, 2, 2]$.

The representation of numbers through continued fractions has been vastly studied for the last three centuries. In 1944 Bissinger established that these are, together with decimal and binary representation, particular cases of what he called f-expansions for real numbers (See [5]).

In general term, an f-expansion scheme yields a representation for a non-negative number x through the iteration of the function $y = f(x)$. Define

$$\begin{aligned} b_0 &= \lfloor x \rfloor \\ x_0 &= x - b_0 \\ b_n &= \lfloor f^{-1}(x_{n-1}) \rfloor \\ x_n &= f^{-1}(x_{n-1}) - b_n \end{aligned} \tag{5.1}$$

where $\lfloor \cdot \rfloor$ represents the integer part. With this information, one should be able to reconstruct x by

$$x = b_0 + f(b_1 + f(b_2 + f(b_3 + \cdots))). \tag{5.2}$$

There are of course several conditions that f should satisfy to obtain a valid f-expansion scheme (See [37]). At the very least, there should be sets $\mathcal{R}$ and $\mathcal{D}$, such

that $\mathcal{R} \subseteq [0,1]$, $\mathcal{D}$ is a subset of the non-negative real numbers and $f : \mathcal{D} \to \mathcal{R}$ is a bijection. Normally, both $\mathcal{D}$ and $\mathcal{R}$ are intervals.

The binary system correspond to $f(x) = x/2$, where $\mathcal{D} = [0,2)$ and $\mathcal{R} = [0,1)$. In the case of the continued functions, $f(x) = x^{-1}$, with $\mathcal{D} = [1,\infty)$ and $\mathcal{R} = (0,1]$ and the additional condition that the iterations should stop the first time that $x_n = 0$ for some n.

In Chapter 6 we will analyze the use of binary representation in A/D conversion, as well as the so called β-expansion (another particular case of f-expansions), their strenghts and potential weaknesses. In Chapter 7 we will introduce the $\beta\alpha$-expansions, a variation of β-expansions that overcomes some of limitations of the latter. Finally, in Chapter 8 we will analyze some of the ergodic properties of the dynamical system introduced by the $\beta\alpha$-expasions.

CHAPTER VI

THE β-ENCODER

The constant need to improve current strategies and technologies to encode images, video or audio in order to obtain a better quality with a lesser cost on the existing resources makes analog-to-digital (A/D) conversion a dynamic area of research.

One of the most basic problems within this area consist in the representation of a signal x coming from a continuous media using a string of characters coming from a finite alphabet, a *digital* expression.

6.1 Imperfect Quantizers

Probably the better known scheme to obtain a digital expresion of a signal is using a binary expansion. On this scheme, a finite or infinite string of binary digits is obtained to represent $x \in [0, 1)$ in the following way

$$\begin{aligned} x_0 &= x \\ b_n &= Q(2x_{n-1}) \\ x_n &= 2x_{n-1} - b_n \end{aligned} \tag{6.1}$$

where the quantization function Q is given by

$$Q(t) = \begin{cases} 0 & \text{if } t < 1, \\ 1 & \text{otherwise.} \end{cases} \tag{6.2}$$

This leads to a perfect reconstruction method given by

$$x = \sum_{n=1}^{\infty} b_n 2^{-n}. \tag{6.3}$$

Furthermore the accuracy improves exponentially as more bits are used.

$$\left| x - \sum_{n=1}^{N} b_n 2^{-n} \right| < 2^{-N}.$$

One important drawback on this scheme is the fact that such representation is unique for almost all x, in the sence that if $\{b_n\}$ is the binary representation of x and $b_{n_k} \neq \tilde{b}_{n_k}$ for a collection $\{n_k\}_k$, then

$$\sum_{n=1}^{\infty} b_n 2^{-n} \neq \sum_{n=1}^{\infty} \tilde{b}_n 2^{-n}.$$

The importance of such drawback comes from the fact that in a practical implementation, the scheme has a quantization threshold. In a practical set up, the quantizer given in (6.2) is unattainable with infinite precision, and instead the available quantizer Q_f has some indetermination

$$Q_f(t) = \begin{cases} 0 & \text{if } t < \nu_1, \\ 0 \text{ or } 1 & \text{if } \nu_1 \leq t \leq \nu_2, \\ 1 & \text{otherwise,} \end{cases} \tag{6.4}$$

where the values of ν_1 and ν_2 are unknown, though, they lie within a known range. If the source of the signal is assumed to be uniformly distributed in $[0, 1]$, and $\nu_1 < \nu_2$, then the scheme would fail to produce a correct encoding with probability 1. Furthermore, if during the encoding, a quantization error is made on the n-th iteration, then, the reconstruction error is at least $2^{-n} \left| x_n - \frac{1}{2} \right|$.

The β-quantization scheme was recently introduced in [10] and studied in more detail in [12] and [13]. Here we introduce a variant of the $\beta-$expansion quantization scheme, where the introduction of a secondary parameter α improves the robustness of the scheme without sacrifizing the exponential acuracy reached by it.

6.2 β-Expansions

The so called β-encoder is based on the β-expansion introduced originally in [37] as a particular case of an f-expansion. There, Renyi introduced the posibility to use non-integral bases to represent real numbers. Then, given a non integer $\beta > 1$, if

$0 < x < 1$ one can express any $0 \leq x \leq 1$ as

$$x = \sum_{n=1}^{\infty} b_n \beta^{-n}. \tag{6.5}$$

The *digits* b_n can be chosen recursively by

$$\begin{aligned} x_0 &= x \\ b_n &= \lfloor \beta x_{n-1} \rfloor \\ x_n &= \beta x_{n-1} - b_n \end{aligned} \tag{6.6}$$

where $\lfloor \cdot \rfloor$ denotes the integer part. At each step, $0 \leq b_i \leq \lfloor \beta \rfloor$. There is an immediate gain using this representation instead of the representation obtained by an integral base: There are many possible choices of $\{b_n\}$ that still yield a valid reconstruction for x with the expansion (6.5). In fact it is proved (see Sidrov [39]) that for almost every $x \in (0, 1)$ there are uncontably many such representations.

Furthermore, in [34], Parry proved the following theorem.

Theorem 6.1 *Let $1 < \beta$ and let $T(x) = \beta x(\mod 1)$. Consider the function*

$$h(x) = \sum_{x < T^n(1)} \frac{1}{\beta^n},$$

where $T^0(1) = 1$, and consider the measure ν on $[0, 1]$ given by

$$\nu(E) = \int_E h(x) d\mu$$

Then ν is a finite positive T-invariant measure that is ergodic with respect to T.

Note that if $\{x_n\}_{n \geq 0}$ are defined as in (6.6), and T as in the theorem above, then $x_{n+1} = T(x_n)$. This function, often denoted as T_β, is generally called *the β-transformation*, and it has been widely studied by Renyi [37], Parry [34, 35], Kopf [27] among others.

6.3 Robustness of the β-Encoder

Even with the vast literature on the β-transformation dating back to the late 1950s, to the best of the knowledge of the author, it was not until 2002 when Daubechies, DeVore, Güntürk and Vaishampayan saw the advantages it could offer for A/D conversion (See [10, 12]). They introduced the idea of a β-encoder, which enables one to overcome the imprecision of the quantizers, i.e. the *flaky quantizer* problem, by introducing redundancy in the representation of the signal.

