QUANTIZATION IN SIGNAL PROCESSING WITH FRAME THEORY

By

JIAYI JIANG

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Mathematics

May, 2016

Nashville, Tennessee

Approved:

Alexander Powell

Akram Aldroubi

Doug Hardin

Brett Byram

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor, Alexander.M.Powell. He taught me too many thing in these five years which are not only on mathematics research but also on how to be a polite and organised person. Thanks a lot for his encouragement, patience and professional advices. I really appreciate for his supporting and helping when I decide to work in industry instead of academic. All of these mean a lot to me!

My time in Vanderbilt has been a lot of fun. Thanks for the faculty of the math department. Thanks for all my teachers, Akram Aldroubi, Doug Hardin, Mark Ellingham, Mike Mihalik, Edward Saff. Each Professors in these list taught me at least one course which I will be benefited from them in my future carer. I really enjoy the different way of teaching from these professors.

Thanks for my Vandy friends, Corey Jones, Yunxiang Ren, Sui tang and Zhengwei Liu. You guys give me too many helps and joys.

I have to say thanks to my wife Suyang Zhao. She fills my life with beauty and makes me become a better man.

Finally, I wish to state my gratitude for the my parents for their love and support. They are the reason that I can be here. They always believe me and are willing for help. Thanks for all my relatives.

OVERVIEW

In signal processing, an essential problem is to encode an analog signal by using finitely many bits in an efficient and robust manner. To achieve this propose, there are two things we are interested in. The first is how to choose a “nice basis”. Finite frames provide redundant, stable and usually non-unique representations of signals. A second important issue is the quantization algorithm. Sigma-Delta($\Sigma\Delta$) quantization is an efficient method for analog-to-digital (A/D) conversion. Daubechies and DeVore in [DD03] established an approximation theoretical framework and provided rigorous results on the relationship between robustness of $\Sigma\Delta$ schemes, redundancy of signal representations and the approximation error. Moreover, another problem is how to reconstruct a signal from the quantized coefficients, which is called digital-to-analog (D/A) conversion. In [BLPY10], the authors use rth-order $\Sigma\Delta$ quantization with finite frame theory to obtain the error bounds equal to $\mathscr{O}(N^{-r})$ in (D/A) conversion for a wide class of finite frames of size N with special dual frames called Sobolev duals. However, observing the definition of Sobolev duals, we can find it is constructed depend the frame we choose. Hence, a natural problem is to ask: Is $\mathscr{O}(N^{-r})$ the best error bound possible for all frames? In chapter 3, we will answer this question.

Finite frame theory is a widely used tool, but it may not be suitable for all applications, see [CKL08]. For instance, see Figure 1, in wireless sensor networks, the sensors which have limited capacity and power are spread in a large area to measure physical quantities, such as temperature, sound, vibration or pressure. Such a sensor system is typically redundant, and there is no orthogonality among sensors, therefore each sensor functions as a frame element in the system. However, for practical and cost reasons, sensors employed in such applications have severe constraints in their processing power and transmission bandwidth. They often have limited power supply as well. Consequently, a typical large sensor network necessarily divides the network into redundant sub-networks forming a set of subspaces(W_i). The primary goal is to have local measurements transmitted to a local sub-station within a subspace. These local sub-stations can transmit a high bits local signal(analog signal u_i) to the other nearby sub-stations and can send a low bit signal(digital signal q_i) to the central processing station. An entire sensor system in such applications could have a number of such local processing centers. They function as relay stations, and have the gathered information further submitted to a central processing station for final assembly.

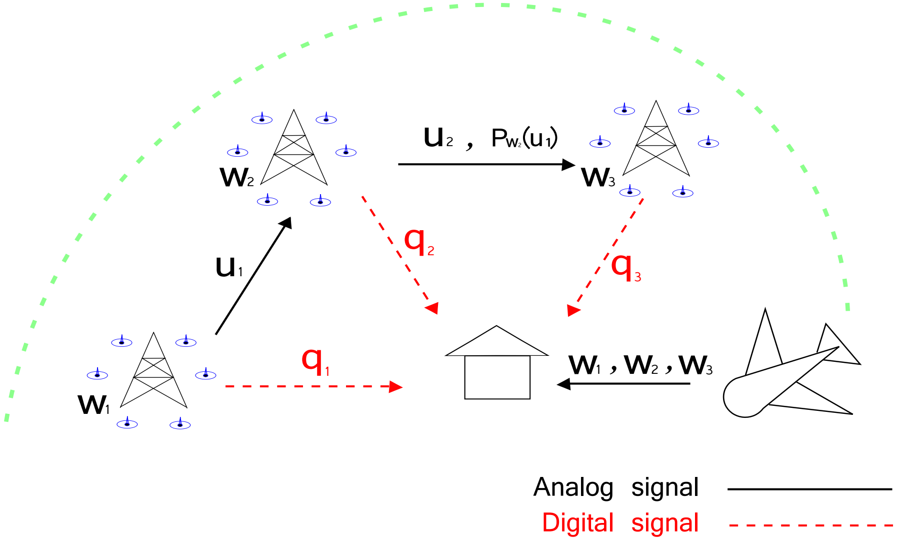


Figure 1: Wireless Sensor Network

For these situations, we have two problems to solve. The first one is how to perform data fusion among a set of overlapping, non-orthogonal and redundant data measurements and the answer is fusion frame systems which are created for the situation, see [CKL08]. Secondly, we need to find an efficient and robust method to do (A/D) conversion in local sub-stations and (D/A) conversion in the central processing station to reconstruct the signal with a fusion frame system. Hence, we get the idea of applying $\Sigma\Delta$ quantization in fusion frames.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Chapter 1

Introduction to frame theory

In this chapter, we give a brief introduction to frame and fusion frame theory. Frames are a system which provide redundant and usually non-unique representations of vectors. Frame theory has been used in filter bank theory, signal and image processing, and wireless communications. Redundant frames are used in these applications because they yield representations that are robust and stable under:

- additive noise [D+92] (in the setting of Gabor and wavelet frames for $L^2(\mathbb{R}^d)$), [BT01] (in the setting of oversampled bandlimited functions), and [Mun92](in the setting of tight Gabor frames)

- quantization [DD03, GLV01, Yil02a] (in the setting of oversampled bandlimited functions), [Yıl03](in the setting of tight Gabor frames) and [LPY10, BLPY10, BYP04, GVT98a, GLP+10] (in the setting of finite frames).

- partial data loss [GKK01, RG02] (in the setting of finite frames)

We shall mainly focus on introducing the basic definitions and properties of finite frames in finite dimensional space $\mathbb{R}^d$. More information can be found in these articles and books, such as [Cas99, Chr02, Hei10, Zim01, CKP13]

For certain applications, such as sensor networks, physical considerations require building frames ”locally” and then piecing them together to obtain frames for the whole space. This idea led to a distributed frame theory known as fusion frames, see [CK03, CKL08]. In this thesis, a main result is constructing a Sigma-Delta Quantization algorithm for fusion frames. Before that, we will state the basic knowledge of fusion frames in this chapter.

1.1 Frame operator and dual frame

Definition 1.1.1. *A finite collection of vectors* $\{e_n\}_{n=1}^N \subseteq \mathbb{R}^d$ *is a frame with frame bounds* $0 < A \leq B < \infty$ *if*

$$\forall x \in \mathbb{R}^d,\ A\|x\|^2 \leq \sum_{n=1}^{N} |\langle x, e_n\rangle|^2 \leq B\|x\|^2. \tag{1.1.1}$$

where $\|\cdot\|$ *denotes the Euclidean norm. The frame bounds are taken to be the respective largest and smallest values of A and B such that* (1.1.1) *holds. If* $A = B$ *then the frame is said to be tight. If* $\|e_n\| = 1$ *holds for each* $n = 1,...,N$*, then the frame is said to be unit-norm.*

Definition 1.1.2. *Given a frame* $\{e_n\}_{n=1}^N \subseteq \mathbb{R}^d$*, we define the analysis operator by:*

$$L : \mathbb{R}^d \longrightarrow l^2(N),\ \ (Lx)_n = \langle x, e_n\rangle.$$

and the synthesis operator which defined to be the adjoint operator by

$$L^* : l^2(N) \longrightarrow \mathbb{R}^d,\ \ L^*(\{x_n\}_{n=1}^N) = \sum_{i=1}^{N} x_n e_n.$$

The associated frame operator $S : \mathbb{R}^d \to \mathbb{R}^d$*, is defined by*

$$S(x) = L^*L = \sum_{n=1}^{N} \langle x, e_n\rangle e_n,$$

and it satisfies

$$AI \leq S \leq BI,$$

where I is the identity operator on $\mathbb{R}^d$*. The inverse of S,* S^{-1} *is called the dual frame operator, and it satisfies:*

$$B^{-1}I \leq S^{-1} \leq A^{-1}I,$$

The following theorem shows that frames can be used to provide signal decompositions in signal processing.

Theorem 1.1.3. *If $\{e_n\}_{n=1}^N \subseteq \mathbb{R}^d$ is a frame with frame bounds A and B and S is the frame operator, then S is positive and invertible. Moreover, there exists a dual frame $\{f_n\}_{n=1}^N \subseteq \mathbb{R}^d$ such that*

$$\forall x \in \mathbb{R}^d,\ x = \sum_{n=1}^N \langle x, e_n\rangle f_n = \sum_{n=1}^N \langle x, f_n\rangle e_n. \tag{1.1.2}$$

In particular one may take $f_n = S^{-1}e_n$ when $\{f_n\}$ is called the canonical dual fame. If $N > d$ then the frame $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ is an overcomplete collection and the choice of dual frame $\{f_n\}_{n=1}^N$ is not unique. If $\{f_n\}_{n=1}^N$ is not the canonical dual frame then we refer to it as an alternative or non-canonical dual frame.

Tight frames have the property that the dual frame can be chosen as $f_n = A^{-1}e_n$, where A is the frame bound. For more background on tight frames see [BF03, GKK01]. There are many examples of unit-norm tight frames. Here is an example for $\mathbb{R}^d$:

Example 1.1.4. *(Harmonic Frames). The harmonic frames are constructed using columns of the Fourier matrix, e.g., see[GKK01, GVT98a, Zim01] The definition of the harmonic frame $H_N^d = \{h_n^N\}_{n=1}^N$, $N \geq d$, depends on whether the dimension d is even or odd.*
If d is even, let

$$h_n^N = \sqrt{\frac{2}{d}} \left[\cos(\frac{2\pi n}{N}), \sin(\frac{2\pi n}{N}), \cdots, \cos(\frac{2\pi \frac{d}{2} n}{N}), \sin(\frac{2\pi \frac{d}{2} n}{N}) \right]^T \tag{1.1.3}$$

for $n = 1, 2, ..., N$.
If d is odd, let

$$h_n^N = \sqrt{\frac{2}{d}} \left[\frac{1}{\sqrt{2}}, \cos(\frac{2\pi n}{N}), \cdots, \cos(\frac{2\pi \frac{d-1}{2} n}{N}), \sin(\frac{2\pi \frac{d-1}{2} n}{N}) \right]^T \tag{1.1.4}$$

for $n = 1, 2, ..., N$.

It is shown in [Zim01] that H_N^d, as defined above, is a unit-norm tight frame for $\mathbb{R}^d$. The frame $\mathbb{R}^d$ is built by uniformly sampling a smooth vector valued function whose components are sines and cosines. Generalizations of this property are important in the study of Sigma-Delta quantization, see [BLPY10, BPA07a]

We say a function $f : [0,1] \to \mathbb{R}^d$ is piecewise C^1 if it is C^1 except at finitely many points in [0,1], and the left and right limits of f and f' exist at all of these point.

Definition 1.1.5. *A vector valued function $E : [0,1] \to \mathbb{R}^d$ given by*

$$E(t) = [e_1(t), e_2(t), ..., e_d(t)]^*,$$

is a piecewise C^1 uniformly-sampled frame path if the following three conditions hold:

1. *For $1 \leq n \leq N$, $e_n : [0,1] \to \mathbb{R}$ is piecewise $- C^1$ on $[0,1]$*
2. *The functions $\{e_n\}_{n=1}^d$ are linearly independent.*
3. *There exists N_0 such that for each $N \geq N_0$ the collection $\{E(n/N)\}_{n=1}^N$ is a frame for $\mathbb{R}^d$.*

Frame vectors generated by a frame path are uniformly bounded in norm which means there exists M such that $\|E(n/N)\| \leq M$ holds for all n and N.

Here are some examples of frame paths:

Example 1.1.6. *(Roots of unity frame path). Consider the frame path defined by $E(t) = [cos(2\pi t), sin(2\pi t)]^*$. For each $N \geq 3$, the collection $U_N = \{E(n/N)\}_{n=1}^N \subset \mathbb{R}^2$ given by:*

$$E(n/N) = [\cos(2\pi n/N), \sin(2\pi n/N)]^*, \ \ 1 \leq n \leq N, \qquad (1.1.5)$$

is a unit-norm tight frame for R^2.

Example 1.1.7. *(Repetition frame path). Consider the frame path defined by*

$$E(t) = [\chi_{[0,\frac{1}{d}]}(t), \chi_{(\frac{1}{d},\frac{2}{d}]}(t), ..., \chi_{(\frac{d-1}{d},1]}(t)]^*, \tag{1.1.6}$$

where χ_S denotes the characteristic function of S. Note that $R_N = \{E(n/N)_{n=1}^N\}$ is a frame for $\mathbb{R}^d$ for all $N \geq d$.

1.2 Finite frames in matrix form

For finite frames in $\mathbb{R}^d$, the basic definitions can be conveniently reformulated in terms of matrices.

Definition 1.2.1. *Given a frame $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ we define the associated frame matrix E to be the $d \times N$ matrix:*

$$E = [e_1 \; e_2 \; ... \; e_N]$$

where e_j is the jth column. Note that the columns of a $d \times N$ matrix E form a frame for $\mathbb{R}^d$ if and only if E has rank d.

If E is a frame matrix then the associated canonical dual frame has frame matrix

$$\widetilde{E} = (EE^*)^{-1}E$$

In particular, $\widetilde{E}E^* = E\widetilde{E}^* = I_d$, where I_d is $d \times d$ identity matrix.

Moreover, an alternative dual frame to $\{e_n\}_{n=1}^N$ is simply a set of frame vectors $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$. Let F be the corresponding $d \times N$ frame matrix. Then the frame expansions (1.1.2) can be expressed in terms of E and F as:

$$FE^* = EF^* = I_d, \tag{1.2.1}$$

see [Li95] and [EC06, LO04] for more information on dual frames.

Definition 1.2.2. *Given a $d \times N$ matrix E with the vectors $\{e_n\}_{n=1}^N$ as its columns, then we respectively define the Frobenius norm and operator norm of E by*

$$\|E\|_{\mathscr{F}} = \sqrt{tr(EE^*)} = \left(\sum_{n=1}^N \|e_n\|^2\right)^{\frac{1}{2}} \quad \text{and} \quad \|E\|_{op} = \sup_{\|x\|=1} \|Ex\|.$$

Here $tr(\dot{})$ denotes the trace of a square matrix.

Theorem 1.2.3. *Let E be a fixed $d \times N$ frame matrix and let F be an arbitrary dual frame to E. The two quantities $\|F\|_{op}, \|F\|_{\mathscr{F}}$ are minimized when F is taken to be the canonical dual frame of E, namely $F = (EE^*)^{-1}E$.*

The theorem stated above is a basic property of canonical duals and the proof follows easily from Theorem 3.6 in [Li95]. Moreover, the property is also used to construct Sobolev duals which we will introduce in next chapter.

1.3 Fusion frames and the fusion frame operator

In frame theory the signal is represented by a collection of scalar coefficients that measure the projection of that signal onto one dimensional subspaces spanned by each frame vector. The representation space employed in this theory equals $\ell^2(I)$. However, in fusion frame theory the signal is represented by collection of vector coefficients that represent the projections onto higher dimensional subspaces. Therefore the representation space employed in this setting is defined as follows.

Definition 1.3.1.

$$\sum_{i\in I} \bigoplus W_i = \{\{f_i\}_{i\in I} : f_i \in W_i, and \{\|f_i\|\}_{i\in I} \in l^2(I)\}. \tag{1.3.1}$$

With the inner product

$$\langle \{f_i\}_{i\in I}, \{g_i\}_{i\in I}\rangle = \sum_{i\in I} \langle f_i, g_i\rangle.$$

the space $\sum_{i\in I} \bigoplus W_i$ can be considered as a Hilbert space.

