

MATHEMATICAL PROPERTIES OF COARSE  
QUANTIZATION SCHEMES IN SIGNAL ANALYSIS  
WITH NEW APPLICATIONS

ÖZGÜR YILMAZ

A DISSERTATION  
PRESENTED TO THE FACULTY  
OF PRINCETON UNIVERSITY  
IN CANDIDACY FOR THE DEGREE  
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE  
BY THE PROGRAM IN  
APPLIED AND COMPUTATIONAL MATHEMATICS

JANUARY 2002

© Copyright by Özgür Yılmaz, 2001.

All Rights Reserved

# Abstract

This thesis consists of two parts. In the first part<sup>1</sup> we investigate stability and robustness properties of a family of algorithms used to “coarsely quantize” bandlimited functions. The algorithms we will consider are one-bit second-order sigma-delta quantization schemes and some modified versions of these. We prove that there exists a bounded region that remains invariant under the two-dimensional piecewise-affine discrete dynamical system associated with each of these quantizers. Moreover, this bounded region can be constructed so that it is robust under small changes in the quantizer. We also show some interesting properties of the resulting binary sequences.

The second part is on coarse quantization of redundant representations, in particular Weyl-Heisenberg frame expansions. We introduce two algorithms –that are inspired by sigma-delta quantization algorithms for bandlimited functions– to quantize Weyl-Heisenberg frame expansions of certain classes of square-integrable functions. One of the two algorithms, TF $\Sigma\Delta$ -I, is translation invariant; however it produces a weak type approximation. The other algorithm, TF $\Sigma\Delta$ -II produces an approximation in  $L^2$ ; however the algorithm is not translation invariant and the class of functions that can be quantized by TF $\Sigma\Delta$ -II is smaller than the class of functions that can be quantized by TF $\Sigma\Delta$ -I. We discuss these and various other properties of each algorithm in detail.

---

<sup>1</sup>The first part of this thesis is submitted for publication [1].

## Acknowledgements

First of all, I would like to thank my advisor Ingrid Daubechies for her endless support during my graduate studies at Princeton University. Working with such an outstanding academic mentor and a great person has been a wonderful experience.

Sinan Güntürk has been there with all his support, both in academic issues and everyday life problems that I encountered during my first months in the Unites States, from the first day I came to Princeton. I thank him.

I would like to thank Ron DeVore for his invaluable comments on my thesis and for the discussions we had, which all inspired me.

I got to know Mete Soner here at Princeton. He has been a great mentor, both academically and personally. He offered advice and friendship whenever I needed and has been a source of inspiration for me.

I would like to thank Phil Holmes for his important comments and suggestions about my thesis. I also would like to thank Erhan Çınlar and René Carmona for their support in many aspects of my academic life. The discussions and email exchanges with Thao Nguyen has been fruitful. I thank him.

The summer internships that I had during my graduate studies have been invaluable. I thank Siemens Corporate Research and Alexander Jourjine, and AT&T Shannon Labs and Robert Calderbank for giving me these opportunities. Working with Zoran Čvetkovic at AT&T has been fruitful and fun. I thank him.

I thank Radu Balan. With his extensive knowledge, he has contributed to my research by pointing out interesting resources and references.

The people at PACM have been a very important part of my life here in Princeton. I thank them all. Especially, I want to thank my friends Toufic Suidan, who has been my office mate for the entire four years, and Scott Rickard, with whom I collaborated during my internship at Siemens.

I thank my friends Sinan, Refet, Taragay, Ömer, Toufic, Cliona and Jorge for all

the fun we had together. Without them life in Princeton would not be the same.

My friends in İstanbul (my “maalle”) have been with me all the time in spite of the physical distance. I thank Ayhan, Burhan, Murat, Tuna, Uğur and Serdar. I also want to thank my dear friend Hasan.

Apart from all, I owe everything to my beloved family. They have always been there with all their support and love. Words come short of expressing my love and gratitude towards them, but “Anne, Baba, Özge... Sevginizi, desteęinizi bana her an hissettirdiniz. Bunun benim için ne kadar önemli olduęunu ben size ne kadar hissetirebildim bugüne kadar, bilmiyorum. Sizleri çok seviyorum ve bütün her şey için teşekkür ediyorum. Sağolun, varolun...”. Thank you, Neriman, Erol and Özge Yılmaz.

Finally, I want to thank my wife, Ipek, who has been my colleague, my friend and my love for more than 10 years. During the creation of this thesis she has been there for me at every moment. We shared the excitement of building a life together, miles away from home, as well as the stressful days of hard work. I thank her for everything –listening to me for hours talking about some detail in my research, making me feel good even in the worst days, and for all the fun we had during these years in Princeton.

To my family and İpek

# Contents

Abstract . . . . .	iii
Acknowledgements . . . . .	iv
<b>I Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions</b>	<b>1</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Sampling & Oversampling . . . . .	2
1.2 Quantization . . . . .	4
1.3 Sigma-Delta quantization . . . . .	4
1.3.1 First-order sigma-delta quantization . . . . .	5
1.3.2 Higher order sigma-delta quantization schemes . . . . .	5
<b>2 Second-order sigma-delta quantizers</b>	<b>8</b>
2.1 Standard second-order sigma-delta quantizer . . . . .	8
2.2 The output sequence $q_n^\lambda$ has infinite memory . . . . .	10
2.3 Defeating the infinite memory: Introducing an enriched alphabet . . .	11
2.4 Defeating the infinite memory: Finite-memory (leaky) sigma-delta quantization . . . . .	13
<b>3 Stability and robustness of the standard second-order sigma-delta quantizer</b>	<b>20</b>

3.1	Stability of the standard second-order scheme . . . . .	20
3.2	Robustness of the standard second-order scheme . . . . .	28
<b>4</b>	<b>Stability and robustness of the tri-level second-order quantizer</b>	<b>34</b>
4.1	Stability of the tri-level quantizer . . . . .	34
4.2	Robustness of the tri-level quantizer . . . . .	38
<b>5</b>	<b>Stability and robustness of the finite-memory second-order sigma-delta quantizer</b>	<b>40</b>
<b>II</b>	<b>Coarse quantization of highly redundant time-frequency representations of square-integrable functions</b>	<b>43</b>
<b>6</b>	<b>Introduction</b>	<b>44</b>
<b>7</b>	<b>The Time-Frequency Sigma-Delta Quantization Algorithm I (TFΣΔ-I)</b>	<b>47</b>
7.1	The Algorithm . . . . .	47
7.2	Translation Invariance . . . . .	59
7.3	Numerical Experiment . . . . .	64
7.4	Higher-order time-frequency sigma-delta schemes . . . . .	72
7.4.1	Numerical Experiment revisited . . . . .	80
<b>8</b>	<b>The Time-Frequency Sigma-Delta Quantization Algorithm II (TFΣΔ-II)</b>	<b>86</b>
8.1	The Algorithm . . . . .	86
8.1.1	Coarse quantization of the Fourier coefficients of certain compactly supported functions . . . . .	92
8.2	Higher-order schemes . . . . .	96
8.3	Numerical Experiment . . . . .	99

# Part I

Stability analysis for several  
sigma-delta methods of coarse  
quantization of bandlimited  
functions

# Chapter 1

## Introduction

### 1.1 Sampling & Oversampling

Suppose we have a function  $f \in L^2(\mathbb{R})$  that is bandlimited, i.e.  $\text{supp } \hat{f} \subseteq [-\Omega, \Omega]$ , for some  $\Omega > 0$ . Then, it is a well known fact that we can reconstruct  $f$  from its sample values,  $f\left(\frac{n\pi}{\Omega}\right)$ :

$$f(t) = \sum_{n \in \mathbb{Z}} f\left(\frac{n\pi}{\Omega}\right) \frac{\sin(\Omega t - n\pi)}{\Omega t - n\pi}. \quad (1.1)$$

Of course, the reconstruction is perfect only if we know the exact values of  $f\left(\frac{n\pi}{\Omega}\right)$ . If we have a maximum error of  $\epsilon$  in the first  $N$  sample values, i.e.  $\tilde{f}_n = f\left(\frac{n\pi}{\Omega}\right) + \epsilon_n$ , with  $|\epsilon_n| \leq \epsilon$  and  $\epsilon_n = 0$  for  $n > N$ , then we have

$$|f(t) - \tilde{f}(t)| \leq C\epsilon \log N, \quad (1.2)$$

where  $\tilde{f}(t)$  is calculated by replacing the sample values of  $f$  in (1.1) by  $\tilde{f}_n$ .

Obviously, this is not good because in practice we always have inaccurate measurements and if  $N$  is also large, we might end up with a substantial reconstruction error.

One way to overcome this problem is **oversampling**: Instead of using  $f_n = f(\frac{n\pi}{\Omega})$ , let us sample more frequently and use  $f_n^\lambda = f(\frac{n\pi}{\lambda\Omega})$ ,  $\lambda > 1$ , to reconstruct. In this case, one can prove that

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f\left(\frac{n\pi}{\lambda\Omega}\right) g\left(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\right), \quad (1.3)$$

if  $g$  satisfies:

$$\hat{g}(\xi) = \begin{cases} \frac{1}{\sqrt{2\pi}} & |\xi| \leq \Omega \\ 0 & |\xi| \geq \lambda\Omega \end{cases} \quad (1.4)$$

$$\hat{g} \in C^\infty. \quad (1.5)$$

Because  $g$  is smooth with fast decay, we expect the reconstruction formula to be more robust. Indeed, we can show that

$$|f(t) - \tilde{f}(t)| \leq \epsilon C_g \frac{\Omega}{\pi}, \quad (1.6)$$

where

$$C_g \leq \frac{\Omega}{\pi} (\|g\|_{L^1} + \frac{1}{\lambda} \|g'\|_{L^1}), \quad (1.7)$$

and

$$\tilde{f}(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} \tilde{f}_n g\left(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\right),$$

with  $\tilde{f}_n = f\left(\frac{n\pi}{\lambda\Omega}\right) + \epsilon_n$ ,  $\epsilon_n \leq \epsilon$ .

## 1.2 Quantization

We have shown that a bandlimited  $L^2$  function,  $f$ , can be perfectly represented by a sequence of real numbers,  $f_n^\lambda = f\left(\frac{n\Omega}{\lambda\pi}\right)$  with  $\lambda \geq 1$ . The important question now is how to represent the real numbers,  $f_n^\lambda$ , by a discrete set of numbers which is possibly finite. In other words, we want to “quantize”  $f_n^\lambda$ .

There are many ways to quantize; most are aimed at quantizing with relatively fine resolution [2]. In the first part of the thesis, i.e. in Chapters 1 through 5, we will restrict ourselves to a particular class of quantization algorithms called sigma-delta ( $\Sigma\Delta$ ) quantization schemes. These schemes are commonly used to quantize oversampled bandlimited functions very coarsely. Moreover, we will restrict ourselves to the very extreme case where we will replace the sample values by just one bit.

## 1.3 Sigma-Delta quantization

We are interested in quantizing an oversampled, bandlimited function,  $f$ . For simplicity, we will assume  $\Omega = \pi$ . Also, we will restrict ourselves to functions  $f$  such that  $\|f\|_{L^\infty} \leq \alpha < 1$ . From (1.3) we know that

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f\left(\frac{n}{\lambda}\right) g\left(t - \frac{n}{\lambda}\right). \quad (1.8)$$

We want to find a sequence  $q_n^\lambda$  such that

$$\tilde{f}(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \quad (1.9)$$

is a “good” approximation of  $f$ .

### 1.3.1 First-order sigma-delta quantization

A first-order sigma-delta quantizer produces  $(q_n^\lambda)_{n \in \mathbb{Z}}$  via the following scheme:

$$\begin{aligned} v_n - v_{n-1} &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(v_{n-1} + f_n^\lambda) \end{aligned} \quad (1.10)$$

where  $v$  is an internal state variable, with  $v_0 \in (-1, 1)$ . In this case, one can show that [3]

$$\bullet |v_n| < 1 \text{ for all } n, \text{ and} \quad (1.11)$$

$$\bullet |f(t) - \tilde{f}(t)| \leq \frac{1}{\lambda} \|g'\|_{L^1}. \quad (1.12)$$

### 1.3.2 Higher order sigma-delta quantization schemes

Define  $(\Delta^k v)_n = \sum_{l=0}^k (-1)^l \binom{k}{l} v_{n-l}$ . Note that  $(\Delta^0 v)_n = v_n$  and  $(\Delta^1 v)_n = v_n - v_{n-1}$ . A  $k^{\text{th}}$  order sigma-delta quantization scheme is defined by the following system of difference equations:

$$\begin{aligned} (\Delta^k v)_n &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(M((\Delta^0 v)_n, \dots, (\Delta^{k-1} v)_n, f_n^\lambda)) \end{aligned} \quad (1.13)$$

where  $M$  is an arbitrary function on  $\mathbb{R}^{k+1}$  constructed so that the sequence  $(v_n)$  stays bounded. In this case we have:

**Theorem 1.** *Let  $f \in L^2(\mathbb{R})$ ,  $\text{supp } f \subset [-\pi, \pi]$ , and  $\|f\|_{L^\infty} \leq \alpha < 1$ . Suppose, for a given  $M$ , that  $(v_n)_{n \in \mathbb{Z}}$ , produced by (1.13), is a bounded sequence. Then, for all*

$t \in \mathbb{R}$ ,

$$\left| f(t) - \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| \leq \frac{1}{\lambda^k} \|v\|_{l^\infty} \|g^{(k)}\|_{L^1} \quad (1.14)$$

The proof of Theorem 1, as well as an explicit construction of a family of  $k^{\text{th}}$  order stable sigma-delta quantizers (i.e. quantizers for which  $(v_n)_{n \in \mathbb{Z}}$  is guaranteed to remain bounded) is presented in [3]. Throughout the rest of Part I, we will mostly discuss properties of second-order sigma-delta schemes, for different rules  $M$ , for both “standard” and modified quantizers. In particular we will introduce schemes where the quantized  $q_n^\lambda$  can take the value 0 as well as  $\pm 1$ ; we also discuss a “finite memory” version of sigma-delta. Similar finite memory  $\Sigma\Delta$ -schemes were considered earlier by other authors, e.g. [7, 8]. These schemes have special advantages that we will discuss later.

Our main concern is the stability and robustness of these various second-order schemes. In practice, since the schemes have to be implemented with analog hardware, the function  $M$  used in the quantizer (1.13) is never known exactly; for instance, if  $M$  is a linear function, then all its coefficients will be specified within a certain tolerance only. In addition, the quantizer itself is not entirely precise, leading to the replacement of  $\text{sign}(M)$  in (1.13) by  $\text{sign}(M + \epsilon)$ , where the exact value of  $\epsilon$  is unknown;  $\epsilon$  is again known within a certain tolerance only. It is important that the scheme is robust for small changes within these tolerances.

The study of this robustness is one of the main topics of the first part of this thesis, both for the standard scheme, and the enriched alphabet and the finite memory schemes. But before tackling this, we have to derive stability results for all schemes; we show that there exists a bounded region  $R$  that is left invariant by the dynamical system underlying the sigma-delta quantizer; moreover, this  $R$  can be constructed so

that it is itself robust under changes in  $M$  and the quantizer.

In Chapter 2 we review several standard second-order sigma-delta quantizers, and we introduce and motivate our enriched alphabet and finite memory modified schemes. Sections 3, 4 and 5 then discuss the stability and robustness for the standard scheme, the enriched alphabet scheme, and the finite memory scheme, respectively.

# Chapter 2

## Second-order sigma-delta quantizers

### 2.1 Standard second-order sigma-delta quantizer

Let us first discuss in some more detail the standard second-order scheme. It corresponds to the following system of difference equations:

$$\begin{aligned}(\Delta^2 v)_n &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(M((\Delta^1 v)_n, (\Delta^0 v)_n, f_n^\lambda)).\end{aligned}\tag{2.1}$$

Let us put  $u_n = v_n - v_{n-1}$ . Then (2.1) becomes:

$$\begin{aligned}u_n - u_{n-1} &= f_n^\lambda - q_n^\lambda \\ v_n - v_{n-1} &= u_n \\ q_n^\lambda &= \text{sign}(M(u_{n-1}, v_{n-1}, f_n^\lambda)).\end{aligned}\tag{2.2}$$

Figure 2.1: The partition of the  $(u,v)$ -plane. In each figure  $C_M$  denotes the curve consisting of points  $(u,v)$  at which  $M(u,v) = 0$ . In the left-most graph  $M(u,v,x) = u + 0.2v$ ; in the middle graph  $M(u,v,x) = u + x + 5\text{sign}(v)$  with  $x = 0$ ; in the right-most graph  $M(u,v,x) = \frac{6|x|-7}{3} + (u + 0.5(x + 3\text{sign}(x)))^2 + 2(\text{sign}(x) - x)v$  with  $x = 0.5$ .

Note that  $M$  determines the way we partition the  $(u,v)$ -space into two regions,  $\Lambda_+(x)$  and  $\Lambda_-(x)$  where

$$\begin{aligned}\Lambda_+(x) &= \{(u,v) : M(u,v,x) \geq 0\} \\ \Lambda_-(x) &= \{(u,v) : M(u,v,x) < 0\}.\end{aligned}$$

This is illustrated in Figure 2.1. Some examples from the literature are [4, 3, 6]:

- $M(u,v,x) = u + \gamma v$  with  $\gamma > 0$ ,
- $M(u,v,x) = u + x + M\text{sign}(v)$  with  $M > 0$ ,
- $M(u,v,x) = \frac{6|x|-7}{3} + (u + \frac{x+3\text{sign}(x)}{2})^2 + 2(\text{sign}(x) - x)v$ .

Note that in either region,  $\Lambda_+(x)$  or  $\Lambda_-(x)$ , the system described in (2.2), is affine.

Indeed, we can write:

$$\begin{pmatrix} u_n \\ v_n \end{pmatrix} = \begin{cases} S_l^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in \Lambda_+ \\ S_r^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in \Lambda_- \end{cases} \quad (2.3)$$

$$:= S(u_{n-1}, v_{n-1}, f_n^\lambda), \quad (2.4)$$

where

$$S_l^x(u, v) = A \begin{pmatrix} u \\ v \end{pmatrix} + (x-1) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and

$$S_r^x(u, v) = A \begin{pmatrix} u \\ v \end{pmatrix} + (x+1) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

with  $A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ .

## 2.2 The output sequence $q_n^\lambda$ has infinite memory

In this section, we want to concentrate on the output sequence  $(q_n^\lambda)_{n \in \mathbb{Z}}$  of a sigma-delta quantizer. By definition of the one-bit sigma-delta quantization,  $(q_n^\lambda)$  is a sequence in  $\{-1, 1\}$  such that  $\sum q_n^\lambda$  follows  $\sum f_n^\lambda$  closely. (This is common to any order sigma-delta.) Indeed, for a stable scheme of arbitrary order  $k$ , we have

$$\left| \sum_{n=1}^N f_n^\lambda - \sum_{n=1}^N q_n^\lambda \right| \leq |u_N - u_0| < 2C, \quad (2.5)$$

where  $u_n = (\Delta^{(k-1)}v)_n$ , and  $C$  is a constant bounding  $(\Delta^{(k-1)}v)_n$  uniformly. Note that  $C$  is independent of  $N$ .

One important question is: What happens if  $f_n^\lambda$  is zero after some  $N$ , i.e.  $|f_n^\lambda| = 0$  for  $n \geq N$ ? Although for true bandlimited functions the samples  $f(\frac{n\Omega}{\lambda})$  cannot really vanish identically for  $n \geq N$ , we may well have  $|f(\frac{n\Omega}{\lambda})| \leq \epsilon$  for  $n \geq N$ . We shall investigate the persistence of the memory of different sigma-delta schemes by investigating their behavior for idealized input that vanishes from one point onwards. (To avoid confusion with sequences that are samples of a bandlimited function, we will denote such idealized sequences by  $(x_n)$ .)

For the first-order scheme, we can answer the question above easily:

**Proposition 1.** *Let  $x := (x_n)$  be a sequence such that  $\|x\|_{l^\infty} < 1$  and  $x_n = 0$  for all  $n \geq 0$ . Suppose  $v_0$  is arbitrary. Then there exists  $K$  such that  $q_n = q_K(-1)^{n-K}$  for all  $n \geq K$ .*

**Proof:** Since the first-order scheme is a contraction with the invariant set  $(-1, 1)$ , there exists  $K > 0$  such that  $v_{K-1} \in (-1, 1)$ . If  $v_{K-1} \in (0, 1)$ ,  $q_K = \text{sign}(v_{K-1}) = 1$ , and  $v_K = v_{K-1} - 1 < 0$ . Therefore  $q_{K+1} = -1$  and  $v_{K+1} = v_{K-1}$  again. The same reasoning also applies when  $v_{K-1} \in (-1, 0)$ . So, by induction, we conclude that  $q_n = \text{sign}(v_{K-1})(-1)^{n-K}$ .  $\square$

For stable higher order schemes, determining the exact asymptotic structure of the sequence  $(q_n^\lambda)$  produced by zero input is an open problem. Typically it is a one-sided periodic sequence in  $\{-1, 1\}$  that sums up to zero over one period.

## 2.3 Defeating the infinite memory: Introducing an enriched alphabet

The one-bit sigma-delta quantizer is very effective for coarse quantization of long lasting signals (e.g. audio). We will be interested in using these coarse quantization schemes in different contexts, where it will be specifically useful to segment zones, where the input is negligible. In particular, we shall introduce a longer alphabet containing 0 as well as 1 and -1, and study constraints under which stretches of zero input translate to stretches of zero output. For such schemes,  $(q_n^\lambda)$  would carry the information on the support of the input in a direct way. Input sequences with finite support would be represented by finite output sequences (i.e.  $q_n^\lambda \neq 0$  for only finitely many times.). Even in audio, when sigma-delta quantizers are used in D/A conversion, the filters used in the reconstruction of the analog signal are such that periodic oscillatory patterns in the  $q_n$  cause “pure tone” oscillatory artifacts. Long

stretches of such patterns automatically arise when the input  $f(\frac{v}{\lambda})$  becomes very small. One can make an ideal abstraction of this phenomenon by studying the behavior of the quantizer for input  $x_n = 0$  for  $n \geq M$ . With a tri-level quantizer that allows  $q_n$  to be zero, it would be interesting and useful to have a scheme that ensures that such tail-vanishing  $x_n$  lead to vanishing  $q_n$  after some point  $N$ .

One way to introduce 0 into the alphabet is by changing the quantizer, i.e. we replace (1.13) by

$$\begin{aligned} (\Delta^k v)_n &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \begin{cases} 0; & \text{if } |M(\cdot)| \leq 0.5 \\ \text{sign}(M(\cdot)); & \text{otherwise} \end{cases} := r(M(\cdot)). \end{aligned} \quad (2.6)$$

Indeed, for the first-order case, the tri-level quantization scheme described in (2.6) is doing what we want:

**Proposition 2.** *Suppose  $v_0 \in (-1, 1)$ , and  $x_n = 0$  for all  $n \geq 0$ . Then,  $(q_n)_{n \geq 2}$ , produced by (2.6) with  $k = 1$  and  $M(v, x) = v + x$ , is identically 0.*

**Proof:** First, note that if  $v_0 \in (-1, 1)$  and  $x_1 = 0$ ,  $v_1 \in (-0.5, 0.5)$ . Now suppose  $v_{n-1} \in (-0.5, 0.5)$ . Then  $v_n = v_{n-1} - q_n = v_{n-1}$  because  $q_n = r(v_{n-1}) = 0$ . By induction we are done.  $\square$

Proposition 2 shows that we reach our goal in the case of first-order quantization. Now let us consider the second-order sigma-delta quantization.