Using the non-uniqueness (redundancy) of β-expansions, they showed that it is possible to implement a quantizer with an unknown and possibly fluctuating threshold (although a such threshold has to be contained within a certain range) that would yield a perfect reconstruction of the original input x. The following theorem is proved in [12]:

Theorem 6.2 *Let $1 < \beta < 2$, $0 \leq x < 1$, $1 \leq \nu_0 < \nu_1 \leq (\beta - 1)^{-1}$ and Q_f as defined in (6.4), and define x_n^f, b_n^f by the algorithm*

$$\begin{aligned} x_0^f &= x, \\ b_n^f &= Q_f(\beta x_{n-1}^f), \\ x_n^f &= \beta x_{n-1}^f - b_n^f. \end{aligned} \tag{6.7}$$

Then, for all $N \in \mathbb{N}$

$$0 \leq x - \sum_{n=1}^{N} b_n^f \beta^{-n} \leq \nu_1 \beta^{-N}.$$

Note that $\nu_1 \geq 1$. This means that even though the β-encoder allows certain imprecision on the quantizer, it does not allow the quantizer to err upward, i.e. reading off a 0 as a 1. The scheme would fail if this occurs. In Figure 1 one can appreciate how the rages where $b_n = 0$ and $b_n = 1$ intersect, but if $x_n < 1$, then the scheme fails if one obtains an output $b_n = 1$.

To overcome this problem we consider an alternative. We introduce the $\beta\alpha$ encoder as a variation of the β-encoder, which allows for precise reconstruction in the case of

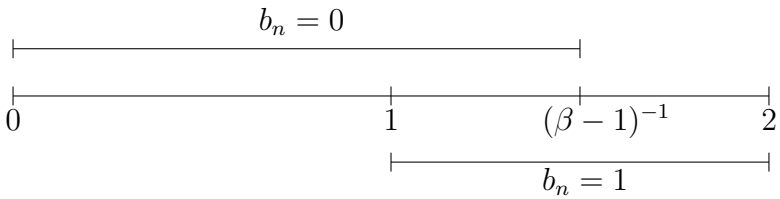

Figure 1: Shown in the image, the rages for x_n producing respectively $b_n = 0$ and $b_n = 1$ for $\beta = 5/3$, where a stable reconstruction is possible.

$\nu_1 < 1$.

CHAPTER VII

$\beta\alpha$-ENCODER

As it has been already discussed, a β-expansion of a real number $x \in [0,1]$ is any collection of *digits* $\{b_n\}_{n\in\mathbb{N}}$ such that

$$x = \sum_{n\in\mathbb{N}} b_n \beta^{-n}.$$

Such expression is far from unique. A very intuitive way to obtain such a collection of digits is described by (6.6), and thus we will call this specific β-expansion of x as its *canonical expansion*. In this chapter we will analyze another way to obtain β-expansions, and will seize on the properties of this alternative method to obtain a stable scalar quantization scheme where the implementation can be given with some freedom unattained by the β-Encoder.

7.1 A Non-Canonical β-Expansion

We will introduce a non-canonical β-expansion, that we will call a $\beta\alpha$-expansion. This one is similar to the β-expansion in that it still uses a possibly non-integer β as the base. However, unlike in the β-expansion the digits b_n are obtained at each stage using an amplification factor α instead of β. More precisely, for any $0 \leq x < 1$ we set $x_0 = x$ and obtain b_n, x_n for $n \geq 1$ using the following scheme:

$$\begin{aligned} b_n &= \lfloor \alpha x_{n-1} \rfloor, \\ x_n &= \beta x_{n-1} - b_n. \end{aligned} \tag{7.1}$$

Observe that $x_{n-1} = \beta^{-1}(x_n + b_n)$ for every $n \geq 1$, and therefore, nesting this identity we obtain for any $N \in \mathbb{N}$ the expression

$$x = \beta^{-N} x_N + \sum_{n=1}^{N} b_n \beta^{-n},$$

or equivalently,

$$x - \sum_{n=1}^{N} b_n \beta^{-n} = \beta^{-N} x_N. \tag{7.2}$$

In order for perfect reconstruction $x = \sum_{n=1}^{\infty} b_n \beta^{-n}$ we will need $\beta^{-N} x_N \to 0$, preferably at an exponential rate. To make it happen, let $\{t\}$ denote the *fractional part* of t. Then $x = \lfloor x \rfloor + \{x\}$, and

$$\begin{aligned} x_N &= \beta x_{N-1} - b_N \\ &= \beta x_{N-1} - \lfloor \alpha x_{N-1} \rfloor \\ &= \beta x_{N-1} - \alpha x_{N-1} + \{\alpha x_{N-1}\} \\ &= (\beta - \alpha) x_{N-1} + \{\alpha x_{N-1}\} \\ &= (\beta - \alpha)^N x + \sum_{n=1}^{N} \{\alpha x_{n-1}\}. \end{aligned}$$

Since $0 \le \{t\} < 1$, it follows that $(\beta - \alpha)^N x \le x_N < (\beta - \alpha)^N x + N$, and

$$\beta^{-N}(\beta - \alpha)^N x \le \beta^{-N} x_N < \beta^{-N}\big((\beta - \alpha)^N x + N\big).$$

Thus if we set $\beta > 1$ and $0 < \alpha \le \beta$ we will ensure a perfect reconstruction with exponential rate convergence. Furthermore, all $x_n \ge 0$ and hence all digits b_n are nonnegative. For quantization applications, the magnitude of x_n matters because it determines the magnitude of b_n. Since these digits b_n must come from a finite alphabet we shall require that x_n be bounded. A necessary condition is $\beta - \alpha < 1$. In what follows we focus on the case $0 \le \beta - \alpha < 1$. We ask the following questions: Are $\{b_n\}$ bounded, and if so, what is the upper bound?

Lemma 7.1 *Let $1 < \beta$, $\alpha \le \beta$ and $\beta - \alpha < 1$. Define $T(x) = \beta x - \lfloor \alpha x \rfloor$ and set $\omega = [1 - (\beta - \alpha)]^{-1}$. Let $K = \lceil \omega(\beta - 1) \rceil$ where $\lceil y \rceil$ denotes the least integer greater than or equal to y. Then the fixed points of T are $\{k(\beta - 1)^{-1} : 0 \le k < K\}$.*

Proof. First we notice that $T(x) \ge (\beta - \alpha)x$ implies that $T(x) > x$ if $x < 0$. So T cannot have a negative fixed point. Now, notice that if $T(x) = x$ then $\beta x - k = x$

where $k = \lfloor \alpha x \rfloor$. Thus $x = k(\beta - 1)^{-1}$. So all fixed points must be in the form of $x = k(\beta-1)^{-1}$ for some integer $k \geq 0$. We shall determine which of these k's actually yield fixed points. To do so, let $x = k(\beta-1)^{-1}$ be a fixed point. Then $\beta x - \lfloor \alpha x \rfloor = x$. It follows that $\lfloor \alpha x \rfloor = (\beta - 1)x = k$.

Now $\lfloor \alpha x \rfloor = \alpha x - \{\alpha x\}$. So we have $\alpha x - k = \{\alpha x\}$. Note that

$$\alpha x - k = \frac{\alpha k}{\beta - 1} - k = \frac{(1 - \beta + \alpha)k}{\beta - 1} = \frac{k}{\omega(\beta - 1)}.$$

Thus we have $k[\omega(\beta-1)]^{-1} = \{\alpha x\} < 1$, which yields $k < \omega(\beta-1)$ or equivalently, $k < K$. Conversely, if $0 \leq k < K$ and $x = \frac{k}{\beta-1}$ the above calculations can be reversed to show that x is a fixed point.