Definition 1.3.2. *Let I be a countable index set, let $\{W_i\}_{i\in I}$ be a family of closed subspaces in Hilbert space $\mathbb{H}$, and let $\{w_i\}_{i\in I}$ be a family of positive weights, i.e., $w_i > 0$ for all $i \in I$. Then $\{(W_i, w_i)\}_{i\in I}$ is a fusion frame for $\mathbb{H}$, if there exist constants $0 < C \leq D < \infty$ such that*

$$\forall f \in \mathbb{H},\ C\|f\|^2 \leq \sum_{n\in I} w_i^2 \|\pi_{W_i}(f)\|^2 \leq D\|f\|^2. \tag{1.3.2}$$

here π_{W_i} is the orthogonal projection onto the subspace W_i. We call C and D the fusion frame bounds. The family $\{(W_i, w_i)\}_{i\in I}$ is called a C-tight fusion frame, if in (1.3.2) the constants $C = D$, and it is called an orthonormal fusion basis if $\mathbb{H}$ is the orthogonal sum of the subspaces W_i. If $\{(W_i, w_i)\}_{i\in I}$ satisfies upper fusion frame bound in (1.3.2), we call it a Bessel fusion sequence with Bessel fusion bound D.

Definition 1.3.3. *If $\{(W_i,w_i)\}_{i\in I}$ is a fusion frame for $\mathbb{H}$, then we can define the bounded linear operator, called the synthesis operator by*

$$T_{W,w}:\sum_{i\in I}\bigoplus W_i \longrightarrow \mathbb{H}, T_{W,w}(\{f_i\}_{i\in I})=\sum_{i\in I} w_i f_i$$

The adjoint of $T_{W,w}$is called the analysis operator,

$$T^*_{W,w}:\mathbb{H}\longrightarrow \sum_{i\in I}\bigoplus W_i,$$

$$T^*_{W,w}(f)=\{w_i\pi_{W_i}(f)\}.$$

The fusion frame operator is

$$S_{W,w}=T_{W,w}T^*_{W,w}:\mathbb{H}\longrightarrow\mathbb{H},$$

$$S_{W,w}(f)=T_{W,w}T^*_{W,w}(f)=\sum_{i\in I} w_i^2\pi_{W_i}(f).$$

$S_{W,w}$ is positive, self-adjoint and invertible. The fusion frame operator has similar properties to the frame operator. If $\{(W_i,w_i)\}_{i\in I}$ is the fusion frame for $\mathbb{H}$ with fusion frame bounds C and D, then

$$CI_{\mathbb{H}}\leq S_{W,w}\leq DI_{\mathbb{H}}.$$

Moreover, if $\{(W_i,w_i)\}_{i\in I}$ is a C-tight fusion frame, then $S_{W,w}=CI_{\mathbb{H}}$.

Fusion frames are closed related to the idea of combining local frames for a collection of subspaces.

Definition 1.3.4. *Let $\{(W_i,w_i)\}_{i\in I}$ be a fusion frame for $\mathbb{H}$, and let $\{f_{ij}\}_{j\in J_i}$ be a frame for W_i for each $i\in I$. Then we call $\{W_i,w_i,\{f_{ij}\}_{j\in J_i}\}_{i\in I}$ a fusion frame system for $\mathbb{H}$. C and D are the associated fusion frame bounds if they are the fusion frame bounds for $\{(W_i,w_i)\}_{i\in I}$, and A_i and B_i are the local frame bounds if these are the common frame bound for the local* *frames $\{f_{ij}\}_{j\in J_i}$. A collection of dual frames $\{\widetilde{f_{ij}}\}_{j\in J_i}$ associated with the local frames will be called local dual frames.*

Now we state the following theorem from [CK03] that provides a relation between properties of the associated fusion frame and the sequence consisting of all local frame vectors.

Theorem 1.3.5. *For each $i \in I$, let $w_i > 0$, let W_i be a closed subspace of $\mathbb{H}$, and let $\{f_{ij}\}_{j\in J_i}$ be a frame for W_i with frame bounds A_i and B_i. Suppose that*

$$0 < A = \inf_{i\in I}(A_i) \leq \sup_{i\in I}(B_i) = B < \infty.$$

Then the following conditions are equivalent.

1. *$\{(W_i, w_i)\}_{i\in I}$ is a fusion frame for $\mathbb{H}$.*

2. *$\{w_i f_{ij}\}_{j\in J_i, i\in I}$ is a frame for $\mathbb{H}$.*

In particular, if $\{W_i, w_i, \{f_{ij}\}_{j\in J_i}\}_{i\in I}$ is a fusion frame system for $\mathbb{H}$ with fusion frame bounds C and D, then $\{w_i f_{ij}\}_{j\in J_i, i\in I}$ is a frame for $\mathbb{H}$ with frame bounds AC and BD. Also if $\{w_i f_{ij}\}_{j\in J_i, i\in I}$ is a frame for $\mathbb{H}$ with frame bounds C and D, then $\{W_i, w_i, \{f_{ij}\}_{j\in J_i}\}_{i\in I}$ is a fusion frame system for $\mathbb{H}$ with fusion frame bounds $\frac{C}{B}$ and $\frac{D}{A}$. Moreover, $\{(W_i, w_i)\}_{i\in I}$ is a C-tight fusion frame for $\mathbb{H}$ if and only if $\{w_i f_{ij}\}_{j\in J_i, i\in I}$ is a C-tight frame for $\mathbb{H}$.

Besides, the following proposition shows that the fusion frame bound C of a C-tight fusion frame can be interpreted as the redundancy of his fusion frame.

Proposition 1.3.6. *Let $\{(W_i, w_i)\}_{i\in I}$ be a C-tight fusion frame for $\mathbb{H}$ with* $\dim(\mathbb{H}) < \infty$. *Then we have:*

$$C = \frac{\sum_{i=1}^{n} w_i^2 \dim(W_i)}{\dim(\mathbb{H})} \tag{1.3.3}$$

A different expression for the fusion frame operator can be defined by the local frame operators as follows, see [CK03].

Proposition 1.3.7. *Let $\{W_i, w_i, F_i\}_{i\in I}$ be a fusion frame system for $\mathbb{H}$ with $F_i = \{f_{ij}\}_{j\in J_i}$, and let $\widetilde{F}_i = \{f_{ij}\}_{j\in J_i}, i \in I$ be associated local dual frames. Then the associated fusion frame operator $S_{W,w}$ can be written as:*

$$S_{W,w} = \sum_{i\in I} w_i^2 L^*_{\widetilde{F}_i} L_{F_i} = \sum_{i\in I} w_i^2 L^*_{F_i} L_{\widetilde{F}_i}$$

Definition 1.3.8. *Assume that $\{(W_i, w_i)\}_{i\in I}$ is fusion frame for $\mathbb{H}$. Then $\{S^{-1}_{W,w} W_i, w_i\}_{i\in I}$ is called the canonical dual fusion frame of $\{(W_i, w_i)\}_{i\in I}$. And we have the following canonical fusion frame representation:*

$$f = S^{-1}_{W,w} S_{W,w} f = \sum_{j\in I} w_j^2 S^{-1}_{W,w} \pi_{W_j}(f), \ for\ all\ f \in \mathbb{H} \tag{1.3.4}$$

If we want to use the expression 1.3.4 in terms of operators as in 1.2.1. We find that the range of $T^*_{W,w}$ is a subset in$\sum_{i\in I} \bigoplus W_i$, but the domain of $T_{S^{-1}_{W,w} W,w}$ is $\sum_{i\in I} \bigoplus S^{-1} W_i$, so $T_{S^{-1}_{W,w} W,w} T^*_{W,w}$ is not always well defined and this requires a different definition in terms of operators. In [HMBZ13], the authors give a new definition of dual fusion frame which can solve the domain problem. Also, if we only want to find a way to reconstruct or represent the signal, in [CKL08], the authors give the definition of distributed reconstruction and global dual frames which depend on the frame system. Moreover, in this thesis, instead of using dual fusion frame or global dual frames, we define the left inverse operator of fusion frame which can be easily applied in quantization problems.

Chapter 2

Introduction to Sigma-Delta ($\Sigma\Delta$) Quantization

In the previous chapter, we reviewed the basic definitions and properties of frame theory. By using frame theory, we can give robust, stable and redundant signal representations. More precisely, one expands a given signal x over a finite dictionary $\{e_n\}_{n=1}^N$ such that:

$$x = \sum_{n=1}^{N} c_n e_n \tag{2.0.1}$$

where c_n are real or complex numbers or vectors in the case of fusion frame represents. Usually, the choice of c_n is not unique since frames are redundant.

However, even (2.0.1) is a discrete representation, it is not a digital representation since the coefficient sequence $\{c_n\}_{n=1}^N$ is real, complex or vector valued. Hence, we have to reduce the continuous range of the sequence to a discrete and finite set. We call this step quantization or A/D (analog to digital) conversion.

Definition 2.0.9. *A quantizer maps each expansion (2.0.1) to an element of:*

$$\Gamma_{\mathscr{A}} = \{\sum_{n=1}^{N} q_n e_n : q_n \in \mathscr{A}\}$$

where the quantization alphabet $\mathscr{A}$ is a given discrete and finite set. The performance of a quantizer is reflected in the approximation error $\|x - \widetilde{x}\|$, where $\|\cdot\|$ is any suitable norm, and:

$$\widetilde{x} = \sum_{n=1}^{N} q_n e_n$$

is the quantized expansion.

There are other more general approaches to quantization, such as consistent reconstruction, e.g., [TV94, GVT98b], using nonlinear reconstruction. However, in this thesis, we only focus on linear reconstruction as in (2.0.1).

A simple method of quantization, for a given expansion 2.0.1 is to choose q_n to be the closest point in the alphabet $\mathscr{A}$ to c_n. Quantizers defined this way are usually called memory less quantization or pulse code modulation(PCM) algorithms. For example, we may use a truncated binary expansion to replace c_n. If we know a priori that $|c_n| \leq A < \infty$ for all n, then we can write:

$$q_n = -A + A\sum_{k=0}^{K} b_k^n 2^{-k}, \tag{2.0.2}$$

with $b_k^n \in \{0,1\}$ for all k. Here we spend K bits per coefficient c_n. Following the constructing of $\widetilde{x}$, we have:

$$\|x - \widetilde{x}\| \leq C2^{-K+1}A,$$

where C is a independent constant. The quantization method we just gave is widely used but also has shortcomings.

- The quantization method we just gave is widely used but also has shortcomings.
- In practice, it is difficult to build analog devices that can divide the amplitude range $[-A,\ A]$ into 2^{-K+1} precisely equal bins.
- If we use a redundant representation of the signal, for example, a frame expansion, then the error will not generally decrease as a function of frame size N.
- The algorithm performs poorly for 1-bit or low-bit quantization.
- Not robust against bit-flips.
- PCM quantization often requires an analysis under the white noise assumption. More details about the noise model can be find in [Ben48, JWW07, Gra90].

Since the reasons we state above, in this chapter, we introduce Sigma-delta quantization schemes which are a popular way to quantize the signal. We will show that the algorithm spend few bits on each quantized coefficient and the overall error $\|x - \widetilde{x}\|$ will decrease as the frame size increases. Information about noise models for $\Sigma\Delta$ algorithms can be find in [ST05, Wan08, BH01, BYP04].

Definition 2.1.1. *Fix $\delta > 0$ and $K \in \mathscr{N}$. Given the $2K$-level midrise quantization alphabet with stepsize δ,*

$$\mathscr{A}_K^\delta = \{(-K+1/2)\delta, (-K+3/2)\delta, ..., (-1/2)\delta, (1/2)\delta, ..., (K-3/2)\delta, (K-1/2)\delta\}$$

define the associated scalar quantizer:

$$\mathscr{Q}(u) = \arg\min_{q \in \mathscr{A}_K^\delta} \|u-q\|. \tag{2.1.1}$$

Now, we can define first order Sigma-Delta quantization. Let $\mathscr{A}_K^\delta$ be the $2K$-level midrise quantization alphabet with stepsize δ, and let $\mathscr{Q}$ be the associated scalar quantizer from 2.1.1. Let $\{e_n\}_{n=1}^N \in \mathbb{R}^d$ be a unit-norm frame for $\mathbb{R}^d$ with frame operator S. Let $\{f_n\}_{n=1}^N \in \mathbb{R}^d$ be any, not necessarily unit-norm, dual frame.

Given $x \in \mathbb{R}^d$ satisfying $\|x\| \leq (K-1/2)\delta$, and having frame coefficients $x_n = \langle x, e_n \rangle$, the first order $\Sigma\Delta$ algorithm quantizes frame coefficients q_n by running the iteration:

$$q_n = \mathscr{Q}(u_{n-1} + x_n),$$

$$u_n = u_{n-1} + x_n - q_n. \tag{2.1.2}$$

for $n = 1, 2, ..., N$, and with $u_0 = 0$. The u_n are internal state variables of the $\Sigma\Delta$ scheme, and q_n are the quantized frame coefficients from which we linearly reconstruct:

$$\widetilde{x} = \sum_{n=1}^N q_n f_n. \tag{2.1.3}$$

The $\Sigma\Delta$ scheme is stable, by [DD03] page 4. In particular,

$$\textit{For any } n \in \{1, 2, ..., N\},\ |x_n| \leq (K-1/2)\delta \Rightarrow |u_n| \leq \delta/2,\ \textit{for } n = 1, 2, ...N. \tag{2.1.4}$$

For unit-norm frames, it is easy to prove that $\|x\| \leq (K-1/2)\delta$ implies $|x_n| = |\langle x, e_n\rangle| \leq (K-1/2)\delta$. Hence, $|u_n| \leq \delta/2$ holds.

Error estimates for $\Sigma\Delta$ quantization in the setting of finite frames are given in [BYP04, BPY06b, BPY06a], see also [BP07, BPA07b]. For instance, if we use the canonical dual frame in reconstruction (2.1.3), then we have:

$$\|x-\widetilde{x}\| \leq \frac{\delta}{2}\|S^{-1}\|_{op}(\sum_{n=1}^{N}\|e_n - e_{n+1}\| + 1) \tag{2.1.5}$$

Here the ordering of the frame $\{e_n\}_{n=1}^{N}$ is quite important. For example, for the frame (1.1.4) in its natural ordering (same as the definition), the frame variation is bounded by 2π and the error (2.1.5) yields:

$$\|x-\widetilde{x}\| \leq \frac{\delta}{N}(2\pi+1)$$

which is clearly showing that the error will decreases by the frame size N.

The following notation will help simplify the error analysis of $\Sigma\Delta$ schemes. Let D be the $N \times N$ first-order difference matrix given by

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \tag{2.1.6}$$

If one linearly reconstructs from the $\Sigma\Delta$ quantized coefficients q_n, obtained via (2.2.1), as in (2.2.2) using a dual frame $\{f_n\}_{n=1}^{N}$, then the reconstruction error equals,

$$\|x-\widetilde{x}\| = \|\sum_{n=1}^{N}(x_n - q_n)f_n\| = \|\sum_{n=1}^{N}(u_n - u_{n-1})f_n\| = \|FD^*(u)\| \tag{2.1.7}$$

where $u = [u_1, u_2, ..., u_N]^*$ and F is the frame matrix associated to $\{f_n\}_{n=1}^{N}$. Observing the error estimate (2.1.7), it will be important to improve the error by choosing a suitable dual frame which we will introduce in section 3.

2.2 High order Sigma-Delta quantization with finite frames

High order $\Sigma\Delta$ schemes works analogously to (2.1.2) by controlling high order difference operators. Suppose that $\{e_n\}_{n=1}^N$ is a frame for $\mathbb{R}^d$ and $x \in \mathbb{R}^d$. Let $x_n = \langle x, e_n \rangle$, then the general rth order sigma-delta scheme with alphabet $\mathscr{A}_K^\delta$ runs the following iteration:

$$q_n = \mathscr{Q}(G(u_{n-1}^1, u_{n-1}^2, \ldots, u_{n-1}^r, x_n)),$$

$$u_n^1 = u_{n-1}^1 + x_n - q_n,$$

$$u_n^2 = u_{n-1}^2 + u_n^1,$$

$$\vdots$$

$$u_n^r = u_{n-1}^r + u_n^{r-1} \tag{2.2.1}$$

where $u_0^1 = u_0^2 = \ldots = u_0^r = 0$ and the iteration runs for $n = 1, 2, \ldots, N$.