**Proposition 3.** *Let  $M$  be chosen such that the system, described in (2.6) with  $k=2$ , is stable, i.e. there exists a constant  $C$  such that for all input sequences  $(x_n)_{n \in \mathbb{Z}}$  satisfying  $|x_n| \leq 1$  for all  $n$ , we have  $|v_n| < C$  for all  $n$ . Now suppose  $x_n = 0$  for all  $n \geq 0$  and  $q_1 = 0$ . Then  $q_n = 0 \forall n \geq 1$  if and only if  $u_0 = 0$ , where  $u_n = v_n - v_{n-1}$ .*

**Proof:** First, suppose that  $u_0 = 0$ . Then, by induction, we have

1.  $u_n = u_0 = 0, \forall n$ : Suppose  $u_{n-1} = 0$  and  $q_n = 0$ . Then  $u_n = u_{n-1} - q_n = 0$ .

2.  $v_n = v_0, \forall n$ :  $v_n = v_{n-1} + u_n = v_{n-1}$ . So put  $n = 1$ .

Therefore,  $q_n = M(u_{n-1}, v_{n-1}, 0) = M(u_0, v_0, 0) = 0$ .

On the other hand, suppose  $q_n = 0$  for all  $n \geq 0$  with  $u_0 \neq 0$ . Then  $v_n = v_0 + nu_0$  which implies that  $|v_n|$  grows unboundedly since  $u_0 \neq 0$ .  $\square$

Proposition 3 implies that, for a second-order scheme, changing the quantizer to (2.6) helps only if the initial value  $u_0$  is zero. In other words, the largest invariant set  $I \subset \mathbb{R}^2$  for zero constant input such that  $r(u, v) = 0$  for every  $(u, v) \in I$ , is a measure zero subset of  $\mathbb{R}^2$ . It follows that if we start running the quantizer with an input function  $f$  that is non-zero but converges to zero, we do **not** expect to have  $q_n = 0$  even though  $f$  becomes negligibly small, because typically  $u_n$  will not vanish.

Note that the reconstruction theorem, Theorem 1, still holds for these tri-level schemes.

## 2.4 Defeating the infinite memory: Finite-memory (leaky) sigma-delta quantization

In the previous section we have shown that we cannot have a stable tri-level second-order scheme that represents a sequence  $(x_n)_{n \geq 0} \equiv 0$  with a sequence  $(q_n)_{n \geq 0} \equiv 0$  for arbitrary initial conditions. This indicates, in some sense, that the system has “infinite memory”. We will now turn our attention to a finite memory version of the above-described sigma-delta schemes to avoid this problem.

Let  $0 < \beta_\lambda < 1$ ,  $f$  as before, and define the first-order finite memory scheme as

follows:

$$\begin{aligned} v_n &= \beta_\lambda v_{n-1} + f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(\beta_\lambda v_{n-1} + f_n^\lambda) \end{aligned} \quad (2.7)$$

The system defined in (2.7) is equivalent to the first-order sigma-delta quantization scheme given in (1.10) if  $\beta_\lambda = 1$ . When  $\beta_\lambda < 1$ , the discrete integrators in our system are leaky, i.e. the storage of a value in memory is not perfect. Instead of  $v_{n-1}$ , after one time unit, we have  $\beta_\lambda v_{n-1}$  in memory. Physically one always encounters some leakage and this is usually considered to be a problem (or an imperfection)[7, 8].

Throughout this chapter we will assume that the integrator leakage depends on the sampling rate (or oversampling ratio). It is reasonable to take

$$\beta_\lambda = e^{-\frac{c}{\lambda}}, \quad (2.8)$$

where  $c$  is some constant, and  $\lambda$  is the oversampling ratio. (If the sigma-delta scheme is built with analog hardware, as in A/D converters, then keeping the  $v_{n-1}$  in memory for one step requires using a capacitor, which is bound to have an exponential leakage for a time interval  $1/\lambda$  as in (2.8); when the scheme is implemented digitally, as in D/A converters, we always have the freedom to choose  $\beta_\lambda$  as in (2.8).)

First of all we want to show that we can reconstruct  $f$  using  $(q_n^\lambda)$  with an error bound of order  $\frac{1}{\lambda}$  in the first-order case.

**Theorem 2.** *Let  $f \in L^2(\mathbb{R})$  be bandlimited with  $\text{supp} \hat{f} \subseteq [-\pi, \pi]$  and  $\|f\|_{L^\infty} \leq 1$ . Let  $g$  be a function satisfying (1.4) and (1.5) with  $\Omega = \pi$ . Let the leakage factor be  $\beta_\lambda = e^{-\frac{c}{\lambda}}$ . Assume that the sequence  $(v_n)$  generated by (2.7) is bounded. If  $(q_n^\lambda)$  is*

the output of the first-order leaky sigma-delta quantizer given in (2.7), then

$$|f(t) - \tilde{f}(t)| \leq \frac{\|v\|_{L^\infty}}{\lambda} (\|g'\|_{L^1} + cC_g), \quad (2.9)$$

where  $C_g$  is as in (1.7) with  $\Omega = \pi$ , and  $\tilde{f}(t) = \frac{1}{\lambda} \sum q_n^\lambda g(t - \frac{n}{\lambda})$ .

**Proof:** We have  $v_n - \beta_\lambda v_{n-1} = f_n^\lambda - q_n^\lambda$ . Therefore

$$\begin{aligned} f(t) - \tilde{f}(t) &= \frac{1}{\lambda} \left( \sum (v_n - v_{n-1}) g(t - \frac{n}{\lambda}) + (1 - \beta_\lambda) \sum v_{n-1} g(t - \frac{n}{\lambda}) \right) \\ &= \frac{1}{\lambda} \left( \sum v_n \left( g(t - \frac{n}{\lambda}) - g(t - \frac{n+1}{\lambda}) \right) + (1 - \beta_\lambda) \sum v_{n-1} g(t - \frac{n}{\lambda}) \right). \end{aligned}$$

Then, substituting  $\beta_\lambda = e^{-\frac{c}{\lambda}}$ , and using the fact that  $e^x \geq 1 + x$  for all  $x$ , we get

$$\begin{aligned} |f(t) - \tilde{f}(t)| &\leq \frac{\|v\|_{L^\infty}}{\lambda} \left( \sum |g(t - \frac{n}{\lambda}) - g(t - \frac{n+1}{\lambda})| + \frac{c}{\lambda} \sum |g(t - \frac{n}{\lambda})| \right) \\ &\leq \frac{\|v\|_{L^\infty}}{\lambda} (\|g'\|_{L^1} + cC_g). \end{aligned}$$

□

A similar result holds for second order. In this case, we define the finite memory scheme as

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda \\ v_n &= \beta_\lambda v_{n-1} + u_n \\ q_n^\lambda &= \text{sign}(M(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})). \end{aligned} \quad (2.10)$$

**Theorem 3.** Let  $f, g$  and  $\beta_\lambda$  be as in Theorem 2. Assume that  $(v_n)$ , generated by (2.10) is bounded. If  $(q_n^\lambda)$  is the output of the second-order leaky sigma-delta quantizer

given in (2.10), then

$$|f(t) - \tilde{f}(t)| \leq \frac{\|v\|_{l^\infty}}{\lambda^2} (\|g''\|_{L^1} + 2c\|g'\|_{L^1} + 2c^2C_g), \quad (2.11)$$

where  $C_g$  is as before and  $\tilde{f}(t) = \frac{1}{\lambda} \sum q_n^\lambda g(t - \frac{n}{\lambda})$ .

**Proof:** We have  $v_n - 2\beta_\lambda v_{n-1} - \beta_\lambda^2 v_{n-2} = f_n^\lambda - q_n^\lambda$ , with  $\beta_\lambda = e^{-\frac{c}{\lambda}}$ . Therefore

$$\begin{aligned} f(t) - \tilde{f}(t) &= \frac{1}{\lambda} \left( \sum (\Delta^2 v)_n g(t - \frac{n}{\lambda}) + 2(1 - \beta_\lambda) \sum v_{n-1} g(t - \frac{n}{\lambda}) \right. \\ &\quad \left. - (1 - \beta_\lambda^2) \sum v_{n-2} g(t - \frac{n}{\lambda}) \right) \\ &= \frac{1}{\lambda} \left( \sum (\Delta^2 v)_n g(t - \frac{n}{\lambda}) + 2(1 - \beta_\lambda) \sum (v_{n-1} - v_{n-2}) g(t - \frac{n}{\lambda}) \right. \\ &\quad \left. + (1 - \beta_\lambda)^2 \sum v_{n-2} g(t - \frac{n}{\lambda}) \right). \end{aligned} \quad (2.12)$$

Using Theorem 1 and the fact that  $e^x \geq 1 + x$  for all  $x$ , we get

$$|f(t) - \tilde{f}(t)| \leq \frac{\|v\|_{l^\infty}}{\lambda} \left( \frac{1}{\lambda} \|g''\|_{L^1} + \frac{2c}{\lambda} \|g'\|_{L^1} + \frac{2c^2}{\lambda} C_g \right). \quad (2.13)$$

□

We define the tri-level finite memory second-order sigma-delta quantizer by replacing  $sign$  in (2.10) by  $r$ , as in (2.6), i.e.

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda \\ v_n &= \beta_\lambda v_{n-1} + u_n \\ q_n^\lambda &= r(M(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})). \end{aligned} \quad (2.14)$$

Note that Theorem 3 still holds (supposing, again, that the  $v_n$  generated by the tri-level finite memory second-order are bounded).

Now we will turn our attention back to the output sequence  $(q_n^\lambda)$  for an input

sequence identically equal to zero after some  $N$ .

**Proposition 4.** *Consider the tri-level finite memory second-order sigma-delta quantizer defined in (2.14) with  $M(u, v) = u + \gamma v$ . Let the input sequence  $(x_n)$  be identically equal to zero for all  $n \geq N$  for some  $N$ , suppose  $q_N = 0$  and*

$$|u_{N-1}| < \frac{(1 - \beta_\lambda)}{2\gamma\beta_\lambda^2}. \quad (2.15)$$

Then  $q_n = 0$  for all  $n \geq N$ .

**Proof:** We use induction. We know that  $q_N^\lambda = 0$ , which implies

$$|\beta_\lambda(u_{N-1} + \gamma v_{N-1})| \leq 0.5, \quad (2.16)$$

since  $q_N^\lambda = r(\beta_\lambda(u_{N-1} + \gamma v_{N-1}))$ . We also know that

$$\begin{aligned} u_N &= \beta_\lambda u_{N-1} \\ v_N &= \beta_\lambda(v_{N-1} + u_{N-1}) \end{aligned} \quad (2.17)$$

since  $x_N = 0$  and  $q_N = 0$ . Then

$$q_{N+1} = r(\beta_\lambda(\beta_\lambda(u_{N-1} + \gamma v_{N-1}) + \gamma\beta_\lambda u_{N-1})). \quad (2.18)$$

But by (2.15) and (2.16) we have

$$\begin{aligned} &|\beta_\lambda(\beta_\lambda(u_{N-1} + \gamma v_{N-1}) + \gamma\beta_\lambda u_{N-1})| \\ &\leq \beta_\lambda|\beta_\lambda(u_{N-1} + \gamma v_{N-1})| + \gamma\beta_\lambda|u_{N-1}| \\ &\leq 0.5, \end{aligned} \quad (2.19)$$

which implies that  $q_{N+1} = 0$ . Since  $0 < \beta_\lambda < 1$ , (2.18) implies  $|u_N| < |u_{N-1}|$  and by

induction we are done. □

Proposition 4 shows that the tri-level finite memory second-order sigma-delta quantizer produces an all-zero output sequence  $(q_n)_{n \geq 1}$  if the input sequence  $(x_n)_{n \geq 1}$  is identically equal to zero, and  $(u_0, v_0) \in \Lambda$  where

$$\Lambda = \{(u, v) : |\beta_\lambda(u + \gamma v)| \leq 0.5, |u| < (1 - \beta_\lambda)/(2\gamma\beta_\lambda^2)\}$$

is a subset of  $\mathbb{R}^2$  with positive measure.

**Remark:**

In Section 3 we explicitly construct a compact set  $R$  in the  $(u, v)$ -plane which is invariant under all second-order sigma-delta schemes described above. Unlike the “non-leaky” tri-level case, in the leaky tri-level case we have  $\Lambda$  with positive measure such that  $q_{n+l} = 0$  for all  $0 \leq l \leq L$  if  $(u_{n-1}, v_{n-1}) \in \Lambda$  and  $x_{n+l} = 0$  for all  $0 \leq l \leq L$ . Note that  $\Lambda$  and all its preimages (under the dynamical system associated with the second-order leaky tri-level sigma-delta quantizer) do not cover all of  $R$ . In fact, there are points in the invariant set  $R$  that have periodic orbits outside  $\Lambda$ . For example, take  $\gamma = 0.2$ ,  $\beta_\lambda = 0.9$  and consider the point  $(u, v) = (1/(1 + \beta_\lambda), 1/(1 + \beta_\lambda)^2)$ . One readily checks that  $S_{LT}(u, v, 0) = (-u, -v)$  and  $S_{LT}^2(u, v, 0) = (u, v)$ , where  $S_{LT}$  is the map that describes the dynamical system corresponding to the finite memory (leaky) tri-level scheme (which is defined in 5.4). Because  $r(\beta_\lambda(u + \gamma v)) = 1$  and  $r(\beta_\lambda(-u - \gamma v)) = -1$ , we see that  $(u, v)$  and  $(-u, -v)$  constitute a periodic orbit outside  $\Lambda$ . Thus, if  $x_n = 0$  for  $n \geq N$  and  $(u_N, v_N) = (1/(1 + \beta_\lambda), 1/(1 + \beta_\lambda)^2) := P$ , then  $q_{N+k} = (-1)^{k+1}$  for  $k \geq 1$ . A similar oscillating tail results if  $(u_N, v_N)$  is any preimage of the point  $P$  under a power of  $S_{LT}(\cdot, \cdot, 0)$ .

Numerical observations suggest that the rule  $M$  can be adjusted to guarantee that

the set of points  $(u', v')$  for which  $S_{LT}^n(u', v', 0) \notin \Lambda$  holds for all  $n$  has measure zero. A more detailed analysis of the fine structure of  $R$  is in progress.

In all the theorems we have proven so far we assume stability, which we define as the existence of a uniform bound for the  $v_n$ . For first-order schemes, it is easy to prove that  $(v_n)_{n \in \mathbb{Z}}$  is an  $l^\infty$  sequence. However, for the higher order schemes, proving boundedness of  $(v_n)$  is harder. We will start by proving stability of the standard second-order scheme for a particular family of the function  $M$  used in the quantizer. Then we will extend this to “non-standard” schemes of interest to us, using the same  $M$ . In particular, we will consider the non-standard schemes with

- a tri-level quantizer, i.e. the scheme described in (2.6) with  $k = 2$ ,
- a finite memory quantizer, i.e. the scheme described in (2.10),
- a finite memory tri-level quantizer, i.e. the scheme described in (2.14).

The quantization rule,  $M$ , will be specified when necessary.

## Chapter 3

# Stability and robustness of the standard second-order sigma-delta quantizer

### 3.1 Stability of the standard second-order scheme

In this section, we will prove the stability of the second-order scheme, defined in (2.2), with the quantization rule  $M(u, v, x) = u + \gamma v$  for a range of  $\gamma$  to be specified later. Since any  $M$  of this form does not depend on  $x$ , we will drop  $x$  from its argument, i.e.  $M = M(u, v)$ . In the next section, we will show that the stability result is valid not only for  $M$  of this form but for a wide range of rules, which will be described in detail.

We will restrict the input sequence  $(x_n)$  to  $|x_n| \leq \alpha < 1$ . Then  $\delta_n = |x_n - q_n|$  can take values from  $\delta_- = 1 - \alpha$  to  $\delta_+ = 1 + \alpha$ . The system defined in (2.2) can be

Figure 3.1:  $\Gamma_{B_1}$  and  $\Gamma_{B_2}$  are the graphs of the functions  $B_1$  and  $B_2$ , respectively.  $L$  is the line consisting of the points  $(u, v)$  for which  $M(u, v) = 0$ .

rewritten as

$$\begin{aligned} (u_n, v_n) &= \begin{cases} S_l^{\delta_n}(u_{n-1}, v_{n-1}) = (u_{n-1} - \delta_n, u_{n-1} + v_{n-1} - \delta_n) & \text{if } q_n = 1 \\ S_r^{\delta_n}(u_{n-1}, v_{n-1}) = (u_{n-1} + \delta_n, u_{n-1} + v_{n-1} + \delta_n) & \text{if } q_n = -1 \end{cases}, \\ q_n &= \text{sign}(M(u_{n-1}, v_{n-1})), \end{aligned} \quad (3.1)$$

In this case we will also write

$$(u_n, v_n) = S(u_{n-1}, v_{n-1}, \delta_n). \quad (3.2)$$

Let us define now the functions

$$B_1(u) = \begin{cases} -\frac{1}{2\delta_-}(u - \frac{\delta_-}{2})^2 + \frac{\delta_-}{8} + C; & \text{if } u \geq 0 \\ -\frac{1}{2\delta_+}(u - \frac{\delta_+}{2})^2 + \frac{\delta_+}{8} + C; & \text{if } u < 0 \end{cases}, \quad (3.3)$$

$$B_2(u) = \begin{cases} \frac{1}{2\delta_+}(u + \frac{\delta_+}{2})^2 - \frac{\delta_+}{8} - C; & \text{if } u \geq 0 \\ \frac{1}{2\delta_-}(u + \frac{\delta_-}{2})^2 - \frac{\delta_-}{8} - C; & \text{if } u < 0 \end{cases}, \quad (3.4)$$

where the constant  $C$  will have to be determined below.

Figure 3.1 illustrates the graphs  $\Gamma_{B_1}$ , respectively  $\Gamma_{B_2}$ , of the function  $B_1$ , respec-

tively  $B_2$ , for one particular choice of  $C$ ; it also shows the region trapped between  $\Gamma_{B_1}$  and  $\Gamma_{B_2}$  which we denote by  $R$ . If we start from  $(u_{n-1}, v_{n-1})$  in  $R$ , then depending on whether  $v_{n-1} \geq l(u_{n-1}) := -\frac{1}{\gamma}u_{n-1}$  or  $v_{n-1} \leq l(u_{n-1})$  (note that the graph  $L$  of  $l$  is exactly the set of  $(u, v)$  where  $M(u, v) = 0$ ), a move  $S_l^\delta$  or  $S_r^\delta$  will be applied to find the next  $(u_n, v_n)$ . We thus split  $R$  into two regions  $R_1$  and  $R_2$ . More precisely,

$$\begin{aligned} R_1 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), v \geq l(u)\} \\ R_2 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), v \leq l(u)\} \\ R &= R_1 \cup R_2. \end{aligned} \tag{3.5}$$

Note that any line  $L$  with a  $v$ -axis intercept between  $-C$  and  $C$  intersects  $\Gamma_{B_1}$  at two points. From this point on we will refer to the intersection point with *smaller* first coordinate whenever we say  $L \cap \Gamma_{B_1}$  for any line  $L$  of this type. Similarly, for any line  $L$  with a  $v$ -axis intercept between  $-C$  and  $C$ ,  $L \cap \Gamma_{B_2}$  will refer to the intersection point of  $L$  and  $\Gamma_{B_2}$  with the *larger* first coordinate.

To fix the notation, let us make another remark. For any point  $P$ ,  $u(P)$  will refer to the first coordinate of  $P$ , and  $v(P)$  will refer to the second coordinate of  $P$ , i.e.  $u(P) = x$  and  $v(P) = y$  for a point  $P = (x, y)$ .

We will denote the left-most intersection point of  $\Gamma_{B_1}$  and  $\Gamma_{B_2}$  by  $P_0 = (u_0, v_0)$ , the intersection of  $L$  and  $\Gamma_{B_1}$  by  $P_1 = (u_1, v_1)$ , and the intersection of  $L$  and  $\Gamma_{B_2}$  by  $P_2 = (u_2, v_2)$ , where  $L$  is the graph of the line consisting of points  $(u, v)$  for which  $M(u, v) = 0$ . Note that  $P_0 = (u_0, v_0)$ , shown in Figure 3.1, is given by

$$\begin{aligned} u_0 &= -[2C(1 - a^2)]^{1/2}, \\ v_0 &= B_1(u_0), \end{aligned} \tag{3.6}$$

and  $(u_2, v_2) = (-u_1, -v_1)$ .

**Lemma 1.** *The region below the graph  $\Gamma_{B_1}$  of  $B_1$  is invariant for all possible moves*

$$S_l^\delta : (u, v) \rightarrow (u - \delta, u + v - \delta),$$

if  $\delta \in [\delta_-, \delta_+]$ . In other words, if  $v \leq B_1(u)$ , then  $u + v - \delta \leq B_1(u - \delta)$  for  $\delta \in [\delta_-, \delta_+]$ .

**Proof:**

1. Case 1:  $u \leq 0$ . By construction of  $B_1$  we have  $B_1(u - \delta_+) = B_1(u) + u - \delta_+$ . Suppose  $v \leq B_1(u)$ . Then it is enough to show that  $B_1(u) + u - \delta \leq B_1(u - \delta)$ . Now,

$$\begin{aligned} B_1(u) + u - \delta &= B_1(u) + u - \delta_+ - \delta + \delta_+ \\ &= B_1(u - \delta_+) - \delta + \delta_+. \end{aligned}$$

In other words, we want to prove

$$\delta_+ - \delta \leq B_1(u - \delta) - B_1(u - \delta_+). \quad (3.7)$$

But,

$$B_1(u - \delta) - B_1(u - \delta_+) = \frac{-1}{2\delta_+}(\delta_+ - \delta)(2u - \delta - 2\delta_+). \quad (3.8)$$

Then (3.7) reduces to  $u \leq \delta/2$ , which is true for  $u \leq 0$ .

2. Case 2:  $u \geq \delta$ . In this case, both  $u$  and  $u - \delta$  are nonnegative. Therefore, by construction of  $B_1$ , we have  $B_1(u - \delta_-) = B_1(u) + u - \delta_-$ . We again want to

prove that  $B_1(u) + u - \delta \leq B_1(u - \delta)$ . We have

$$\begin{aligned} B_1(u) + u - \delta &= B_1(u) + u - \delta_- - \delta + \delta_- \\ &= B_1(u - \delta_-) - \delta + \delta_-. \end{aligned}$$

Proceeding as before, we want to show

$$-(\delta - \delta_-) \leq B_1(u - \delta) - B_1(u - \delta_-), \quad (3.9)$$

which reduces to showing that

$$B_1(u - \delta) - B_1(u - \delta_-) = \frac{1}{2\delta_-}(\delta_- - \delta)(2u - \delta - 2\delta_-) \geq -(\delta - \delta_-). \quad (3.10)$$

But (3.10) is true if and only if  $u \geq \delta/2$ , which is true since we are considering the case  $u \geq \delta$ .