■

Proposition 7.2 *Let $1 < \alpha \leq \beta$ and $\beta - \alpha < 1$. Define $T(x) = \beta x - \lfloor \alpha x \rfloor$ and set*

$$M = \left\lceil \frac{\alpha(\beta - 1)}{\beta[1 - (\beta - \alpha)]} \right\rceil, \tag{7.3}$$

where $\lceil t \rceil$ denotes the least integer greater than or equal to t. Set $\tau = M(\beta\alpha^{-1}-1)+1$. For any $0 \leq x \leq \tau$ we have $0 \leq T^n(x) < \tau$ for all $n \geq 1$.

Proof. Note that $T(x) = (\beta-\alpha)x+\{\alpha x\}$ we have $T(x) \geq 0$ for $x \geq 0$. Furthermore, as $\alpha < \beta$, then, $\alpha\beta^{-1}\omega(\beta - 1) < \omega(\beta - 1)$, where $\omega = [1 - (\beta - \alpha)]^{-1}$, thus $M \leq \lceil \omega(\beta - 1) \rceil$. Hence $(M - 1)(\beta - 1)^{-1}$ is a fixed point. First, for $x < M\alpha^{-1}$,

$$T(x) < (\beta - \alpha)x + 1 < (\beta - \alpha)M\alpha^{-1} + 1 = \tau.$$

If $M < \lceil \omega(\beta - 1) \rceil$, then, by Lemma 7.1, $M(\beta - 1)^{-1}$ would be also a fixed point, besides

$$\frac{\alpha(\beta - 1)}{\beta[1 - (\beta - \alpha)]} < M \Rightarrow M(\beta\alpha^{-1} - 1) + 1 \leq \frac{M}{\beta - 1}$$

and therefore, for every $x \in [M\alpha^{-1}, \tau)$, $T(x) \leq x$, thus $T(x) < \tau$.

If $M = \lceil \omega(\beta - 1) \rceil$, then, $(M - 1)(\beta - 1)^{-1}$ is the largest fixed point of T, and thus, for every $x > M\alpha^{-1}$, $T(x) < x$.

As it was just proven, $0 \leq x \leq \tau$ implies $0 \leq T(x) \leq \tau$. The iteration step is trivial.

■

Proposition 7.3 *Let $1 < \alpha \leq \beta$ and $\beta - \alpha < 1$. Set M and τ as in Proposition 7.2. For any $x \in [0, \tau)$ define $x_0 = x$ and x_n, b_n for $n \geq 1$ by $b_n = \lfloor \alpha x_{n-1} \rfloor$ and $x_n = x_{n-1} - b_n$. Then $0 \leq x_n < \tau$ and $b_n \in \{0, 1, \ldots, M\}$.*

Proof. Notice that $x_n = T^n(x_0)$. By Proposition 7.2 we have $0 < x_n < \tau$. Also, $b_n = \lfloor \alpha x_{n-1} \rfloor$, then, it is enough to prove that $\tau\alpha \leq M + 1$. Now, as $\alpha < \beta$, then $\alpha(\beta - 1) > \beta(\alpha - 1)$, and thus

$$\frac{\alpha - 1}{1 - (\beta - \alpha)} < \frac{\alpha(\beta - 1)}{\beta[1 - (\beta - \alpha)]} \leq M,$$

hence $\alpha - 1 < M[1 - (\beta - \alpha)] \Rightarrow \tau = M(\beta - \alpha) + \alpha < (M + 1)\alpha^{-1} \Rightarrow b_n \leq M$.

■

We shall point out that the map T is a piece-wise linear so the dynamical system given by T has an invariant measure, see Lasota and York [28] (see also [29]). However there are a few questions that remain to be answered. For example, what are the invariant sets, and what more can we say about the invariant measures? These are interesting mathematical questions that are relevant to the theme of this study. The invariant sets will determine the number of digits in the quantization schemes. It is possible that fewer digits than what we have shown here will be enough. The next question we face is how robust is this scheme, that is, how tolerant is such a scheme to quantizer imperfections. This question will be answered in the next section.

7.2 The $\beta\alpha$-Encoder vs. the β-Encoder

The $\beta\alpha$-expansion described in the previous section leads naturally to a quantization scheme assuming a perfect quantizer. When a *flaky* quantizer is used, it can still yield a perfect reconstruction with suitable chioces of the parameters.

Bounding ourselves to the conditions $1 < \alpha \leq \beta$, $\beta - \alpha < 1$ and the quantization scheme $x_0 = x$, $b_n = Q_f(\alpha x_{n-1})$ and $x_n = \beta x_{n-1} - b_n$ where set of all possible outputs of Q_f is $\{0, 1, \ldots, B-1\}$ for some integer D, our main concern is to keep x_N bounded for every N.

A first natural question is: what bounds should x_N have to preserve a robust scheme? Note that if $x_0 < 0$, then, $x_1 = \beta x_0 - Q_f(\alpha x_0) \leq \beta x_0$, and thus $x_N \leq \beta^{-N} x_0$, making the sequence diverge to negative infinity. From here that x_n should be positive. On the other hand, note that if x_N is bounded for all N, then, by (7.2), one has that

$$x_N = \lim_{K \to \infty} \sum_{n=1}^{K} b_{N+n} \beta^{-n} \leq \sum_{n=1}^{\infty} (B-1)\beta^{-n} = \frac{B-1}{\beta - 1}$$

and thus, $x_N \leq (B-1)(\beta-1)^{-1}$ for all N. We prove the following theorem.

Theorem 7.4 *Let B be a given positive integer, $1 < \beta < B$, $0 \leq x < 1$, $0 < \beta - \alpha < 1$, let $\mu = (B-1)(\beta-1)^{-1}$ and let Q_f be defined such that $Q_f(t) \in \{0, 1, \ldots, B-1\}$, where $Q_f(t) = j \Rightarrow t \in [j\alpha\beta^{-1}, \alpha\beta^{-1}(\mu + j)]$, and define x_n^f, b_n^f by the algorithm*

$$\begin{aligned} x_0^f &= x, \\ b_n^f &= Q_f(\alpha x_{n-1}^f), \\ x_n^f &= \beta x_{n-1}^f - b_n^f. \end{aligned} \tag{7.4}$$

Then, for all $n \in \mathbb{N}$, $0 \leq x_n^f \leq \mu$ and

$$0 \leq x - \sum_{n=1}^{N} b_n^f \beta^{-n} \leq \mu \beta^{-N}.$$

Proof. Note that as $\beta < B$ then $\mu > 1$, and therefore $x_0^f < \mu$. Note that as $x_n^f = \beta x_{n-1}^f - b_n^f$, then (7.2) is valid regardless of how b_n^f are chosen, therefore it is sufficient to prove that $0 \leq x_n^f \leq \mu$. Let's now consider the respective subintervals.

Assume that $j\beta^{-1} \leq x_n^f \leq \beta^{-1}(\mu + j)$ and $Q(\alpha x_n^f) = j$, then $0 = \beta(j\beta^{-1}) - j \leq x_{n+1}^f \leq \beta[\beta^{-1}(\mu + j)] - j = \mu$.