We may equivalent by define the iteration using:

$$u_n^j = \Delta u_n^{j+1}, j = 1, 2, ..., r-1$$

and

$$x_n - q_n = \Delta^r u_n^r,$$

where Δ^r is the rth order backwards difference operator defined by

$$\Delta\omega_n = \omega_n - \omega_{n-1} \;\; and \;\; \Delta^r = \Delta^{r-1}\Delta$$

Here $G : \mathbb{R}^{r+1} \to \mathbb{R}$ is a fixed function called the quantization rule. The algorithm above gives $q_n \in \mathscr{A}$ as the output coefficients. One can linearly reconstruct a signal $\widetilde{x}$ from the q_n

with a dual frame $\{f_n\}_{n=1}^{N}$ by

$$\widetilde{x} = \sum_{n=1}^{N} q_n f_n. \tag{2.2.2}$$

A main issue of Sigma-Delta quantization with finite frame is to make the reconstruction error $\|x-\widetilde{x}\|$ decay faster as the frame size N increase.

It is important for Sigma-Delta algorithms to be stable in the following sense:

$$\exists C_1, C_2, such\ that\ for\ any\ N > 0\ and\ any\ \{x_n\}_{n=1}^{N} \in \mathbb{R}$$

we have,

$$\forall 1 \leq n \leq N,\ x_n \leq C_1 \Longrightarrow \forall 1 \leq n \leq N,\ \forall j = 1,2,\dots,r,\ |u_n^j| \leq C_2$$

Here the constants C_1, C_2 depend on the quantization alphabet $\mathscr{A}_K^{\delta}$ and the quantization rule G.

The construction of high order 1-bit $\Sigma\Delta$ schemes (when $\mathscr{A}_K^{\delta}$ has $K = 1$) is a difficult problem. In fact, the existence of arbitrary order stable 1-bit $\Sigma\Delta$ schemes was only recently proven by Daubechies and DeVore in [DD03]. Also there is other related work on 1-bit $\Sigma\Delta$ such as [Gün03, PL93, Yil02b].

Example 2.2.1. *The following 1-bit second order $\Sigma\Delta$ scheme is stable, in [Yil02a]*

$$q_n = sign(u_{n-1}^1 + \frac{1}{2}u_{n-1}^2),$$

$$u_n^1 = u_{n-1}^1 + x_n - q_n,$$

$$u_n^2 = u_{n-1}^2 + u_n^1,$$

where $u_0^1 = u_0^2 = 0$ and $n = 1, 2, ..., N$. Here,

$$sign(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

2.3 Sobolev duals

The main result in this section is according to [BLPY10]. If we define the discrete Laplacian $\nabla = D^*D$, note that D and ∇ are invertible, obtained via 2.1.7, as in 2.2.2 using a dual frame $\{f_n\}_{n=1}^N$, then the rth order $\Sigma\Delta$ reconstruction error equals:

$$\|x - \widetilde{x}\| = \| \sum_{n=1}^{N} (x_n - q_n) f_n \| = \| \sum_{n=1}^{N} \triangle^r u_n^r f_n \| = \|FD^{r*}(u)\| \tag{2.3.1}$$

where $u = [u_1^r, u_2^r, ..., u_N^r]^*$ and F is the frame matrix associated to $\{f_n\}_{n=1}^N$.

Definition 2.3.1. *Let F be an $d \times N$ matrix. Define*

$$\|F\|_{r,op} = \|FD^{r*}\|_{op} = \|D^r F^*\|_{op}$$

$$\|F\|_{r,\mathscr{F}} = \|FD^{r*}\|_{\mathscr{F}} = \|D^r F^*\|_{\mathscr{F}}$$

Now we introduce the class of Sobolev dual frames:

Definition 2.3.2. *(Sobolev dual)Fix a positive integer r. Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a frame for $\mathbb{R}^d$ and let E be the associated $d \times N$ frame matrix. The r-th order Sobolev dual $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ of E is defined so that f_j is the j-th column of the matrix:*

$$\widetilde{F} = (ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r}(D^*)^{-r}$$

Theorem 2.3.3. *Let E be an $d \times N$ frame matrix. The r-th order Sobolev dual $\widetilde{F}$ is the dual frame F of E for which $\|F\|_{r,op}, \|F\|_{r,\mathscr{F}}$ are minimal.*

Theorem 2.3.3 shows that Sobolev duals lead the minimal value of the error bound of r-th order $\Sigma\Delta$ quantization for a fix frame. The proof can be found in section 4 in [BLPY10]. Also in section 5, the authors gave the specific bound for frames constructed by a frame path:

Theorem 2.3.4. *Let r be a positive integer, and suppose that one is given an r-th order $\Sigma\Delta$ scheme, with quantization alphabet $\mathscr{A}_K^\delta$, that is stable for all inputs $x \in \mathbb{R}^d$ with $\|x\| \leq \eta$ for some $\eta > 0$. Let $E(t)$ be a frame path for $\mathbb{R}^d$ defined as 1.1.5 and $E_N = E(n/N)_{n=1}^N$ is a frame for $\mathbb{R}^d$. Given $x \in \mathbb{R}^d$, $\|x\| \leq \eta$, with frame coefficients $\{\langle x, E(n/N)\rangle\}_{n=1}^N$, let $\{q_n^N\}_{n=1}^N \subset \mathscr{A}$ be the sequence of quantized frame coefficients that are generated by the r-th order $\Sigma\Delta$ scheme. If one uses the r-th order Sobolev dual frame $\widetilde{F}_N$ of E_N to linearly reconstruct an approximation $\widetilde{x}$ to x from the quantized frame coefficients via:*

$$\widetilde{x} = \widetilde{F}_N q$$

where $q = [q_1^N, q_2^N, ..., q_N^N]^*$, then

$$\|x - \widetilde{x}\| \leq \|FD^{r*}(u)\| = \mathscr{O}(N^{-r}). \tag{2.3.2}$$

The implicit constant may be taken independent of x.

Chapter 3

Optimizing ΣΔ quantization error with finite unit-norm frames

In previous chapter, we introduced Sigma-Delta(ΣΔ) quantization for finite frames. Given a fixed frame, the definition 2.3.2 and Theorem 2.3.3 shows that the Sobolev dual gives the minimal value of the error bound, the Sobolev dual depends on the frame itself. Here, we consider three questions:

1. Is $\|x-\widetilde{x}\| = \mathscr{O}(N^{-r})$ the best error bound we can get for rth order ΣΔ quantization?
2. If no, then what is the best order of the error bound?
3. If yes, then can we get the same error bound by using other frames? Can one optimize the constant?

We will explore how to optimize quantization error bounds for unit-norm finite frames related to Sobolev duals and will answer the first two questions.

3.1 Optimizing the operator norm

Since in Theorem 2.3.4. we assume that the $\Sigma\Delta$ scheme is stable, there exists $C > 0$ such that $\Sigma\Delta$ state variables satisfy $\|u_n^j\| \leq C$ for $1 \leq n \leq N$. Letting $u = [u_1^r, u_2^r, ..., u_N^r]^*$ gives $\|u\|_2 \leq C\sqrt{N}$ and it follows from 2.3.1 that:

$$\|x - \widetilde{x}\| = \|FD^{r*}(u)\| \leq \|FD^{r*}\|_{op}\|u\| \leq C\sqrt{N}\|FD^{r*}\|_{op} \tag{3.1.1}$$

Thus,we shall consider the problem of minimizing $\|FD^{r*}\|_{op}$ instead of the original error bound.

Lemma 3.1.1. *Fix a positive integer r. Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a unit-norm frame and let E be the associated $d \times N$ frame matrix. Moreover, let $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ be the r-th order Sobolev dual of E and let F be the associated $d \times N$ frame matrix. Then:*

$$\|FD^{r*}\|_{op}^2 = \|((ED^{-r})(ED^{-r})^*)^{-1}\|_{op}. \tag{3.1.2}$$

Proof. By the definition of Soblev dual, we have:
$\|FD^{r*}\|_{op}^2$
$= \|(FD^{r*})(FD^{r*})^*\|_{op}$
$= \|(ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r}((ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r})^*\|_{op}$
$= \|(ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r}(D^*)^{-r}E^*(ED^{-r}(D^*)^{-r}E^*)^{-1})^*\|_{op}$
$= \|(ED^{-r}(D^*)^{-r}E^*)^{-1})^*\|_{op}$
$= \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op}$ □

Lemma 3.1.2. *Let A be a invertible matrix with inverse A^{-1}. If λ is an eigenvalue of A, then $1/\lambda$ is an eigenvalue of A^{-1}.*

Proof. Since λ is a eigenvalue of A, then there are exist a nonezero vector u such that,

$$Au = \lambda u$$

which is equivalent to

$$A^{-1}Au = A^{-1}\lambda u$$

Thus

$$A^{-1}u = (1/\lambda)u$$

Hence, $1/\lambda$ is a eigenvalue of A^{-1}. □

Let $\{\lambda_i\}_{i=1}^{d}$ be the eigenvalues of $ED^{-r}(ED^{-r})^*$. Since $ED^{-r}(ED^{-r})^*$ is the inverse matrix of $(ED^{-r}(ED^{-r})^*)^{-1}$, by Lemma 3.1.2, we have $\{1/\lambda_i\}_{i=1}^{d}$ are the eigenvalues of $(ED^{-r}(ED^{-r})^*)^{-1}$.

Moreover, by Lemma 3.1.1, we know:

$$\|FD^{r*}\|_{op}^2 = \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} = \max_{i=1,2,\ldots,d}\{1/\lambda_i\}$$

which means, if we fix the order r, and we choose Sobolev dual, then the error $\|x-\widetilde{x}\|$ only depends on the frame matrix E. To answer question 1 in the beginning of this chapter, now we focus on

$$\min_E \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} = \min_E\{\max_{i=1,2,\ldots,d}\{1/\lambda_i\}\} \tag{3.1.3}$$

where E is the frame matrix associated to a unit-norm frames $\{e_n\}_{n=1}^{N} \subset \mathbb{R}^d$.

Lemma 3.1.3. *Let E be the $d \times N$ frame matrix associated to any unit-norm frames $\{e_n\}_{n=1}^{N} \subset \mathbb{R}^d$, D is defined as in (2.1.6), then:*

$$\min_E \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} \geq 1/(\frac{1}{d}\max_E\{tr((ED^{-r}(ED^{-r})^*)\}$$

Proof. Since

$$\min_E \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} = \min_E\{\max_{i=1,2,\ldots,d}\{1/\lambda_i\}\}$$

Besides,

$$\min_E\{\max_{i=1,2,\ldots,d}\{1/\lambda_i\}\} = \min_E\{1/\min_{i=1,2,\ldots,d}\{\lambda_i\}\} = 1/\max_E\{\min_{i=1,2,\ldots,d}\{\lambda_i\}\}.$$

Moreover,

$$\max_E\{\min_{i=1,2,\ldots,d}\{\lambda_i\}\} \leq \max_E\{\frac{1}{d}\sum_{i=1}^{d}\lambda_i\} = \frac{1}{d}\max_E\{tr((ED^{-r}(ED^{-r})^*)\}$$

Hence,

$$\min_E \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} \geq 1/(\frac{1}{d}\max_E\{tr((ED^{-r}(ED^{-r})^*)\}.$$

□

By Lemma 3.1.3, instead of focusing on $\min\|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op}$, we start to look at $\max\{tr((ED^{-r}(ED^{-r})^*)\}$. Firstly, recall the definition of D as:

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix} = I - J$$

here I is the $N \times N$ identity matrix and

$$J = \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & \cdots & 0 & 0 & -1 \\ 0 & \cdots & 0 & 0 & 0 \end{pmatrix}$$

Now, for a fixed r, we define:

$$D^{-r} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N-1} & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N-1} & a_{2,N} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{d-1,1} & a_{d-1,2} & \cdots & a_{d-1,N-1} & a_{d-1,N} \\ a_{d,1} & a_{d,2} & \cdots & a_{d,N-1} & a_{d,N} \end{pmatrix}$$

Then we state:

Lemma 3.1.4. *For a fixed r, N is the size of the unit-norm frame and $N \gg r$. Let matric E and D be defined as above, then:*

$$\max_E \{tr((ED^{-r}(ED^{-r})^*)\} \leq C_r \frac{N^{2r+1}}{(r!)^2(2r+1)}.$$

where $C_r \ll (r!)^2$ is a constant related to r.

Proof. Since $E = [e_1\ e_2\ ...\ e_N]$ and following the definition of D^{-r} we gave above, we have:

$$ED^{-r} = [\sum_{i=1}^{N} a_{1,i}e_i\ \sum_{i=1}^{N} a_{2,i}e_i\ ...\ \sum_{i=1}^{N} a_{N,i}e_i] \tag{3.1.4}$$

Hence,

$$tr(ED^{-r}(ED^{-r})^*) = \sum_{j=1}^{N} \| \sum_{i=1}^{N} a_{i,j}e_i\|^2 \leq \sum_{j=1}^{N}\sum_{i=1}^{N} \|a_{i,j}e_i\|^2 = \sum_{j=1}^{N}\sum_{i=1}^{N} (a_{i,j})^2$$

The last equality holds since $\{e_i\}_{i=1}^N$ are unit-norm frames. The inequality follows from the triangle inequality for vectors and the inequality holds with equality if and only if all of the vectors $a_{i,j}e_i$ are in the same direction. Since $\{e_i\}_{i=1}^N$ are unit-norm frames, then equality holds in 3.1.4 if and only if:

$$e_i = e_j\ for\ any\ i,j \in \{1,2,...,N\} \tag{3.1.5}$$

We shall use that

$$(1-x)^{-r} = 1 + rx + \binom{r+1}{2}x^2 + \binom{r+2}{3}x^3 + ...$$

The sequence above converges if and only if when $|x| < 1$, and if we switch 1 to identify matrix I and switch x to any other matrix X. Then the matrix sequence converges if and only if the spectral radius of X is less then 1.

By induction, for any N, we have the spectral radius of J is 0. Moreover, since $J^N = 0$, then:

$$D^{-r} = (I-J)^{-r} = I + rJ + \binom{r+1}{2}J^2 + ... + \binom{r+N-2}{N-1}J^{N-1}$$

Let $J^k_{i,j}$ denote the elements of J^k, where $k = 1, 2, ..., N-1$. If $i > j$, then $J^k_{i,j} = 0$. Moreover, if $i \leq j$, then,

$$J^k_{i,i+l} = \begin{cases} 1, & l = k; \\ 0, & l \neq k. \end{cases}$$

Thus, we have:

$$a_{i,j} = \begin{cases} 0, & i > j; \\ \binom{r+j-i-1}{j-i}, & i \leq j. \end{cases}$$

Hence,

$$\sum_{i=1}^{N} a_{i,j} = \sum_{i=1}^{j} \binom{r+j-i-1}{j-i} = \binom{r+j-1}{j-1} = \binom{r+j-1}{r}.$$

Now, we get:

$$\max_E\{tr((ED^{-r}(ED^{-r})^*)\} \leq \sum_{j=1}^{N}\sum_{i=1}^{N}(a_{i,j})^2 = \sum_{j=1}^{N}\sum_{i=1}^{N}(\binom{r+j-1}{r})^2$$

Since we fix r and we always assume $N \gg r$, we have:

$$\sum_{j=1}^{N}\sum_{i=1}^{N}\binom{r+j-1}{r}^2 = \frac{1}{(r!)^2}\sum_{j=1}^{N}(j(j+1)\cdots(j+r-1))^2$$

Moreover, for a small positive integer j, we have

$$(j+r-1) < j(r-1),$$

which implies,

$$(j(j+1)\cdots(j+r-1))^2 < (r!)^2 j^{2r},$$

and when $j \gg r$, we have

$$(j+r-1) < c_j j,$$

here c_j is constant related to j and greater but close to 1, Thus

$$(j(j+1)\cdots(j+r-1))^2 < (c_j)^{2r} j^{2r},$$

Hence, there exists a constant C_r related to r, such that,

$$\frac{1}{(r!)^2}\sum_{j=1}^{N}(j(j+1)\cdots(j+r-1))^2 \leq C_r \frac{1}{(r!)^2}\sum_{j=1}^{N} j^{2r},$$

and if $N \gg r$, for most j, we have $(c_j)^{2r} \ll (r!)^2$, then $C_r \ll (r!)^2$. Then,

$$C_r \frac{1}{(r!)^2}\sum_{j=1}^{N} j^{2r} \leq C_r \frac{1}{(r!)^2}\int_0^N x^{2r}dx = C_r \frac{N^{2r+1}}{(r!)^2(2r+1)}$$

□

Theorem 3.1.5. *Fix a positive integer r. Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a unit-norm frame and let E be the associated $d \times N$ frame matrix. Moreover, Let $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ is the r-th order*

Sobolev dual of E and let F be the associated $d \times N$ frame matrix. Then:

$$\min_E\{\|FD^{r*}\|_{op}\} = \mathcal{O}(N^{-(r+\frac{1}{2})})$$

Proof. By the proof for Theorem 2.3.4 in [BLPY10], we have:

$$\min_E\{\|FD^{r*}\|_{op}\} \leq \mathcal{O}(N^{-(r+\frac{1}{2})})$$

On the other hand, by Lemma 3.1.3, we have :

$$\min_E\{\|FD^{r*}\|_{op}\} = \min_E \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} \geq 1/(\frac{1}{d}\max_E\{tr((ED^{-r}(ED^{-r})^*)\}$$

Since in Lemma 3.1.4, we get:

$$\max_E\{tr((ED^{-r}(ED^{-r})^*)\} \leq C_r\frac{N^{2r+1}}{(r!)^2(2r+1)}$$

Then we have:

$$1/(\frac{1}{d}\max_E\{tr((ED^{-r}(ED^{-r})^*)\} \geq \frac{(r!)^2(2r+1)}{dC_rN^{2r+1}}$$

Since we assume $N \gg r$ and $N \gg d$, we have:

$$\frac{(r!)^2(2r+1)}{dC_rN^{2r+1}} = \mathcal{O}(N^{-(r+\frac{1}{2})})$$

Which implies:

$$\min_E\{\|FD^{r*}\|_{op}\} \geq \mathcal{O}(N^{-(r+\frac{1}{2})})$$

Hence,

$$\min_E\{\|FD^{r*}\|_{op}\} = \mathcal{O}(N^{-(r+\frac{1}{2})})$$

□

3.2 Optimizing the Frobenius norm

If the $\Sigma\Delta$ state variables are modeled as i.i.d. random variables with zero mean and variance σ^2 than the expected error squared in (2.3.1) become

$$\mathbb{E}\|F(D^*)^r(u)\|^2 = \sigma^2\|FD^{r*}\|^2_{\mathscr{F}}.$$

Motivated by this, in this section we shall analyze $\|F\|_{r,\mathscr{F}}$ is a similar manner as $\|F\|_{r,op}$ in previous section.