3. Case 3: It remains to check  $0 \leq u \leq \delta$ . In this case,

$$B_1(u) = -\frac{1}{2\delta_-}\left(u - \frac{\delta_-}{2}\right)^2 + \frac{\delta_-}{8} + C, \quad (3.11)$$

and

$$B_1(u - \delta) = -\frac{1}{2\delta_+}\left(u - \frac{\delta_+}{2}\right)^2 + \frac{\delta_+}{8} + C. \quad (3.12)$$

Again we want to show that  $B_1(u) - B_1(u - \delta) \leq \delta - u$ , which reduces to

$$\frac{1}{2}\left(\frac{1}{\delta_+} - \frac{1}{\delta_-}\right) + u\left(1 - \frac{\delta}{\delta_+}\right) + \frac{\delta}{2}\left(\frac{\delta}{\delta_+} - 1\right) \leq 0. \quad (3.13)$$

But the left hand side equals

$$\left(1 - \frac{\delta}{\delta_+}\right)\left(-\frac{1}{2\delta}u^2 + u - \frac{\delta}{2}\right) + \frac{1}{2}u^2\left(\frac{1}{\delta} - \frac{1}{\delta_-}\right). \quad (3.14)$$

Since  $(1 - \frac{\delta}{\delta_+}) \geq 0$ ,  $(-\frac{1}{2\delta}u^2 + u - \frac{\delta}{2}) \leq 0$  and  $(\frac{1}{\delta} - \frac{1}{\delta_-}) \leq 0$  we indeed have (3.13).

□

**Lemma 2.** *The region above the graph  $\Gamma_{B_2}$  of  $B_2$  is invariant for all possible moves*

$$S_r^\delta : (u, v) \rightarrow (u + \delta, u + v + \delta),$$

if  $\delta \in [\delta_-, \delta_+]$ .

**Proof:** Similar to the proof of the previous lemma. □

We shall now determine the conditions on the function  $M(u, v) = u + \gamma v$  and the constant  $C$  ensuring that  $S_l^\delta(R_1) \subset R$  and similarly  $S_r^\delta(R_2) \subset R$ .

**Theorem 4.** *Let  $P_1 = (u_1, v_1)$  be the intersection point of the line  $L$ , defined by  $M(u, v) = u + \gamma v = 0$ , and  $\Gamma_{B_1}$ , as shown in Figure 3.1. Suppose*

$$u_0 + \delta_+ \leq u_1 \leq -\delta_+. \quad (3.15)$$

*Then  $S_l^\delta(R_1) \subseteq R$ , for any  $\delta \in [\delta_-, \delta_+]$ .*

**Proof:** By Lemma 1, we know that  $S_l^\delta(R_1)$  lies under  $\Gamma_{B_1}$ . Therefore, we need to prove only that  $S_l^\delta(R_1)$  stays above  $\Gamma_{B_2}$ .

Note that if  $v_1 \geq v_2$ , then  $(u, v_1)$  and  $(u, v_2)$  get mapped to  $(u', v'_1)$  and  $(u', v'_2)$  with  $v'_1 \geq v'_2$ . Hence, we need to check only that the map of the line segment connecting  $P_1$  to  $P_2$ , and the map of  $\Lambda$ , a piece of  $\Gamma_{B_2}$  shown in Figure 3.1, stay above  $\Gamma_{B_2}$ . (More precisely,  $\Lambda = \{(u, v) : v = B_2(u), \text{ and } u_2 \leq u \leq -u_0\}$ .) Moreover, since the region above  $\Gamma_{B_2}$  is convex, and the map  $S_l^\delta$  is linear in  $u, v$ , it suffices, for the line segment, to check only the end points  $P_1$  and  $P_2$ . Also, for each end point, since the map  $S_l^\delta$

is linear in  $\delta$ , we only need to check  $\delta = \delta_-$  and  $\delta = \delta_+$ . For the curved piece,  $\Lambda$ , we similarly need to check only for  $\delta = \delta_-$  and  $\delta = \delta_+$ .

1. Since  $P_1$  is in the left half plane,  $S_l^{\delta+}$  maps  $P_1$  to a point on  $\Gamma_{B_1}$  by construction. Moreover,  $u(S_l^{\delta+}(P_1)) = u_1 - \delta_+ \geq u_0$ . Therefore,  $S_l^{\delta+}(P_1)$  is above  $\Gamma_{B_2}$ .  
 $S_l^{\delta-}(P_1)$  is on the line through  $S_l^{\delta+}(P_1)$  with slope 1 in the increasing direction. Since  $B_2'(u) \leq 1$  for  $u \leq 0$ , and  $u(S_l^{\delta-}(P_1)) \leq 0$ , it follows that  $S_l^{\delta-}(P_1)$  is above  $\Gamma_{B_2}$ .
2. We know that  $u_2 = u(P_2) = -u_1$ . Then,  $u(S_l^{\delta+}(P_2)) = -u_1 - \delta_+ \geq 0$  by our condition on  $u_1$ . But  $B_2$  is increasing in  $u$  for  $u \geq 0$ , thus  $B_2(u(S_l^{\delta+}(P_2))) < B_2(u_2)$ . We also know that  $v(S_l^{\delta+}(P_2)) > v_2 = v(P_2)$ . Hence we have that  $S_l^{\delta+}(P_2)$  is above  $\Gamma_{B_2}$ . Because  $0 < \delta_- < \delta_+$ , we also conclude that  $u(S_l^{\delta-}(P_2)) \geq 0$ , and hence  $S_l^{\delta-}(P_2)$  is above  $\Gamma_{B_2}$ .
3. Finally, we want to show that  $S_l^\delta(\Lambda)$  lies above  $\Gamma_{B_2}$ . But by our condition this is obvious:  $u(S_l^\delta(P))$  will be positive for any  $P$  on  $\Lambda$  for  $\delta \in [\delta_-, \delta_+]$ . Therefore, since  $v(S_l^\delta(P)) \geq v(P)$  for any point  $P$  on  $\Lambda$  ( $u$  value of any point on  $\Lambda$  is greater than  $|u_1|$ , and thus greater than  $\delta_+$ .) and since  $B_2(u)$  is increasing for  $u \geq 0$ , we will have  $S_l^\delta(\Lambda)$  above  $\Gamma_{B_2}$  for any  $\delta \in [\delta_-, \delta_+]$ .

□

**Remarks:**

1. The condition  $u_0 + \delta_+ \leq u_1 \leq -\delta_+$  makes sense only if  $u_0 = -[2C(1 - a^2)]^{1/2} \leq -2\delta_+$  which is equivalent to the condition

$$C \geq 2 \frac{1 + \alpha}{1 - \alpha}. \tag{3.16}$$

2. The range of  $\gamma$  for a given  $C \geq 2\frac{1+\alpha}{1-\alpha}$  is:

$$\frac{1}{\gamma} \geq \frac{[2C(1-a^2)]^{1/2} + 2\alpha C}{2\{[2C(1-a^2)]^{1/2} - (1+\alpha)\}}, \quad (3.17)$$

and

$$\frac{1}{\gamma} \leq \frac{C - (1+\alpha)}{1+\alpha}. \quad (3.18)$$

For the minimum allowed value of  $C$ , i.e. if  $C = 2\frac{1+\alpha}{1-\alpha}$ , we have

$$\frac{1}{\gamma} = 1 + \frac{2\alpha}{1-\alpha}. \quad (3.19)$$

Similarly one can prove that  $S_r^\delta(R_2)$  is a subset of  $R$ . Hence we will conclude:

**Theorem 5.** *Let  $S$  be the mapping defined by (3.2) with the rule  $M(u, v) = u + \gamma v$ . Suppose  $C$  and  $\gamma$  satisfy (3.16), (3.17) and (3.18) for some  $\alpha < 1$ . Then the set  $R$  is positively invariant under  $S(\cdot, \cdot, \delta)$  for any  $\delta \in [1 - \alpha, 1 + \alpha]$ . In other words,  $S(u, v, \delta) \in R$  for any  $(u, v) \in R$  and  $\delta \in [1 - \alpha, 1 + \alpha]$ .*

**Corollary 1.** *Let  $(x_n)$  be an arbitrary sequence such that  $|x_n| \leq \alpha < 1$ . Suppose  $(u_0, v_0) \in R$  and  $(u_n, v_n)$  are obtained via the recursion defined in (2.1) with  $M(u, v) = u + \gamma v$ . If  $C$  and  $\gamma$  satisfy (3.16), (3.17) and (3.18),  $(u_n, v_n) \in R$  for all  $n$ ; thus  $|v_n| < C$  for all  $n$ .*

This shows that the second-order sigma-delta scheme is stable for the range of quantizers described in Theorem 5. There are similarities between our invariant region  $R$  and the trapping region  $R_c$  described by Pinault and Lopresti in [5]. One difference is that we did not impose any conditions on the input sequence  $(x_n)$ , except that it remains bounded. Pinault and Lopresti, on the other hand, consider input sequences of the form  $x_n = x_c + \tilde{x}_n$ , where  $|x_c| < 1$  and  $\tilde{x}_n$  is such that the partial

sums  $a_n = \sum_{k=1}^n \tilde{x}_k$  and  $b_n = \sum_{k=1}^n a_k$  remain bounded. Since for the signals in which one is interested in practice the low frequency content is negligible, such conditions are very reasonable as long as the oversampling ratio  $\lambda$  remains fixed. We will be interested in studying the asymptotic behavior for a wide range of  $\lambda$ ; in that case the bounds on  $a_n$  and  $b_n$  would increase as  $O(\lambda)$  and  $O(\lambda^2)$ , respectively, leading to increasingly large trapping regions  $R_c$ . There is no such dependence on  $\lambda$  for our invariant set  $R$ , because it is completely determined by  $C$ ,  $\gamma$  and  $\alpha$ , the upper bound on the input sequence. As long as  $|f| < \alpha$ , and  $C$  and  $\gamma$  satisfy (3.16), (3.17), and (3.18), the set  $R$ , defined by (3.5), is invariant for any input sequence  $(f(\frac{n}{\lambda}))$ , so that unlike the trapping region  $R_c$  in [5], our invariant region  $R$  stays fixed when we change the oversampling ratio.

Note that the choice of the quantization rule may affect the robustness of the scheme which we will discuss next.

## 3.2 Robustness of the standard second-order scheme

Theorem 5 implies robustness of the second-order sigma-delta scheme with respect to certain variations of  $\gamma$ . Indeed, we have the same bound on the reconstruction error, defined in (1.14), for all  $\gamma$  within the allowed range specified in (3.17) and (3.18). Moreover, our analysis still holds even if  $\gamma$  does not remain fixed, but varies with  $n$ , i.e. if in (3.1) we replace  $\gamma$  in  $M$  by  $\gamma_n$  during the iteration, where  $\gamma_n$  all satisfy (3.17) and (3.18), for some fixed  $C$ . This is because the bound on the reconstruction error depends only on  $\|g''\|_{L^1}$  and on the uniform bound on  $|v_n|$ , as shown in Theorem 1; by Theorem 5, the set  $R$  remains invariant as long as  $\gamma_n$  at each step satisfies (3.17) and (3.18), leading to the same uniform bound on  $|v_n|$ .

In this section we will show that the second-order sigma-delta scheme in (3.1) with  $M(u, v) = u + \gamma v$  is also robust with respect to shifts in offset of the line  $L$ .

This is very important for practical applications. In A/D conversion, for example, where sigma-delta schemes are widely used, we are in the world of analog signals and equipment until we obtain the sequence  $(q_n)$ . Therefore, it is impossible to know what the exact value of  $\gamma$  is in the quantization rule  $M$ , and it is impossible to know what the “toggle point” of our quantizer really is. More precisely, a perfect one-bit quantizer is supposed to compare its input with zero, and decide whether it is greater than zero or not. A practical (analog) quantizer, however, can be modeled as a comparator whose output is 1 if the input is greater than some  $\epsilon_1$ , -1 if the input is less than some  $\epsilon_2$ , and 1 or -1 if the input is between  $\epsilon_1$  and  $\epsilon_2$ , where we assume  $\epsilon_2 \leq \epsilon_1$ . The value of  $\epsilon_i$  is not known, and it can also change in time, depending on external factors like temperature, oversampling ratio, etc.. So, we would like to have a scheme that has a fixed invariant set  $R$ , and hence a fixed bound  $C$  on  $v_n$ , as long as  $|\epsilon_i| < \epsilon$ , for some  $\epsilon > 0$  whose value we can control. If we have this, then the estimate for the reconstruction error will remain unchanged by Theorem 1 for reasons we have explained in the previous paragraph.

Now we will prove that the second-order sigma-delta scheme is indeed robust for such imprecisions in the quantizer. Suppose that we have a stable second-order sigma-delta scheme, given in (3.1) with  $M(u, v) = u + \gamma v$ , with the invariant set  $R$  corresponding to some  $C$  and  $\gamma$ , where  $C$  satisfies (3.16) with strict inequality, and  $\gamma$  satisfies (3.17) and (3.18). In this case, we will show that there exists  $\epsilon_0 > 0$  such that  $R$  is also invariant with respect to the second-order sigma-delta scheme with  $M^\epsilon(u, v) = u + \gamma v + \epsilon$ , as long as  $|\epsilon| < \epsilon_0$ .

**Proposition 5.** *Let  $M^\epsilon(u, v) = u + \gamma v + \epsilon$  with  $|\epsilon| < \gamma C$  be the quantization rule used in the second-order sigma-delta scheme, described in (3.1). Let  $u_0$  be as in (3.7). Let  $L^\epsilon$  be the line consisting of points  $(u, v)$  that satisfy  $M^\epsilon(u, v) = 0$ , and define  $P_1^\epsilon = L^\epsilon \cap \Gamma_{B_1}$  and  $P_2^\epsilon = L^\epsilon \cap \Gamma_{B_2}$ . Suppose the input sequence  $\|x_n\|_{l^\infty}$  is*

bounded by  $\alpha$ , as before. Then the set  $R$ , as in (3.5), is positively invariant if both

$$u_0 + 1 + \alpha < u(P_1^\epsilon) < -(1 + \alpha), \quad (3.20)$$

$$1 + \alpha < u(P_2^\epsilon) < -u_0 - (1 + \alpha), \quad (3.21)$$

hold.

**Proof:** Let us define

$$\begin{aligned} \tilde{R}_1 &= R \cap \{(u, v) : v > -\frac{1}{\gamma}(u + \epsilon)\} \\ \tilde{R}_2 &= R \setminus \tilde{R}_1. \end{aligned}$$

We will first show that  $S_i^\delta(\tilde{R}_1) \subset R$ , for any  $\delta \in [\delta_-, \delta_+]$ . Note that  $\tilde{R}_1$  is convex, and  $S_i^\delta$  is linear in its arguments and in  $\delta$ . Moreover if  $v_1 \geq v_2$ , then  $(u, v_1)$  and  $(u, v_2)$  get mapped to  $(u', v'_1)$  and  $(u', v'_2)$  with  $v'_1 \geq v'_2$ . Therefore it is enough to show that  $S_i^\delta(P_1^\epsilon)$ ,  $S_i^\delta(P_2^\epsilon)$ , and  $S_i^\delta(\Lambda(P_2^\epsilon))$ , where we define

$$\Lambda(P) = \{(u, v) : v = B_2(u), u(P) < u < -u_0\}, \quad (3.22)$$

are in  $R$  for  $\delta = \delta_+$  and  $\delta = \delta_-$ .

Suppose that  $P_1^\epsilon$  and  $P_2^\epsilon$  satisfy (3.20) and (3.21), respectively. Then, clearly, both  $P_1^\epsilon$  and  $-P_2^\epsilon$  satisfy (3.15); by Theorem 5  $S_i^{\delta_+}(P_i^\epsilon)$ , and  $S_i^{\delta_-}(P_i^\epsilon)$ ,  $i = 1, 2$ , are in  $R$ . Moreover, again by Theorem 5, if any point  $P$  satisfies (3.15), then both  $S_i^{\delta_+}$  and  $S_i^{\delta_-}$  map  $\Lambda(-P)$  into  $R$ . Since  $-P_2^\epsilon$  satisfies (3.15), we conclude that both  $S_i^{\delta_+}(\Lambda(P_2^\epsilon))$  and  $S_i^{\delta_-}(\Lambda(P_2^\epsilon))$  are in  $R$ .

This shows that  $S_i^\delta(\tilde{R}_1) \subset R$  for any  $\delta$  in  $[\delta_-, \delta_+]$ . Finally, by symmetry (replace  $P_1^\epsilon$  by  $-P_2^\epsilon$ , and  $P_2^\epsilon$  by  $-P_1^\epsilon$ , and note that  $-P_2^\epsilon$  satisfies (3.20), and  $-P_1^\epsilon$  satisfies (3.21)), we conclude that  $S_i^\delta(R \setminus \tilde{R}_1) \subset R$ , and hence  $S_i^\delta(R) \subseteq R$  for any  $\delta \in [\delta_-, \delta_+]$ .  $\square$

We now investigate under what condition on  $\epsilon$ , for fixed  $\gamma$  and  $C$ , (3.20) and (3.21) will be satisfied.

**Theorem 6.** *Let  $M(u, v) = u + \gamma v$  be a given quantization rule with  $\gamma$  satisfying (3.17) and (3.18) with strict inequalities for some  $C > 2\frac{1+\alpha}{1-\alpha}$ . Let  $u_0$  be as in (3.7) and let  $(u_1, v_1)$  be the intersection of  $L$  with  $\Gamma_{B_1}$ , as in Theorem 4. Take  $\epsilon$  such that*

$$|\epsilon| < \min\{u_1(\alpha, \gamma, C) - u_0(\alpha, C) - (1 + \alpha); -u_1(\alpha, \gamma, C) - (1 + \alpha)\}, \quad (3.23)$$

with

$$\begin{aligned} u_0(\alpha, C) &= -[2C(1 - a^2)]^{1/2}, \quad \text{as in (3.7),} \\ u_1(\alpha, \gamma, C) &= (1 + \alpha) \left( \frac{\gamma+2}{2\gamma} - \left[ \left( \frac{\gamma+2}{2\gamma} \right)^2 + \frac{2C}{1+\alpha} \right]^{1/2} \right). \end{aligned} \quad (3.24)$$

Then the second-order sigma-delta scheme with the quantization rule  $M^\epsilon(u, v) = u + \gamma v + \epsilon$  is stable with the invariant set  $R$ , where  $R$  is as in (3.5).

**Proof:** If  $C > 2\frac{1+\alpha}{1-\alpha}$  and if  $\gamma$  satisfies (3.17) and (3.18) with strict inequalities, both  $u_1 - u_0 - (1 + \alpha)$ , and  $-u_1 - (1 + \alpha)$  are positive; therefore (3.23) makes sense.

Note that (3.23) can be rewritten as

$$|\epsilon| + u_0 + \delta_+ < u_1 < -|\epsilon| - \delta_+. \quad (3.25)$$

Since the line  $L$  is passing through the origin,  $P_2 = L \cap \Gamma_{B_2}$  is equal to  $-P_1$ , and therefore we also have

$$|\epsilon| + \delta_+ < u_2 < -|\epsilon| - u_0 - \delta_+. \quad (3.26)$$

One checks by explicit calculation that  $u_1$  is as in (3.24). Let us denote the line consisting of the points  $(u, v)$  such that  $M^\epsilon(u, v) = 0$  by  $L^\epsilon$ . Let  $P_1^\epsilon = L^\epsilon \cap \Gamma_{B_1}$  and

$P_2^\epsilon = L^\epsilon \cap \Gamma_{B_2}$ . Note that these points are well-defined since (3.23) guarantees that the  $v$ -axis intercept of  $L^\epsilon$  is between  $-C$  and  $C$ . We only need to show that  $P_1^\epsilon$  and  $P_2^\epsilon$  satisfy (3.20) and (3.21), respectively, if they satisfy (3.25) and (3.26), respectively, since then by Proposition 5 we will be done.

Assume (3.25) and (3.26) are true. Then clearly we know that  $u_1 < 0$  and  $u_2 > 0$ . Also,

$$|u(P_i^\epsilon) - u_i| \leq |\epsilon|, \quad (3.27)$$

because  $B_1(u)$  is increasing for negative  $u$  and  $B_2(u)$  is increasing for positive  $u$ , and  $L$  and  $L_\epsilon$  have identical negative slopes. But then, since we have

$$u_i - |\epsilon| < u(P_i^\epsilon) < u_i + |\epsilon|, \quad i = 1, 2. \quad (3.28)$$

The combination of (3.28) and (3.25) implies that  $P_1^\epsilon$  satisfies (3.20); similarly combining (3.28) and (3.26) we have that  $P_2^\epsilon$  satisfies (3.21). Hence we conclude that  $R$  is positively invariant under the second-order sigma-delta scheme with the rule  $M^\epsilon$ .  $\square$

**Remarks:**

1. Proposition 5 and Theorem 6, along with Theorem 5 show that the second-order sigma-delta scheme with the family of quantization rules we are considering is not only stable, but provides us a range of parameters for which we have a fixed positively invariant set. The invariant set also remains fixed if we replace  $M(u, v) = u + \gamma v$  in (3.1) with  $M_n(u, v) = u + \gamma_n v + \epsilon_n$  at each step of the iteration defined in (3.1), as long as  $\gamma_n$  satisfies (3.17) and (3.18) for all  $n$ , and  $\epsilon_n$  satisfies (3.23) for all  $n$ , when we replace  $\gamma$  in (3.23) with  $\gamma_n$ .
2. We do not have to partition the plane by a line. Define  $L_r$  to be the line

Figure 3.2:  $P_{r,i}$  and  $P_{l,i}$  are the points defined in the second remark above.  $\Gamma_{\tilde{M}}$  the curve consisting of the points  $(u, v)$  for which  $\tilde{M}(u, v) = 0$ ;  $\tilde{M}$  is such that the conditions described in the second remark above are satisfied.

segment connecting  $P_{r,1} = (-\delta_+, B_1(-\delta_+))$  and  $P_{r,2} = (-u_0 - \delta_+, B_2(-u_0 - \delta_+))$ ; likewise  $L_l$  to be the line segment connecting  $P_{l,1} = (u_0 + \delta_+, B_1(u_0 + \delta_+))$  and  $P_{l,2} = (\delta_+, B_2(\delta_+))$ . Then, with little effort, one can see that the set  $R$  is positively invariant under the mapping  $S(\cdot, \cdot, \delta)$  with the rule  $\tilde{M}$  as long as the set of points  $(u, v) \in R$  such that  $\tilde{M}(u, v) = 0$  constitutes a continuous curve that stays between the line segments  $L_r$  and  $L_l$ . An example is illustrated in Figure 3.2.

3. By above remark we observe that the set  $R$  corresponding to a sufficiently large  $C$  is also invariant under the second-order sigma-delta quantization scheme introduced by [3].

# Chapter 4

## Stability and robustness of the tri-level second-order quantizer

### 4.1 Stability of the tri-level quantizer

In this section we will consider the second-order sigma-delta scheme given in (2.6) with  $k = 2$ , i.e.

$$\begin{aligned} u_n - u_{n-1} &= f_n^\lambda - q_n^\lambda \\ v_n - v_{n-1} &= u_n \\ q_n^\lambda &= \begin{cases} 1; & \text{if } M(u_{n-1}, v_{n-1}, f_n^\lambda) > 0.5 \\ 0; & \text{if } |M(u_{n-1}, v_{n-1}, f_n^\lambda)| \leq 0.5 \\ -1; & \text{if } M(u_{n-1}, v_{n-1}, f_n^\lambda) < -0.5 \end{cases}, \end{aligned} \quad (4.1)$$

with the same  $M$  we used in the previous section, i.e.

$$M(u, v) = u + \gamma v \quad (4.2)$$

for some range of  $\gamma$ . We will prove that, under some additional constraints, this system is stable with the same invariant set  $R$  as in Theorem 5. Let  $L : M(u, v) = 0$ ,  $L_1 : M(u, v) = 0.5$  and  $L_2 : M(u, v) = -0.5$  be the lines whose graphs are shown in Figure 4.1. Define

$$\begin{aligned} R'_1 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), M(u, v) > 0.5\} \\ R'_2 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), M(u, v) < -0.5\} \\ R'_0 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), |M(u, v)| \leq 0.5\}, \end{aligned} \quad (4.3)$$

such that

$$R'_0 \cup R'_1 \cup R'_2 = R, \quad (4.4)$$

where  $R$  is identical to the invariant set in Theorem 5. Note that the scheme described in (4.1) is equivalent to

$$(u_n, v_n) = \begin{cases} S_l^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in R'_1 \\ S_r^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in R'_2 \\ S_0^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in R'_0 \end{cases}, \quad (4.5)$$

$$:= S_T(u_{n-1}, v_{n-1}, f_n^\lambda), \quad (4.6)$$

where

$$S_0^\delta : (u, v) \rightarrow (u + \delta, u + v + \delta). \quad (4.7)$$

To prove stability, we will show that  $S_0^\delta$  maps  $R'_0$  into  $R$ . We already know from Theorem 5 that  $S_l^\delta(R'_1) \subset S_l^\delta(R_1) \subset R$  since  $R'_1 \subset R_1$ , and similarly  $S_r^\delta(R'_2) \subset S_r^\delta(R_2) \subset R$  since  $R'_2 \subset R_2$ , assuming that the conditions given in Theorem 5 are satisfied. We

Figure 4.1:  $\Gamma_{B_1}$  and  $\Gamma_{B_2}$  are the graphs of  $B_1$  and  $B_2$ .  $L$ ,  $L_1$  and  $L_2$  are lines consisting of the points  $(u, v)$  such that  $M(u, v) = 0$ ,  $M(u, v) = 0.5$  and  $M(u, v) = -0.5$ , respectively.

therefore need to show only that  $S_0^\delta$  maps  $R'_0$  into  $R$  to conclude that the tri-level quantizer described in (4.1) is stable.