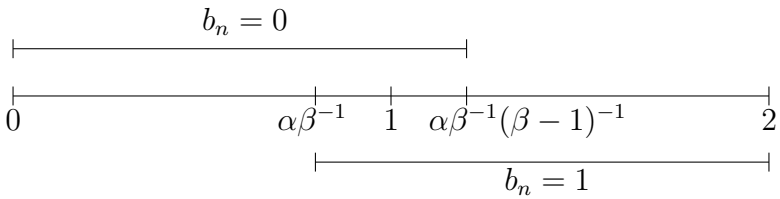

Figure 2: Shown in the image, the rages for x_n producing respectively $b_n = 0$ and $b_n = 1$ for $\beta = 5/3$ and $\alpha = 4/3$, where a stable one-bit reconstruction is possible.

An important remark is that as $\mu > 1$, then $\beta^{-1}(j+1) < \beta^{-1}(\mu+j)$, therefore, the collection of intervals $\{j\beta^{-1}, \beta^{-1}(\mu+j) | 0 \leq j \leq B-1\}$ actually covers the interval $[0, \mu]$.

■

A rephrasing of this theorem for the 1-bit $\beta\alpha$ that resembles Theorem 6.2 would be the following.

Theorem 7.5 *Let $1 < \beta < 2$, $0 \leq x < 1$, $\beta(\beta-1) < \alpha < \beta$, $\alpha\beta^{-1} \leq \nu_0 < \nu_1 \leq \alpha\beta^{-1}(\beta-1)^{-1}$ and Q_f as defined in (6.4), and define x_n^f, b_n^f by the algorithm*

$$\begin{aligned} x_0^f &= x, \\ b_n^f &= Q_f(\alpha x_{n-1}^f), \\ x_n^f &= \beta x_{n-1}^f - b_n^f \end{aligned} \tag{7.5}$$

Then, for all $N \in \mathbb{N}$

$$0 \leq x - \sum_{n=1}^{N} b_n^f \beta^{-n} \leq (\beta-1)^{-1}\beta^{-N}.$$

In Figure 2 one can appreciate how the rages where $b_n = 0$ and $b_n = 1$ intersect, allowing quantizer erros both by excess and by defect, for a one-bit quantization case.

7.3 Imprecise α-Multiplication

An imperfect quantizer is not the only problem that can arise in a real application. The multiplication via analog circuits can potentially be another source of inaccuracy. Thus, by performing two multiplications in the $\beta\alpha$-encoder we introduce an extra source for potential errors. In this section, we show that the α-multiplication in the $\beta\alpha$-encoder does not have to be very accurate. We prove the following theorem.

Theorem 7.6 *Let B be a given positive integer, $1 < \beta < B$, $0 \leq x < 1$. Let $(\alpha_n)_{n\in\mathbb{N}}$ be a sequence such that*

$$\max\left(0, \frac{\beta[(\beta-1) - M(1-c)]}{(\beta-1)(M+1)}\right) < \beta - \alpha_n \leq c < 1.$$

Let $\mu = (B-1)(\beta-1)^{-1}$ and let Q_f be defined such that $Q(t) \in \{0, 1, \ldots, M\}$, $Q_f(t) = j \Rightarrow t \in [j(\sup \alpha_k)\beta^{-1}, (\inf \alpha_k)\beta^{-1}(\mu + j)]$ and $Q_f(t) = B-1$ if $t \geq (\inf \alpha_k)\mu$. Define x_n^f, b_n^f by the algorithm

$$\begin{aligned} x_0^f &= x, \\ b_n^f &= Q_f(\alpha_n x_{n-1}^f), \\ x_n^f &= \beta x_{n-1}^f - b_n^f. \end{aligned} \tag{7.6}$$

Then, for all $n \in \mathbb{N}$, $0 \leq x_n^f \leq \mu$ and

$$0 \leq x - \sum_{n=1}^{N} b_n^f \beta^{-n} \leq \mu\beta^{-N}.$$

Proof. Note that trivally, for any integer n and $0 \leq j < B$ one has that

$$[j(\sup \alpha_k)\beta^{-1}, (\inf \alpha_k)\beta^{-1}(\mu + j)] \subseteq [j\alpha_n\beta^{-1}, \alpha_n\beta^{-1}(\mu + j)].$$

Then, the only difference from the proof of Theorem 7.4 is to prove that the set of intervals $I_j = [j(\sup \alpha_k), (\inf \alpha_k)(\mu + j)]$ cover $[0, (\inf \alpha_k)\mu\beta]$. As $0 \in I_0$ and $(\inf \alpha_k)\mu \in I_{B-1}$, and as the lower endpoints (as well as the upper endpoints) are in

increasing order, then the only thing left to prove is that $I_j \cap I_{j+1} \neq \emptyset$. For this, it suffies that $(j+1)\sup \alpha_k \leq (\mu + j)\inf \alpha_k$, and

$$\frac{\mu + j}{j+1}\inf \alpha_k \geq \frac{\mu + (B-1)}{B}(\beta - c) = \frac{(B-1)\beta(\beta - c)}{B(\beta - 1)} \geq \sup \alpha_k$$

by hypothesis.

■

CHAPTER VIII

ERGODIC PROPERTIES OF THE $\beta\alpha$-ENCODER

In the previous chapter we discussed the $\beta\alpha$-Encoder. The scheme defined in (7.1) gives rise to the dynamical system $x_{n+1} = T(x_n)$, where $T(x) = \beta x - \lfloor \alpha x \rfloor$. Beyond its practical applications, this system is tremendously interesting from a mathematical point of view, specifically, the ergodicity of T, on which we will focus on this chapter.

8.1 Some Invariant Sets for T

We will use the notation introduced in Lemma 7.1, this is, $1 < \beta$, $\alpha \leq \beta$ and $\beta - \alpha < 1$. $T(x) = \beta x - \lfloor \alpha x \rfloor$, $\omega = [1 - (\beta - \alpha)]^{-1}$and $K = \lceil \omega(\beta - 1) \rceil$ where $\lceil y \rceil$.

For simplicity we will introduce the following additional notation. For $0 < k \leq K$,

$$\lambda_k = \frac{k}{(\beta - 1)}, \quad \xi_k = k\left(\frac{\beta - \alpha}{\alpha}\right) + 1, \quad \zeta_k = k\left(\frac{\beta - \alpha}{\alpha}\right). \tag{8.1}$$

By Lemma 7.1, for $k < K$, λ_k are all the fixed points of T other than 0. For $1 \leq k \leq K$, ξ_k the upper extreme of the discontinuity jumps of T as defined in Lemma 7.1, and ζ_k the lower extremes, for those discontinuities immediately before and immediately after the fixed points.

Proposition 8.1 *If i and j are indices such that $\lambda_{i-1} \leq \zeta_i$ and $\xi_j \leq \lambda_j$, then $\zeta_i < \xi_j$, and $T(x) = \beta x - \lfloor \alpha x \rfloor = (\beta - \alpha)x + \{\alpha x\}$. Consider $\Psi = [\zeta_i, \xi_j]$, then Ψ is invariant, i.e. $\overline{T(\Psi)} = \Psi$.*

Proof. Note that

$$\begin{aligned}\lambda_{i-1} &\leq \zeta_i \\ &< \zeta_i + 1 - (\beta - \alpha) \\ &= \zeta_{i-1} + 1 \\ &= \xi_{i-1},\end{aligned}$$

therefore $j > i - 1$, i.e. $i \leq j$, and therefore $\zeta_i < \xi_j$.