Lemma 3.2.1. *Fix a positive integer r. Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a unit-norm frame and let E be the associated $d \times N$ frame matrix. Moreover, let $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ be the r-th order Sobolev dual of E and let F be the associated $d \times N$ frame matrix. Then:*

$$\|FD^{r*}\|^2_{\mathscr{F}} = tr((ED^{-r}(ED^{-r})^*)^{-1}) \tag{3.2.1}$$

Proof. By the definition of Soblev dual, we have:

$\|FD^{r*}\|^2_{\mathscr{F}}$

$= tr((FD^{r*})(FD^{r*})^*)$

$= tr((ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r}((ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r})^*)$

$= tr((ED^{-r}(D^*)^{-r}E^*)^{-1})^*)$

$= tr((ED^{-r}(ED^{-r})^*)^{-1})$ □

We still let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $ED^{-r}(ED^{-r})^*$. By lemma 3.1.2, we have $\{1/\lambda_i\}_{i=1}^d$ are the eigenvalues of $(ED^{-r}(ED^{-r})^*)^{-1}$.

Moreover, by lemma 3.2.1, we know:

$$\|FD^{r*}\|^2_{\mathscr{F}} = tr((ED^{-r}(ED^{-r})^*)^{-1}) = \sum_{i=1}^d 1/\lambda_i$$

Hence, now we focus on

$$\min_E\{tr((ED^{-r}(ED^{-r})^*)^{-1}\} = \min_E\{\sum_{i=1}^{d} 1/\lambda_i\} \tag{3.2.2}$$

where E is the frame matrix associated to a unit-norm frames $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$.

Lemma 3.2.2. *Let E be the $d \times N$ frame matrix associated to any unit-norm frame $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$, D is defined as in (2.1.6), then:*

$$\min_E\{tr((ED^{-r}(ED^{-r})^*)^{-1}\} \geq 1/(\frac{1}{d}\max_E\{tr((ED^{-r}(ED^{-r})^*)\}.$$

The proof follows from

$$\sum_{i=1}^{d} 1/\lambda_i \geq \max_{i=1,2,\ldots,d}\{1/\lambda_i\}$$

Hence,

$$\min_E\{tr((ED^{-r}(ED^{-r})^*)^{-1}\} \geq \min_E \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} \geq 1/(\frac{1}{d}\max_E\{tr((ED^{-r}(ED^{-r})^*)\}$$

Theorem 3.2.3. *Fix a positive integer r. Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a unit-norm frame and let E be the associated $d \times N$ frame matrix. Moreover, Let $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ is the r-th order Sobolev dual of E and let F be the associated $d \times N$ frame matrix. Then:*

$$\min_E\{\|FD^{r*}\|_{\mathscr{F}}\} = \mathscr{O}(N^{-(r+\frac{1}{2})})$$

Proof. Since,

$$\|FD^{r*}\|_{op}^2 = \|(ED^{-r}(ED^{-r})^*)^{-1}\|_{op} = \max_{i=1,2,\ldots,d}\{1/\lambda_i\}$$

Then following the proof for Theorem 2.3.4 in [BLPY10], we have:

$$\min_{E}\{\|FD^{r*}\|_{op}\} \leq \mathscr{O}(N^{-(r+\frac{1}{2})})$$

which means,

$$\min_{E}\{\max_{i=1,2,\ldots,d}\{1/\lambda_i\}\} \leq \mathscr{O}(N^{-(r+\frac{1}{2})})$$

Since,

$$\min_{E}\{\sum_{i=1}^{d} 1/\lambda_i\} \leq d\min_{E}\{\max_{i=1,2,\ldots,d}\{1/\lambda_i\}\}$$

Following the equation3.2.2 and as we assume $d \ll N$, we have

$$\min_{E}\{\|FD^{r*}\|_{\mathscr{F}}\} \leq d\mathscr{O}(N^{-(r+\frac{1}{2})}) = \mathscr{O}(N^{-(r+\frac{1}{2})})$$

Then, combining the result of Lemma 3.2.2 and Lemma 3.1.4, we get,

$$\min_{E}\{\|FD^{r*}\|_{\mathscr{F}}\} \geq \mathscr{O}(N^{-(r+\frac{1}{2})}).$$

Hence,

$$\min_{E}\{\|FD^{r*}\|_{\mathscr{F}}\} = \mathscr{O}(N^{-(r+\frac{1}{2})}).$$

□

Chapter 4

Sigma-Delta Quantization with Fusion Frames

In Chapter 1, we gave a review for finite frame theory and fusion frame theory. Then, in Chapter 2 we gave motivation for using Sigma-Delta ($\Sigma\Delta$) quantization in A/D conversion instead of PCM quantization. We also introduced how to apply $\Sigma\Delta$ quantization in the setting of finite frames. In this chapter, we shall develop and analyze Sigma-Delta algorithms for fusion frames.

In [CKL08], the authors gave the following distributed reconstruction formula that depends on the equation (1.3.4).

Definition 4.0.4. *Let $\{W_i, w_i, \{f_{ij}\}_{j\in J_i}\}_{i\in I}$ be a fusion frame system for $\mathbb{H}$ with fusion frame bounds C and D, and let $\{\{\widetilde{f}_{ij}\}_{j\in J_i}\}_{i\in I}$ be associated local dual frames. Then for any signal $x \in \mathbb{H}$, we could reconstruct the signal as:*

$$x = \sum_{i\in I}\sum_{j\in J_i} \langle x, w_i f_{i,j}\rangle (S_{W,w}^{-1} w_i \widetilde{f}_{ij}). \tag{4.0.1}$$

For the distributed reconstruction in (4.0.1), we can easily apply $\Sigma\Delta$ quantization since it is essentially the same we using $\Sigma\Delta$ quantization with the frame $\{\{w_i f_{ij}\}_{j\in J_i}\}_{i\in I}$ and the associated dual frame $\{\{S_{W,w}^{-1} w_i \widetilde{f}_{ij}\}_{j\in J_i}\}_{i\in I}$.

However, sometimes it is difficult to construct the fusion frame system in practical problems. We might only have access to the vector $\pi_{W_i}(x)$ instead of its frame representation. In these situations, we do not have the analog scalar coefficients $\langle x, w_i f_{i,j}\rangle$ and are not able to directly apply $\Sigma\Delta$ quantization to fusion frames as we did with finite frames.

An important property of $\Sigma\Delta$ quantization is that there exist stable 1-bit $\Sigma\Delta$ algorithms. This issue become more complicated in the setting of fusion frames. For each subspace W_i, suppose we apply 1-bit $\Sigma\Delta$ quantization to a frame for W_i, then at least $\dim(W_i)$ bits are needed. Then for the whole fusion frame, we have to spend $\sum_{i\in I}\dim(W_i)$ bits. But actually, we can spend less bits if we quantize the vector coefficients $\pi_{W_i}(x)$.

Following the reasons above, in this chapter, instead of using dual fusion frame or distributed reconstruction, we define the left inverse operator of a fusion frame and will apply this to the quantization problem. By defining the left inverse and the canonical left inverse for fusion frames, we prove the property that the canonical left inverse has the minimal $\|\cdot\|_{OP}$ and $\|\cdot\|_{\mathscr{F}}$. Then we construct a stable first-order and high-order $\Sigma\Delta$ quantization algorithm to quantize vector valued fusion frame coefficients. Then we will prove our first-order $\Sigma\Delta$ quantization algorithm can offer an improvement in providing low bit representation for each subspace. Besides, we will give a algorithm to calculate the Kashin representations for fusion frames to improve the performance of the high-order $\Sigma\Delta$ quantization algorithm. In the last subsection, we will give the definition of Sobolev left inverse and prove it leads to the minimal squared error.

4.1 Left inverse operator of fusion frames

Let $\{(W_i, w_i)\}_{i\in I}$ be a fusion frames. Recall that $S_{W,w}$ is a positive, selfadjoint invertible operator on $\mathbb{H}$, see [CK03]. Also since $S_{W,w}^{-1}S_{W,w} = I_{\mathbb{H}}$, we have the following reconstruction formula (1.3.4).

$$f = \sum_{j\in I} w_j^2 S_{W,w}^{-1}\pi_{W_j}(f),\ for\ all\ f \in \mathbb{H}.$$

It can be also written in this way:

$$(S_{W,w}^{-1}T_{W,w})T_{W,w}^* = I_{\mathbb{H}}.$$

Here $S_{W,w}^{-1}T_{W,w}$ is a left inverse of $T_{W,w}^*$.

Definition 4.1.1. *Let $\{(W_i, w_i)\}_{i\in I}$ be a fusion frame for $\mathbb{H}$. $T_{W,w}^*$ is the analysis operator. Let F be a bounded linear operator from $\sum_{i\in I}\bigoplus W_i$ to $\mathbb{H}$. If F satisfies:*

$$FT_{W,w}^* = I_{\mathbb{H}},$$

then we call F a left inverse of the fusion frame $\{(W_i, w_i)\}$. Also, let $\widetilde{F} = (T_{W,w}T_{W,w}^*)^{-1}T_{W,w}$. Then we call $\widetilde{F}$ the canonical left inverse of $T_{W,w}^*$.

About the canonical left inverse of $T_{W,w}^*$, there are some properties about it can be checked in [HMBZ13].

Theorem 4.1.2. *Let $\{(W_i, w_i)\}_{i\in I}$ be a fusion frames for $\mathbb{H}$. $T_{W,w}^*$ is the analysis operator. A general formula for F which is the left inverse of $T_{W,w}^*$ is given by*

$$F = (T_{W,w}T_{W,w}^*)^{-1}T_{W,w} + G^*(I - T_{W,w}^*(T_{W,w}T_{W,w}^*)^{-1}T_{W,w}), \tag{4.1.1}$$

where G^* is the adjoint operator of an arbitrary bounded linear operator

$$G^* : \sum_{i\in I}\bigoplus W_i \longrightarrow \mathbb{H}$$

.

Proof. Since $T_{W,w}T^*_{W,w} = S$ and as we know $S^{-1}_{W,w}T_{W,w}$ is the Left Inverse of $T^*_{W,w}$, we multiply $T^*_{W,w}$ on the right of $(4.1.1)$, then we obtain

$$FT^*_{W,w} = (T_{W,w}T^*_{W,w})^{-1}T_{W,w}T^*_{W,w} + G^*(T^*_{W,w} - T^*_{W,w}(T_{W,w}T^*_{W,w})^{-1}T_{W,w}T^*_{W,w})$$

which implies,

$$FT^*_{W,w} = I_{\mathbb{H}} + G^*(T^*_{W,w} - T^*_{W,w}) = I_{\mathbb{H}}.$$

And to see why $(4.1.1)$ is the most general formula for the left inverse of $T^*_{W,w}$. Let F_0 be any bounded left inverse of $T^*_{W,w}$, we shall show that there is a G^* such that $(4.1.1)$ yields F_0. For doing this, take $G^* = F_0$, then the right hand side of $(4.1.1)$ becomes

$$(T_{W,w}T^*_{W,w})^{-1}T_{W,w} + F_0(I - T^*_{W,w}(T_{W,w}T^*_{W,w})^{-1}T_{W,w}) = F_0$$

□

Theorem 4.1.3. *Let $\{(W_i,w_i)\}_{i\in I}$ be a fusion frame for $\mathbb{H}$. $T^*_{W,w}$ is the analysis operator. $\widetilde{F} = (T_{W,w}T^*_{W,w})^{-1}T_{W,w}$ is the canonical left inverse of $T^*_{W,w}$. Then for any left inverse F of $T^*_{W,w}$, We have $\|\widetilde{F}\|_{OP} \leq \|F\|_{OP}$. Here $\|F\|_{OP} = \sup_{\|x\|=1}\|Fx\|$*

Proof. By Theorem 4.1.2, F has the general form

$$F = (T_{W,w}T^*_{W,w})^{-1}T_{W,w} + G^*(I - T^*_{W,w}(T_{W,w}T^*_{W,w})^{-1}T_{W,w}).$$

We can write F in the form $F = \widetilde{F} + Z$, where Z satisfies $ZT^*_{W,w} = 0$. It also follows that

$$\widetilde{F}Z^* = 0 \;\; and \;\; Z\widetilde{F}^* = 0.$$

Thus, by Theorem 2.13 in [Hei10]

$$\|F\|_{OP}^2 = \|(\widetilde{F}+Z)(\widetilde{F}^*+Z^*)\|_{OP} = \|\widetilde{F}\widetilde{F}^*+Z\widetilde{F}^*+\widetilde{F}Z^*+ZZ^*\|_{OP} = \|\widetilde{F}\widetilde{F}^*+ZZ^*\|_{OP}.$$

Therefore,

$$\|F\|_{OP}^2 = \|\widetilde{F}\widetilde{F}^*+ZZ^*\|_{OP} \leq \|\widetilde{F}\widetilde{F}^*\|_{OP}+\|ZZ^*\|_{OP}.$$

Moreover, since $\widetilde{F}\widetilde{F}^*$ and ZZ^* are both positive operator, then by Theorem 2.15 in [Hei10], we know

$$\|\widetilde{F}\widetilde{F}^*+ZZ^*\|_{OP} = \sup_{\|x\|=1} |\langle(\widetilde{F}\widetilde{F}^*+ZZ^*)(x),x\rangle|$$

Let $\|\widetilde{F}\widetilde{F}^*\|_{OP} = |\langle\widetilde{F}\widetilde{F}^*(x_0),x_0\rangle|$, where $\|x_0\| = 1$ and $x_0 \in \mathbb{H}$, then

$$\|F\|_{OP}^2 = \sup_{\|x\|=1} |\langle(\widetilde{F}\widetilde{F}^*+ZZ^*)(x_0),x\rangle| \geq |\langle\widetilde{F}\widetilde{F}^*(x_0),x_0\rangle+\langle ZZ^*(x_0),x_0\rangle| \geq \|\widetilde{F}\widetilde{F}^*\|_{OP}.$$

Hence,$\|F\|_{OP}$ is minimal when $ZZ^*(x_0) = 0$ and $F = \widetilde{F}$ is the canonical left inverse. Thus the minimized left inverse for operator norm is not unique. □

Following the similar idea, we state the theorem for the Frobenius norm of canonical left inverse

Theorem 4.1.4. *Let $\{(W_i,w_i)\}_{i\in I}$ be a fusion frame for $\mathbb{H}$. $T_{W,w}^*$ is the analysis operator. $\widetilde{F} = (T_{W,w}T_{W,w}^*)^{-1}T_{W,w}$ is the canonical left inverse of $T_{W,w}^*$. Then for any left inverse F of $T_{W,w}^*$, we have $\|\widetilde{F}\|_{\mathscr{F}} \leq \|F\|_{\mathscr{F}}$. Here the Frobenius norm if defined by*

$$\|F\|_{\mathscr{F}} = \sqrt{trace(F^*F)} = \sqrt{trace(FF^*)}.$$

Proof. By Theorem 4.1.2, F has the form

$$F = (T_{W,w}T_{W,w}^*)^{-1}T_{W,w}+G^*(I-T_{W,w}^*(T_{W,w}T_{W,w}^*)^{-1}T_{W,w}).$$

We can write F by the form $F = \widetilde{F} + Z$, where Z satisfies $ZT^*_{W,w} = 0$. It also follows that

$$\widetilde{F}Z^* = 0 \;\; and \;\; Z\widetilde{F}^* = 0.$$

Thus,

$$\|F\|^2_{\mathscr{F}} = trace(FF^*) = trace\left((\widetilde{F} + Z)(\widetilde{F}^* + Z^*)\right) = trace(\widetilde{F}\widetilde{F}^* + Z\widetilde{F}^* + \widetilde{F}Z^* + ZZ^*)$$

which implies,

$$\|F\|^2_{\mathscr{F}} = trace(\widetilde{F}\widetilde{F}^* + ZZ^*).$$

Therefore, following the property of Frobenius norm, we have

$$\|F\|^2_{\mathscr{F}} = trace(\widetilde{F}\widetilde{F}^* + ZZ^*) = \|\widetilde{F}\|_{\mathscr{F}} + \|Z\|_{\mathscr{F}}$$

Hence,$\|F\|_{\mathscr{F}}$ is minimal when $Z = 0$ and $F = \widetilde{F}$ is the canonical dual left inverse. Thus the minimized left inverse for Frobenius norm is unique. □

4.2 First-order sigma-delta quantization for fusion frames

After we define the left inverse operator of fusion frame, now we can start to consider a nontrivial $\Sigma\Delta$ quantization with fusion frame. Here we always assume the set I is finite and $\mathbb{H}$ is a finite Hilbert space($\mathbb{R}^d$).