First of all, let  $P_0 = (u_0, v_0)$ ,  $P_1 = (u_1, v_1)$  and  $P_2 = (u_2, v_2)$  be as in Theorem 4. Denote the point  $L_1 \cap \Gamma_{B_1}$  by  $P_3 = (u_3, v_3)$ , the point  $L_2 \cap \Gamma_{B_1}$  by  $P_5 = (u_5, v_5)$ . Denote the point  $L_1 \cap \Gamma_{B_2}$  by  $P_4 = (u_4, v_4)$  and the point  $L_2 \cap \Gamma_{B_2}$  by  $P_6 = (u_6, v_6)$ .

**Theorem 7.** *Let  $P_1 = (u_1, v_1) = L \cap \Gamma_{B_1}$ , where  $L$  is the line consisting of points  $(u, v)$  that satisfy  $M(u, v) = 0$ . Suppose  $C$ , satisfying (3.16), and  $\alpha < 1$  are such that*

$$u_0 + \delta_+ < u_1 < -\delta, \tag{4.8}$$

*with  $\delta = 1 + \delta_+/2$  and  $\delta_+ = 1 + \alpha$ , for some  $\gamma$  satisfying (3.17) and (3.18). Then  $S_0^\delta(R'_0) \subset R$ , and the system defined in (4.1) is stable with the invariant set  $R$ , where  $R$  is as in (3.5).*

**Proof:** We need to check only that  $S_0^\delta(R'_0) \subset R$  for reasons explained above. Since  $S_0^\delta$  is linear in its arguments and in  $\delta$ , and since  $R'_0$  is convex, it is enough to check

whether  $S_0^\delta(P_i)$  is in  $R$  for  $i = 3, 4, 5, 6$ , and  $S_0^\delta(\Lambda_i) \subset R$  for  $i = 1, 2$ , where

$$\begin{aligned}\Lambda_1 &= \{(u, v) : v = B_1(u), u_5 < u < u_3\} \\ \Lambda_1 &= \{(u, v) : v = B_1(u), u_5 < u < u_3\}\end{aligned}\tag{4.9}$$

Clearly,  $u_3 \leq u_1 + 0.5 \leq -\delta_+/2 - 1/2$  by construction. This implies that  $P'_3 = (u_3 + 1, v_3 + 1)$  is in  $R$  (It is above  $B_2$  because  $u'_3 < 0$ ,  $v'_3 > v_3$ , and  $B_2(u)$  is decreasing for negative  $u$ ; and one can easily check that it is under  $B_1$ , because we have an explicit expression for the derivative of  $B_1$ ).  $P'_5 = (u_5 + 1, v_5 + 1)$  is in  $R$  by the same argument. Moreover, we claim that both  $P'_3$  and  $P'_5$  are on  $R_1$ , that is above the line  $L_1$ . This is clear for  $P'_3$  since  $P_3$  itself is on  $R_1$ . We know that  $P'_5$  is above the line  $L_1$ :  $u'_5 > u_5 + 0.5$  and  $v'_5 > v_5$ ; also  $B'_1(u) > 1$  for  $u_5 < u < u'_5$ ,  $P'_5$  is in  $R_1$ . Finally, any point  $P$  on  $\Lambda_1$  with  $u_5 < u(P) < u_3$  will be staying in  $R_1$  when translated by  $(1,1)$ , because the arguments for  $P'_3$  and  $P'_5$  will hold also for  $P$ , i.e.

$$\Lambda'_1 = \{(u + 1, v + 1) : (u, v) \in \Lambda_1\} \subset R_1.$$

But by Theorem 5 we have  $S_0^\delta(P_3) = S_l^\delta(P'_3) \in R$ ,  $S_0^\delta(P_5) = S_l^\delta(P'_5) \in R$  and  $S_0^\delta(\Lambda_1) = S_l^\delta(\Lambda'_1) \subset R$ .

Similarly, by symmetry,

$$\Lambda'_2 = \{(u - 1, v - 1) : (u, v) \in \Lambda_2\},$$

will be contained in  $R_2$ , i.e.  $S_0^\delta(\Lambda_2) = S_l^\delta(\Lambda'_2) \subset R$ , and so will the points  $P'_4 = (u_4 - 1, v_4 - 1)$  and  $P'_6 = (u_6 - 1, v_6 - 1)$ . Thus the proof is complete.  $\square$

### Remarks:

1. The condition (4.8) makes sense only if  $u_0 \leq -\delta_+ - \delta = -\frac{3}{2}\delta_+ - 1$ , which is

equivalent to the condition

$$C \geq \frac{1}{2(1-\alpha^2)} + \frac{3(4+3(1+\alpha))}{8(1-\alpha)}. \quad (4.10)$$

2. For  $C$  satisfying (4.10) the set  $R$ , as in (3.5), is invariant for the tri-level scheme if

$$\frac{1}{\gamma} \geq \frac{B_1(u_0 + \delta_+)}{|u_0 + \delta_+|}, \quad (4.11)$$

which is the same as (3.17), and

$$\frac{1}{\gamma} \leq \frac{B_1(-\delta)}{\delta} = \frac{8C(1+\alpha) - (1+\alpha) - 4}{4(1+\alpha)((1+\alpha) + 2)}, \quad (4.12)$$

with  $\delta_+ = 1 + \alpha$ , and  $\delta$  as in Theorem 7.

## 4.2 Robustness of the tri-level quantizer

Like the standard second-order sigma-delta quantizer, the tri-level second-order quantizer is robust in many different ways. Let us rewrite (4.5) as follows.

$$(u_n, v_n) = \begin{cases} S_l^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } u_{n-1} + \gamma v_{n-1} > 0.5 \\ S_r^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } u_{n-1} + \gamma v_{n-1} < -0.5 \\ S_0^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } |u_{n-1} + \gamma v_{n-1}| < 0.5 \end{cases}, \quad (4.13)$$

Like the case with the standard scheme, the invariant set  $R$  of the tri-level scheme remains fixed for all  $\gamma$  that satisfy (4.11) and (4.12) by Theorem 7.

Now let us replace  $\gamma$  in (4.13) by  $\gamma_n$ , i.e. at every step of the iteration  $\gamma$  changes. Theorem 7 shows that as long as  $C$ , in the definitions of the functions  $B_1$  and  $B_2$ , satisfies (4.10),  $\gamma_n$  satisfies (4.11) and (4.12) for all  $n$ , the set  $R$ , as in (3.5), is invariant

under the scheme in (4.13).

The tri-level scheme is also robust with respect to small shifts of the offset of the line defined by  $L = \{(u, v) : M(u, v) = 0\}$ , i.e. there exists  $\epsilon_0 > 0$  such that the scheme obtained by replacing  $M$  in (4.1) by  $M^\epsilon(u, v) = u + \gamma v + \epsilon$  is stable with the same invariant set  $R$  if  $|\epsilon| < \epsilon_0$ . Note that the proof of Theorem 7 is valid for any line  $L'$  if  $u(L' \cap \Gamma_{B_1})$  and  $-u(L' \cap \Gamma_{B_2})$  satisfy (4.8). Let us replace the rule  $M$  in Theorem 7 by  $M^\epsilon(u, v) = u + \gamma v + \epsilon$  with  $|\epsilon| < \gamma C$ . Let  $L^\epsilon$  be as in Proposition 5. Then if  $u(L^\epsilon \cap \Gamma_{B_1})$  and  $-u(L^\epsilon \cap \Gamma_{B_2})$  satisfy (4.8),  $R$ , as in (3.5), will be invariant for the tri-level quantizer with the quantization rule  $M^\epsilon$ . Similar to Theorem 6, if  $|\epsilon| < \min\{u_1(\alpha, \gamma, C) - u_0(\alpha, C) - (1 + \alpha); -u_1(\alpha, \gamma, C) - \delta\}$ , with  $\delta$  as in Theorem 7,  $u(L^\epsilon \cap \Gamma_{B_1})$  and  $-u(L^\epsilon \cap \Gamma_{B_2})$  will satisfy (4.8). Hence we have proven:

**Theorem 8.** *Let  $M(u, v) = u + \gamma v$  be a given quantization rule with  $\gamma$  satisfying (4.11) and (4.12) for some  $C$  satisfying (4.10). Let  $u_0$  be as in (3.7) and let  $(u_1, v_1)$  be the intersection of  $L$  with  $\Gamma_{B_1}$ , as in Theorem 4. Take  $\epsilon$  such that*

$$|\epsilon| < \min\{u_1(\alpha, \gamma, C) - u_0(\alpha, C) - (1 + \alpha); -u_1(\alpha, \gamma, C) - \delta\}, \quad (4.14)$$

*with  $\delta$  as in Theorem 7,  $u_0(\alpha, C)$  and  $u_1(\alpha, \gamma, C)$  as in (3.24). Then the **tri-level** second-order sigma-delta scheme obtained by replacing  $M$  in (4.1) by  $M^\epsilon(u, v) = u + \gamma v + \epsilon$  is stable with the invariant set  $R$ , where  $R$  is as in (3.5).*

## Chapter 5

# Stability and robustness of the finite-memory second-order sigma-delta quantizer

In this section we will consider the finite memory versions of the standard and tri-level second-order sigma-delta schemes. More precisely, we will consider the standard second-order finite-memory (leaky) sigma-delta scheme which is described by

$$\begin{aligned}u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\v_n &= \beta_\lambda v_{n-1} + \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\q_n &= \text{sign}(M(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})),\end{aligned}\tag{5.1}$$

with  $\beta_\lambda = e^{-\frac{c}{\lambda}}$  where  $c$  some constant; and the tri-level second-order finite-memory (leaky) sigma-delta scheme which is given by

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\ v_n &= \beta_\lambda v_{n-1} + \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\ q_n &= r(M(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})). \end{aligned} \tag{5.2}$$

with  $\beta_\lambda = e^{-\frac{c}{\lambda}}$  where  $c$  is a constant.  $M$  in the above equations will be specified when necessary. We will write

$$(u_n, v_n) = S_{LS}(u_{n-1}, v_{n-1}, f_n^\lambda), \tag{5.3}$$

for the scheme in (5.1), and

$$(u_n, v_n) = S_{LT}(u_{n-1}, v_{n-1}, f_n^\lambda), \tag{5.4}$$

for the scheme in (5.2). Note that if  $S$  and  $S_T$  are as in (3.2) and (4.6), respectively, then

$$S_{LS}(u_{n-1}, v_{n-1}, f_n^\lambda) = S(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1}, f_n^\lambda), \tag{5.5}$$

$$S_{LT}(u_{n-1}, v_{n-1}, f_n^\lambda) = S_T(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1}, f_n^\lambda). \tag{5.6}$$

Then we will have:

**Theorem 9.** *Fix  $0 < \alpha < 1$ . Let  $M(u, v) = u + \gamma v + \epsilon$  be such that the standard second-order sigma-delta scheme, as in (3.1), is stable with the invariant set  $R$  as in (3.5) for some  $C > 0$ . Then the standard second-order finite-memory sigma-delta scheme defined by (5.1) is also stable with the same invariant set  $R$ .*

**Proof:** Let  $(u, v)$  be in  $R$ ,  $\delta \in [\delta_-, \delta_+]$ . We want to show that  $S_{LS}(u, v, \delta) \in R$ . But

by (5.6),  $S_{LS}(u, v, \delta) = S(\beta_\lambda u, \beta_\lambda v, \delta)$ . Since  $R$  is by construction a convex set such that  $(0, 0) \in R$ , and since  $\beta_\lambda < 1$ ,  $(\beta_\lambda u, \beta_\lambda v) \in R$ . Since  $R$  is invariant under  $S(\cdot, \cdot, \delta)$  for  $\delta \in [\delta_-, \delta_+]$ , we have  $S(\beta_\lambda u, \beta_\lambda v, \delta) \in R$ .  $\square$

Similarly, we have:

**Theorem 10.** *Fix  $0 < \alpha < 1$ . Let  $M(u, v) = u + \gamma v + \epsilon$  be such that the tri-level second-order sigma-delta scheme, as in (4.1), is stable with the invariant set  $R$  as in (3.5) for some  $C$  satisfying (4.10). Then the tri-level second-order finite-memory sigma-delta scheme defined by (5.2) is also stable with the same invariant set  $R$ .*

**Proof:** Similar to the proof of the previous theorem.  $\square$

**Remark:**

Theorem 9 implies that all the robustness results proven in Section 3.2 for the standard second-order sigma-delta quantizer are valid for the standard finite-memory second-order sigma-delta quantizer, too. Similarly by Theorem 10 all the robustness results in Section 4.2 for tri-level second-order sigma-delta quantizer are also true for the tri-level finite-memory second-order sigma-delta quantizer.

## Part II

Coarse quantization of highly  
redundant time-frequency  
representations of  
square-integrable functions

# Chapter 6

## Introduction

In the second part of this thesis we will introduce two algorithms to ‘coarsely quantize’ redundant time-frequency representations of certain classes of functions in  $L^2(\mathbb{R})$ . The algorithms are inspired by sigma-delta quantization which is discussed in detail in Part I. However the transposition of the algorithm to the time-frequency representations is nontrivial. The first difficulty arises from the norm of interest. In Part I we concentrated on  $L^\infty$  estimates of the approximation error, which were appropriate for the application purposes we considered. Here, our goal is to have an approximation in  $L^2$  that is achieved for certain classes of functions. Another important difference is that the redundancy in the ‘frequency direction’ will have to be exploited in a different way; sigma-delta methods will have to be adapted in this case because the complex exponential function and its derivative are not integrable. We shall explain this in more detail below, and show how to overcome these problems.

Throughout the second part of the thesis we will be discussing methods to quantize certain frame expansions of functions in  $L^2(\mathbb{R})$ . In particular we are interested in frames of  $L^2(\mathbb{R})$  that are generated by shifting a fixed function  $\varphi \in L^2(\mathbb{R})$  along a lattice  $\Gamma = \tau_0\mathbb{Z} \times \xi_0\mathbb{Z}$  in the time-frequency plane, i.e.  $\{\varphi_{n,m} : n, m \in \mathbb{Z}\}$  will be a

frame in  $L^2(\mathbb{R})$  with

$$\varphi_{n,m}(t) = \varphi(t - n\tau_0)e^{im\xi_0 t}. \quad (6.1)$$

Frames of this form are called *Weyl-Heisenberg Frames*. There are certain conditions that the function  $\varphi$ , and the parameters  $\tau_0$  and  $\xi_0$  must satisfy if the  $\{\varphi_{n,m}\}$  constitute a frame. A detailed discussion can be found in [9, 10]. We will now discuss some properties of frames in general and Weyl-Heisenberg frames in particular which we will need throughout the rest of the thesis. (A detailed analysis of Weyl-Heisenberg Frames can be found in [11].)

**Definition 1.** *A family of functions  $\{\varphi_n : n \in \mathbb{Z}\}$  in a Hilbert space  $H$  is called a frame if there exist constants  $A > 0$  and  $B < \infty$  such that for any  $f \in H$*

$$A\|f\|^2 \leq \sum_n |\langle f, \varphi_n \rangle|^2 \leq B\|f\|^2. \quad (6.2)$$

If  $A = B$ , the frame is said to be a *tight frame*.

**Theorem 11.** *Suppose  $\{\varphi_n\}$  is a frame of a Hilbert space  $H$  with frame bounds  $A$  and  $B$ . Then there exists a frame  $\{\bar{\varphi}_n\}$  with frame bounds  $B^{-1}$  and  $A^{-1}$  such that any  $f \in H$  can be written as*

$$f = \sum \langle f, \varphi_n \rangle \bar{\varphi}_n. \quad (6.3)$$

$\{\bar{\varphi}_n\}$  is called the *dual* of  $\{\varphi_n\}$ . The dual of  $\{\bar{\varphi}_n\}$  is  $\{\varphi_n\}$ , thus we also have

$$f = \sum \langle f, \bar{\varphi}_n \rangle \varphi_n. \quad (6.4)$$

Finally, if  $A = B$  then we have  $\bar{\varphi}_n = \frac{1}{A}\varphi_n$ .

The proof of Theorem 11 can be found in various standard books on the subject,

e.g. [9, 10].

Next let us go back to discussing Weyl-Heisenberg frames. Let  $(\varphi, \tau_0, \xi_0)$  denote the collection  $\{\varphi_{n,m}\}_{(n,m) \in \mathbb{Z}^2}$  with  $\varphi_{n,m}(t) = \varphi(t - n\tau_0)e^{im\xi_0 t}$ , where  $\varphi$  is a fixed function. Suppose  $(\varphi, \tau_0, \xi_0)$  is a tight Weyl-Heisenberg frame of  $L^2(\mathbb{R})$  where  $\varphi$  is a smooth and well-localized function that is normalized in  $L^2$ ,  $x\varphi \in L^2$ , and  $\xi\hat{\varphi} \in L^2$ . If the frame bound is  $A$ , it is a standard result that  $A > 1$  (necessary to have a frame) and  $A = \frac{2\pi}{\tau_0\xi_0}$ . By Theorem 11 we have

$$f = \frac{1}{A} \sum \langle f, \varphi_{n,m} \rangle \varphi_{n,m}, \quad (6.5)$$

where equality is in the sense of  $L^2$ . Note that the continuous windowed Fourier transform of  $f$ , denoted by  $F$ , is given by  $F = \langle f, \varphi_{\tau,\xi} \rangle$ , where  $\varphi_{\tau,\xi} = \varphi(t - \tau)e^{i\xi t}$ . Combining this with (6.5) implies

$$F(\tau, \xi) = \frac{1}{A} \sum_{n,m} \langle f, \varphi_{n,m} \rangle \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle, \quad (6.6)$$

where the convergence is pointwise as well as in  $L^2$ .

Now we are ready to state the problem. Equation (6.5) essentially tells us how to reconstruct  $f$  from its frame coefficients  $\langle f, \varphi_{n,m} \rangle$ . Clearly the  $c_{n,m} := \langle f, \varphi_{n,m} \rangle$  are complex numbers. Our goal is to construct an algorithm to replace the  $c_{n,m}$  by some  $q_{n,m} \in \{d_1, d_2, \dots, d_K\}$ , with  $d_i \in \mathbb{C}$ , (i.e. to quantize  $c_{n,m}$ ) such that

$$\tilde{f}_A = \frac{1}{A} \sum q_{n,m} \varphi_{n,m} \quad (6.7)$$

is a ‘good’ approximation of  $f$  in some norm, preferably in  $L^2$ -norm.

In the next chapter we will introduce a quantization algorithm to obtain  $(q_{n,m})$ . We will discuss how and in what sense we can reconstruct the original functions using the  $(q_{n,m})$ .

# Chapter 7

## The Time-Frequency Sigma-Delta Quantization Algorithm I (TFΣΔ-I)

### 7.1 The Algorithm

We will consider functions  $f \in L^2(\mathbb{R})$  that satisfy  $|\langle f, \varphi_{n,m} \rangle| < 1$  for all integers  $n$  and  $m$ . Denote the collection of such functions by  $\mathcal{B}^\varphi$ . Let  $a_{n,m}$  and  $b_{n,m}$  be the real and imaginary parts of  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$  respectively. Now consider the recursions:

$$\begin{aligned} u_{n,m}^R - u_{n-1,m}^R &= a_{n,m} - p_{n,m}^R \\ p_{n,m}^R &= \text{sign}(u_{n-1,m}^R + a_{n,m}) \\ \\ v_{n,m}^R - v_{n,m-1}^R &= u_{n,m}^R - r_{n,m}^R \\ r_{n,m}^R &= \text{sign}(v_{n,m-1}^R + u_{n,m}^R) \end{aligned} \tag{7.1}$$

and

$$\begin{aligned}
u_{n,m}^I - u_{n-1,m}^I &= b_{n,m} - p_{n,m}^I \\
p_{n,m}^I &= \text{sign}(u_{n-1,m}^I + b_{n,m}) \\
v_{n,m}^I - v_{n,m-1}^I &= u_{n,m}^I - r_{n,m}^I \\
r_{n,m}^I &= \text{sign}(v_{n,m-1}^I + u_{n,m}^I),
\end{aligned} \tag{7.2}$$

where

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}.$$

The difference equations given in (7.1) will be used to quantize the real part of the frame coefficients  $c_{n,m}$  and (7.2) produces the bit sequence we will use to quantize the imaginary part of  $c_{n,m}$ . Denote the sequences  $(u_{n,m}^R), (v_{n,m}^R), (u_{n,m}^I)$  and  $(v_{n,m}^I)$  by  $u^R, v^R, u^I$  and  $v^I$  respectively. Similarly  $p^R, r^R, p^I$  and  $r^I$  will denote  $(p_{n,m}^R), (r_{n,m}^R), (p_{n,m}^I)$  and  $(r_{n,m}^I)$  respectively. Note that

$$(\Delta_1 \Delta_2 v^R)_{n,m} = a_{n,m} - (p_{n,m}^R + (\Delta_1 r^R)_{n,m}), \tag{7.3}$$

and

$$(\Delta_1 \Delta_2 v^I)_{n,m} = b_{n,m} - (p_{n,m}^I + (\Delta_1 r^I)_{n,m}), \tag{7.4}$$

where  $(\Delta_1 v)_{n,m} := v_{n,m} - v_{n-1,m}$  and  $(\Delta_2 v)_{n,m} := v_{n,m} - v_{n,m-1}$ . We will define the sequences  $q^R$  and  $q^I$  by  $q_{n,m}^R := p_{n,m}^R + (\Delta_1 r^R)_{n,m}$  and  $q_{n,m}^I := p_{n,m}^I + (\Delta_1 r^I)_{n,m}$ ,

respectively. Let  $c := (c_{n,m})$  and define the mapping  $T_{TF}$  from  $l^2(\mathbb{C})$  to  $\mathcal{Q}$  by

$$T_{TF}(c) = q := q^R + iq^I; \quad (7.5)$$

where  $\mathcal{Q}$  denotes the collection of all sequences  $(x_{n,m} + iy_{n,m})$  where both  $x_{n,m}$  and  $y_{n,m}$  take values in  $\{-3, -1, 1, 3\}$ .