Now, for any $0 \leq i \leq n$, $\zeta_i \leq i\alpha^{-1} < (i+1)\alpha^{-1} < \xi_{i+1}$, and also

$$T\left([i\alpha^{-1}, (i+1)\alpha^{-1}]\right) = [\zeta_i, \xi_{i+1}),$$

therefore, regardless of how i and j are chosen, as long as $0 \leq i \leq j \leq n$ we would have $\overline{T(\Psi)} \supseteq \Psi$. Now, as $\zeta_i < \zeta_{i+1}$ and $\xi_i < \xi_{i+1}$ for any i, we only have left to prove that for the i and j described in the statement, $T([\zeta_i, i\alpha^{-1}, \xi_j]) \subseteq \Psi$ and $T([j\alpha^{-1}, \xi_j]) \subseteq \Psi$.

Note that

$$\sup_{x < i\alpha^{-1}} T(x) = \xi_i \leq \xi_j.$$

Also, as $\lambda_{i-1} \leq \zeta_i \leq i\alpha^{-1}$, then, if one takes $\zeta_i \leq \tilde{x} < i\alpha^{-1}$, then $T(\zeta_i) \leq T(\tilde{x}) < \xi_i$. Note that $T(x) - x$ is continuous and increasing in such interval, and as $\lambda_{i-1} \leq \tilde{x}$ and λ_{i-1} is a fixed point, one has that $T(\zeta_i) > \zeta_i$, therefore if $\zeta_i \leq \tilde{x} \leq i\alpha^{-1}$ then $T(\tilde{x}) \in \Psi$. A basically analogous argument proves that $T([j\alpha^{-1}, \xi_j]) \subseteq \Psi$. Note that by definition, Ψ is a closed set and we have $T(\Psi) \subseteq \Psi \subseteq \overline{T(\Psi)}$, therefore $\overline{T(\Psi)} = \Psi$.

■

From the sets described by Proposition 8.1, the smallest of them, this is $[\zeta_m, \xi_n]$ where $m = \max\{i : \lambda_{i-1} \leq \zeta_i\}$, and $n = \min\{i : \xi_i \leq \lambda_i\}$, will be call $\Omega_{\beta\alpha}$ or Ω where the choice of α and β is clear by context.

8.2 Li-Yorke Theorem and Ergodicity of T for $K = 1$

As it has already been proved, given α and β with $\beta > 1$, $\alpha \leq \beta$ and $\beta - \alpha < 1$, and Ω the smallest of the sets described by Proposition 8.1, we have proven that $\overline{T(\Omega)} = \Omega$. Note that T is a piecewise monotone C^∞ function. Furthermore, if we call Ω^* the set where both T and dT/dx are continuous,

$$\inf_{x \in \Omega^*} \left| \frac{d}{dx} T(x) \right| > 1.$$

In [28], Lasota and Yorke proved that under these conditions there exist at least one non-negative function f of bounded variation such that the measure μ with $d\mu = f dm$ (where m is the Lebesgue measure) is invariant under T, in the sense that

$$\mu(E) = \int_E f dm = \int_{T^{-1}(E)} f dm = \mu\left(T^{-1}(E)\right).$$

In a more general setting, let's $\tau : I \to I$ is a piecewise continuous and piecewise twice continuous differentiable. Call I^* the set of points where $d\tau/dx$ exists, and let

$$\inf_{x \in I^*} \left| \frac{d}{dx} \tau(x) \right| > 1. \tag{8.2}$$

We will refer to the points of $I - I^* = \{x_1, \ldots, x_k\}$ as the *points of discontinuity*. For $x \in I$, let $\Lambda(x)$ be the set of limit points of $\tau^n(x)$, that is

$$\Lambda(x) = \bigcap_{N=1}^{\infty} \overline{\{\tau^n(x)\}_{n=N}^{\infty}}. \tag{8.3}$$

An important property of this set is that it is a fixed set of τ. This means, $\tau(\Lambda(x)) = \Lambda(x)$. As said before, Lasota-Yorke proves that there are densities invariant under τ. Let $\mathcal{F}$ be the set of $f \in L^1(I)$, such that f is an invariant density under τ. In [30], Li and Yorke proved the following theorem.

Theorem 8.2 *Let $\tau : I \to I$ be a piecewise continuous and twice continuous differentiable interval map satisfying (8.2). Then, there exists a finite collection of sets $L_1, L_2, \ldots, L_n$ and a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that*

(1) Each L_i is a finite union of closed intervals,

(2) $L_i \cap L_j$ contains at most a finite number of points when $i \neq j$;

(3) each L_i contains at least one point of discontinuity x_j, $1 \leq j \leq k$ on its interior; hence $n \leq k$;

(4) $f_i(x) = 0$ for $x \notin L_i$ and $f(x) > 0$ for almost all x in L_i;

(5) $\int_{L_i} f_i(x)dx = 1$ for $1 \leq i \leq n$;

(6) if $g \in \mathcal{F}$ satisfy both (4) and (5), then $g = f_i$ almost everywhere;

(7) every $f \in \mathcal{F}$ can be written as $f = \sum_{i=1}^{n} a_i f_i$ for suitable chosen $\{a_i\}_{i=1}^{n}$;

(8) for almost every $x \in I$ there is an index i such that $\Lambda(x) = L_i$.

It has been discussed, if $1 < \beta < 2$ and $\beta(\beta - 1) < \alpha < \beta$, $T(x) = \beta x - \lfloor \alpha x \rfloor$ generates a one bit quantization for every $x \in [0, 1]$. Now, by Proposition 8.1, T restricted to $\Omega = [\alpha^{-1}\beta - 1, \alpha^{-1}\beta]$. This interval contains a unique point of discontinuity (both of T and its first derivative), and therefore, by Theorem 8.2, up to normalization, there exists a unique non-negative function $f \in L^1$ that generates measure μ that is invariant under T. As this measure is unique, T is ergodic with respect to μ.

Indeed, the density of this function can be given in a closed form. In [35], Parry proved that if τ is a linear transformation $\mod 1$, (i.e. $\tau(x) = bx + a \mod 1$ for real numbers a and b), then the function

$$h(x) = \sum_{x < \tau^n(1)} \frac{1}{\beta^n} - \sum_{x < \tau^n(0)} \frac{1}{\beta^n},$$

where $\tau^0(x) = x$ by definition, is the density of a, potentially signed (but not null) measure.

Note that if α and β are the parameters of a one bit quantization scheme, then, we can define $b = \beta$, $a = (\beta - 1)(\beta - \alpha)\alpha^{-1}$, and $f(x) = x - (\beta - \alpha)\alpha^{-1}$, then $T(x) = f^{-1}(\tau(f(x)))$. By Parry's theorem, we know then that the function

$$g(x) = \sum_{x<T^n(\beta\alpha^{-1})} \frac{1}{\beta^n} - \sum_{x<T^n((\beta-\alpha)\alpha^{-1})} \frac{1}{\beta^n}$$

is the density of an absolutely continuous signed measure on Ω, and by Li-Yorke's Theorem, such measure is necessarily unique up to a re-scalling factor, therefore, the density of the unique normalized invariant measure under T is

$$f(x) = \left(\sum_{n=0}^{\infty} \frac{T^n(\beta\alpha^{-1}) - T^n((\beta-\alpha)\alpha^{-1})}{\beta^n} \right)^{-1} \left(\sum_{x<T^n(\beta\alpha^{-1})} \frac{1}{\beta^n} - \sum_{x<T^n((\beta-\alpha)\alpha^{-1})} \frac{1}{\beta^n} \right).$$

8.3 Ergodicity of T for $K > 1$

We have already proved that, if $K = 1$, then Ω is a fixed set that pocesses an absolutely continuous measure that is invariant under T. The natural question at this point is: If $K > 1$, is there a measure μ, that is absolutely continuous with respect to the Lebesgue measure and ergodic with respect to T? The answer in general is no.