Let $\{(W_n, w_n)\}_{n\in I}$ be a fusion frame for $\mathbb{H}$, π_{W_n} is the orthogonal projection from $\mathbb{H}$ to the subspace W_n and $f \in \mathbb{H}$ is the signal we want to quantize. We give the algorithm for first-order $\Sigma\Delta$ quantization of fusion frames as following:

$$q_n = Q_n(\pi_{W_n}(u_{n-1}) + x_n),$$

$$u_n = \pi_{W_n}(u_{n-1}) + x_n - q_n, \tag{4.2.1}$$

Here, $x_n = \pi_{W_n}(f)$ and $Q_n : W_n \to \mathscr{A}_n$ is a vector quantizer, where $\mathscr{A}_n \in W_n$ is the quantization alphabet of vectors, and $u_0 = 0$.

For $\Sigma\Delta$ quantization of fusion frames, the question is how to define the scalar quantizer Q_n. Our goal is to define Q_n so that u_n is uniform bounded and so that $|R(Q_n)|$as small as possible ($|R(\cdot)|$ means cardinality of the range of the image of Q_n). Suppose the subspace W_n is an M_n dimension subspace of $\mathbb{H}$. We firstly prove that when $|R(Q_n)| \approx 1.57^{M_n}$, we can keep u_n uniformly bounded. For proving this, we need to give the concept of epsilon-nets.

Definition 4.2.1. *Let $(\mathscr{X}, d)$ be a metric space and let $\varepsilon > 0$. A subset $\mathscr{N}_\varepsilon$ of $\mathscr{X}$ is called an $\varepsilon - net$ of $\mathscr{X}$ if every point $x \in \mathscr{X}$ can be approximated to within ε by some point $y \in \mathscr{N}_\varepsilon$, i.e so that $d(x,y) \leq \varepsilon$.*

Theorem 4.2.2. *Given $\varepsilon > 0$, the unit Euclidean sphere $\mathbb{S}^{n-1}$ equipped with the Euclidean metric has an $\varepsilon - net$ of cardinality at most $(1+\frac{2}{\varepsilon})^n$.*

The proof and more details about nets can be checked in [Ver10].

Corollary 4.2.3. *In $\mathbb{R}^n$, for any angle $\theta > 0$, there exists a set of unit vectors $\mathscr{N}_\theta$, such that* *for any vector $v \in \mathbb{R}^n$, there exits a vector $u \in \mathscr{N}_\theta$, such that the angle between v and u is less than θ, and the cardinality of $\mathscr{N}_\theta$ is at most $(1+\frac{2}{\sin(\theta)})^n$.*

Proof. Let $\varepsilon = \sin(\theta)$, then by Theorem 4.2.2 there exists an $\varepsilon - net$ $\mathscr{N}_\varepsilon$ with cardinality $\mathscr{N}(\mathbb{R}^N, \theta)$ at most $(1+\frac{2}{\sin(\theta)})^n$. Then let

$$\mathscr{N}_\theta = \{u | u = \overrightarrow{x-0}, x \in \mathscr{N}_\varepsilon\}.$$

By the definition of $\varepsilon - net$, we know for any vector $v \in \mathbb{R}^n$ and any $u \in \mathscr{N}_\theta$, we have

$$|u-v| \leq \sin(\theta),$$

which means the angle between between v and u at most θ. □

After we have the concept of nets, we can give the first definition of quantizer Q_n as following:

Theorem 4.2.4. *Let $\{(W_n, w_n)\}_{n\in I}$ be a fusion frame for $\mathbb{H}$ and $\mathscr{N}_{\frac{\pi}{3},n} = \{e_{i,n}\}_{i=1}^{\mathscr{N}(\mathbb{R}^n,\frac{\pi}{3})}$ be unit vectors set in W_n constructed in Corollary 4.2.3. The algorithm is defined by (4.2.1). For any $x \in W_n$, we define:*

$$Q_n(x) = e_{i,n},\ i = \min\{k \in \mathscr{N}(\mathbb{R}^n, \frac{\pi}{3}) : \langle x/|x|, e_{k,n}\rangle \geq \langle x/|x|, e_{j,n}\rangle\ for\ all\ j \neq k\}$$

and

$$Q_n(x) = e_{1,n},\ if\ x = 0.$$

Then if $\|x\| \leq \delta < \frac{1}{2}$, for all n, we have:

$$\|u_n\| \leq C = \max\{\frac{\delta^2 - \delta + 1}{1-2\delta}, 1\}$$

Proof. Let $a_k = \|\pi_{W_k}(u_{k-1}) + x_k\|$, $b_k = \|q_k\|$, $c_k = \|u_k\|$, θ_k is the angle between vector

$\pi_{W_k}(u_{k-1}) + x_k$ and vector q_k. Then by the definition of $\mathcal{N}_{\frac{\pi}{3},n}$, we know $\theta_k \leq \frac{\pi}{3}$.

We prove the theorem by induction. When $n = 1$, since $u_0 = 0$ and $\theta_1 < \frac{\pi}{3}$, then by the law of cosines, we have:

$$c_1^2 = a_1^2 + b_1^2 - 2\cos(\theta_1)a_1b_1 \leq a_1^2 + b_1^2 - a_1b_1.$$

Since $|a_1| \leq |\delta + 0| = \delta < \frac{1}{2}$ and $b_1 = 1$. Thus:

$$a_1^2 + b_1^2 - a_1b_1 = a_1^2 + 1 - a_1 = (a_1 - \frac{1}{2})^2 + \frac{3}{4} \leq \frac{1}{4} + \frac{3}{4} = 1$$

then we get

$$|u_1| \leq C = \max\{\frac{\delta^2 - \delta + 1}{1 - 2\delta}, 1\}$$

so when $n = 1$ the result is true.

Suppose for all $n \leq k - 1$, the result is true, then when $n = k$, by Law of cosines, we have:

$$c_k^2 = a_k^2 + b_k^2 - 2\cos(\theta_k)a_kb_k.$$

Since $\theta_k \leq \frac{\pi}{3}$ and $b_k = 1$, so we have:

$$c_k^2 = a_k^2 + 1 - 2\cos(\theta_k)a_k \leq a_k^2 + 1 - 2\cos(\frac{\pi}{3})a_k \leq a_k^2 + 1 - a_k.$$

Consider about the function:$f(x) = x^2 + 1 - x$, since $f'(x) = 2x - 1$ and $f''(x) = 2 > 0$, then f(x) has minimum at $x = \frac{1}{2}$. By the inductive assumption, $0 < a_k \leq C + \delta$, so for $x \in [0, C + \delta]$, $f(x)$ has its maximum value at $f(0)$ or $f(C + \delta)$. Now, we claim

$$f(x) \leq C = \max\{\frac{\delta^2 - \delta + 1}{1 - 2\delta}, 1\}, for\, x \in (0, C + \delta). \qquad (4.2.2)$$

For $f(0)$, we have $f(0) = 1$.

For $f(C+\delta)$, if we can prove:

$$C^2 \geq (C+\delta)^2 + 1 - (C+\delta) \tag{4.2.3}$$

then we know our claim (4.2.2) is true. Since the inequality (4.2.3) is equivalent to:

$$C^2 + \delta^2 + 2C\delta + 1 - (C+\delta) \leq C^2$$

which is equivalent to:

$$2C(\delta - \frac{1}{2}) \leq -\delta^2 + \delta - 1$$

we know $(\delta - \frac{1}{2}) < 0$, then the inequality (4.2.3) becomes:

$$C \geq \frac{\delta^2 - \delta + 1}{1 - 2\delta},$$

which follows from the definition of C:

$$C = \max\{\frac{\delta^2 - \delta + 1}{1 - 2\delta}, 1\}.$$

Thus we have $C \geq |u_k|$. Hence by induction, we prove the Theorem. □

By Corollary 4.2.3, we know

$$|R_{Q_n}| = \mathscr{N}(\mathbb{R}^{M_n}, \frac{\pi}{3}) \leq (1 + \frac{2}{\frac{\sqrt{3}}{2}})^{M_n} \approx 3.3^{M_n}.$$

Thus for each subspace we need $\log_2(3.3^{M_n}) \preceq 1.7M_n$ bits to quantize the signal. Actually, we can show that to keep $|u_n|$ have uniform bound, we just need $\log_2(M_n + 1)$ bits. Moreover, it is the best we can get. For proving this, we need several lemmas.

Lemma 4.2.5. *In $\mathbb{R}^d$, there exists a set of vectors $\{e_i\}_{i=1}^{d+1}$ such that:*

$$\langle e_i, e_i \rangle = 1, \ \ \langle e_i, e_j \rangle = c = \cos\theta, \ for\ any\ i \neq j, \ \ \sum_{i=1}^{d+1} e_i = 0$$

Moreover, $\theta = \arccos(-\frac{1}{d})$.

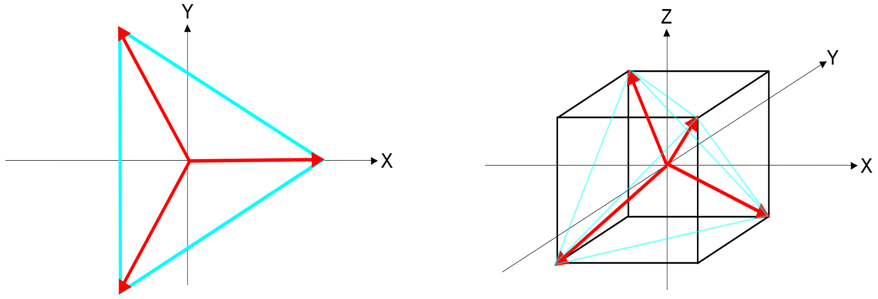


Figure 4.1: Example in $\mathbb{R}^2$ and $\mathbb{R}^3$

Proof. Consider the subspace

$$A = \{(x_1, x_2, \ldots, x_{d+1}) | \sum_{i=1}^{d+1} x_i = 0\} \in \mathbb{R}^{d+1}.$$

There exists a 1-to-1 invertible mapping from A to $\mathbb{R}^d$ which preserves the Euclidean metric. Hence, A is isomorphic to $\mathbb{R}^d$. Then let

$$V_i = (0, 0, \ldots, 1, \ldots, 0) \ \ for \ \ i = 1, 2, \ldots, d+1$$

where the i-th coordinate is 1 and the rest are all 0, and let

$$V_0 = (\frac{1}{d+1}, \frac{1}{d+1}, \ldots, \frac{1}{d+1}).$$

Let $u_i = V_i - V_0$ for $i = 1, 2, \ldots, d+1$, then we know $\{u_i\}$ is a subset in A.

For any $i \neq j$,

$$\langle u_i, u_j \rangle = \langle V_i - V_0, V_j - V_0 \rangle = \langle V_i, V_j \rangle - \langle V_i, V_0 \rangle - \langle V_0, V_j \rangle + \langle V_0, V_0 \rangle = -\frac{1}{d+1} = C_1,$$

here C_1 is a constant.

For all $i = 1, 2, \ldots, d+1$,

$$\langle u_i, u_i \rangle = (1 - \frac{1}{d+1})^2 + \frac{d}{(d+1)^2} = \frac{d}{d+1} = C_2,$$

here C_2 is a constant.

Also,

$$\sum_{i=1}^{d+1} u_i = \sum_{i=1}^{d+1} V_i - (d+1)V_0 = (1, 1, \ldots, 1) - (1, 1, \ldots, 1) = 0.$$

Let $e_i = \frac{u_i}{|u_i|} = \frac{u_i}{\sqrt{C_2}}$.

Firstly, $\{e_i\}$ is also a subset in A, and for all $i = 1, 2, \ldots, d+1$, we know $|e_i| = 1$.

Secondly, for any $i \neq j$,

$$\langle e_i, e_j \rangle = \langle \frac{u_i}{|u_i|}, \frac{u_j}{|u_j|} \rangle = \frac{C_1}{C_2} = C.$$

Lastly,

$$\sum_{i=1}^{d+1} e_i = \sum_{i=1}^{d+1} \frac{u_i}{\sqrt{C_2}} = 0 * \sqrt{C_2} = 0$$

Hence, $\{e_i\}$ exists has the designed properties. Additionally, since $\sum_{i=1}^{d+1} e_i = 0$, then for any e_j, $\sum_{i=1}^{d+1} \langle e_i, e_j \rangle = 0$ which is equivalent to:

$$dc + 1 = 0$$

.

Hence $c = -\frac{1}{d}$, which implies $\theta = \arccos(-\frac{1}{d})$. □

Lemma 4.2.6. *Let $\{e_i\}_{i=1}^{d+1} \in \mathbb{R}^d$ be the vectors set in Lemma 4.2.5. For any unit vector*

$u \in \mathbb{R}^d$, we define the vector function $Q(u)$ as following:

$$Q(u) = e_i,\ if\ \langle x, e_i \rangle\ \geq\ \langle x, e_j \rangle\ for\ all\ i \neq j$$

Let $\theta = \arccos\langle u, Q(u)\rangle$, then $\theta \leq \pi - \arccos(-\frac{1}{d})$.

Proof. Let $u = (x_1, x_2, \ldots, x_{d+1})$ be a unit vector in A, here A is the same as in Lemma 4.2.5. As we know, there exists a 1-to-1 invertible mapping from A to $\mathbb{R}^d$ which preserves the Euclidean metric, thus we can consider u as a unit vector in $\mathbb{R}^d$. Then we have:

$$\sum_{i=1}^{d+1} x_i = 0;\ \ \sum_{i=1}^{d+1} x_i^2 = 1 \tag{4.2.4}$$

Let $\{e_i\}$ be the same as in Lemma 4.2.5. For $i = 1, 2, \ldots, d+1$, let

$$\langle u, e_i \rangle = (-\frac{1}{d+1}\sum_{j=1}^{d+1} x_j + x_i)/C_2 = x_i/C_2.$$

Then by definition of $Q(u)$, we have

$$\langle u, Q(u) \rangle = \max\{x_i\}/C_2.$$

Since u is a unit vector, then

$$\theta = \arccos\langle u, Q(u)\rangle \leq \pi - \arccos(-\frac{1}{d}),$$

if and only if,

$$\langle u, Q(u) \rangle = \max\{x_i\}/C_2 \geq \frac{1}{d}.$$

Hence, to prove $\theta \leq \pi - \arccos(-\frac{1}{d})$ is equivalent to prove

$$\min\max\{x_i\} = \frac{C_2}{d} = \sqrt{\frac{1}{d(d+1)}},$$

here $\{x_i\}$ satisfy equation (4.2.4).