**Theorem 12.** *Let  $(\varphi, \tau_0, \xi_0)$  be a tight Weyl-Heisenberg frame of  $L^2(\mathbb{R})$  with frame bound  $A$ . Let  $f$  be in  $\mathcal{B}^\varphi$  and set  $q = T_{TF}(c)$  where  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$ . Consider the function*

$$\tilde{F}_A(\tau, \xi) = \frac{1}{A} \sum_{n,m} q_{n,m} \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle. \quad (7.6)$$

Suppose  $\varphi$  is chosen such that  $\Phi(\tau, \xi) = \langle \varphi, \varphi_{\tau,\xi} \rangle$  satisfies

- $\tau\xi\Phi(\tau, \xi) \in L^1(\mathbb{R}^2)$ ,
- $\tau\partial_1\Phi(\tau, \xi) \in L^1(\mathbb{R}^2)$ ,
- $\xi\partial_2\Phi(\tau, \xi) \in L^1(\mathbb{R}^2)$ , and
- $\partial_1\partial_2\Phi(\tau, \xi) \in L^1(\mathbb{R}^2)$ .

where  $\partial_i\Phi$  is the  $i^{\text{th}}$  partial derivative of  $\Phi$ . Then

$$|F(\tau, \xi) - \tilde{F}_A(\tau, \xi)| \leq \frac{1}{A}(C_{\varphi,1} + |\tau|C_{\varphi,2}), \quad (7.7)$$

where  $C_{\varphi,1}$  and  $C_{\varphi,2}$  depend only on  $\varphi$ . We will call  $\tilde{F}_A$  the time-frequency sigma-delta approximation of  $F$ .

Before we proceed to prove this theorem we observe that the discussion in Section 1.3.1 implies:

**Lemma 3.** For each  $u^R, v^R, u^I, v^I$ , defined as in (7.1) and (7.2), the  $l_\infty$ -norm is bounded by 1.

**Proof:** Note that  $u^R$  is the state variable of a first-order sigma-delta quantizer, described in (1.10), where the sequence  $(a_{n,m})$  is the input and the sigma-delta quantization is over the index  $n$ . Since  $f \in \mathcal{B}^\varphi$ ,  $|a_{n,m}|$  is bounded by 1. Then by (1.11)  $u_{n,m}^R$  is bounded by 1. Similarly,  $v_{n,m}^R$  are the state variables of a first-order sigma-delta quantizer with the input  $(u_{n,m})$ , where sigma-delta quantization is over  $m$ ; again since  $u_{n,m}^R$  is bounded by 1, so is  $v_{n,m}^R$ . The proof also applies for  $u_{n,m}^I$  and  $v_{n,m}^I$  since  $b_{n,m}$  is bounded by 1, too.  $\square$

Now, we are ready to prove Theorem 12.

**Proof of Theorem 12:** First note that

$$\langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle = \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi), \quad (7.8)$$

where  $\Phi_{n,m}(\tau, \xi) := \Phi(\tau - n\tau_0, \xi - m\xi_0)$  and  $\alpha_{n,m}(\xi) = e^{-in\tau_0(\xi - m\xi_0)}$ . Now let us write the error term, i.e.

$$F(\tau, \xi) - \tilde{F}_A(\tau, \xi) = \frac{1}{A} \sum_{n,m} (c_{n,m} - q_{n,m}) \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi), \quad (7.9)$$

$$= \frac{1}{A} \sum_{n,m} (\Delta_1 \Delta_2 v)_{n,m} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi), \quad (7.10)$$

$$= \frac{1}{A} \sum_{n,m} v_{n,m} (\bar{\Delta}_2 \bar{\Delta}_1 \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi))_{n,m}, \quad (7.11)$$

where, for any  $x = (x_{n,m})$ ,  $(\bar{\Delta}_1 x)_{n,m} := x_{n,m} - x_{n+1,m}$  and  $(\bar{\Delta}_2 x)_{n,m} := x_{n,m} - x_{n,m+1}$ . (To avoid unnecessarily complicated notation, sometimes we will write  $(\Delta_i x_{n,m})$  instead of  $(\Delta_i x)_{n,m}$ , and  $(\bar{\Delta}_i x_{n,m})$  instead of  $(\bar{\Delta}_i x)_{n,m}$ .) The first equality is obvious,

the second comes directly from the quantization algorithm by setting

$$v_{n,k} = v_{n,k}^R + iw_{n,k}^I. \quad (7.12)$$

The third equality is the result of summing (7.10) by parts; note that the boundary values disappear since  $\alpha_{n,m}(\xi)\Phi_{n,m}(\tau, \xi)$  vanishes as  $n$  and/or  $m$  tends to infinity for any  $\tau, \xi$ . Let us define  $I$  by  $I := (\bar{\Delta}_2 \bar{\Delta}_1 \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi))_{n,m}$ . Then

$$I = \bar{\Delta}_2 \bar{\Delta}_1 \left( e^{-in\tau_0(\xi - m\xi)} \Phi(\tau - n\tau_0, \xi - m\xi_0) \right), \quad (7.13)$$

$$= e^{-i\tau\xi} \bar{\Delta}_2 \bar{\Delta}_1 e^{i\tau m\xi_0} \Gamma(\tau - n\tau_0, \xi - m\xi_0), \quad (7.14)$$

where  $\Gamma(t, z) := e^{itz} \Phi(t, z)$ . Now set  $\Omega_{\tau, \xi}(t, z) := e^{iz\tau} \Gamma(t, \xi - z)$  to get

$$I = e^{-i\tau\xi} \bar{\Delta}_2 \bar{\Delta}_1 \Omega_{\tau, \xi}(\tau - n\tau_0, m\xi_0). \quad (7.15)$$

Since  $\Omega_{\tau, \xi}$  is smooth, we can rewrite (7.15) as

$$\begin{aligned} I &= e^{-i\tau\xi} \left( \bar{\Delta}_2 \int_{\tau - (n+1)\tau_0}^{\tau - n\tau_0} \partial_1 \Omega_{\tau, \xi}(t, m\xi_0) dt \right) \\ &= e^{-i\tau\xi} \int_{\tau - (n+1)\tau_0}^{\tau - n\tau_0} [\partial_1 \Omega_{\tau, \xi}(t, m\xi_0) - \partial_1 \Omega_{\tau, \xi}(t, (m+1)\xi_0)] dt \\ &= e^{-i\tau\xi} \int_{\tau - (n+1)\tau_0}^{\tau - n\tau_0} \int_{(m+1)\xi_0}^{m\xi_0} \partial_2 \partial_1 \Omega_{\tau, \xi}(t, z) dt dz \end{aligned} \quad (7.16)$$

Substituting (7.16) into (7.11) we obtain

$$F(\tau, \xi) - \tilde{F}_A(\tau, \xi) = \frac{1}{A} \sum_{n,m} v_{n,m} e^{-i\tau\xi} \int_{(\tau - n+1)\tau_0}^{\tau - n\tau_0} \int_{(m+1)\xi_0}^{m\xi_0} \partial_2 \partial_1 \Omega_{\tau, \xi}(t, z) dt dz, \quad (7.17)$$

which yields

$$\begin{aligned}
|F(\tau, \xi) - \tilde{F}_A(\tau, \xi)| &\leq \frac{1}{A} \sum_{n,m} |v_{n,m} e^{-i\tau\xi}| \int_{(\tau-n+1)\tau_0}^{\tau-n\tau_0} \int_{(m+1)\xi_0}^{m\xi_0} |\partial_2 \partial_1 \Omega_{\tau,\xi}(t, z)| dt dz \\
&\leq \frac{\sqrt{2}}{A} \|\partial_2 \partial_1 \Omega_{\tau,\xi}(t, z)\|_{L^1(\mathbb{R}^2)}. \tag{7.18}
\end{aligned}$$

Note that in the second inequality we used Lemma 3 to bound  $\|v\|_{l^\infty}$  by  $\sqrt{2}$ . We complete the proof by estimating the  $L^1$ -norm of  $\partial_2 \partial_1 \Omega_{\tau,\xi}(t, z)$ : Clearly,

$$\partial_2 \partial_1 \Omega_{\tau,\xi}(t, z) = i\tau e^{iz\tau} \partial_1 \Gamma(t, \xi - z) - e^{iz\tau} \partial_2 \partial_1 \Gamma(t, \xi - z).$$

Then we have

$$\|\partial_2 \partial_1 \Omega_{\tau,\xi}(t, z)\|_{L^1(\mathbb{R}^2)} \leq \|\partial_2 \partial_1 \Gamma\|_{L^1(\mathbb{R}^2)} + |\tau| \|\partial_1 \Gamma\|_{L^1(\mathbb{R}^2)},$$

which yields the desired bound by setting

$$C_{\varphi,1} := \sqrt{2} \|\partial_2 \partial_1 \Gamma\|_{L^1(\mathbb{R}^2)} \tag{7.19}$$

and

$$C_{\varphi,2} := \sqrt{2} \|\partial_1 \Gamma\|_{L^1(\mathbb{R}^2)}. \tag{7.20}$$

□

Now we want to raise the question of whether we can approximate  $f$  using  $\tilde{F}_A$ , and if yes, in what sense. Consider the following space of ‘test functions’; let

$$\mathcal{G} = \{g \in L^2(\mathbb{R}) : (1 + \tau) \langle g, \varphi_{\tau,\xi} \rangle \in L^1(\mathbb{R}^2)\}.$$

Clearly any function  $f \in L^2(\mathbb{R})$  defines a linear functional  $L_f$  on  $\mathcal{G}$  by  $L_f g := \langle f, g \rangle$ .

By the Parseval identity we also have  $L_f g = (2\pi)^{-1} \langle F, G \rangle$ , where  $F$  and  $G$  are the continuous windowed Fourier transforms of  $f$  and  $g$ , respectively. Let  $\tilde{F}_A$  be as above and define  $\langle \tilde{F}_A, G \rangle$  as

$$\langle \tilde{F}_A, G \rangle = \int \tilde{F}_A(\tau, \xi) \overline{G(\tau, \xi)} d\tau d\xi. \quad (7.21)$$

Note that (7.21) makes sense since

$$\begin{aligned} \left| \int \tilde{F}_A(\tau, \xi) \overline{G(\tau, \xi)} d\tau d\xi \right| &\leq |\langle F, G \rangle| + \left| \int (\tilde{F}_A - F)(\tau, \xi) \overline{G(\tau, \xi)} d\tau d\xi \right| \\ &\leq |\langle F, G \rangle| + \frac{C_{\varphi,1}}{A} \|G\|_{L^1} + \frac{C_{\varphi,2}}{A} \|\tau G(\tau, \xi)\|_{L^1} \\ &< \infty \end{aligned} \quad (7.22)$$

This suggests that we define  $\tilde{f}_A$  as the linear functional that maps  $g \in \mathcal{G}$  to  $(2\pi)^{-1} \langle \tilde{F}_A, G \rangle$ .

Thus we have

**Theorem 13.** *Let  $\tilde{f}_A$  be defined as above, i.e.*

$$\tilde{f}_A : g \in \mathcal{G} \rightarrow \langle \tilde{f}_A, g \rangle := (2\pi)^{-1} \langle \tilde{F}_A, G \rangle. \quad (7.23)$$

Then  $\tilde{f}_A$  converges to  $f$  on  $\mathcal{G}$  as  $A$  tends to infinity, in the sense that

$$|\langle \tilde{f}_A, g \rangle - \langle f, g \rangle| \leq \frac{1}{2\pi A} (C_{\varphi,1} \|G\|_{L^1} + C_{\varphi,2} \|\tau G(\tau, \xi)\|_{L^1}). \quad (7.24)$$

Note that  $A = \frac{2\pi}{\tau_0 \xi_0}$ ; thus increasing  $A$  means decreasing the time and/or frequency translation steps,  $\tau_0$  and  $\xi_0$ , so increasing the redundancy of the expansion.

**Proof:** Let  $g \in \mathcal{G}$  be arbitrary. Then

$$\langle \tilde{f}_A, g \rangle = (2\pi)^{-1} \int \tilde{F}_A(\tau, \xi) \overline{G(\tau, \xi)} d\tau d\xi, \quad (7.25)$$

$$\langle f, g \rangle = (2\pi)^{-1} \int F(\tau, \xi) \overline{G(\tau, \xi)} d\tau d\xi, \quad (7.26)$$

where (7.25) is by definition true, and (7.26) follows from the Parseval identity for windowed Fourier transform. Thus

$$|\langle \tilde{f}_A, g \rangle - \langle f, g \rangle| = (2\pi)^{-1} \left| \int (\tilde{F}_A - F)(\tau, \xi) \overline{G(\tau, \xi)} d\tau d\xi \right|, \quad (7.27)$$

$$\leq (2\pi)^{-1} \int |\tilde{F}_A - F|(\tau, \xi) |G|(\tau, \xi) d\tau d\xi \quad (7.28)$$

$$\leq \frac{1}{2\pi A} (C_{\varphi,1} \|G\|_{L^1} + C_{\varphi,2} \|\tau G(\tau, \xi)\|_{L^1}), \quad (7.29)$$

where to obtain (7.29) we use Theorem 12. □

Now we have a way of approximating  $f$  using the discrete sequence  $(q_{n,m})$ ; of course the approximation is in the above described sense and we do not even know whether  $\tilde{f}_A$  is a function. However, one can observe that this way of approximation is particularly useful for ‘comparing’ two functions (thus leading to applications such as pattern recognition); next we will show how one can ‘compare’ two functions in  $L^2$  using their approximations which are obtained via this time-frequency sigma-delta quantization algorithm.

First let us focus on how to calculate the inner product  $\langle \tilde{F}_A, G \rangle$ ; note that

$$\langle \tilde{F}_A, G \rangle = \left\langle \frac{1}{A} \sum_{n,m} q_{n,m} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi), G(\tau, \xi) \right\rangle \quad (7.30)$$

$$= \frac{1}{A} \sum_{n,m} q_{n,m} \langle \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi), G(\tau, \xi) \rangle. \quad (7.31)$$

But by the Parseval identity for windowed Fourier transform,

$$\langle \alpha_{n,m}(\xi) \tilde{\Phi}_{n,m}(\tau, \xi), G(\tau, \xi) \rangle = 2\pi \langle \varphi_{n,m}, g \rangle. \quad (7.32)$$

Let us denote the frame coefficients  $\langle g, \varphi_{n,m} \rangle$  of  $g$  by  $d_{n,m}$ . After substituting (7.32) in (7.31), we get

$$\langle \tilde{F}_A, G \rangle = \frac{2\pi}{A} \sum_{n,m} q_{n,m} \overline{d_{n,m}}. \quad (7.33)$$

Hence we have proved:

**Theorem 14.** *Let  $f \in \mathcal{B}^\varphi$ ,  $g \in \mathcal{G}$ ,  $F = \langle f, \varphi_{\tau,\xi} \rangle$ ,  $G = \langle g, \varphi_{\tau,\xi} \rangle$  with  $\varphi$  such that  $(\varphi, \tau_0, \xi_0)$  is a tight Weyl-Heisenberg frame of  $L^2(\mathbb{R})$  for some fixed  $\tau_0$  and  $\xi_0$ . Suppose that  $\varphi$  also fulfills the assumptions of Theorem 12. Then  $\tilde{F}_A$ , the time-frequency sigma-delta approximation of  $F$ , satisfies*

$$\langle \tilde{F}_A, G \rangle = \frac{2\pi}{A} \sum_{n,m} q_{n,m} \overline{d_{n,m}}, \quad (7.34)$$

where  $d_{n,m} = \langle g, \varphi_{n,m} \rangle$ . Moreover since we choose  $g$  such that  $\langle g, \varphi_{n,m} \rangle$  is absolutely summable, we have:

(i)

$$\langle F - \tilde{F}_A, G \rangle = \frac{2\pi}{A} \sum_{n,m} (c_{n,m} - q_{n,m}) \overline{d_{n,m}}, \quad (7.35)$$

where  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$ ,  $d_{n,m} = \langle g, \varphi_{n,m} \rangle$  and the sequence  $q$  is given by  $q = T_{TF}(c)$ ;

and

(ii)

$$\langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle = \frac{2\pi}{A} \sum_{n,m} (q_{n,m}^1 - q_{n,m}^2) \overline{d_{n,m}}, \quad (7.36)$$

where  $\tilde{F}_A^j$  is the time-frequency sigma-delta approximation of  $F^j = \langle f_j, \varphi_{\tau,\xi} \rangle$  for some  $f_j$  in  $\mathcal{B}^\varphi$  and  $q^j = T_{TF}(c^j)$  with  $c_{n,m}^j = \langle f^j, \varphi_{n,m} \rangle$ .

**Remarks:**

1. Note that (7.34) is an explicit formula to calculate the inner product  $\langle \tilde{F}_A, G \rangle$ ; the only terms in (7.34) that do depend on the function  $f$  are the  $q_{n,m}$ . In other words, one can calculate the  $d_{n,m}$  just once and store them in memory.
2. The second part of the theorem, in particular (7.36), specifies a simple way of determining how ‘similar’ two functions are by using only the corresponding bit sequences; next we shall make clear what we mean by ‘similar’.

**Theorem 15.** *Let  $f_1, f_2$  be in  $\mathcal{B}^\varphi$ ,  $F^j = \langle f_j, \varphi_{\tau,\xi} \rangle$  for  $j = 1, 2$ , and  $\tilde{F}_A^j$  the time-frequency sigma-delta approximation of  $F^j$ . Then*

$$|\langle F^1 - F^2, G \rangle - \langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle| \leq \frac{4\pi}{A} (C_{\varphi,1} \|G\|_{L^1} + C_{\varphi,2} \|\tau G(\tau, \xi)\|_{L^1}). \quad (7.37)$$

where  $C_{\varphi,i}$ ,  $i = 1, 2$ , is defined as in (7.19) and (7.20) respectively.

**Proof:** Note that

$$\langle F^1 - F^2, G \rangle - \langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle = \langle F^1 - \tilde{F}_A^1, G \rangle - \langle F^2 - \tilde{F}_A^2, G \rangle. \quad (7.38)$$

Thus,

$$|\langle F^1 - F^2, G \rangle - \langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle| \leq |\langle F^1 - \tilde{F}_A^1, G \rangle| + |\langle F^2 - \tilde{F}_A^2, G \rangle| \quad (7.39)$$

$$\leq \frac{4\pi}{A}(C_{\varphi,1}\|G\|_{L^1} + C_{\varphi,2}\|\tau G(\tau, \xi)\|_{L^1}), \quad (7.40)$$

where the second inequality is due to Theorem 13.  $\square$

Theorem 15 clearly shows that  $\langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle$  is an estimate of  $f_1 - f_2$  in the direction of  $g$ . In other words, our measure of similarity of  $f_1$  and  $f_2$ , i.e.  $\langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle$ , is completely insensitive to functions that are orthogonal to  $g$ . However if two functions are close to each other in  $L^2$ , clearly  $\langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle$  will also be small. In other words,

**Corollary 2.** *Let  $g$  be in  $\mathcal{G}$  and  $f_1, f_2$  be in  $\mathcal{B}^\varphi$ . Then*

$$1. |\langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle| \leq 2\pi\|f_1 - f_2\|_{L^2}\|g\|_{L^2} + \frac{4\pi}{A}(C_{\varphi,1}\|G\|_{L^1} + C_{\varphi,2}\|\tau G(\tau, \xi)\|_{L^1}).$$

$$2. |\langle F^1 - F^2, G \rangle| \leq |\langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle| + \frac{4\pi}{A}(C_{\varphi,1}\|G\|_{L^1} + C_{\varphi,2}\|\tau G(\tau, \xi)\|_{L^1}),$$

where  $\tilde{F}_A^j$  is the time-frequency sigma-delta approximation of  $f_j$ , and  $C_{\varphi,i}$ ,  $i = 1, 2$ , is defined as in (7.19) and (7.20) respectively.

We now generalize the above discussion in the following way. Let  $g_1, \dots, g_K$  be functions in  $\mathcal{G}$  such that  $\|g_j\|_{L^2} = 1$  and  $\langle g_i, g_j \rangle = \delta_{i,j}$ .

**Theorem 16.** *Let  $P$  be the projection operator defined by*

$$P(F) = \sum_{j=1}^K \langle F, G_j \rangle G_j, \quad (7.41)$$

where  $G_i = \langle g_i, \varphi_{\tau, \xi} \rangle$ . Let  $F$  be given by  $F(\tau, \xi) = \langle f, \varphi_{\tau, \xi} \rangle$  for some  $f \in \mathcal{B}^\varphi$ . Let  $c$  be the sequence  $(\langle f, \varphi_{n,m} \rangle)$  and  $q = T_{TF}(c)$ . Suppose  $\tilde{F}_A$  is the time-frequency sigma-

delta approximation of  $F$ . Then

$$\|P(F - \tilde{F}_A)\|^2 = \frac{4\pi^2}{A^2} \sum_{n,m,n',m'} (c_{n,m} - q_{n,m}) \overline{(c_{n',m'} - q_{n',m'})} \langle \tilde{P}\varphi_{n,m}, \varphi_{n',m'} \rangle, \quad (7.42)$$

where  $\tilde{P}$  is defined by  $\tilde{P}(f) := \sum_{i=1}^K \langle f, g_i \rangle g_i$  for  $f \in \mathcal{B}^\varphi$ .

**Proof:** By (7.22),  $P(\tilde{F}_A)$  is well-defined and thus in the span of  $\{G_1, \dots, G_K\}$ . Then we can write

$$\begin{aligned} \|P(F - \tilde{F}_A)\|^2 &= \sum_{i=1}^K |\langle F - \tilde{F}_A, G_i \rangle|^2 \\ &= \frac{4\pi^2}{A^2} \sum_{i=1}^K \left( \sum_{n,m} (c_{n,m} - q_{n,m}) \overline{d_{n,m}^i} \right) \left( \sum_{n',m'} \overline{(c_{n',m'} - q_{n',m'})} d_{n',m'}^i \right) \\ &= \frac{4\pi^2}{A^2} \sum_{n,m,n',m'} (c_{n,m} - q_{n,m}) \overline{(c_{n',m'} - q_{n',m'})} \sum_{i=1}^K \langle \varphi_{n,m}, g_i \rangle \langle g_i, \varphi_{n',m'} \rangle \\ &= \sum_{n,m,n',m'} (c_{n,m} - q_{n,m}) \overline{(c_{n',m'} - q_{n',m'})} \langle \tilde{P}\varphi_{n,m}, \varphi_{n',m'} \rangle, \end{aligned} \quad (7.43)$$

where  $d_{n,m}^i := \langle g_i, \varphi_{n,m} \rangle$ . The first equality is due to the definition of  $P$ ; the second equality follows from Theorem 14; the third and fourth equalities are obvious.  $\square$

**Remarks:**

1. Let  $F^1$  and  $F^2$  be the windowed Fourier transforms of two functions  $f^1$  and  $f^2$  in  $\mathcal{B}^\varphi$ . Denote the sequence  $(\langle f^i, \varphi_{n,m} \rangle)$  by  $c^i$  and let  $q^i = T_{TF}(c^i)$ . Suppose  $\tilde{F}_A^1$  and  $\tilde{F}_A^2$  are the time-frequency sigma-delta approximations of  $F^1$  and  $F^2$  respectively. Then replacing  $F$  and  $\tilde{F}_A$  in the proof of the previous theorem by  $\tilde{F}_A^1$  and  $\tilde{F}_A^2$ , respectively, yields

$$\|P(\tilde{F}_A^1 - \tilde{F}_A^2)\|^2 = \frac{4\pi^2}{A^2} \sum_{n,m,n',m'} (q_{n,m}^1 - q_{n,m}^2) \overline{(q_{n',m'}^1 - q_{n',m'}^2)} \langle \tilde{P}\varphi_{n,m}, \varphi_{n',m'} \rangle. \quad (7.44)$$

2. By Corollary 2 we have

$$\|P(\tilde{F}_A^1 - \tilde{F}_A^2)\| \leq \|f^1 - f^2\|_{L^2} \sum_{i=1}^K \|g_i\|_{L^2} + \frac{4\pi}{A} (C_{\varphi,1} \sum_{i=1}^K \|G_i\|_{L^1} + C_{\varphi,2} \sum_{i=1}^K \|\tau G_i(\tau, \xi)\|_{L^1}). \quad (7.45)$$

## 7.2 Translation Invariance

As mentioned before, one possible application area for the time-frequency sigma-delta quantization scheme described in this chapter is pattern recognition. We have shown above that we can measure how similar two functions  $f_1$  and  $f_2$  are by calculating  $\langle \tilde{F}_A^1 - \tilde{F}_A^2, G \rangle$ . The next important question is whether the quantization scheme is robust with respect to translation in both arguments; in this section we shall investigate how shifts in the bit-sequence affect the approximation.