In Figure 3 we can appreciate the graph of the $\beta\alpha$ encoding fucntion with $\alpha = 3/4$ and $\beta = 3/2$. Under this conditions, $K = 2$ and $\Omega = [1, 3]$. Neverhteless, T has two different invariant sets, namely $[1, 2]$ and $[2, 3]$. Therefore, by Li-Yorke, there is a measure with respect invariant under T for each of these intervals, each one independent on the other, and therefore, for this election of parameters, T is not ergodic.

Indeed, by Li-York, there is one invariant measure for each of those two sets, and their respective densities can be computed in a closed form. Notice that in this case, λ_1, ξ_1 and ζ_2, as defined in (8.1), are all equal. Our simulations suggesst that if for every index i, the three numbers λ_i, ξ_i and ζ_{i+1} are all different, then the system is indeed ergodic, leading us to conjecture the following.

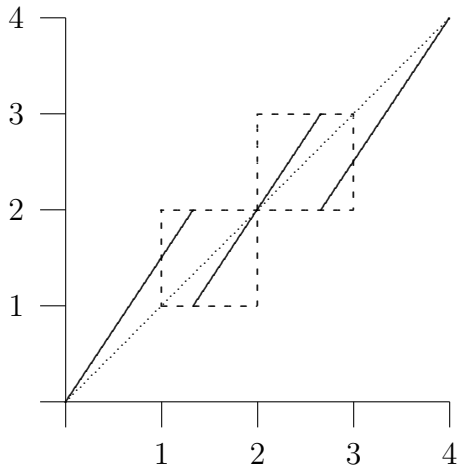

Figure 3: Graph of T withm for $\beta = 3/2$ and $\alpha = 3/4$. Example of T non-ergodic.

Conjecture 8.1 *Let α and β be two positive real numbers such that $1 < \beta$, $\alpha < \beta$, $\beta - \alpha < 1$. Let $\omega = [1 - (\beta - \alpha)]^{-1}$, $K = \lceil \omega(\beta - 1) \rceil > 1$. Furthermore, asume that neither $\beta^{-1}\alpha\omega$ nor $\beta^{-1}(\beta - 1)\alpha\omega$ are integers and $\beta \neq 2\alpha$. Call $M = \lfloor \beta^{-1}\alpha\omega \rfloor$ and $N = \lceil \beta^{-1}(\beta - 1)\alpha\omega \rceil$, and let $\Omega = [M\alpha^{-1}(\beta - \alpha), N\alpha^{-1}(\beta - \alpha) + 1]$. Then, $\overline{T(\Omega)} = \Omega$ and there exists, up to normalization, a unique measure μ absolutely continuous with respect to the Lebesgue measure that is invariant under T. Furthermore, if f is the density of such function, then $\overline{\operatorname{supp} f} = \Omega$.*

APPENDIX A

SAMPLING THEOREM

The need to convey all the information contained in a function into a discrete set of values produced what is called today the *Sampling Theorem* or *Shannon-Nyquist Theorem*, that was, to the best of the knowledge of the author, first stated and formally proved in [38], by Claude Shannon, although, he does not claim authorship of the result, and writes below the statement: *This is a fact that is common knowledge in the communication art.* In [33], published in 1928, Harry Nyquist clearly implies the same result, although this is not stated nor formally proved.

On Shannon's paper, the statement of this theorem reads

Theorem A.1 *If a function $f(t)$ contains no frequencies higher than W cps*[1]*, it is completely determined by giving its ordinates at a series of points spaced $1/2W$ seconds apart.*

A more modern paraphrasing of the same result would be

Theorem A.2 *If $f(t)$ is a bandlimited function with maximum frequency no higher than Ω, then*

$$f(t) = \sum_{k\in\mathbb{Z}} f\left(\frac{k\pi}{\Omega}\right) \operatorname{sinc}\left(\frac{t\Omega}{\pi} - k\right), \tag{A.1}$$

and thus, $f(t)$ can be fully reconstructed from its samples $f(k\pi\Omega^{-1})$.

The following proof, although not exactly that given in [38], is based in the same ideas.

[1] Cycles per second*, now known as hertz, the SI unit for frequency*

Proof. To simplify notation, it will be assumed that $\Omega = \pi$. This is just a scaling factor. Consider the *Dirac Comb*:

$$\Delta(t) = \sum_{k\in\mathbb{Z}} \delta(t-k) = \sum_{k\in\mathbb{Z}} e^{2k\pi it}$$

where δ is the Dirac delta function. Define the *sampled* function $f_s(t) = f(t)\Delta(t)$. Note that

$$f_s(t) = \sum_{k\in\mathbb{Z}} f(t)e^{2k\pi it} \text{ and thus } \hat{f}_s(\xi) = \sum_{k\in\mathbb{Z}} \hat{f}(\xi - 2\pi k).$$

By definition, $\hat{f}(\xi) = 0$ if $|\xi| > \pi$, therefore, if $\hat{h}(\xi)$ is the characteristic function of the interval $[-\pi, \pi]$, then, $\hat{f}(\xi) = \hat{f}_s(\xi)\hat{h}(\xi)$. Now

$$h(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{it\xi} d\xi = \frac{1}{2\pi}\cdot\frac{e^{i\pi t} - e^{-i\pi t}}{it} = \frac{\sin \pi t}{\pi t} = \operatorname{sinc} t.$$

and thus, we can write

$$f(t) = (f_s * \operatorname{sinc})(t) = \sum_{k\in\mathbb{Z}} f(k)\operatorname{sinc}(t-k).$$

■

The formula in (A.1) is known as the Whittaker-Shannon Interpolation Formula. Although useful, it does not come without drawbacks. As sinc $\notin L^1(\mathbb{R})$, then the right hand side of (A.1) is not, in general, absolutely convergent, and this introduces questions about the proper sumation strategy to apply in practice. A way to get around this problem is to introduce a finer sampling rate. The statement, as stated in Chapter 1, reads as follows.

Theorem A.3 (Sampling Theorem) *Let $f : \mathbb{R} \to \mathbb{R}$ be a bandlimited function such that $\hat{f}$ is supported in $[-\pi, \pi]$. Let $\lambda > 1$, and $\varphi \in L^1(\mathbb{R})$ such that $\hat{\varphi}$ satisfies*

$$\hat{\varphi}(\xi) = \begin{cases} 1 & \text{if } |\xi| \le \pi, \text{ and} \\ 0 & \text{if } |\xi| \ge \lambda\pi. \end{cases} \tag{A.2}$$

Then, the following equality holds in the Cesàro mean for all t.

$$f(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\left(\frac{n}{\lambda}\right)\varphi\left(t - \frac{n}{\lambda}\right). \tag{A.3}$$

The proof of this version of the theorem is a simple modification of previous one. Note that if we call

$$\Delta_\lambda(t) = \sum_{k\in\mathbb{Z}} \delta\left(t - \frac{k}{\lambda}\right) = \sum_{k\in\mathbb{Z}} e^{2k\lambda\pi it}$$

and we modify $f_s(t) = f(t)\Delta_\lambda(t)$, then

$$f_s(t) = \sum_{k\in\mathbb{Z}} f(t)e^{2k\lambda\pi it} \text{ and thus } \hat{f}_s(\xi) = \sum_{k\in\mathbb{Z}} \hat{f}(\xi - 2\lambda\pi k).$$

Under this condition we still have that $\hat{f} = \hat{f}_s\hat{\varphi}$, and therefore

$$f(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\left(\frac{n}{\lambda}\right)\varphi\left(t - \frac{n}{\lambda}\right).$$

The Cesàro mean convergence is given by Parseval's formula (see [26, p.35]) for every t. There is another gain of this approach. Let's consider the space

$$\mathcal{S}(\mathbb{R}) = \left\{ f \in C^\infty(\mathbb{R}) \,\middle|\, \lim_{x\to\infty} x^\alpha \frac{d^n f}{dx^n}(x) = 0,\ \forall \alpha \in \mathbb{R}, n \in \mathbb{Z}, n \geq 0 \right\}.$$

This is called the *Schwartz space*. It is a known fact that the Fourier transform is an isomorphism of $\mathcal{S}(\mathbb{R})$ to itself (see [16, p.74]). With this approach, it is possible to choose $\hat{\varphi}$ in (A.2) to be C^∞, and thus, an element of $\mathcal{S}(\mathbb{R})$, and therefore so would be φ, ensuring the absolute convergence of the right hand side of (A.3).