Without loss of generality, we suppose $x_1 \geq x_2 \geq \ldots \geq x_{d+1}$.

First step, if $x_1 \neq x_2$, then we let

$$x_1' = \frac{x_1 + x_2}{2},\ x_2' = \frac{x_1 + x_2}{2},\ x_i' = x_i,\ for\ i = 3, 4, \ldots, d+1.$$

Firstly we know

$$\sum_{i=1}^{d+1} x_i' = \sum_{i=1}^{d+1} x_i = 0.$$

Also since $x_1^2 + x_2^2 \geq x_1'^2 + x_2'^2$, we get

$$\sum_{i=1}^{d+1} x_i'^2 \leq \sum_{i=1}^{d+1} x_i^2 = 1.$$

Let

$$c = \sum_{i=1}^{d+1} x_i'^2\ and\ let\ x_i^1 = x_i'/c\ for\ i = 1, 2, \ldots, d+1.$$

Then for $\{x_i^1\}$, we know $\sum_{i=1}^{d+1} x_i^1 = \frac{1}{c} \sum_{i=1}^{d+1} x_i^1 = 0$ and $\sum_{i=1}^{d+1} {x_i^1}^2 = 1$. Since $c \leq 1$, also we still have $x_1^1 \geq x_2^1 \geq \ldots \geq x_{d+1}^1$, then we get $x_i^1 \geq x_i$ for all $i = 2, 3, \ldots, d+1$, Hence $x_1^1 \leq x_1$. Thus $\max\{x_i\} \geq \max\{x_i^1\}$.

If $x_1 = x_2$, we just let $x_i^1 = x_i$ for all $i = 1, 2, \ldots, d+1$. Thus we have

$$\max\{x_i\} \geq \max\{x_i^1\}.$$

Second step, by same method we can get $\{x_i^2\}_{i=1}^{d+1}$ satisfy that equation(4.2.4), $x_1^2 = x_2^2 = x_3^2 \geq x_4^2 \geq \ldots \geq x_{d+1}^2$ and $\max\{x_i^1\} \geq \max\{x_i^2\}$.

We can keep doing the same thing for $d-1$ times until we get $\{x_i^{d-1}\}_{i=1}^{d+1}$ satisfy that equation(4.2.4), $x_1^{d-1} = x_2^{d-1} = \ldots = x_d^{d-1} \geq x_{d+1}^2$ and $\max\{x_i^{d-1}\} \leq \max\{x_i^j\}$ for all $j = 1, 2, \ldots, d-2$. Since $\sum_{i=1}^{d+1} x_i^{d-1} = 0$, then we can not do the next step. Thus $\min \max\{x_i\} = \max\{x_i^{d-1}\}$.

Let $a = x_1^{d-1}$ and $b = x_{d+1}^{d-1}$, by equation (4.2.4), we have:

$$da^2 + b^2 = 1, \quad ad + b = 0$$

By solving the equations above, we get $a = \sqrt{\frac{1}{d(d+1)}}$ which is we want.

□

After all the preparation, now we give the definition of the new quantizer Q_n as following and prove the stability of the scheme.

Theorem 4.2.7. *Let $\{(W_n, w_n)\}_{n \in I}$ be a fusion frame for $\mathbb{H}$ where $\dim(W_n) = M_n$ and let $\{e_{i,n}\}_{i=1}^{M_n+1}$ be the set of unit vectors in W_n which we constructed in Lemma 4.2.5. The algorithm is defined as (4.2.1). For any $x \in W_n$, we define:*

$$Q_n(x) = e_{i,n},\ i = \min\{k \in \{1,2,...,M_n+1\} : \langle x/\|x\|, e_{k,n}\rangle \geq \langle x/\|x\|, e_{j,n}\rangle\ for\ all\ j \neq k\}$$

and

$$Q_n(x) = e_{1,n},\ if\ x = 0$$

Let $d = \max\{\dim(W_n)\}$, $\theta = \pi - \arccos(-\frac{1}{d})$. If $\|x\| \leq \delta < \cos(\theta) = \frac{1}{d}$. Then for all n, we have:

$$\|u_n\| \leq C = \max\{\frac{\delta^2 - 2\cos(\theta)\delta + 1}{2(\cos(\theta) - \delta)}, 1\}$$

Proof. Let $a_k = \|\pi_{W_k}(u_{k-1}) + x_k\|$, $b_k = \|q_k\|$,$c_k = \|u_k\|$, θ_k is the angle between vector $\pi_{W_k}(u_{k-1}) + x_k$ and vector q_k. Then by Lemma 4.2.6, we know $\theta_k \leq \theta = \pi - \arccos(-\frac{1}{d})$. Since $d \geq 2$, thus $\theta \in [\frac{\pi}{3}, \frac{\pi}{2})$

We prove the theorem by induction, when $n = 1$, since $u_0 = 0$ and $\theta_1 < \frac{\pi}{2}$,then by the law of cosines,we have:

$$c_1^2 = a_1^2 + b_1^2 - 2\cos(\theta_1)a_1b_1 \leq a_1^2 + b_1^2 - 2\cos(\theta)a_1b_1$$

Since $|a_1| \leq |\delta + 0| = \delta < \cos(\theta)$ and $b_1 = 1$, then

$$a_1^2 + b_1^2 - 2\cos(\theta)a_1b_1 = a_1^2 + 1 - 2\cos(\theta)a_1 = (a_1 - \cos(\theta))^2 + 1 - \cos^2(\theta)$$

which implies,

$$a_1^2 + b_1^2 - 2\cos(\theta)a_1b_1 \leq \cos^2(\theta) + 1 - \cos^2(\theta) = 1$$

which means:

$$|u_1| \leq C = \max\{\frac{\delta^2 - 2\cos(\theta)\delta + 1}{2(\cos(\theta) - \delta)}, 1\}$$

so when $n = 1$ the result is true.

Suppose for all $n \leq k - 1$, the result is true, then when $n = k$, by the Law of cosines, we have:

$$c_k^2 = a_k^2 + b_k^2 - 2\cos(\theta_k)a_kb_k.$$

Since $\theta_k \leq \theta < \frac{\pi}{2}$ and $b_k = 1$, then we get:

$$|c_k|^2 = a_k^2 + 1 - 2\cos(\theta_k)a_k \leq a_k^2 + 1 - 2\cos(\theta)a_k$$

Consider the function: $f(x) = x^2 + 1 - 2\cos(\theta)x$, since $f'(x) = 2x - 2\cos(\theta)$ and $f''(x) = \cos(\theta) > 0$, then f(x) has its minimum at $x = \cos(\theta)$. Since $0 < a_k \leq C + \delta$, then for $x \in [0, C + \delta]$, f is maximal at $f(0)$ or $f(C + \delta)$. Now we claim:

$$f(x) \leq C^2,\ for\ x \in [0, C + \delta] \tag{4.2.5}$$

For $f(0)$, we have $f(0) = 1$.

For $f(C + \delta)$, if we can prove:

$$C^2 \geq (C + \delta)^2 + 1 - 2\cos(\theta)(C + \delta), \tag{4.2.6}$$

then we know our claim (4.2.5) is true. The inequality (4.2.6) is equivalent to:

$$C^2 + \delta^2 + 2C\delta + 1 - 2\cos(\theta)(C + \delta) \le C^2,$$

which is equivalent to:

$$2C(\delta - \cos(\theta)) \le -\delta^2 + 2\cos(\theta)\delta - 1.$$

Since $\delta < \cos(\theta)$, then $(\delta - \cos(\theta)) < 0$, hence the inequality (4.2.6) is equivalent to:

$$C \ge \frac{\delta^2 - 2\cos(\theta)\delta + 1}{2(\cos(\theta) - \delta)}$$

which holds by the definition of C. Hence by induction, we have proven the Theorem. □

Remark 4.2.8. *Suppose for every subspace W_n, the cardinality $|R(Q_n)|$ is less then $M_n + 1$, for example, $R(Q_n) = M_n$. Then no matter how we define the quantize Q_n, the situation $\theta_k \ge \pi/2$ will happen, then $\cos(\theta_k) < 0$, so for any $\delta > 0$ we have $2(\delta - \cos(\theta_k)) > 0$. Hence we can only get*

$$C \le \frac{\delta^2 - 2\cos(\theta_k)\delta + 1}{2(\cos(\theta_k) - \delta)}$$

which is always false since the right part of the inequality is negative, so we cannot get a positive uniform bound C.

Actually, if we suppose all $\theta_k = \pi/2$, by the law of cosine we can get that $|u_{k+1}|^2 = |\pi_{W_{k+1}}(u_k) + x_{k+1}|^2 + 1$ for all k, and we can easily get $|u_{k+1}| \ge |u_k| + \sigma$, where σ is a positive constant. Hence $\{|u_n|\}$ is increasing with n, which means $\{|u_n|\}$ does not have uniform bound. Hence $|R(Q_n)| = M_n + 1$ is the best we can hope for Theorem 4.2.7.

However, in practice, quantizers are never perfect. In [DD03], the authors assume the quantizer $q(x) = sign(x+\rho)$, where ρ is unknown noise except for the specification $|\rho| < \tau$.

Then the authors define the non-ideal Q as following:

$$Q(x) = sign(x),\ for\ \|x\| \geq \tau;$$

or

$$Q(x) \leq 1,\ for\ \|x\| \leq \tau.$$

and proved the 1-bit $\Sigma\Delta$ quantization scheme is still stable.

Here, we will show the robustness of the quantizers Q_n in Theorem 4.2.7. Observe the definition of the quantizer Q_n in Theorem 4.2.7, not same as the quantizer Q above, Q_n mapping the vector $x \in \mathbb{R}^d$ to alphabet vectors not depend on the $\|x\|$ but on the $\langle x/|x|, e_{k,n}\rangle$ which is the angle between x and each alphabet vector $e_{i,n}$. Now, define the non-ideal quantizer Q_n with a noise ρ as following:

$$Q_n(x) = e_{i,n},\ i = \min\{k \in \{1,2,...,M_n+1\} : \langle x/|x|, e_{k,n}\rangle - \langle x/|x|, e_{j,n}\rangle \geq \rho\ for\ all\ j \neq k\}, \tag{4.2.7}$$

$$Q_n(x) = e_{i,n},\ i = \{k \in \{1,2,...,M_n+1\} : 0 < \langle x/|x|, e_{k,n}\rangle - \langle x/|x|, e_{j,n}\rangle \leq \rho\ for\ all\ j \neq k\}. \tag{4.2.8}$$

In equation (4.2.8), we just let i be any vector which satisfies the condition:

$$0 < \langle x/|x|, e_{k,n}\rangle - \langle x/|x|, e_{j,n}\rangle \leq \rho\ for\ all\ j \neq k.$$

Then we have the proposition for our new quantizer Q_n as:

Proposition 4.2.9. *Let $\{(W_n, w_n)\}_{n\in I}$ be fusion frame for $\mathbb{H}$ and $\{e_{i,n}\}_{i=1}^{M_n+1}$ be the set of unit vectors in W_n which we constructed in Lemma 4.2.5. The algorithm is defined as 4.2.1. For any $x \in W_n$, we define Q_n by equation (4.2.7) and (4.2.8). Let $d = \max\{\dim(W_n)\}$, $\theta = \pi - \arccos(-\frac{1}{d})$. If $\|x\| \leq \delta < \cos(\theta + \arccos(\rho))$ and $\arccos(\rho) < \arccos(-\frac{1}{d}) - \frac{\pi}{2}$.*

Then for all n, we have:

$$\|u_n\| \leq \max\{\frac{\delta^2 - 2\cos(\theta + \arccos(\rho))\delta + 1}{2(\cos(\theta + \arccos(\rho)) - \delta)}, 1\}.$$

Proof. Let $a_k = \|\pi_{W_k}(u_{k-1}) + x_k\|$, $b_k = \|q_k\|$, $c_k = \|u_k\|$, θ_k is the angle between vector $\pi_{W_k}(u_{k-1}) + x_k$ and vector q_k. Notice that we assume $\arccos(\rho) < \arccos(-\frac{1}{d}) - \frac{\pi}{2}$. Then by Lemma 4.2.6 and the definition of Q_n, we have

$$\theta_k \leq \theta + \arccos(\rho) = \pi - \arccos(-\frac{1}{d}) + \arccos(\rho) < \frac{\pi}{2}$$

thus we still have $\theta + \arccos(\rho) \in [\frac{\pi}{3}, \frac{\pi}{2})$. Since every thing else is the same as Theorem4.2.7, following the same method, we have:

$$\|u_n\| \leq C = \max\{\frac{\delta^2 - 2\cos(\theta + \arccos(\rho))\delta + 1}{2(\cos(\theta + \arccos(\rho)) - \delta)}, 1\}.$$

□

In [DD03], the noise ρ do not have any limitation. Unlike as [DD03], in Proposition 4.2.9, we assume the noise ρ should satisfy $\arccos(\rho) < \arccos(-\frac{1}{d}) - \frac{\pi}{2}$ which is equivalent to $\rho < |\cos(\arccos(-\frac{1}{d}) - \frac{\pi}{2})|$. However, since $d = \max\{\dim(W_n)\}$, in practice, the dimension of the subspace W_n is always far less the the dimension of the whole space $\mathbb{H}$, for example, when $d = 10$, we only need $\arccos(\rho) < \frac{\pi}{15}$ which is a reasonable noise. Hence, Q_n is robust for a acceptable noise.

In Theorem 4.2.7, we know

$$|R_{Q_n}| = M_n + 1.$$

Thus for each subspace we need $\log_2(M_n + 1) \ll M_n$ bit to quantize the signal. Compared to the algorithm (4.0.4) we gave in the beginning of this section and the first algorithm (4.2.4) we gave before, it is a big improvement.

Now we give a example to show the error bound of the fisrt-order Sigma-Delta quantization with fusion frame:

Example 4.2.10. *Let $\{E(n/N)\}_{n=1}^{N} \subset \mathbb{R}^3$ be defined as following:*

$$E(n/N) = [\cos(2\pi n/N), \sin(2\pi n/N), 0]^*, \;\; 1 \leq n \leq N, \qquad (4.2.9)$$

Let W_n be the two dimensional subspace of $\mathbb{R}^3$ with normal vector $E(n/N)$. By Theorem 1.3.5, we know there exists a constant w, such that $\{W_n, w\}$ is a $C-tight$ fusion frame. Moreover, by Proposition 1.3.6, we have:

$$\|S_{W,w}\|_{op} = C = \frac{\sum_{n=1}^{N} w_n^2 \dim(W_n)}{\dim(\mathbb{R}^3)} = \frac{w^2 2N}{3}$$

Suppose f is a signal in $\mathbb{R}^3$ that satisfies $\|f\| \leq \frac{1}{2}$, then we know,

$$f = S_{W,w}^{-1} S_{W,w} f = \sum_{n=1}^{N} w^2 S_{W,w}^{-1} \pi_{W_n}(f)$$

Now we use the fisrt-order Sigma-Delta quantization as in Theorem 4.2.7, then we have the reconstructed signal,

$$\widetilde{f} = \sum_{n=1}^{N} w^2 S_{W,w}^{-1} q_n$$

Thus, the error is equal to:

$$\begin{aligned}\|f-\widetilde{f}\| &= \|\sum_{n=1}^{N} w^2 S_{W,w}^{-1}\pi_{W_n}(f) - \sum_{n=1}^{N} w^2 S_{W,w}^{-1} q_n\| \\ &= w^2\|\sum_{n=1}^{N} S_{W,w}^{-1}(\pi_{W_n}(f)-q_n)\| \\ &= w^2\|\sum_{n=1}^{N} S_{W,w}^{-1}(u_n-\pi_{W_n}(u_{n-1}))\| \\ &\leq w^2\|S_{W,w}^{-1}\|_{op}\|\sum_{n=1}^{N}(u_n-\pi_{W_n}(u_{n-1}))\| \\ &= w^2\|S_{W,w}^{-1}\|_{op}\|\sum_{n=1}^{N-1}(u_n-\pi_{W_{n+1}}(u_n))+u_N\|\end{aligned}$$

By the definition of W_n, we have:

$$u_n-\pi_{W_{n+1}}(u_n) = u_n\sin(\frac{2\pi}{N}),$$

and by Theorem 4.2.7 we have $\|u_n\| \leq C$. Thus,

$$\|\sum_{n=1}^{N-1}(u_n-\pi_{W_{n+1}}(u_n))+u_N\| = \|\sum_{n=1}^{N-1} C\sin(\frac{2\pi}{N})+C\| = \|C(N-1)\sin(\frac{2\pi}{N})+C\|$$

Since when N is large enough, we have $\sin(\frac{N}{2\pi}) \approx \frac{2\pi}{N}$. Then,

$$\|C(N-1)\sin(\frac{N}{2\pi})+C\| \approx \|C(N-1)\frac{2\pi}{N}+C\| = \mathscr{O}(1),$$

which means $\|\sum_{n=1}^{N-1}(u_n-\pi_{W_{n+1}}(u_n))+u_N\|$ is equivalent to a constant C'. Hence,

$$w^2\|S_{W,w}^{-1}\|_{op}\|\sum_{n=1}^{N-1}(u_n-\pi_{W_{n+1}}(u_n))+u_N\| = C'w^2\frac{3}{w^2 2N} = \mathscr{O}(N^{-1}).$$

So the error $\|f-\widetilde{f}\|$ is equal to $\mathscr{O}(N^{-1})$.