Let us start by noting that

$$\langle f(\cdot + N\tau_0), \varphi_{n,m} \rangle = e^{imN\frac{2\pi}{A}} \langle f, \varphi_{n+N,m} \rangle, \quad (7.46)$$

where  $A = \frac{2\pi}{\tau_0\xi_0}$  is the frame bound. Let us denote  $\langle f, \varphi_{n,m} \rangle$  by  $c_{n,m}$  and  $e^{imN\frac{2\pi}{A}}$  by  $\gamma_N$  and rewrite (7.46) as

$$\langle f(\cdot + N\tau_0), \varphi_{n,m} \rangle = (\gamma_N)^m c_{n+N,m}. \quad (7.47)$$

Thus we conclude

$$f(\cdot + N\tau_0) = \sum_{n,m} (\gamma_N)^m c_{n+N,m} \varphi_{n,m}. \quad (7.48)$$

From the previous section we know that

$$\tilde{F}_A = \frac{1}{A} \sum q_{n,m} \alpha_{n,m} \Phi_{n,m} \quad (7.49)$$

approximates  $F = \langle f, \varphi_{\tau, \xi} \rangle$  as in (7.7). In (7.49)  $q = (q_{n,m}) = T_{TF}(c)$  with  $c = (c_{n,m}) = (\langle f, \varphi_{n,m} \rangle)$ . We also know by (7.48) that

$$H(\tau, \xi) = \frac{1}{A} \sum_{n,m} (\gamma_N)^m c_{n+N,m} \alpha_{n,m}(\xi) \Phi(\tau, \xi) \quad (7.50)$$

is the windowed Fourier transform of  $f(\cdot + N\tau_0)$ . One important question to ask is whether

$$\tilde{H}_A(\tau, \xi) = \frac{1}{A} \sum_{n,m} (\gamma_N)^m q_{n+N,m} \alpha_{n,m}(\xi) \Phi(\tau, \xi) \quad (7.51)$$

approximates  $H(\tau, \xi)$  in a way similar to the unshifted (7.7), i.e. whether  $|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{\tilde{C}_{\varphi,1}}{A} + |\tau| \frac{\tilde{C}_{\varphi,2}}{A}$  for some  $\tilde{C}_{\varphi,1}$  and  $\tilde{C}_{\varphi,2}$ . The next theorem shows that the answer to this question is affirmative.

**Theorem 17.** *Let  $q = T_{TF}(c)$ , where  $c = (c_{n,m})$  with  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$  for some  $f$  in  $\mathcal{B}^\varphi$ . Suppose  $H$  and  $\tilde{H}_A$  are defined as in (7.50) and (7.51) respectively. Then*

$$|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{\tilde{C}_{\varphi,1}}{A} + |\tau| \frac{\tilde{C}_{\varphi,2}}{A}, \quad (7.52)$$

with  $\tilde{C}_{\varphi,1} = \sqrt{2} \|\partial_2 \partial_1 \Gamma\|_{L^1(\mathbb{R}^2)} + N\tau_0 \|\partial_1 \Gamma\|_{L^1(\mathbb{R}^2)}$  and  $\sqrt{2} \tilde{C}_{\varphi,2} = \|\partial_1 \Gamma\|_{L^1(\mathbb{R}^2)}$ .

**Proof:** We want to show that

$$\frac{1}{A} \left| \sum_{n,m} (\gamma_N)^m q_{n+N,m} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) - \sum_{n,m} (\gamma_N)^m c_{n+N,m} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) \right| \quad (7.53)$$

$$= \left| \frac{1}{A} \sum_{n,m} (\gamma_N)^m (\Delta_1 \Delta_2 v)_{n+N,m} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) \right| \quad (7.54)$$

$$\leq \frac{\tilde{C}_{\varphi,1}}{A} + |\tau| \frac{\tilde{C}_{\varphi,2}}{A}, \quad (7.55)$$

for some  $\tilde{C}_{\varphi,1}$  and  $\tilde{C}_{\varphi,2}$  where  $v_{n,m}$  is as in (7.12). Define

$$D := \frac{1}{A} \sum_{n,m} (\Delta_1 \Delta_2 v)_{n+N,m} (\gamma_N)^m \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi).$$

Then clearly

$$D = \frac{1}{A} \sum_{n,m} (\Delta_1 \Delta_2 v)_{n+N,m} e^{-i\tau\xi} e^{im\xi_0(N\tau_0+\tau)} \Gamma(\tau - n\tau_0, \xi - m\xi_0), \quad (7.56)$$

with  $\Gamma(t, z) = e^{itz} \Phi(t, z)$ . Let  $\Omega_{N,\tau,\xi}(t, z) = e^{iz(N\tau_0+\tau)} \Gamma(t, \xi - z)$ . After summing the left hand side of (7.56) by parts we get

$$D = \frac{1}{A} \sum_{n,m} v_{n+N,m} e^{-i\tau\xi} \Delta_1 \Delta_2 \Omega_{N,\tau,\xi}(\tau - n\tau_0, m\xi_0). \quad (7.57)$$

Since  $\Omega_{N,\tau,\xi}$  is smooth, we have

$$D = \frac{1}{A} \sum_{n,m} v_{n+N,m} e^{-i\tau\xi} \int_{(\tau-n+1)\tau_0}^{\tau-n\tau_0} \int_{(m+1)\xi_0}^{m\xi_0} \partial_2 \partial_1 \Omega_{N,\tau,\xi}(t, z) dt dz, \quad (7.58)$$

which yields

$$\begin{aligned} |D| &\leq \frac{\sqrt{2}}{A} \sum_{n,m} \int_{(\tau-n+1)\tau_0}^{\tau-n\tau_0} \int_{(m+1)\xi_0}^{m\xi_0} |\partial_2 \partial_1 \Omega_{N,\tau,\xi}(t, z)| dt dz, \\ &\leq \frac{\sqrt{2}}{A} \|\partial_2 \partial_1 \Omega_{N,\tau,\xi}\|_{L^1(\mathbb{R}^2)}. \end{aligned} \quad (7.59)$$

Finally, after estimating  $\|\partial_2 \partial_1 \Omega_{N,\tau,\xi}\|_{L^1(\mathbb{R}^2)}$  we get

$$|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{1}{A} (\tilde{C}_{\varphi,1} + |\tau| \tilde{C}_{\varphi,2}) \quad (7.60)$$

with

$$\tilde{C}_{\varphi,1} = \sqrt{2} \|\partial_2 \partial_1 \Gamma\|_{L^1(\mathbb{R}^2)} + \sqrt{2} N \tau_0 \|\partial_1 \Gamma\|_{L^1(\mathbb{R}^2)}, \quad (7.61)$$

and

$$\tilde{C}_{\varphi,2} = \sqrt{2}\|\partial_1\Gamma\|_{L^1(\mathbb{R}^2)}. \quad (7.62)$$

□

**Remarks:**

1. Combining Theorem 17 with Theorem 15, we can conclude that

$$\left| \sum_{n,m} ((\gamma_N)^m q_{n+N,m} - \bar{q}_{n,m}) d_{n,m} \right| \leq \frac{\tilde{C}_{\varphi,1}}{A} + |\tau| \frac{\tilde{C}_{\varphi,2}}{A}, \quad (7.63)$$

where  $\bar{q} := (\bar{q}_{n,m}) = T_{TF}(\bar{c})$  with  $\bar{c} := (\langle f(\cdot + N\tau_0), \varphi_{n,m} \rangle)$ .

2. Note that the constant  $\tilde{C}_{\varphi,2}$  given in (7.62) is the same as  $C_{\varphi,2}$  given in (7.20);  $\tilde{C}_{\varphi,1}$ , given in (7.61), has an extra summand proportional to  $N$ , the amount of translation, and  $\tau_0$ , the time translation step, when compared to  $C_{\varphi,1}$ , given in (7.20). Thus, for  $N = 0$ , i.e. there is no shift in the quantizer output  $(q_{n,m})$ , both estimates yield the same upper bound on the approximation error.
3. The time-frequency sigma-delta quantization scheme is translation invariant up to the adjustment factor  $(\gamma_N)^m$ ; the approximation of  $f(\cdot + N\tau_0)$  obtained using  $((\gamma_N)^m q_{n+N,m})$  is (almost) as good as that obtained by quantizing the translated version separately.

Finally, let us investigate shifts in the other index of the bit sequence produced by the time-frequency sigma-delta scheme.

**Theorem 18.** *Let  $f$  be in  $\mathcal{B}^\varphi$ ,  $c = (\langle f, \varphi_{n,m} \rangle)$  and  $q = (q_{n,m}) = T_{TF}(c)$ . Suppose  $H(\tau, \xi)$  is the windowed Fourier transform of  $e^{-iM\xi_0} f(\cdot)$ . Then*

$$|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{C_{\varphi,1}}{A} + |\tau| \frac{C_{\varphi,2}}{A} \quad (7.64)$$

where  $C_{\varphi,1}$  and  $C_{\varphi,2}$  are as in (7.19) and (7.20) respectively, and

$$\tilde{H}_A = \frac{1}{A} \sum_{n,m} q_{n,m+M} \alpha_{n,m} \Phi_{n,m}.$$

**Proof:** Note that

$$\begin{aligned} \langle e^{-iM\xi_0 \cdot} f(\cdot), \varphi_{n,m} \rangle &= \int f(t) \varphi(t - n\tau_0) e^{-i(m+M)\xi_0 t} dt \\ &= \langle f, \varphi_{n,m+M} \rangle, \end{aligned} \quad (7.65)$$

which yields

$$H = \frac{1}{A} \sum_{n,m} c_{n,m+M} \alpha_{n,m} \Phi_{n,m}.$$

Then

$$\begin{aligned} H(\tau, \xi) - \tilde{H}_A(\tau, \xi) &= \frac{1}{A} \sum_{n,m} (q_{n,m+M} - c_{n,m+M}) \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) \\ &= \frac{1}{A} \sum_{n,m} (\Delta_1 \Delta_2 v)_{n,m+M} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi), \end{aligned} \quad (7.66)$$

where  $v_{n,m}$  is as in (7.12). As in the proof of Theorem 12 summing by parts yields the result.  $\square$

Now we can combine these two results: Let  $(\varphi, \tau_0, \xi_0)$  be a tight Weyl-Heisenberg frame of  $L^2$  with frame bound  $A$ ,  $c = (\langle f, \varphi_{n,m} \rangle)$  for some  $f \in \mathcal{B}^\varphi$ , and  $q = T_{TF}(c)$ .

Then

$$H(\tau, \xi) = \frac{1}{A} \sum_{n,m} \gamma_N^{m+M} c_{n+N,m+M} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) \quad (7.67)$$

is clearly the windowed Fourier transform of  $e^{-iM\xi_0} f(\cdot + N\tau_0)$ . Now define  $\tilde{H}_A$  by

$$\tilde{H}_A(\tau, \xi) := \frac{1}{A} \sum_{n,m} \gamma_N^{m+M} q_{n+N, m+M} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi). \quad (7.68)$$

Note that  $H^1(\tau, \xi) := \frac{1}{A} \sum_{n,m} c_{n, m+M} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) = \langle e^{-iM\xi_0} f(\cdot), \varphi_{\tau, \xi} \rangle$ . We then have by Theorem 17,

$$\left| \sum_{n,m} (\gamma_N)^m q_{n+N, m+M} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) - \langle e^{-iM\xi_0(\cdot + N\tau)} f(\cdot + N\tau_0), \varphi_{\tau, \xi} \rangle \right| \leq \frac{\tilde{C}_{\varphi,1}}{A} + |\tau| \frac{\tilde{C}_{\varphi,2}}{A}, \quad (7.69)$$

where  $\tilde{C}_{\varphi,1}$  and  $\tilde{C}_{\varphi,2}$  are as in (7.61) and (7.62) respectively. Finally, since  $|\gamma_N| = 1$ , we can write

$$\left| \sum_{n,m} (\gamma_N)^{(m+M)} q_{n+N, m+M} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) - \langle e^{-iM\xi_0} f(\cdot + N\tau_0), \varphi_{\tau, \xi} \rangle \right| \leq \frac{\tilde{C}_{\varphi,1}}{A} + |\tau| \frac{\tilde{C}_{\varphi,2}}{A}. \quad (7.70)$$

Thus we proved:

**Theorem 19.**  $|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{\tilde{C}_{\varphi,1}}{A} + |\tau| \frac{\tilde{C}_{\varphi,2}}{A}$ , where  $H$  is as in (7.67),  $\tilde{H}_A$  is as in (7.68), and  $\tilde{C}_{\varphi,1}$  and  $\tilde{C}_{\varphi,2}$  are as in (7.61) and (7.62) respectively.

## 7.3 Numerical Experiment

In this section, we will present some experimental results: We will fix a Weyl-Heisenberg frame and quantize the frame expansions of a function  $f$  using the algorithm TFΣΔ-I. We choose  $\varphi(t) = \pi^{1/4} e^{-t^2/2}$ . One can show that  $(\varphi, \tau_0, \xi_0)$  is a frame of  $L^2(\mathbb{R})$  if  $\tau_0$  and  $\xi_0$  are sufficiently small. Moreover, the frame is almost tight<sup>1</sup> (with

<sup>1</sup>A frame is called “almost tight” if the ratio of the frame bounds is close to 1. Suppose  $(\varphi, \tau_0, \xi_0)$  is a frame with frame bounds  $A$  and  $B$ . If we denote the quantity  $B/A - 1$  by  $r$ , then any function  $f \in L^2$  can be written as  $f = \frac{2}{A(2+r)} \sum \langle f, \varphi_{n,m} \rangle \varphi_{n,m} + Rf$  where  $\|R\| \leq \frac{r}{2+r}$  [9]. Hence reconstructing  $f$  by (6.5) (with  $\frac{A(2+r)}{2}$  instead of  $A$ ) introduces an error which is bounded in  $L^2$  by  $\frac{r}{2+r} \|f\|_{L^2}$ .

both frame bounds approximately equal to  $\frac{2\pi}{\tau_0\xi_0}$ ) if one chooses sufficiently small  $\tau_0$  and  $\xi_0$  such that  $\tau_0 \approx \xi_0$ .

Let us now consider the function<sup>2</sup>

$$f(t) = 0.4e^{-i0.9t^3} e^{-0.05t^2}. \quad (7.71)$$

First we compute the frame coefficients of  $f$ ,  $\langle f, \varphi_{n,m} \rangle$ , for different values of  $\tau_0$  and  $\xi_0$ . We use an FFT-based algorithm to compute the frame coefficients using the samples of  $f$ : Let  $\tau_1$  be the period at which we sample  $f$ . (It is convenient to choose  $\tau_1 = \tau_0$ .) We will use the sequence  $(f(k\tau_1))_{k=-K}^K$  for some sufficiently large  $K$  to compute the frame coefficients of  $f$ . Of course  $K$  has to be finite for all practical purposes; however that does not introduce a large error if both  $f$  and  $\varphi$  are well-localized in time and frequency, which is true for our example. Figure 7.1 shows the coefficients of  $f$  for the frame  $(\varphi, 0.1, 0.1)$ . In Figure 7.2, we show the quantized values of the frame coefficients of  $f$ , obtained via the time-frequency sigma-delta quantization scheme. Next, we consider the frame expansions of  $f$  with frames  $(\varphi, \tau_0, \xi_0)$  where  $\tau_0$  and  $\xi_0$  take values between 0.1 and 0.25; thus the frame bound  $A$  ranges from approximately 100.53 to 904.78. We fix  $G(\tau, \xi) = e^{-0.2(\tau^2 + \xi^2)}$  and we use

$$G_{tot} = \sum_{k=-2}^2 \sum_{l=-2}^2 T_{l,k} G, \quad (7.72)$$

where  $T_{l,k} G := G(\cdot + l, \cdot + k)$ , as our test function. Clearly the inverse windowed Fourier transform of  $G_{tot}$  is in  $\mathcal{G}$ ; Figure 7.3 shows the graph of  $G_{tot}$ .

Next, we compute  $\langle F - \tilde{F}_A, G_{tot} \rangle$  via (7.35). Figure 7.4 shows the value of this inner product as the frame bound increases. Theorem 12 bounds the decay of  $|\langle F - \tilde{F}_A, G_{tot} \rangle|$  by  $A^{-1}$ ; however experimental evidence, e.g. Figure 7.4, suggests a faster decay

Therefore, if  $r \approx 0$ , we can assume the frame is tight and reconstruct  $f$  using (6.5). For all the frames we will use in this section  $|r|$  is smaller than the arithmetical precision of the computer.

<sup>2</sup>The function  $f$  is clearly in  $\mathcal{B}^\varphi$ .

Figure 7.1: The frame coefficients  $\langle f, \varphi_{n,m} \rangle$ . The upper left figure shows the real part of the coefficients –black and white correspond to  $-0.42$  and  $0.27$ , respectively ; the upper right figure shows the imaginary parts of the coefficients –black and white correspond to  $-0.27$  and  $0.40$ , respectively. The lower figure shows the absolute value of the coefficients. In this graph, black corresponds to  $0$  and white corresponds to  $0.56$ . The artifacts seen in the figures are introduced by FFT and the total energy that is introduced by the artifacts is less than  $1/100$  of the total energy of  $f$ .

Figure 7.2: The quantized frame coefficients  $\langle f, \varphi_{n,m} \rangle$ . The upper left figure shows the real part of the quantized coefficients; the upper right figure shows the imaginary parts of the quantized coefficients; the lower figure shows the absolute value of the quantized coefficients. In the upper figures black and white correspond to  $-3$  and  $3$  respectively. In the lower figure black corresponds to  $\sqrt{2}$  and white corresponds  $3\sqrt{2}$ .

Figure 7.3: The graph of the test function  $G_{tot}$ . Black corresponds to 0 and white corresponds to 4.49.

Figure 7.4: The ‘approximation error’  $|\langle F - \tilde{F}_A, G_{tot} \rangle|$  vs. the frame bound  $A$ . Both axes are logarithmic. The solid line seen in the figure is the graph  $\{(A, 1/A) : 100 < A < 905\}$ ; the dashed line is the graph  $\{(A, 10A^{-3/2}) : 100 < A < 905\}$ .

Figure 7.5: The value  $\langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.1$  and  $\Omega = 0.9$ ; the minimum is obtained at  $N = 11$  and  $M = 8$ , which means that the algorithm predicts  $T = 1.1$  and  $\Omega = 0.8$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A} - I_{11,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A} - I_{N,8}, G_{tot} \rangle$  vs.  $N$ .

Figure 7.6: The value  $\langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.17$  and  $\Omega = 0.93$ ; the minimum is obtained at  $N = 12$  and  $M = 8$ , which means that the algorithm predicts  $T = 1.2$  and  $\Omega = 0.8$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A} - I_{12,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A} - I_{N,8}, G_{tot} \rangle$  vs.  $N$ .

Figure 7.7: The value  $\langle \tilde{F}_{T,\Omega,A}^\nu - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.17$  and  $\Omega = 0.93$ ; SNR= 23.48 dB; the minimum is obtained at  $N = 12$  and  $M = 6$ , which means that the algorithm predicts  $T = 1.2$  and  $\Omega = 0.6$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A} - I_{12,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A} - I_{N,6}, G_{tot} \rangle$  vs.  $N$ .

Figure 7.8: The value  $\langle \tilde{F}_{T,\Omega,A}^\nu - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.17$  and  $\Omega = 0.93$ ; SNR= 10.73 dB; the minimum is obtained at  $N = 12$  and  $M = 6$ , which means that the algorithm predicts  $T = 1.2$  and  $\Omega = 0.6$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A} - I_{12,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A} - I_{N,6}, G_{tot} \rangle$  vs.  $N$ .

Figure 7.9: The value  $\langle \tilde{F}_{T,\Omega,A}^\nu - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.17$  and  $\Omega = 0.93$ ; SNR= 1.06 dB; the minimum is obtained at  $N = 14$  and  $M = 7$ , which means that the algorithm predicts  $T = 1.4$  and  $\Omega = 0.7$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A} - I_{14,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A} - I_{N,7}, G_{tot} \rangle$  vs.  $N$ .

rate. This is similar to the first-order standard sigma-delta scheme for which the analogous estimate yields a bound of  $O(\lambda^{-1})$  [3] ( $\lambda$  is the oversampling ratio) whereas the empirically expected decay rate is  $\lambda^{-3/2}$ . In [12], S. Güntürk proved that the error can be bounded pointwise by  $C\lambda^{-4/3+\eta}$  where  $C$  depends on  $\eta$  and on the value of the derivative of the original function at the corresponding point; the conjecture is that the error can be bounded pointwise by  $C\lambda^{-3/2+\eta}$ . (A detailed discussion of various types of improved estimates can be found in [13].) Whether there is a similar theorem for our case is an open problem; Figure 7.4 suggests there may well be.

Now, we want to observe the translation invariance of our algorithm. Let  $f$  be as in (7.71). Fix the frame  $(\varphi, 0.1, 0.1)$  and compute  $q = T_{TF}(c)$  where  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$ . Now, define  $f_{T,\Omega}$  by  $f_{T,\Omega}(t) := e^{-i\Omega t} f(t + T)$ . Let  $c_{T,\Omega}$  be the sequence  $(\langle f_{T,\Omega}, \varphi_{n,m} \rangle)$  and  $q_{T,\Omega} := T_{TF}(c_{T,\Omega})$ . Using  $q$  as a template, we will estimate what  $T$  and  $\Omega$  are when we are only given the sequence  $q_{T,\Omega}$ . To accomplish this, we will compare  $\tilde{F}_{T,\Omega,A} := \sum (q_{T,\Omega})_{n,m} \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle$  with  $I_{N,M} := \sum (\gamma_N)^{m+M} q_{n+N,m+M} \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle$  for various  $N$  and  $M$  by comparing the inner products  $\langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$ . We will

calculate these inner products via (7.36). Since the frame constant  $A$  is large ( $A \approx 628$  in this case), we expect according to the Theorem 19, although it is not guaranteed, to have  $T \approx 0.1\bar{N}$  and  $\Omega \approx 0.1\bar{M}$  where  $(\bar{N}, \bar{M}) = \arg \inf_{(N,M) \in \mathbb{Z}^2} \langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  if  $T$  and  $\Omega$  are integer multiples of  $\tau_0 = 0.1$  and  $\xi_0 = 0.1$  respectively.