APPENDIX B

NUMERICAL RESULTS FOR WNH

Here we present data from our computer experiments comparing the ideal **MSE** to the actual **MSE**. We have performed Monte Carlo simulations for several different sets of frames. We also experimented with various distributions for $\mathbf{x} \in \mathbb{R}^d$. As it turns out, we get very similar results for the distributions we used for most of the frames we tried. In the examples shown, the random vectors X are all chosen to be uniformly distributed in $[-5, 5]^d$.

Example B.1 *Let $\{\mathbf{v}_j\}_{j=1}^N$ be the harmonic frame in $\mathbb{R}^2$, with*

$$\mathbf{v}_j = \left[\cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}\right]^T.$$

*This is a tight frame with frame constant $\lambda = \dfrac{N}{2}$. The ideal **MSE** is $\dfrac{\Delta^2}{3N}$ for N odd. Taking $\Delta = \dfrac{1}{2}$, Table 1 displays the actual **MSE**, the ideal **MSE** and the ratio between them. It shows that as N gets larger than 129, the actual **MSE** does not improve, which shows that the WNH is invalid for large Δ.*

Example B.2 *Let $\{\mathbf{v}_j\}_{j=1}^N$ be N independently and randomly generated vectors uniformly distributed on the unit sphere in $\mathbb{R}^4$. Table 2 shows the ratio between the actual **MSE** and the ideal **MSE**, where $\mathbf{MSE}_{ideal} = \frac{\Delta^2}{12}(\sum_{j=1}^d \lambda_j^{-1})$, with $\Delta = 2^{-k}$.*

Example B.3 *Let $\{\mathbf{v}_j\}_{j=0}^{N-1}$ be the harmonic frame in $\mathbb{R}^4$, with*

$$\mathbf{v}_j = \sqrt{\frac{1}{2}}\left[\cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}, \cos\frac{4\pi j}{N}, \sin\frac{4\pi j}{N}\right]^T.$$

*This is a tight frame with frame constant $\lambda = \dfrac{N}{4}$, and the ideal **MSE** is $\dfrac{4\Delta^2}{3N}$. Table 3 shows the ratio between the actual **MSE** and the ideal **MSE** where $\Delta = 2^{-k}$.*

Table 1: The Harmonic frame in $\mathbb{R}^2$

N	Actual **MSE**	Ideal **MSE**	*Ratio*
9	0.00934342	0.00925926	1.009090
17	0.00479521	0.00252525	0.976808
33	0.00246669	0.00490196	0.978223
65	0.00122499	0.00128205	0.955496
129	0.00065858	0.000645995	1.019480
257	0.00057971	0.00032425	1.787810
513	0.00056039	0.00016244	3.449740
1025	0.00052914	0.00008130	6.508450
2049	0.00053895	0.00004067	13.25180
4097	0.00058846	0.00002034	28.93090

Example B.4 *Let* $\{\mathbf{v}_j\}_{j=1}^5$ *be a frame in* $\mathbb{R}^3$*, with the corresponding matrix*

$$F = \begin{pmatrix} 1 & 1 & 1 & 2 & -3 \\ 1 & -1 & -1 & 2 & -3 \\ 1 & 0 & -1 & 2 & -3 \end{pmatrix}$$

Note that the set contains many parallel vectors. The ***MSE*** *under the WNH is* $0.2946\Delta^2$ *and by our result, the ideal* ***MSE*** *is* $0.2959\Delta^2$*. The difference is not significant, as with most of such cases. It is rather intuitive to see that the second part in (4.11) contributes only a small portion of the whole* **MSE***. Table 4 shows the actual* ***MSE****, the ideal* ***MSE****, and the ratio between the actual* ***MSE*** *and the ideal* ***MSE****, where* $\Delta = 2^{-k}$*.*

Table 2: The randomly generated frame in $\mathbb{R}^4$

k/N	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
k= 0	1.581960	2.232260	3.697160	6.497800	12.20670
k= 1	1.076590	1.130510	1.397840	1.649530	2.480920
k= 2	1.003680	0.995214	1.008370	1.033280	1.196680
k= 3	0.967138	0.990876	0.999648	0.981633	1.010090
k= 4	0.989295	1.009840	1.032110	1.002630	1.002260
k= 5	1.011720	1.035590	1.020870	1.002350	1.022250
k= 6	0.978712	1.006760	0.992207	1.001490	0.979342
k= 7	0.997524	1.017840	0.995852	0.972120	0.976273
k= 8	0.998725	1.011380	1.040270	0.978204	0.973284
k= 9	0.982450	1.038580	0.994463	1.021580	1.037800
k=10	0.993099	1.002340	1.009930	1.009870	0.974017
k=11	0.981428	0.998280	0.975881	1.049010	1.009570

Table 3: The Harmonic frame in $\mathbb{R}^4$

k/N	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
k= 0	0.997218	0.928318	1.287990	2.312710	4.497050
k= 1	1.005460	1.004720	0.950783	1.339810	2.395180
k= 2	0.990253	1.001070	0.977474	0.960994	1.354320
k= 3	0.995848	0.993963	0.981683	0.992655	0.955345
k= 4	0.987371	1.007310	1.028120	1.016760	1.002570
k= 5	0.993840	1.015230	1.026680	1.003770	1.023820
k= 6	1.012230	1.012280	0.996363	0.999742	1.004120
k= 7	1.020450	1.025820	1.031120	1.003770	1.004770
k= 8	1.004710	1.010820	0.999289	0.973596	0.970415
k= 9	0.993542	1.003380	0.981550	0.984594	0.981001
k=10	1.015610	1.008740	0.997469	0.986705	1.004360
k=11	1.010690	1.009080	0.994975	1.010510	0.998485

Table 4: The frame of Example 5.4 in $\mathbb{R}^3$

k	Actual **MSE**	Ideal **MSE**	*Ratio*
2	0.00466163000	0.004623720000	1.008200
3	0.00116029000	0.001155930000	1.003770
4	0.00029280000	0.000288983000	1.013220
5	0.00007111000	0.000072246000	0.984317
6	0.00001799100	0.000001806000	0.996100
7	0.00000438600	0.000004515000	0.971450
8	0.00000109200	0.000011288000	0.967129
9	0.00000028070	0.000000280000	0.994956
10	0.00000007063	0.000000070550	1.001090
11	0.00000001776	0.000000017638	1.006860

REFERENCES

[1] Benedetto, J. J. and Fickus, M., "Finite normalized tight frames," *Adv. Comput. Math.*, vol. 18, no. 2-4, pp. 357–385, 2003. Frames.