4.3 High-order sigma-delta quantization for fusion frames

Definition 4.3.1. *For showing the definition clearly, we first give the basic idea for second-order* $\Sigma\Delta$ *quantization of fusion frames. Consider the following recursion:*

$$v_n = u_n - \pi_{W_n}(u_{n-1}),$$

$$v_n - \pi_{W_n}(v_{n-1}) = x_n - q_n,$$

$$q_n = Q_n(U(\pi_{W_n}(u_{n-1}), \pi_{W_n}(v_{n-2}), x_n)), \tag{4.3.1}$$

where vector function $U : (u,v,x) \to W_n$ *is called the quantization rule,* $Q_n : W_n \to \mathscr{A}_n$ *is a vector valued quantizer and* $\mathscr{A}_n \in W_n$ *is the alphabet of vectors.*

Equation (4.3.1) is the definition for second order $\Sigma\Delta$ quantization for fusion frames. As for the finite frame $\Sigma\Delta$ problem, constructing a stable $\Sigma\Delta$ scheme requires carefully choosing the quantization rule U in (4.3.1). Stability analysis of $\Sigma\Delta$ schemes can be quite complicated, especially for lower bit schemes in higher order schemes (even for second order). Moreover, for fusion frames, the problem becomes even more complicated since we need to make the vectors bounded. Hence for higher order $\Sigma\Delta$ fusion frames, we now just deal with the general case which $\mathscr{A}$ is allow to have $\mathscr{O}(2^{(M_n+2)r})$ alphabet, where r is the order. Hence we next give the definition for the alphabet $\mathscr{A}_\delta$, where δ is our step length:

Definition 4.3.2. *For rth-order* $\Sigma\Delta$ *quantization, let* $\theta = \arcsin(\frac{1}{2^{r+1}})$*.Then by Corollary 4.2.3, let vectors set* $\mathscr{L} = \mathscr{N}_\theta \subsetneqq \mathbb{R}^n$ *and* $\mathscr{A}_{\delta,n} = \{u \in \mathbb{R}^n | u = k\delta v, v \in \mathscr{L}, k = 1,2,\ldots,2^r\}$.

Next, we give the definition for high order $\Sigma\Delta$ quantization of fusion frames, then we give the proof of the scheme stability:

Definition 4.3.3.

$$\Delta_n^0 = u_n, \;\; \Delta_n^1 = \Delta_n^0 - \pi_{W_n}(\Delta_{n-1}^0)$$

$$\Delta_n^r = \Delta_n^{r-1} - \pi_{W_n}(\Delta_{n-1}^{r-1})$$

or we can define,

$$\Delta_n^r = \sum_{i=0}^{r} (-1)^{i-1} \binom{r}{i} \pi_{W_n} \pi_{W_{n-1}} \dots \pi_{W_{n-i+1}}(u_{n-i})$$

$$\Delta_n^r = x_n - q_n$$

$$q_n = \arg \min_{a \in \mathscr{A}_{\delta,n}} \{ | \sum_{i=0}^{r} (-1)^{i-1} \binom{r}{i} \pi_{W_n} \pi_{W_{n-1}} \dots \pi_{W_{n-i+1}}(u_{n-i}) + x_n - a | \} \qquad (4.3.2)$$

Theorem 4.3.4. *Let $\{(W_n, w_n)\}_{n \in I}$ be a fusion frame for $\mathbb{H}$ and $\mathscr{A}_{\delta,M_n}$ be the set of alphabet vectors for W_n which we constructed in definition 4.3.2, where* $\dim W_n = M_n$. *The algorithm is defined by (4.3.2). If $\|x_n\| \leq \delta$ for all n, then we have $\|u_n\| \leq \delta$, for all n.*

Proof. We prove the theorem by induction, when $n = 1$, since $u_0 = 0$ and $x_1 \leq \delta$, by definition of $\mathscr{A}_{\delta,M_1}$, we know

$$\|u_1\| = \|u_0 + x_1 - q_1\| \leq \delta$$

Hence for $n = 1$ the result holds.

Suppose for all $n \leq k-1$, the result is true, and consider $n = k$. Firstly we know:

$$\|u_k\| = \| \sum_{i=0}^{r} (-1)^{i-1} \binom{r}{i} u_{n-i} + x_k - q_k \|.$$

Let $a_k = \sum_{i=0}^{r} (-1)^{i-1} \binom{r}{i} u_{n-i} + x_k$, $b_k = q_k$ and $c_k = u_k$, θ_k is the angle between a_k and b_k. Then by law of cosine we know:

$$\|c_k\|^2 = a_k^2 + b_k^2 - 2\cos(\theta_k) a_k b_k = (\cos(\theta_k) a_k)^2 + b_k^2 - 2\cos(\theta_k) a_k b_k + (\sin(\theta_k) a_k)^2,$$

which implies,

$$\|c_k\|^2 = (\cos(\theta_k) a_k - b_k)^2 + (\sin(\theta_k) a_k)^2.$$

By our assumption, since $|a_k| \leq |(2^r - 1) * \delta| + |\delta| = 2^r\delta$ and by definition of $\mathscr{A}_{\delta,M_k}$, we know $|\cos(\theta_k)a_k - b_k| \leq \frac{\delta}{2}$, then

$$|c_k|^2 \leq (\sin(\theta_k)2^r\delta)^2 + (\frac{\delta}{2})^2$$

Now, we claim:

$$(\sin(\theta_k)2^r\delta)^2 + (\frac{\delta}{2})^2 \leq \delta^2 \tag{4.3.3}$$

Inequality (4.3.3) is equivalent to:

$$\sin(\theta_k) \leq \frac{\sqrt{3}}{2^{r+1}}$$

which means we need inequality $\theta_k \leq \arcsin(\frac{\sqrt{3}}{2^{r+1}})$ to be true. By definition of $\mathscr{A}_{\delta,M_k}$ we know:

$$\theta_k \leq \arcsin(\frac{1}{2^{r+1}}) \leq \arcsin(\frac{\sqrt{3}}{2^{r+1}}).$$

Hence $|u_k| = |c_k| \leq \delta$. Thus by induction we have proven the Theorem. □

4.4 Converting canonical representations to Kashin representations

By Theorem 4.3.4, we have proven the stability for high-order $\Sigma\Delta$ quantization with fusion frames. However, unlike 1st-order, in high order, we need $\log_2((1+2^{r+1})^{M_n}2^r) \approx \mathscr{O}((M_n+2)r)$ bits for each subspace. Moreover, in Theorem 4.3.4, we assume $\|x_n\| \leq \delta$ for all n. But in practice, $\|x_n\| \leq \delta$ may not be always true. Generally, suppose $\|x_n\| \leq \alpha = C\delta$ for all n, then to keep the alphabet $\mathscr{A}_{\delta,n}$ has the same "step" δ, we have to spend C times more bits in each subspace. Hence, having the coefficients $\|x_n\|$ with a smaller uninform bound is important. We call this representation a Kashin's representation in frame theory, see [LV10, Pis99, Kas77]. According to the algorithm in [LV10], we will give a similar algorithm to find a Kashin's representation for fusion frames.

Without loss of generality, suppose $I = \{1,2,...N\}$, $\mathbb{H} = \mathbb{R}^d$ and $w_i = 1$ for $i = 1,2,...,N$. Since the canonical expansion $x = \sum_{n=1}^N S^{-1}\pi_{W_n}(x)$ need not be unique, and there will generally exist other representations $x = \sum_{n=1}^N f_n$ where $f_n \in W_n$. We will be interested in balanced or democratic representations where $\max_{1\leq n\leq N}\|f_n\|$ is as small as possible.

Suppose $\{W_n\}_{n=1}^N$ is an A-tight fusion frame. Note that if the canonical fusion frame representation is used, then taking $x \in W_n$ gives

$$\exists x \in \mathbb{R}^d, \quad \max_{1\leq n\leq N} S^{-1}\|\pi_{W_n}(x)\| = A^{-1}\|x\| = \left(\frac{d}{\sum_{n=1}^N \dim(W_n)}\right)\|x\|. \tag{4.4.1}$$

Lemma 4.4.1. *If $\{W_n\}_{n=1}^N$ is an A-tight fusion frame for $\mathbb{R}^d$ and $x = \sum_{n=1}^N f_n$ with $f_n \in W_n$ then*

$$\max_{1\leq n\leq N}\|f_n\| \geq \left(\frac{d}{N\sum_{n=1}^N \dim(W_n)}\right)^{1/2}\|x\|. \tag{4.4.2}$$

Proof. Let $\{b_j^n\}_{j=1}^{\dim(W_n)} \subset W_n$ be an orthonormal basis for W_n. So

$$x = \sum_{n=1}^N f_n = \sum_{n=1}^N \sum_{j=1}^{\dim(W_n)} \langle f_n, b_j^n\rangle b_j^n.$$

Since $\{b_j^n : 1 \leq j \leq \dim(W_n) \text{ and } 1 \leq n \leq N\}$ is an A-tight frame for $\mathbb{R}^d$ we have

$$\begin{aligned}
\|x\|^2 &= A \sum_{n=1}^{N} \sum_{j=1}^{\dim(W_n)} |\langle x, A^{-1} b_j^n \rangle|^2 \\
&\leq A \sum_{n=1}^{N} \sum_{j=1}^{\dim(W_n)} |\langle f_n, b_j^n \rangle|^2 \\
&= A \sum_{n=1}^{N} \|P_{W_n}(f_n)\|^2 = A \sum_{n=1}^{N} \|f_n\|^2. \qquad (4.4.3)
\end{aligned}$$

Here we have used that if $\{\varphi_n\}_{n=1}^N \subset \mathbb{R}^d$ is an A-tight frame and $x = \sum_{n=1}^N a_n \varphi_n$ then $\sum_{n=1}^N |\langle x, A^{-1}\varphi_n \rangle|^2 = \sum_{n=1}^N |a_n|^2$. The conclusion (4.4.1) now follows from (4.4.3). □

Remark 4.4.2. *1. To compare the bounds in (4.4.1) and (4.4.2), note that since* $0 \leq \dim(W_n) \leq d$, *we have:*

$$\left(\frac{d}{\sum_{n=1}^N \dim(W_n)} \right) \geq \left(\frac{d}{N \sum_{n=1}^N \dim(W_n)} \right)^{1/2}.$$

2. In the special case when each subspace W_n has the same dimension, $\dim(W_n) = k$, *the right side of* (4.4.2) *becomes* $\|x\| \left(\frac{d}{Nk}\right)$, *whereas the lower bound in* (4.4.2) *becomes* $\|x\| \frac{\sqrt{d}}{N\sqrt{k}}$.

Definition 4.4.3. *Given subspaces $\{W_n\}_{n=1}^N \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$ we shall say that the representation $x = \sum_{n=1}^N f_n$ with $f_n \in W_n$ is a Kashin representation with level $K \geq 1$ if*

$$\max_{1 \leq n \leq N} \|f_n\| \leq K \|x\| \left(\frac{d}{N \sum_{n=1}^N \dim(W_n)} \right)^{1/2}.$$

Definition 4.4.4. *The collection of subspaces $\{W_n\}_{n=1}^N \subset \mathbb{R}^d$ satisfies the uncertainty principle with parameters* $0 < \delta < 1$ *and* $0 < \eta$, *if*

$$\| \sum_{n \in \Omega} f_n \|^2 \leq \eta^2 \sum_{n \in \Omega} \|f_n\|^2,$$

whenever $|\Omega| \leq \delta N$ *and* $f_n \in W_n$.

Definition 4.4.5. *Let* $\{W_n\}_{n=1}^N$ *be an A-tight fusion frame for* $\mathbb{R}^d$. *Given* $C > 0$, *define*

$$T(x) = T_C(x) = \sum_{n=1}^{N} g_n$$

where

$$g_n = \begin{cases} A^{-1}P_{W_n}(x), & \text{if } \|A^{-1}P_{W_n}(x)\| \leq C\|x\|; \\ \frac{C\|x\|}{\|P_{W_n}(x)\|}P_{W_n}(x), & \text{if } \|A^{-1}P_{W_n}(x)\| > C\|x\|. \end{cases}$$

Lemma 4.4.6. *If* $\{W_n\}_{n=1}^N$ *is an A-tight fusion frame and satisfies the uncertainty principle with parameters* $0 < \delta < 1$ *and* $0 < \eta < \sqrt{A}$ *and* $C = (\delta NA)^{-1/2}$ *then*

$$\|x - T_C(x)\| \leq (\eta/\sqrt{A})\|x\|.$$

Proof. Let $\Omega = \{1 \leq n \leq N : \|A^{-1}P_{W_n}(x)\| > C\|x\|\}$. First, note that $|\Omega| \leq \delta N$ since

$$\|x\|^2 = A\sum_{n=1}^{N} \|A^{-1}P_{W_n}(x)\|^2 \geq A|\Omega|C^2\|x\|^2 = \frac{|\Omega|\ \|x\|^2}{\delta N}.$$

Next, note that

$$\begin{aligned} x - T_C(x) &= \sum_{n\in\Omega} \left(A^{-1}P_{W_n}(x) - \frac{C\|x\|}{\|P_{W_n}(x)\|}P_{W_n}(x)\right) \\ &= \sum_{n\in\Omega} \left(A^{-1} - \frac{C\|x\|}{\|P_{W_n}(x)\|}\right)P_{W_n}(x). \end{aligned}$$

Since $|\Omega| \leq \delta N$, the uncertainty principle gives

$$\begin{aligned}
\|x - T_C(x)\|^2 &\leq \eta^2 \sum_{n\in\Omega} \| \left(A^{-1} - \frac{C\|x\|}{\|P_{W_n}(x)\|} \right) P_{W_n}(x)\|^2 \\
&= \eta^2 \sum_{n\in\Omega} \left(A^{-1} - \frac{C\|x\|}{\|P_{W_n}(x)\|} \right)^2 \|P_{W_n}(x)\|^2 \\
&\leq \eta^2 \sum_{n\in\Omega} A^{-2}\|P_{W_n}(x)\|^2 \\
&\leq \eta^2 A^{-2} \sum_{n=1}^{N} \|P_{W_n}(x)\|^2 \\
&\leq (\eta)^2 A^{-2} A\|x\|^2 \\
&= (\eta^2/A)\|x\|^2.
\end{aligned}$$

□

Theorem 4.4.7. *Let $\{W_n\}_{n=1}^N$ be an A-tight fusion frame that satisfies the uncertainty principle with parameters $0 < \delta < 1$ and $0 < \eta < \sqrt{A}$ and $C = (\delta N A)^{-1/2}$. Then each vector $x \in \mathbb{R}^d$ admits a Kashin representation of level $K = (1-\eta/\sqrt{A})^{-1}\,(\delta A)^{-1/2}$.*

Proof. Let $x_0 = x$ and $x_k = x_{k-1} - T_C(x_{k-1})$. This gives

$$x = \sum_{k=0}^{r} Tx_k + x_{r+1}.$$

It follows from Lemma 4.4.6 by induction that $\|x_k\|^2 \leq (\eta^2/A)^k$, thus

$$x = \sum_{k=0}^{\infty} Tx_k.$$

Moreover, by the definition of the operator T, each vecotr Tx_k has an expansion $\{g_n\}_{n=1}^N$ in the subspaces $\{W_n\}_{n=1}^N$ and $\|g_n\|^2 \leq C\|x_k\|^2 \leq C(\eta^2/A)^k\|x\|^2$. Summing up these expansions for $k = 0, 1, 2, \ldots$, we obtain an expansion of x with coefficients bounded by $(1-\eta/\sqrt{A})^{-1}\,(\delta N A)^{-1/2}\,\|x\|$. Thus x admits Ksshin's representation with level $K = (1-\eta/\sqrt{A})^{-1}\,(\delta A)^{-1/2}$. □

Now we give the algorithm to compute Kashin's Representations:

Input:

1. Let $\{W_n\}_{n=1}^N$ be an A-tight fusion frame that satisfies uncertainty principle with parameters $0 < \delta < 1$ and $0 < \eta < \sqrt{A}$

2. A vector $x \in \mathbb{R}^d$ and a number of iterations r.

Initialize the coefficients and truncation level

$f_i = 0,\ i = 1, \cdots, N;\ C = (\delta NA)^{-1/2}; x_0 = x$

Repeat the following r times:

1.Compute the $\{g_i\}_{i=1}^N$ with truncation level C for x_k.