For  $T = 1.1 = 11 * \tau_0$  and  $\Omega = 0.9 = 9 * \tau_0$ , we observe in Figure 7.5 that the minimum is attained at  $(N, M) = (11, 8)$ . In other words, we estimate the amount translation  $T$  correctly, and we make an error of 0.1 when we estimate  $\Omega$ , the amount of modulation.<sup>3</sup>Figure 7.6 shows the value of  $\langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  as a function of  $N$  and  $M$  for  $T = 1.17$  and  $\Omega = 0.93$ . In this case  $\langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  attains its minimum at  $N = 12$  and  $M = 8$ , i.e. the estimated values of  $T$  and  $\Omega$  are 1.2 and 0.8 respectively. This indicates that even the original function is translated and modulated by amounts that are non-integer multiples of the time and frequency translation steps  $\tau_0$  and  $\xi_0$  (both equal to 0.1 in this example), the algorithm can still estimate these amounts (with the resolution of integer multiples of  $\tau_0$  and  $\xi_0$ ).

Finally, we want to observe the effects of noise. We consider the case where  $f_{T,\Omega}$  is defined as above with  $T = 1.17$  and  $\Omega = 0.93$ . We will add independent identically distributed Gaussian random variables  $\nu_k$  to each sample of  $f_{T,\Omega}(k\tau_1)$  ( $\tau_1$  is period at which  $f_{T,\Omega}$  is sampled; we choose  $\tau_1 = \tau_0$ ) before computing the frame coefficients. We then compute the frame coefficients  $\tilde{c}_{n,m}$  using  $(f_{T,\Omega}(k\tau_1) + \nu_k)_{k=-K}^K$  and via the time-frequency sigma-delta scheme we quantize  $\tilde{c}_{n,m}$  to obtain  $\tilde{F}_{T,\Omega}^\nu$ . Figure 7.7 shows the  $\langle \tilde{F}_{T,\Omega,A}^\nu - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  where the signal-to-noise ratio is approximately 23 dB. The *signal-to-noise ratio* (SNR) will be defined as

$$\text{SNR} = 10 \log \frac{\sum_{k=-K}^K |f_{T,\Omega}(k\tau_1)|^2}{(2K+1)\sigma^2} \text{dB}, \quad (7.73)$$

---

<sup>3</sup>Several other experiments we conducted suggest that this is the case most of the time. One possible reason why we cannot detect the amount of modulation as accurately as we can detect the amount of translation is that  $\|f_{T,\Omega} - f_{T+0.1,\Omega}\|_{L^2}$  is much larger than  $\|f_{T,\Omega} - f_{T,\Omega+0.1}\|_{L^2}$  because  $f$  has fast decay and  $\xi_0$  is small.

where  $\sigma^2$  is the variance of  $\nu_k$ ;  $2K + 1$  samples  $f_{T,\Omega}$  is used to compute the frame coefficients. Figure 7.8 and Figure 7.9 show the  $\langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  where the SNR is 10.73 dB and 1.06 dB respectively. We observe that the algorithm does reasonably well for the two cases where the signal-to-noise ratio is larger; however for SNR= 1.7 dB, the minimum value of  $\langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  is much larger than the other two cases where the SNR is larger and so is the error in the estimation of  $T$  and  $\Omega$ .

## 7.4 Higher-order time-frequency sigma-delta schemes

In this section we will introduce higher-order time-frequency sigma-delta schemes to quantize the frame expansions of functions in  $\mathcal{B}^\varphi$  for tight Weyl-Heisenberg frames. We will show that the approximation error is of order  $(1/A)^k$  with a  $k^{th}$ -order scheme when the frame bound is  $A$ . Let  $(\varphi, \tau_0, \xi_0)$  be a tight Weyl-Heisenberg frame with frame bound  $A$ . Let  $f$  be in  $\mathcal{B}^\varphi$ ;  $c = (c_{n,m})$  with  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$  as before. Denote the real and imaginary parts of  $c_{n,m}$  by  $a_{n,m}$  and  $b_{n,m}$  respectively. Let  $(\Delta_1^{(k)} x)_{n,m} := \sum_{l=0}^k (-1)^l \binom{k}{l} x_{n-l,m}$  and  $(\Delta_2^{(k)} x)_{n,m} := \sum_{l=0}^k (-1)^l \binom{k}{l} x_{n,m-l}$  for any sequence  $x$ . To define the  $k^{th}$ -order time-frequency sigma-delta quantization scheme, consider the recursion relations:

$$\begin{aligned} (\Delta_1^{(k)} u^R)_{n,m} &= a_{n,m} - p_{n,m}^R \\ p_{n,m}^R &= \text{sign}(M((\Delta_1^{(0)} u^R)_{n-1,m}, \dots, (\Delta_1^{(k-1)} u^R)_{n-1,m}, a_{n,m})) \end{aligned} \quad (7.74)$$

$$\begin{aligned} (\Delta_2^{(k)} v^R)_{n,m} &= \bar{u}_{n,m}^R - r_{n,m}^R \\ r_{n,m}^R &= \text{sign}(M((\Delta_2^{(0)} v^R)_{n,m-1}, \dots, (\Delta_2^{(k-1)} v^R)_{n,m-1}, \bar{u}_{n,m}^R)), \end{aligned} \quad (7.75)$$

and

$$\begin{aligned} (\Delta_1^{(k)} u^I)_{n,m} &= b_{n,m} - p_{n,m}^I \\ p_{n,m}^I &= \text{sign}(M((\Delta_1^{(0)} u^I)_{n-1,m}, \dots, (\Delta_1^{(k-1)} u^I)_{n-1,m}, a_{n,m})) \end{aligned} \quad (7.76)$$

$$\begin{aligned} (\Delta_2^{(k)} v^I)_{n,m} &= \bar{u}_{n,m}^I - r_{n,m}^I \\ r_{n,m}^I &= \text{sign}(M((\Delta_2^{(0)} v^I)_{n,m-1}, \dots, (\Delta_2^{(k-1)} v^I)_{n,m-1}, \bar{u}_{n,m}^I)), \end{aligned} \quad (7.77)$$

where  $\bar{u}^R := u^R/C_{k,M}$ ,  $\bar{u}^I := u^I/C_{k,M}$  and  $M$  is a function which guarantees that  $u^R$ ,  $v^R$ ,  $u^I$  and  $v^I$  are uniformly bounded in  $l^\infty$  by  $C_{k,M}$ . Note that the recursion relations (7.74), (7.75), (7.76) and (7.77) all correspond to  $k^{\text{th}}$ -order standard sigma-delta quantizers, as in (1.13), with  $a_{n,k}$ ,  $\bar{u}_{n,k}^R$ ,  $b_{n,k}$  and  $\bar{u}_{n,k}^I$  respectively as their input. Thus, since all these sequences are bounded in  $l^\infty$  by 1, such an  $M$  exists due to [3]. Note that

$$C_{k,M}(\Delta_1^{(k)} \Delta_2^{(k)} v^R)_{n,m} = a_{n,m} - (p_{n,m}^R + C_{k,M}(\Delta_1^{(k)} r^R)_{n,m}), \quad (7.78)$$

and similarly

$$C_{k,M}(\Delta_1^{(k)} \Delta_2^{(k)} v^I)_{n,m} = b_{n,m} - (p_{n,m}^I + C_{k,M}(\Delta_1^{(k)} r^I)_{n,m}). \quad (7.79)$$

We will now define the sequences  $q^R$  and  $q^I$  by  $q_{n,m}^R = p_{n,m}^R + C_{k,M}(\Delta_1^{(k)} r^R)_{n,m}$  and  $q_{n,m}^I = p_{n,m}^I + C_{k,M}(\Delta_1^{(k)} r^I)_{n,m}$ . Finally, let us define  $T_{TF_k}$  by

$$T_{TF_k}(c) := q, \quad (7.80)$$

where  $q_{n,m} := q_{n,m}^R + iq_{n,m}^I$ .

**Theorem 20.** *Let  $(\varphi, \tau_0, \xi_0)$  be a tight Weyl-Heisenberg frame with frame bound  $A$ .*

Let  $f$  be in  $\mathcal{B}^\varphi$  and define  $q$  by (7.80). Consider the function

$$\tilde{F}_{A,k}(\tau, \xi) = \frac{1}{A} \sum_{n,m} q_{n,m} \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle. \quad (7.81)$$

Suppose  $\varphi$  is chosen such that  $\Phi(\tau, \xi) = \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle$  satisfies

$$\frac{\partial^k}{\partial \tau^k} \frac{\partial^k}{\partial \xi^k} (e^{i\tau\xi} \Phi(\tau, \xi)) \in L^1(\mathbb{R}^2). \quad (7.82)$$

Then

$$|F(\tau, \xi) - \tilde{F}_{A,k}(\tau, \xi)| \leq \frac{1}{A^k} \sum_{l=0}^k C_{k,\varphi,l} |\tau|^l \quad (7.83)$$

with

$$C_{k,\varphi,l} = (2\pi)^{k-1} C_{k,M} \|v\|_{l^\infty} \binom{k}{l} \|\partial_2^{(k-l)} \partial_1^k \Gamma\|_{L^1(\mathbb{R}^2)} \quad (7.84)$$

where  $k$  is the order of the quantizer and  $\Gamma(t, z) = e^{itz} \Phi(t, z)$ . We will call  $\tilde{F}_{A,k}$  the  $k^{\text{th}}$ -order time-frequency sigma-delta approximation of  $F$ .

We need the following lemma to prove Theorem 20, which we shall prove for the sake of completeness.

**Lemma 4.** Let  $\bar{\Delta}$  denote the forward difference operator, i.e.  $(\bar{\Delta}x)_n := x_n - x_{n+1}$ , as before. The following equality holds for any function  $f \in C^k$ :

$$\bar{\Delta}^k f(x - n\alpha) = \alpha^{k-1} \int_0^{k\alpha} f^{(k)}(x - (n+k)\alpha + t) \rho_k\left(\frac{t}{\alpha}\right) dt \quad (7.85)$$

for any  $\alpha$ . In (7.85),  $\rho_k$  is the  $k^{\text{th}}$ -order B-spline,  $\rho_k = \chi_{[0,1]} * \cdots * \chi_{[0,1]}$  ( $k$  convolution factors). (Note that the support of  $\rho_k$  is on  $[0, k]$ , and  $\sum_n \rho_k(x+n) = 1$  for all  $y \in \mathbb{R}$ .)

**Proof:** We use induction. Equation (7.85) certainly holds for  $k = 0$ . Suppose (7.85)

is true for  $k$ . Then

$$\begin{aligned} \bar{\Delta}^{k+1}f(x - n\alpha) &= \bar{\Delta}^k f(x - n\alpha) - \bar{\Delta}^k f(x - (n+1)\alpha) \\ &= \alpha^{k-1} \int_0^{k\alpha} (f^{(k)}(x - (n+k)\alpha + t) - f^{(k)}(x - (n+k+1)\alpha + t)) \rho_k\left(\frac{t}{\alpha}\right) dt \end{aligned} \quad (7.86)$$

Note that  $I := f^{(k)}(x - (n+k)\alpha + t) - f^{(k)}(x - (n+k+1)\alpha + t)$  can be rewritten as

$$I = \int_0^\alpha f^{(k+1)}(x - (n+k+1)\alpha + s) \rho_1\left(\frac{s}{\alpha}\right) ds. \quad (7.87)$$

Substituting (7.87) in (7.86) we get

$$\bar{\Delta}^{k+1}f(x - n\alpha) = \int_0^{k\alpha} \left( \int_0^\alpha f^{(k+1)}(x - (n+k+1)\alpha + s) \rho_1\left(\frac{s}{\alpha}\right) ds \right) \rho_k\left(\frac{t}{\alpha}\right) dt. \quad (7.88)$$

Finally a change of variables yields the result: Let  $u = t + s$ . Then

$$\bar{\Delta}^{k+1}f(x - n\alpha) = \int_0^{(k+1)\alpha} f^{(k+1)}(x - (n+k+1)\alpha + u) \left( \int_0^{k\alpha} \rho_1\left(\frac{u-t}{\alpha}\right) \rho_k\left(\frac{t}{\alpha}\right) dt \right) du, \quad (7.89)$$

which completes the proof since

$$\int_0^{k\alpha} \rho_1\left(\frac{u-t}{\alpha}\right) \rho_k\left(\frac{t}{\alpha}\right) dt = \alpha \rho_{k+1}\left(\frac{u}{\alpha}\right). \quad (7.90)$$

The last equation follows because  $\varphi_{k+1} = \rho_k * \rho_1$  and the support of  $\rho_k$  is on  $[0, k]$ .  $\square$

**Corollary 3.** *For any  $f$  in  $C^k$ , we have*

$$\bar{\Delta}^k f(n\alpha) = \alpha^{k-1} \int_0^{k\alpha} f^{(k)}((n-k)\alpha + t) \rho_k\left(\frac{t}{\alpha}\right) dt. \quad (7.91)$$

**Proof:** Set  $x = 0$  in (7.85) and switch  $n$  with  $-n$   $\square$

Now we are ready to prove Theorem 20.

**Proof of Theorem 20:** As in the proof of Theorem 12, let us write

$$\langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle = \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi)$$

where  $\alpha_{n,m}(\tau, \xi) = e^{-in\tau_0(\xi - m\xi_0)}$  and  $\Phi_{n,m}(\tau, \xi) = \Phi(\tau - n\tau_0, \xi - m\xi_0)$  with  $\Phi(\tau, \xi) = \langle \varphi, \varphi_{\tau,\xi} \rangle$ . Then the error term is

$$F(\tau, \xi) - \tilde{F}_{A,k}(\tau, \xi) = \frac{1}{A} \sum_{n,m} (c_{n,m} - q_{n,m}) \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) \quad (7.92)$$

$$= \frac{C_{k,M}}{A} \sum_{n,m} (\Delta_1^k \Delta_2^k v)_{n,m} \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) \quad (7.93)$$

$$= \frac{C_{k,M}}{A} \sum_{n,m} v_{n,m} (\bar{\Delta}_2^k \bar{\Delta}_1^k \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi))_{n,m}. \quad (7.94)$$

Now let us define  $I_{n,m} = \bar{\Delta}_2^k \bar{\Delta}_1^k \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi)$  which we can write also as

$$I_{n,m} = e^{-i\tau\xi} \bar{\Delta}_2^k \bar{\Delta}_1^k e^{i\tau m\xi_0} \Gamma(\tau - n\tau_0, \xi - m\xi_0). \quad (7.95)$$

As usual,  $\Gamma(\tau, \xi) = e^{i\tau\xi} \Phi(\tau, \xi)$ . Now let  $\Omega_{\tau,\xi}(t, z) := e^{iz\tau} \Gamma(t, \xi - z)$ . Then

$$I_{n,m} = e^{-i\tau\xi} \bar{\Delta}_2^k \bar{\Delta}_1^k \Omega_{\tau,\xi}(\tau - n\tau_0, m\xi_0). \quad (7.96)$$

By Lemma 4 we can write (7.96) as

$$I_{n,m} = e^{-i\tau\xi} \bar{\Delta}_2^k \tau_0^{k-1} \int_0^{k\tau_0} \partial_1^{(k)} \Omega_{\tau,\xi}(\tau - (n+k)\tau_0 + t, m\xi_0) \rho_k\left(\frac{t}{\tau_0}\right) dt \quad (7.97)$$

$$= e^{-i\tau\xi} \tau_0^{k-1} \int_0^{k\tau_0} \left( \bar{\Delta}_2^k \partial_1^{(k)} \Omega_{\tau,\xi}(\tau - (n+k)\tau_0 + t, m\xi_0) \right) \rho_k\left(\frac{t}{\tau_0}\right) dt \quad (7.98)$$

$$= e^{-i\tau\xi} \frac{(2\pi)^{k-1}}{A^{k-1}} \int_0^{k\tau_0} \int_0^{k\xi_0} \partial_2^{(k)} \partial_1^{(k)} \Omega_{\tau,\xi}(\tau - (n+k)\tau_0 + t, (m-k)\xi_0 + z) \rho_k\left(\frac{z}{\xi_0}\right) \rho_k\left(\frac{t}{\tau_0}\right) dz dt. \quad (7.99)$$

In the last equality we use the fact that  $A = \frac{2\pi i}{\tau_0 \xi_0}$ . By Lemma 4 the support of  $\rho_k$  is on  $[0, k]$ ; therefore we can replace the integration limits of both integrals in (7.99) by

$-\infty$  and  $\infty$  Thus after the appropriate change of variables in both integrals we get

$$\begin{aligned} & \bar{\Delta}_2^k \bar{\Delta}_1^k \alpha_{n,m}(\xi) \Phi_{n,m}(\tau, \xi) \\ &= \frac{2\pi^{k-1}}{A^{k-1}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \partial_2^{(k)} \partial_1^{(k)} \Omega_{\tau,\xi}(p, s) \rho_k\left(\frac{p}{\tau_0} - \frac{p}{\tau} + n + k\right) \rho_k\left(\frac{s}{\xi_0} - m + k\right) dp ds \end{aligned} \quad (7.100)$$

Finally let us substitute (7.100) into (7.94) and take the absolute value of the resulting expression:

$$\begin{aligned} & |F(\tau, \xi) - \tilde{F}_{A,k}(\tau, \xi)| \\ & \leq \frac{C_{k,M} \|v\|_{l^\infty} (2\pi)^{(k-1)}}{A^k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\partial_2^{(k)} \partial_1^{(k)} \Omega_{\tau,\xi}(p, s)| \sum_{n,m} \rho_k\left(\frac{p}{\tau_0} - \frac{p}{\tau} + n + k\right) \rho_k\left(\frac{s}{\xi_0} - m + k\right) dp ds \\ & \leq \frac{C_{k,M} \|v\|_{l^\infty} (2\pi)^{(k-1)}}{A^k} \|\partial_2^{(k)} \partial_1^{(k)} \Omega_{\tau,\xi}\|_{L^1} \end{aligned} \quad (7.101)$$

Note that we have (7.101) since  $\rho_k \geq 0$  and

$$\sum_{n,m} \rho_k\left(\frac{p}{\tau_0} - \frac{p}{\tau} + n + k\right) \rho_k\left(\frac{s}{\xi_0} - m + k\right) = 1. \quad (7.102)$$

Finally, using

$$\partial_2^{(k)} \partial_1^{(k)} \Omega_{\tau,\xi}(t, z) = \sum_{l=0}^k \binom{k}{l} (i\tau)^l e^{iz\tau} \partial_2^{(k-l)} \partial_1^{(k)} \Gamma(t, \xi - z), \quad (7.103)$$

we get the result.  $\square$

### Remarks:

1. We will again approximate  $f$  as a linear functional on a test function space. For a  $k^{th}$ -order time-frequency sigma-delta quantization scheme an appropriate test function space is

$$\mathcal{G}_k = \{g \in L^2(\mathbb{R}) : (1 + \tau + \dots + \tau^k) \langle g, \varphi_{\tau,\xi} \rangle \in L^1(\mathbb{R}^2)\}. \quad (7.104)$$

Let  $g \in \mathcal{G}_k$  and define  $G := \langle g, \varphi_{\tau,\xi} \rangle$ . For  $f \in \mathcal{B}^\varphi$ , let  $\tilde{F}_{A,k}$  be defined as in

(7.81). Then

$$\langle F - \tilde{F}_{A,k}, G \rangle := \int (F(\tau, \xi) - \tilde{F}_{A,k}(\tau, \xi))G(\tau, \xi)d\tau d\xi \quad (7.105)$$

is finite; thus  $\langle \tilde{F}_{A,k}, G \rangle$  is well-defined. We now define  $\tilde{f}_{A,k}$  as a linear functional on  $\mathcal{G}_k$  such that

$$\langle \tilde{f}_{A,k}, g \rangle := \langle \tilde{F}_{A,k}, G \rangle. \quad (7.106)$$

By Theorem 20 we can conclude

$$|\langle \tilde{f}_{A,k}, g \rangle| \leq \frac{1}{A^k} \sum_{l=0}^k C_{\varphi,l} \|\tau^l G(\tau, \xi)\|_{L^1 \mathbb{R}^2}, \quad (7.107)$$

where  $C_{k,\varphi,l}$  is as in (7.84).

2. Let  $f_1$  and  $f_2$  be two functions in  $\mathcal{B}^\varphi$ ,  $q^1$  and  $q^2$  the corresponding sequences produced by the  $k^{th}$ -order time-frequency sigma-delta scheme, and let  $\tilde{F}_{A,k}^1$  and  $\tilde{F}_{A,k}^2$  be the  $k^{th}$ -order time-frequency sigma-delta approximations of  $f_1$  and  $f_2$ , respectively. Then, regardless of the order of the approximation, we have

$$\langle \tilde{F}_{A,k}^1 - \tilde{F}_{A,k}^2, G \rangle = \frac{1}{A} \sum_{n,m} (q_{n,m}^1 - q_{n,m}^2) \overline{\langle g, \varphi_{n,m} \rangle}. \quad (7.108)$$

Similarly, for any  $f$  in  $\mathcal{B}^\varphi$ , let  $q = T_{TF_k}(c)$  where  $c$  denotes the sequence of the frame coefficients of  $f$ ; suppose  $\tilde{F}_{A,k}$  is the  $k^{th}$ -order time-frequency sigma-delta approximation of  $f$ . Then we have

$$\langle F - \tilde{F}_{A,k}, G \rangle = \frac{1}{A} \sum_{n,m} (c_{n,m} - q_{n,m}) \overline{\langle g, \varphi_{n,m} \rangle}. \quad (7.109)$$

3. Theorem 14 and Theorem 16 are true regardless of the order  $k$  of the time-

frequency sigma-delta scheme that is used to approximate a given function  $f \in \mathcal{B}^\varphi$ , as long as  $\varphi$  satisfies the conditions stated in Theorem 20 and the test functions are in the appropriate test function space. Theorems 15, 17, 18 and 19 need some modification to be true for the case where the quantizer is of  $k^{\text{th}}$ -order. We state these modified versions below: Theorems 21, 22, 23 and 24 are the generalized versions of the aforementioned theorems respectively. The proofs are similar to the first order case and will be omitted.

**Theorem 21.** *Let  $f_1, f_2$  be in  $\mathcal{B}^\varphi$ ,  $F^j = \langle f_j, \varphi_{\tau, \xi} \rangle$  for  $j = 1, 2$ , and  $\tilde{F}_{A,k}^j$  the  $k^{\text{th}}$ -order time-frequency sigma-delta approximation of  $F^j$ . Then*

$$|\langle F^1 - F^2, G \rangle - \langle \tilde{F}_{A,k}^1 - \tilde{F}_{A,k}^2, G \rangle| \leq \frac{4\pi}{A^k} \sum_{l=0}^k C_{k,\varphi,l} \|\tau^l G(\tau, \xi)\|_{L^1(\mathbb{R}^2)}. \quad (7.110)$$

where  $C_{k,\varphi,l}$  is defined as in (7.84).