[2] Benedetto, J. J., Powell, A. M., and Yı lmaz, Ö., "Second-order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions," *Appl. Comput. Harmon. Anal.*, vol. 20, no. 1, pp. 126–148, 2006.

[3] Benedetto, J. J., Powell, A. M., and Yı lmaz, Ö., "Sigma-Delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 1990–2005, 2006.

[4] Bennett, W. R., "Spectra of quantized signals," *Bell System Tech. J.*, vol. 27, pp. 446–472, 1948.

[5] Bissinger, B. H., "A generalization of continued fractions," *Bull. Amer. Math. Soc.*, vol. 50, pp. 868–876, 1944.

[6] Brown, G. and Yin, Q., "β-transformation, natural extension and invariant measure," *Ergodic Theory Dynam. Systems*, vol. 20, no. 5, pp. 1271–1285, 2000.

[7] Casazza, P. G., "The art of frame theory," *Taiwanese J. Math.*, vol. 4, no. 2, pp. 129–201, 2000.

[8] Casazza, P. G. and Kovačević, J., "Equal-norm tight frames with erasures," *Adv. Comput. Math.*, vol. 18, no. 2-4, pp. 387–430, 2003. Frames.

[9] Christensen, O., *An introduction to frames and Riesz bases*. Applied and Numerical Harmonic Analysis, Boston, MA: Birkhäuser Boston Inc., 2003.

[10] Daubechies, I., DeVore, R., Gunturk, C., and Vaishampayan, V., "Beta expansions: a new approach to digitally corrected a/d conversion," vol. vol.2, (Phoenix-Scottsdale, AZ, USA), pp. 784 – 7, 2002//.

[11] Daubechies, I. and DeVore, R., "Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order," *Ann. of Math. (2)*, vol. 158, no. 2, pp. 679–710, 2003.

[12] Daubechies, I., DeVore, R. A., Güntürk, C. S., and Vaishampayan, V. A., "A/D conversion with imperfect quantizers," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 874–885, 2006.

[13] Daubechies, I. and Yı lmaz, Ö., "Robust and practical analog-to-digital conversion with exponential precision," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3533–3545, 2006.

[14] Duffin, R. J. and Schaeffer, A. C., "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, pp. 341–366, 1952.

[15] Eldar, Y. C. and Forney, Jr., G. D., "Optimal tight frames and quantum measurement," *IEEE Trans. Inform. Theory*, vol. 48, no. 3, pp. 599–610, 2002.

[16] Estrada, R. and Kanwal, R. P., *Asymptotic analysis*. Boston, MA: Birkhäuser Boston Inc., 1994. A distributional approach.

[17] Feng, D.-J., Wang, L., and Wang, Y., "Generation of finite tight frames by Householder transformations," *Adv. Comput. Math.*, vol. 24, no. 1-4, pp. 297–309, 2006.

[18] Gersho, A. and Gray, R., *Vector quantization and signal compression*. 1992.

[19] Goyal, V. K., Kovačević, J., and Kelner, J. A., "Quantized frame expansions with erasures," *Appl. Comput. Harmon. Anal.*, vol. 10, no. 3, pp. 203–233, 2001.

[20] Goyal, V. K., Vetterli, M., and Thao, N. T., "Quantized overcomplete expansions in $\mathbf{R}^N$: analysis, synthesis, and algorithms," *IEEE Trans. Inform. Theory*, vol. 44, no. 1, pp. 16–31, 1998.

[21] Gray, R., "Quantization noise spectra," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1220 – 44, 1990/11/.

[22] Gray, R. M. and Neuhoff, D. L., "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998. Information theory: 1948–1998.

[23] Güntürk, C. S., "Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation," *J. Amer. Math. Soc.*, vol. 17, no. 1, pp. 229–242 (electronic), 2004.

[24] Ifrah, G., *The universal history of numbers*. New York: John Wiley & Sons Inc., 2000. From prehistory to the invention of the computer, Translated from the 1994 French original by David Bellos, E. F. Harding, Sophie Wood and Ian Monk.

[25] Jiménez, D., Wang, L., and Wang, Y., "White noise hypothesis for uniform quantization errors," *SIAM J. Math. Anal.*, vol. 38, no. 6, pp. 2042–2056 (electronic), 2007.

[26] Katznelson, Y., *An introduction to harmonic analysis*. New York: John Wiley & Sons Inc., 1968.

[27] Kopf, C., "Invariant measures for piecewise linear transformations of the interval," *Appl. Math. Comput.*, vol. 39, no. 2, part II, pp. 123–144, 1990.

[28] Lasota, A. and Yorke, J. A., "On the existence of invariant measures for piecewise monotonic transformations," *Trans. Amer. Math. Soc.*, vol. 186, pp. 481–488 (1974), 1973.

[29] Lasota, A. and Mackey, M. C., *Chaos, fractals, and noise*, vol. 97 of *Applied Mathematical Sciences*. New York: Springer-Verlag, second ed., 1994. Stochastic aspects of dynamics.

[30] Li, T. Y. and Yorke, J. A., "Ergodic transformations from an interval into itself," *Trans. Amer. Math. Soc.*, vol. 235, pp. 183–192, 1978.

[31] Linder, T., Zamir, R., and Zeger, K., "High-resolution source coding for non-difference distortion measures: multidimensional companding," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 548–561, 1999.

[32] Na, S. and Neuhoff, D. L., "Bennett's integral for vector quantizers," *IEEE Trans. Inform. Theory*, vol. 41, no. 4, pp. 886–900, 1995.

[33] Nyquist, H., "Certain topics in telegraph transmission," *Journal of the American Institute of Electrical Engineers*, vol. 47, pp. 214 – 216, 1928/03/.

[34] Parry, W., "On the β-expansions of real numbers," *Acta Math. Acad. Sci. Hungar.*, vol. 11, pp. 401–416, 1960.

[35] Parry, W., "Representations for real numbers," *Acta Math. Acad. Sci. Hungar.*, vol. 15, pp. 95–105, 1964.

[36] Rath, G. and Guillemot, C., "Recent advances in dft codes based quantized frame expansions for erasure channels," *Digital Signal Processing*, vol. 14, no. 4, pp. 332 – 54, 2004/07/.

[37] Rényi, A., "Representations for real numbers and their ergodic properties," *Acta Math. Acad. Sci. Hungar*, vol. 8, pp. 477–493, 1957.

[38] Shannon, C. E., "Communication in the presence of noise," *Proc. I.R.E.*, vol. 37, pp. 10–21, 1949.

[39] Sidorov, N., "Almost every number has a continuum of β-expansions," *Amer. Math. Monthly*, vol. 110, no. 9, pp. 838–842, 2003.

[40] Thao, N. and Vetterli, M., "Deterministic analysis of oversampled a/d conversion and decoding improvement based on consistent estimates," *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 519 – 31, 1994/03/.

[41] Tsujii, M., "Absolutely continuous invariant measures for expanding piecewise linear maps," *Invent. Math.*, vol. 143, no. 2, pp. 349–373, 2001.

[42] Unser, M., "Sampling-50 years after shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569 – 87, 2000/04/.

[43] Viswanathan, H. and Zamir, R., "On the whiteness of high-resolution quantization errors," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 2029–2038, 2001.

[44] Zamir, R. and Feder, M., "On lattice quantization noise," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1152 – 9, 1996/07/.

[45] Zygmund, A., *Trigonometrical series*. New York: Chelsea Publishing Co., 1950.