2.Reconstruct and compute the error.

$$T(x_k) \leftarrow \sum_{i=1}^{N} g_i; x_{k+1} \leftarrow x_k - T(x_k)$$

3.Update Kashin's coefficients

$$f_i \leftarrow f_i + g_i, i = 1, 2, \ldots, N$$
$$C \leftarrow \eta/\sqrt{A}C$$

Output:

Kashin's decomposition of x with level $K = (1 - \eta/\sqrt{A})^{-1}(\delta A)^{-1/2}$ and with accuracy $(\eta/\sqrt{A})^r \|x\|$. Thus the algorithm finds coefficients $f_1, f_2, \ldots, f_N$ such that:

$$\|x - \sum_{i=1}^{N} f_i\| \leq (\eta/\sqrt{A})^r \|x\|$$

with,

$$\max \|f_i\| \leq K/\sqrt{N}\|x\|$$

4.5 Sobolev Left inverse operator for fusion frames

Definition 4.5.1. *Let $\{(W_n, w_n)\}_{n=1}^{N}$ be a fusion frame for $\mathbb{H}$, $T_{W,w}^{*}$ is the analysis operator, $T_{W,w}$ is the synthesis operator, D is the operator matrix defined in (4.5.1). The Sobolev left inverse operator of the fusion frame is define as:*

$$F = (T_{W,w}D^{-r}(D^{*})^{-r}T_{W,w}^{*})^{-1}T_{W,w}D^{-r}(D^{*})^{-r}.$$

Theorem 4.5.2. *Let $\{(W_n, w_n)\}_{n=1}^{N}$ be a fusion frame for $\mathbb{H}$. Then the Sobolev left inverse operator F is the left inverse of $T_{W,w}^{*}$ for which operator norm $\|F(D^{*})^{r}\|_{OP}$ and Frobenius norm $\|F(D^{*})^{r}\|_{\mathscr{F}}$ are minimal.*

Proof. Note that $FT_{W,w}^{*} = (T_{W,w}D^{-r}(D^{*})^{-r}T_{W,w}^{*})^{-1}T_{W,w}D^{-r}(D^{*})^{-r}T_{W,w}^{*} = I_{\mathbb{H}}$, Thus F is left inverse of $T_{W,w}^{*}$.

Since D is invertible, U is left inverse of $T_{W,w}^{*}$ if and only if $U(D^{*})^{r}(D^{*})^{-r}T_{W,w}^{*} = UT_{W,w}^{*} = I_{\mathbb{H}}$. It means U is left inverse of $T_{W,w}^{*}$ if and only if $U(D^{*})^{r}$ is left inverse of $T_{W,w}D^{-r}$. Since F is Sobolev left inverse operator of $T_{W,w}^{*}$,then

$$F(D^{*})^{r} = (T_{W,w}D^{-r}(D^{*})^{-r}T_{W,w}^{*})^{-1}T_{W,w}D^{-r}$$

is the canonical Left Inverse of $T_{W,w}D^{-r}$. Following the theorem 4.1.3 and 4.1.4, $F(D^{*})^{r}$ is the left inverse of $T_{W,w}^{*}D^{-r}$ with minimal operator norm and Frobenius norm. Hence, the Sobolev left inverse operator $F(D^{*})^{r}$ is the left inverse of $T_{W,w}^{*}D^{-r}$ which minimizes $\|\cdot\|_{OP}$ and $\|\cdot\|_{\mathscr{F}}$. □

Let $\{(W_n, w_n)\}_{n=1}^{N}$ be a fusion frame for $\mathbb{H}$, π_{W_n} is the projection from $\mathbb{H}$ to subspace W_n and $f \in \mathbb{H}$ is the signal. $T_{W,w}^{*}$ is the analysis operator and F is the left inverse Operator. Here we suppose all $w_n = 1$. Then we reconstruct a signal $\widetilde{f}$ from the q_n which are produced by rth-order ΣΔ quantization with F by:

$$\widetilde{f} = F(\{q_n\}).$$

And we also know that,

$$f = FT^*_{W,w}(f) = F(\{\pi_{W_n}(f)\}) = F(\{x_n\}).$$

Hence, the error can be written as:

$$\|f - \widetilde{f}\|_2 = \|F(\{x_n\} - F(\{q_n\})\|_2 = \|F(\{x_n - q_n\})\|_2 = \|F(\Delta_n^r)\|_2 = \|F(D^*)^r(u)\|_2.$$

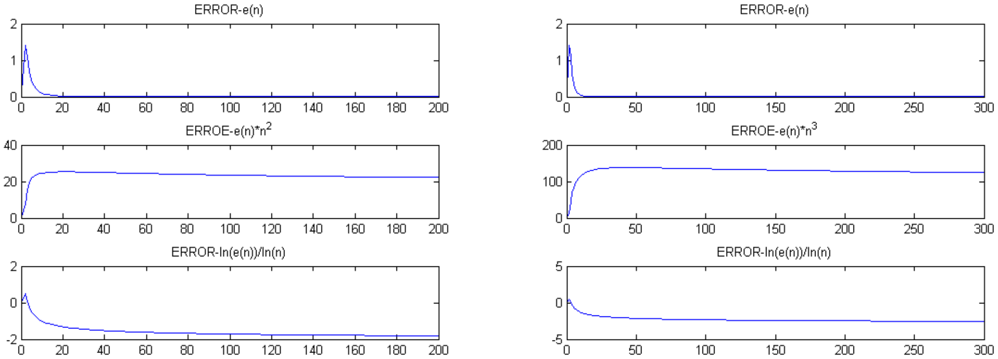
Here, $u = [u1, u2, \ldots, u_n]^*$ and

$$D = \begin{pmatrix} I & -\pi_{W_2} & 0 & \cdots & 0 \\ 0 & I & -\pi_{W_3} & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & \cdots & 0 & I & -\pi_{W_n} \\ 0 & \cdots & 0 & 0 & I \end{pmatrix}. \tag{4.5.1}$$

Note that D is invertible and

$$D^{-1} = \begin{pmatrix} I & \pi_{W_2} & \pi_{W_2}\pi_{W_3} & \cdots & \pi_{W_2}\pi_{W_3}\cdots\pi_{W_n} \\ 0 & I & \pi_{W_3} & \cdots & \pi_{W_3}\pi_{W_4}\cdots\pi_{W_n} \\ & & \ddots & \ddots & \\ 0 & \cdots & 0 & I & \pi_{W_n} \\ 0 & \cdots & 0 & 0 & I \end{pmatrix}.$$

Hence we get $\|F(D^*)^r(u)\|_2 \leq \|F(D^*)^r\|_{OP}\|u\|_2 \leq C\sqrt{N}\|F(D^*)^r\|_{OP}$, where C is a positive constant by Theorem 4.2.7 and Theorem 4.3.4. Thus by Theorem 4.5.2 we know



(a) (b)

Figure 4.2: Numerical experiment

Sobolev left inverse leads the minimal squared error.

For now we can not give a accurate estimation for the error bound $\|F(D^*)^r\|_{OP}$. Instead, we give the following numerical experiment for the error performance of rth order $\Sigma\Delta$ quantization with Sobolev left inverse. The figure (a) and (b) shows the result when $r = 2$ and $r = 3$. In each figure, the x-coordinate is always the size N of the fusion frame. In the top graph, the y-coordinate is the operator norm of $F(D^*)^r$ which is the upper bound of the error. In the middle graph, the y-coordinate is the operator norm of $F(D^*)^r$ times N^{2r}, and in the bottom graph, y-coordinate is $\ln((F(D^*)^r))$ divided by $\ln(N)$, which shows the order of decay speed. We can observe the error is almost equal to $\mathscr{O}(N^{-r})$.

BIBLIOGRAPHY

[Ben48] William Ralph Bennett, *Spectra of quantized signals*, Bell System Technical Journal **27** (1948), no. 3, 446–472.

[BF03] John J Benedetto and Matthew Fickus, *Finite normalized tight frames*, Advances in Computational Mathematics **18** (2003), no. 2-4, 357–385.

[BH01] Helmut Bölcskei and Franz Hlawatsch, *Noise reduction in oversampled filter banks using predictive quantization*, Information Theory, IEEE Transactions on **47** (2001), no. 1, 155–172.

[BLPY10] James Blum, Mark Lammers, Alexander M Powell, and Özgür Yılmaz, *Sobolev duals in frame theory and sigma-delta quantization*, Journal of Fourier Analysis and Applications **16** (2010), no. 3, 365–381.

[BP07] Bernhard G Bodmann and Vern I Paulsen, *Frame paths and error bounds for sigma–delta quantization*, Applied and Computational Harmonic Analysis **22** (2007), no. 2, 176–197.

[BPA07a] Bernhard G Bodmann, Vern I Paulsen, and Soha A Abdulbaki, *Smooth frame-path termination for higher order sigma-delta quantization*, Journal of Fourier Analysis and Applications **13** (2007), no. 3, 285–307.

[BPA07b] ______, *Smooth frame-path termination for higher order sigma-delta quantization*, Journal of Fourier Analysis and Applications **13** (2007), no. 3, 285–307.

[BPY06a] John J Benedetto, Alexander M Powell, and Özgür Yılmaz, *Second-order sigma–delta ($\sigma\delta$) quantization of finite frame expansions*, Applied and Computational Harmonic Analysis **20** (2006), no. 1, 126–148.

[BPY06b] ______, *Sigma-delta ($\sigma\delta$) quantization and finite frames*, Information Theory, IEEE Transactions on **52** (2006), no. 5, 1990–2005.

[BT01] John J Benedetto and Oliver M Treiber, *Wavelet frames: multiresolution analysis and extension principles*, Wavelet transforms and time-frequency signal analysis, Springer, 2001, pp. 3–36.

[BYP04] John J Benedetto, Özgür Yilmaz, and Alexander M Powell, *Sigma-delta quantization and finite frames.*, ICASSP (3), 2004, pp. 937–940.

[Cas99] Peter G Casazza, *The art of frame theory*, arXiv preprint math/9910168 (1999).

[Chr02] Ole Christensen, *An introduction to frames and riesz bases*, Springer, 2002.

[CK03] Peter G Casazza and Gitta Kutyniok, *Frames of subspaces*, arXiv preprint math/0311384 (2003).

[CKL08] Peter G Casazza, Gitta Kutyniok, and Shidong Li, *Fusion frames and distributed processing*, Applied and computational harmonic analysis **25** (2008), no. 1, 114–132.

[CKP13] Peter G Casazza, Gitta Kutyniok, and Friedrich Philipp, *Introduction to finite frame theory*, Finite Frames, Springer, 2013, pp. 1–53.

[D$^+$92] Ingrid Daubechies et al., *Ten lectures on wavelets*, vol. 61, SIAM, 1992.

[DD03] Ingrid Daubechies and Ron DeVore, *Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order*, Annals of mathematics (2003), 679–710.

[EC06] Yonina C Eldar and Ole Christensen, *Characterization of oblique dual frame pairs*, EURASIP Journal on Advances in Signal Processing **2006** (2006), no. 1, 1–11.

[GKK01] Vivek K Goyal, Jelena Kovačević, and Jonathan A Kelner, *Quantized frame expansions with erasures*, Applied and Computational Harmonic Analysis **10** (2001), no. 3, 203–233.

[GLP$^+$10] C Sinan Güntürk, Mark Lammers, Alex Powell, Rayan Saab, and Özgür Yilmaz, *Sigma delta quantization for compressed sensing*, Information Sciences and Systems (CISS), 2010 44th Annual Conference on, IEEE, 2010, pp. 1–6.

[GLV01] C Sinan Güntürk, Jeffrey C Lagarias, and Vinay A Vaishampayan, *On the robustness of single-loop sigma-delta modulation*, Information Theory, IEEE Transactions on **47** (2001), no. 5, 1735–1744.

[Gra90] Robert M Gray, *Quantization noise spectra*, Information Theory, IEEE Transactions on **36** (1990), no. 6, 1220–1244.

[Gün03] C Sinan Güntürk, *One-bit sigma-delta quantization with exponential accuracy*, Communications on Pure and Applied Mathematics **56** (2003), no. 11, 1608–1630.

[GVT98a] Vivek K Goyal, Martin Vetterli, and Nguyen T Thao, *Quantized overcomplete expansions in ir n: analysis, synthesis, and algorithms*, Information Theory, IEEE Transactions on **44** (1998), no. 1, 16–31.

[GVT98b] ______, *Quantized overcomplete expansions in ir n: analysis, synthesis, and algorithms*, Information Theory, IEEE Transactions on **44** (1998), no. 1, 16–31.

[Hei10] Christopher Heil, *A basis theory primer: expanded edition*, Springer, 2010.

[HMBZ13] Sigrid Heineken, Patricia Morillas, Ana Benavente, and María Zakowicz, *Dual fusion frames*, arXiv preprint arXiv:1308.4595 (2013).

[JWW07] David Jimenez, Long Wang, and Yang Wang, *White noise hypothesis for uniform quantization errors*, SIAM journal on mathematical analysis **38** (2007), no. 6, 2042–2056.

[Kas77] BS Kashin, *Sections of some finite dimensional sets and classes of smooth functions*, Izv. Acad. Nauk SSSR **41** (1977), no. 2, 334–351.

[Li95] Shidong Li, *On general frame decompositions*, Numerical functional analysis and optimization **16** (1995), no. 9-10, 1181–1191.

[LO04] Shidong Li and Hidemitsu Ogawa, *Pseudoframes for subspaces with applications*, Journal of Fourier Analysis and Applications **10** (2004), no. 4, 409–431.

[LPY10] Mark Lammers, Alexander M Powell, and Özgür Yılmaz, *Alternative dual frames for digital-to-analog conversion in sigma–delta quantization*, Advances in Computational Mathematics **32** (2010), no. 1, 73–102.

[LV10] Yurii Lyubarskii and Roman Vershynin, *Uncertainty principles and vector quantization*, Information Theory, IEEE Transactions on **56** (2010), no. 7, 3491–3501.

[Mun92] Niels Juul Munch, *Noise reduction in tight weyl-heisenberg frames*, Information Theory, IEEE Transactions on **38** (1992), no. 2, 608–616.

[Pis99] Gilles Pisier, *The volume of convex bodies and banach space geometry*, vol. 94, Cambridge University Press, 1999.

[PL93] Steven C Pinault and Philip V Lopresti, *On the behavior of the double-loop sigma-delta modulator*, Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on **40** (1993), no. 8, 467–479.

[RG02] Gagan Rath and Christine Guillemot, *Syndrome decoding and performance* *analysis of dft codes with bursty erasures*, Data Compression Conference, 2002. Proceedings. DCC 2002, IEEE, 2002, pp. 282–291.

[ST05] Richard Schreier and Gabor C Temes, *Understanding delta-sigma data converters*, vol. 74, IEEE press Piscataway, NJ, 2005.

[TV94] Nguyen T Thao and Martin Vetterli, *Deterministic analysis of oversampled a/d conversion and decoding improvement based on consistent estimates*, Signal Processing, IEEE Transactions on **42** (1994), no. 3, 519–531.

[Ver10] Roman Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, arXiv preprint arXiv:1011.3027 (2010).

[Wan08] Yang Wang, *Sigma–delta quantization errors and the traveling salesman problem*, Advances in Computational Mathematics **28** (2008), no. 2, 101–118.

[Yil02a] Özgür Yilmaz, *Stability analysis for several second-order sigmałdelta method-s of coarse quantization of bandlimited functions*, Constructive approximation **18** (2002), no. 4, 599–623.

[Yil02b] ______, *Stability analysis for several second-order sigmałdelta methods of coarse quantization of bandlimited functions*, Constructive approximation **18** (2002), no. 4, 599–623.

[Yıl03] Özgür Yılmaz, *Coarse quantization of highly redundant time–frequency representations of square-integrable functions*, Applied and Computational Harmonic Analysis **14** (2003), no. 2, 107–132.

[Zim01] Georg Zimmermann, *Normalized tight frames in finite dimensions*, Recent Progress in Multivariate Approximation, Springer, 2001, pp. 249–252.