**Theorem 22.** *Let  $q = T_{TF_k}(c)$ , (i.e. the quantization scheme is of order  $k$ ), where  $c = (c_{n,m})_{(n,m) \in \mathbb{Z}^2}$  with  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$  for some  $f$  in  $\mathcal{B}^\varphi$ . Suppose  $H$  and  $\tilde{H}_A$  are defined as in (7.50) and (7.51) respectively. Then*

$$|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{1}{A^k} \sum_{l=0}^k \tilde{C}_{k,\varphi,l} |\tau|^l \quad (7.111)$$

with

$$\tilde{C}_{k,\varphi,l} = (2\pi)^{k-1} C_{k,M} \|v\|_{l^\infty} \sum_{j=l}^k \binom{k}{j} \binom{j}{l} (N\tau_0)^{j-l} \|\partial_2^{(k-j)} \partial_1^{(k)} \Gamma\|. \quad (7.112)$$

**Theorem 23.** *Let  $f$  be in  $\mathcal{B}^\varphi$ ,  $c = (\langle f, \varphi_{n,m} \rangle)$  and  $q = (q_{n,m}) = T_{TF}(c)$ . Suppose*

$H(\tau, \xi)$  is the windowed Fourier transform of  $e^{-iM\xi_0} f(\cdot)$ . Then

$$|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{1}{A^k} \sum_{l=0}^k C_{k,\varphi,l} |\tau|^l \quad (7.113)$$

where  $C_{k,\varphi,l}$  is as in (7.84), and

$$\tilde{H}_A = \frac{1}{A} \sum_{n,m} q_{n,m+M} \alpha_{n,m} \Phi_{n,m}.$$

**Theorem 24.** Let  $f$  be in  $\mathcal{B}^\varphi$ ,  $c = (\langle f, \varphi_{n,m} \rangle)$  and  $q = (q_{n,m}) = T_{TF_k}(c)$ . Suppose  $H(\tau, \xi)$  is the windowed Fourier transform of  $e^{-iM\xi_0} f(\cdot + N\tau_0)$ . Then

$$|H(\tau, \xi) - \tilde{H}_A(\tau, \xi)| \leq \frac{1}{A^k} \sum_{l=0}^k \tilde{C}_{k,\varphi,l} |\tau|^l, \quad (7.114)$$

where  $\tilde{C}_{k,\varphi,l}$  is as in (7.112) and

$$\tilde{H}_A(\tau, \xi) = \frac{1}{A} \sum_{n,m} (\gamma_N)^{m+M} q_{n+N,m+M} \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle. \quad (7.115)$$

### 7.4.1 Numerical Experiment revisited

In this section, we will present the results of numerical experiments for the second-order TF $\Sigma\Delta$ -I quantizer analogous to those discussed in Section 7.3 for the first-order quantizer. We choose  $\varphi(t) = \pi^{1/4} e^{-\frac{t^2}{2}}$ . As we have discussed before,  $(\varphi, \tau_0, \xi_0)$  constitutes a frame if  $\tau_0$  and  $\xi_0$  is sufficiently small; moreover the frame is almost tight with the frame bound  $A \approx \frac{2\pi}{\tau_0 \xi_0}$  if  $\tau_0$  and  $\xi_0$  are sufficiently small and  $\tau_0 \approx \xi_0$ .

We will quantize the frame expansion of the function  $f(t) = 0.4e^{-(i0.9t^3 + 0.05t^2)}$ , which is the same function we have used in Section 7.3. We have already computed the frame coefficients  $\langle f, \varphi_{n,m} \rangle$  of  $f$ . Using the algorithm described in (7.74)-(7.77) with  $k = 2$  and  $M(u, v, x) = u + 0.5v$  we obtain the quantized frame coefficients

$q_{n,m}$  of  $f$ ; these are shown in Figure 7.10. Next, we fix the function  $G_{tot}$ , defined as in (7.72), as our test function and compute the inner product  $\langle F - \tilde{F}_{A,2}, G_{tot} \rangle$  via (7.109) for various values of the frame bound  $A$ . Figure 7.11 shows the value of  $\langle F - \tilde{F}_{A,2}, G_{tot} \rangle$  while  $A$  takes values between 100 and 628. Similar to the first-order case, the decay of the approximation error is faster than the predicted rate, i.e. instead of being  $O(A^{-2})$ , the approximation error seems to be of order  $A^{-5/2}$ . This again matches the empirical error decay rate observed for the standard second-order sigma-delta quantizers.

Next, we want to observe the translation invariance of the second-order quantizers. To this end, we repeat the experiment we did in Section 7.3 : Fix the frame  $(\varphi, 0.1, 0.1)$  and compute  $q = T_{TF_2}(c)$ , i.e. use a second order quantizer, where  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$ . Now, as in Section 7.3, define  $f_{T,\Omega}$  by  $f_{T,\Omega}(t) := e^{-i\Omega t} f(t + T)$ . Let  $c_{T,\Omega}$  be the sequence  $(\langle f_{T,\Omega}, \varphi_{n,m} \rangle)$  and  $q_{T,\Omega} := T_{TF_2}(c_{T,\Omega})$ . Using  $q$  as a template, we will estimate what  $T$  and  $\Omega$  are when we are only given the sequence  $q_{T,\Omega}$ . To accomplish this, we will compare  $\tilde{F}_{T,\Omega,A,2} := \sum (q_{T,\Omega})_{n,m} \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle$  with  $I_{N,M} := \sum (\gamma_N)^{m+M} q_{n+N,m+M} \langle \varphi_{n,m}, \varphi_{\tau,\xi} \rangle$  for various  $N$  and  $M$  by comparing the inner products  $\langle \tilde{F}_{T,\Omega,A,2} - I_{N,M}, G_{tot} \rangle$ . We will calculate these inner products via (7.108). Since the frame constant  $A$  is large ( $A \approx 628$  in this case), we expect according to Theorem 24 (although it is not guaranteed) to have  $T \approx 0.1\bar{N}$  and  $\Omega \approx 0.1\bar{M}$  where  $(\bar{N}, \bar{M}) = \arg \inf_{(N,M) \in \mathbb{Z}^2} \langle \tilde{F}_{T,\Omega,A} - I_{N,M}, G_{tot} \rangle$  if  $T$  and  $\Omega$  are integer multiples of  $\tau_0 = 0.1$  and  $\xi_0 = 0.1$  respectively.

For  $T = 1.1 = 11 * \tau_0$  and  $\Omega = 0.9 = 9 * \tau_0$ , we observe in Figure 7.12 that the minimum is attained at  $(N, M) = (11, 9)$ , in other words our algorithm estimated the translation amounts  $T$  and  $\Omega$  correctly. Next we test whether the algorithm can detect translation and modulation amounts that are **not** integer multiples of  $\tau_0$  and  $\xi_0$  (of course with the resolution given by  $\tau_0$  and  $\xi_0$ ). Figure 7.13 shows the result when  $T = 1.17$  and  $\Omega = 0.93$ . One observes that the algorithm has estimated  $T$  and

$\Omega$  as well as the resolution allows.

Finally, we add noise to our signal the way we described in Section 7.3, and again we use our algorithm to estimate the translation and modulation amounts  $T$  and  $\Omega$ . Figures 7.14-7.15 show the plots of  $\langle \tilde{F}_{T,\Omega,A,2}^\nu - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for different of SNR. (We define  $\tilde{F}_{T,\Omega,A,2}^\nu$  is defined the same way we defined  $\tilde{F}_{T,\Omega,A}^\nu$  just above (7.73), only this time using the  $q$  produced by the second-order quantizer.) One observes that the algorithm seems to be working well as long as the SNR is reasonably large.

Figure 7.10: The quantized frame coefficients  $q_{n,m}$  –obtained via the second-order scheme. The upper left figure shows the real part of the quantized coefficients; the upper right figure shows the imaginary parts of the quantized coefficients –black corresponds to  $-10$  and white corresponds to  $10$  in these figures. The lower figure shows the absolute value of the quantized coefficients; in this figure black corresponds to  $0$  and white corresponds to  $10\sqrt{2}$ .

Figure 7.11: The ‘approximation error’  $|\langle F - \tilde{F}_{A,2}, G_{tot} \rangle|$  vs. the frame bound  $A$  for the second-order case. Both axes are logarithmic. The solid line seen in the figure is the graph  $\{(A, A^{-2}) : 100 < A < 628\}$ ; the dashed line is the graph  $\{(A, 10A^{-5/2}) : 100 < A < 628\}$ .

Figure 7.12: The value  $\langle \tilde{F}_{T,\Omega,A,2} - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.1$  and  $\Omega = 0.9$ ; the minimum is obtained at  $N = 11$  and  $M = 9$ , which means that the algorithm predicts  $T = 1.1$  and  $\Omega = 0.9$ , i.e. the correct values of  $T$  and  $\Omega$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A,2} - I_{11,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A} - I_{N,9}, G_{tot} \rangle$  vs.  $N$ .

Figure 7.13: The value  $\langle \tilde{F}_{T,\Omega,A,2} - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.17$  and  $\Omega = 0.93$ ; the minimum is obtained at  $N = 12$  and  $M = 9$ , which means that the algorithm predicts  $T = 1.2$  and  $\Omega = 0.9$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A} - I_{12,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A} - I_{N,9}, G_{tot} \rangle$  vs.  $N$ .

Figure 7.14: The value  $\langle \tilde{F}_{T,\Omega,A,2}^\nu - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.1$  and  $\Omega = 0.9$ ; SNR= 10.94 dB; the minimum is obtained at  $N = 11$  and  $M = 10$ , which means that the algorithm predicts  $T = 1.1$  and  $\Omega = 1$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A,2}^\nu - I_{12,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A,2}^\nu - I_{N,9}, G_{tot} \rangle$  vs.  $N$ .

Figure 7.15: The value  $\langle \tilde{F}^{\nu}_{T,\Omega,A,2} - I_{N,M}, G_{tot} \rangle$  versus  $N$  and  $M$  for  $T = 1.1$  and  $\Omega = 0.9$ ; SNR= 1.65 dB; the minimum is obtained at  $N = 13$  and  $M = 12$ , which means that the algorithm predicts  $T = 1.3$  and  $\Omega = 1.2$ . The upper left figure shows  $\langle \tilde{F}_{T,\Omega,A,2} - I_{13,M}, G_{tot} \rangle$  vs.  $M$ ; the upper right one shows  $\langle \tilde{F}_{T,\Omega,A,2} - I_{N,12}, G_{tot} \rangle$  vs.  $N$ .

# Chapter 8

## The Time-Frequency Sigma-Delta Quantization Algorithm II (TFΣΔ-II)

In this chapter we will introduce an alternative algorithm to quantize the frame coefficients of certain functions in  $L^2(\mathbb{R})$ . This time we will reconstruct the original function in  $L^2$  and we will prove that the  $L^2$ -approximation error is of order  $A^{-k}$  for a  $k^{\text{th}}$ -order scheme. Although the above stated results are much stronger than the analogous results we obtained in the previous chapter, the algorithm which we will introduce in this chapter is **not** translation invariant, unlike the *time-frequency sigma-delta quantization algorithm I* (TFΣΔ-I) of the last chapter. Also, the class of functions which can be quantized using TFΣΔ-II is smaller than the class of functions that can be quantized using TFΣΔ-I.

### 8.1 The Algorithm

Let  $G$  be a fixed function in  $L^2(\mathbb{R}^2)$  that is smooth with nice decay. Let  $(\varphi, \tau_0, \xi_0)$  be a tight Weyl-Heisenberg frame with frame bound  $A$ . Denote the collection of functions

in  $L^2(\mathbb{R})$  whose frame expansion converges almost everywhere (as well as in  $L^2$ ) by  $\mathcal{B}_{ae}^\varphi$ <sup>1</sup> and consider those functions  $f$  in  $\mathcal{B}_{ae}^\varphi$  which satisfy

$$|\langle f, \varphi_{\tau, \xi} \rangle| < G(\tau, \xi), \quad (8.1)$$

where  $\varphi_{\tau, \xi} = \varphi(t - \tau)e^{i\xi t}$  as before. Denote the collection of such functions by  $\mathcal{B}_G^\varphi$ .

Let  $f$  be in  $\mathcal{B}_G^\varphi$ . Denote the frame coefficients of  $f$ ,  $\langle f, \varphi_{n, m} \rangle$  by  $c_{n, m}$ ; define  $a_{n, m}$  and  $b_{n, m}$  as the real and imaginary parts of  $c_{n, m}$  respectively. Since  $\langle f, \varphi_{n, m} \rangle = F(n\tau_0, m\xi_0)$  with  $F(\tau, \xi) = \langle f, \varphi_{\tau, \xi} \rangle$  and since  $f \in \mathcal{B}_G^\varphi$ , we clearly have

$$\begin{aligned} |a_{n, m}| &< G(n\tau_0, m\xi_0), \quad \text{and} \\ |b_{n, m}| &< G(n\tau_0, m\xi_0). \end{aligned}$$

In other words, if we define  $\tilde{a}_{n, m} := \frac{a_{n, m}}{G(n\tau_0, m\xi_0)}$  and  $\tilde{b}_{n, m} := \frac{b_{n, m}}{G(n\tau_0, m\xi_0)}$ , both  $|\tilde{a}_{n, m}|$  and  $|\tilde{b}_{n, m}|$  will be bounded by 1. Now, define the first-order TF $\sigma\Delta$ -II scheme as follows: Consider the recursion relations

$$\begin{aligned} u_{n, m}^R - u_{n-1, m}^R &= \tilde{a}_{n, m} - p_{n, m}^R \\ p_{n, m}^R &= \text{sign}(u_{n-1, m}^R + a_{n, m}) \\ \\ v_{n, m}^R - v_{n, m-1}^R &= u_{n, m}^R - r_{n, m}^R \\ r_{n, m}^R &= \text{sign}(v_{n, m-1}^R + u_{n, m}^R) \end{aligned} \quad (8.2)$$

---

<sup>1</sup>Clearly, the frame expansion of  $f$ ,  $\sum_{n, m} \langle f, \varphi_{n, m} \rangle \varphi_{n, m}$ , converges to  $f$  in  $L^2$ . Thus, to have almost everywhere convergence as well, it is sufficient, for example, that the frame coefficients of  $f$  are in  $l^1$ .

and

$$\begin{aligned}
u_{n,m}^I - u_{n-1,m}^I &= \tilde{b}_{n,m} - p_{n,m}^I \\
p_{n,m}^I &= \text{sign}(u_{n-1,m}^I + b_{n,m}) \\
v_{n,m}^I - v_{n,m-1}^I &= u_{n,m}^I - r_{n,m}^I \\
r_{n,m}^I &= \text{sign}(v_{n,m-1}^I + u_{n,m}^I).
\end{aligned} \tag{8.3}$$

By (1.11), the scheme described above is stable, i.e. the internal state variables stay uniformly bounded: since  $|\tilde{a}_{n,m}|$  is bounded by 1,  $|u_{n,m}^R|$  is bounded by 1; but this implies that  $v_{n,m}^R$  is also bounded by 1. Similarly,  $u_{n,m}^I$  and  $v_{n,m}^I$  are uniformly bounded by 1.

Note that

$$(\Delta_1 \Delta_2 v^R)_{n,m} = \tilde{a}_{n,m} - (p_{n,m}^R + \Delta_1 r_{n,m}^R), \tag{8.4}$$

and similarly

$$(\Delta_1 \Delta_2 v^I)_{n,m} = \tilde{b}_{n,m} - (p_{n,m}^I + \Delta_1 r_{n,m}^I); \tag{8.5}$$

thus we have

$$(\Delta_1 \Delta_2 v)_{n,m} = \tilde{c}_{n,m} - (p_{n,m} + \Delta_1 r_{n,m}), \tag{8.6}$$

where  $v_{n,m} = v_{n,m}^R + iv_{n,m}^I$ ,  $p_{n,m} = p_{n,m}^R + ip_{n,m}^I$ ,  $r_{n,m} = r_{n,m}^R + ir_{n,m}^I$ , and  $\tilde{c}_{n,m} := \frac{c_{n,m}}{G(n\tau_0, m\xi_0)}$ . If we then multiply both sides of (8.6) by  $G(n\tau_0, m\xi_0)$ , we get

$$G(n\tau_0, m\xi_0)(\Delta_1 \Delta_2 v)_{n,m} = c_{n,m} - G(n\tau_0, m\xi_0)(p_{n,m} + \Delta_1 r_{n,m}). \tag{8.7}$$

This suggests us to define  $T_{TFII}$  as the mapping that maps the sequence  $c = (c_{n,m})$  to the sequence  $q = (q_{n,m})$  with

$$q_{n,m} = G(n\tau_0, m\xi_0)(p_{n,m} + \Delta_1 r_{n,m}). \quad (8.8)$$

**Theorem 25.** *Let  $(\varphi, \tau_0, \xi_0)$ , with  $\varphi$  in  $L^\infty$ , be a tight Weyl-Heisenberg frame of  $L^2(\mathbb{R})$  with frame bound  $A$ . Let  $f \in \mathcal{B}_G^\varphi$ . Suppose  $\varphi$  and  $G$  satisfy i-iv:*

- i.  $|\varphi| \star G_1 \in L^2$  where  $\star$  stands for convolution and  $G_1(s) = \int |\partial_1 \partial_2 G(s, z)| dz$ .
- ii.  $|\varphi'| \star G_2 \in L^2$  where  $G_2(s) = \int |\partial_2 G(s, z)| dz$ .
- iii.  $t(|\varphi| \star G_3)(t) \in L^2$  where  $G_3(s) = \int |\partial_1 G(s, z)| dz$ , and
- iv.  $t(|\varphi'| \star G_4)(t) \in L^2$  where  $G_4(s) = \int |G(s, z)| dz$ .

Set  $q := T_{TFII}(c)$  and define

$$\tilde{f}_A = \frac{1}{A} \sum_{n,m} q_{n,m} \varphi_{n,m}. \quad (8.9)$$

Then

$$\|f - \tilde{f}_A\|_{L^2} \leq \frac{C}{A}, \quad (8.10)$$

with  $C = \|\varphi| \star G_1\|_{L^2} + \|\varphi'| \star G_2\|_{L^2} + \|t(|\varphi| \star G_3)(t)\|_{L^2} + \|t(|\varphi'| \star G_4)(t)\|_{L^2}$ .

**Proof:** Since  $f \in \mathcal{B}_G^\varphi$  (thus the frame expansion of  $f$  converges almost everywhere, too), the error term can be written as

$$f(t) - \tilde{f}_A(t) = \frac{1}{A} \sum_{n,m} (c_{n,m} - q_{n,m}) \varphi_{n,m}(t) \quad (8.11)$$

for almost every  $t$ . Then,

$$f(t) - \tilde{f}_A(t) = \frac{1}{A} \sum_{n,m} G(n\tau_0, m\xi_0) (\Delta_1 \Delta_2 v^I)_{n,m} \varphi_{n,m}(t) \quad (8.12)$$

$$= \frac{1}{A} \sum_{n,m} v_{n,m} \bar{\Delta}_2 \bar{\Delta}_1 G(n\tau_0, m\xi_0) \varphi_{n,m}(t), \quad (8.13)$$

where the first equality is due to (8.7); the second equality is the result of summation by parts. Note that the boundary terms disappear since  $G$  has nice decay in both its arguments. Now, denote  $\bar{\Delta}_2 \bar{\Delta}_1 G(n\tau_0, m\xi_0) \varphi_{n,m}$  by  $I_{n,m}$ . Then

$$I_{n,m}(t) = \bar{\Delta}_2 \bar{\Delta}_1 \Gamma(n\tau_0, m\xi_0, t), \quad (8.14)$$

where

$$\Gamma(s, z, t) = G(s, z) \varphi(t - s) e^{izt}. \quad (8.15)$$

Let us rewrite  $I_{n,m}$  as

$$I_{n,m}(t) = \bar{\Delta}_2 (\Gamma(n\tau_0, m\xi_0, t) - ((n+1)\tau_0, m\xi_0, t)) \quad (8.16)$$

$$= \bar{\Delta}_2 \int_{(n+1)\tau_0}^{n\tau_0} \partial_1 \Gamma(s, m\xi_0, t) ds \quad (8.17)$$

$$= \int_{(n+1)\tau_0}^{n\tau_0} (\partial_1 \Gamma(s, m\xi_0, t) - \partial_1 \Gamma(s, (m+1)\xi_0, t)) ds \quad (8.18)$$

$$= \int_{(n+1)\tau_0}^{n\tau_0} \int_{(m+1)\xi_0}^{m\xi_0} \partial_2 \partial_1 \Gamma(s, z, t) dz ds. \quad (8.19)$$

Note that since both  $\varphi$  and  $G$  are smooth with nice decay, so is  $\Gamma$ ; thus all the steps above are justified. Next, we substitute (8.19) in (8.13) and take the absolute value

of both sides to get

$$|f(t) - \tilde{f}_A(t)| \leq \frac{1}{A} \sum_{n,m} |v_{n,m}| \left| \int_{(n+1)\tau_0}^{n\tau_0} \int_{(m+1)\xi_0}^{m\xi_0} \partial_2 \partial_1 \Gamma(s, z, t) dz ds \right| \quad (8.20)$$

$$\leq \frac{1}{A} \|v\|_{l^\infty} \|\partial_2 \partial_1 \Gamma(\cdot, \cdot, t)\|_{L^1(\mathbb{R}^2)}, \quad (8.21)$$

for almost every  $t$ . To complete the proof, we will write  $\Gamma$  in terms of  $\varphi$  and  $G$  and show that  $\|\partial_2 \partial_1 \Gamma(\cdot, \cdot, t)\|_{L^2}$  is uniformly bounded, i.e. it does not depend on the particular choice of  $f$ , which implies that the error  $\|f - \tilde{f}_A\|_{L^2}$  is  $O(A^{-1})$ . A simple calculation yields

$$\begin{aligned} \partial_2 \partial_1 \Gamma(s, z, t) &= \varphi(s-t) e^{izt} (\partial_1 \partial_2 G(s, z) + it \partial_1 G(s, z)) \\ &\quad - \varphi'(s-t) e^{izt} (\partial_2 G(s, z) + it G(s, z)). \end{aligned} \quad (8.22)$$

Thus,

$$\begin{aligned} \|\partial_2 \partial_1 \Gamma(\cdot, \cdot, t)\|_{L^2} &\leq (G_1 \star |\varphi|)(t) + (G_2 \star |\varphi'|)(t) \\ &\quad + |t|((G_3 \star |\varphi|)(t) + (G_4 \star |\varphi'|)(t)), \end{aligned} \quad (8.23)$$

where  $G_i$  are defined as in the statement of the theorem. This means that

$$|f(t) - \tilde{f}_A(t)| \leq \frac{\|v\|_{l^\infty}}{A} D(t), \quad (8.24)$$

where  $D(t) = \|\partial_2 \partial_1 \Gamma(\cdot, \cdot, t)\|_{L^1(\mathbb{R}^2)}$  does not depend on  $f$ . Finally since  $\varphi$  and  $G$  are chosen such that the conditions stated in Theorem 25 are satisfied,  $\|D\|_{L^2}$  is finite and

$$\|D\|_{L^2} \leq \|\varphi \star G_1\|_{L^2} + \|\varphi' \star G_2\|_{L^2} + \|t(|\varphi| \star G_3)(t)\|_{L^2} + \|t(|\varphi'| \star G_4)(t)\|_{L^2}, \quad (8.25)$$

thus we conclude

$$\|f - \tilde{f}_A\|_{L^2} \leq \frac{\|v\|_{l^\infty}}{A} \|D\|_{L^2}. \quad (8.26)$$

□

**Remarks:**

1. Theorem 25 lists several conditions on  $\varphi$  and  $G$ ; these are all fulfilled if, for example,  $\varphi$  and  $G$  are in the Schwartz spaces  $\mathcal{S}(\mathbb{R})$  and  $\mathcal{S}(\mathbb{R}^2)$ , respectively.
2. Clearly, the algorithm is not shift-invariant: knowing the exact index of a quantizer output is essential, because to construct  $\tilde{f}_A$  we need to multiply the quantizer output with the index  $(n, m)$ , i.e.  $p_{n,m} + \Delta_1 r_{n,m}$ , by  $G(n\tau_0, m\xi_0)$ .
3. The TF $\Sigma\Delta$ -II algorithm suggests an algorithm to “coarsely quantize” the Fourier coefficients of a function with compact support; this will be explained in the next section.

### 8.1.1 Coarse quantization of the Fourier coefficients of certain compactly supported functions

Let  $f$  be a function in  $L^1(\mathbb{R})$  such that its support is on  $[-\pi, \pi]$ , and its Fourier transform  $\hat{f}(\xi)$  is  $O(1/\xi)$ . Clearly we can extend  $f$  to a periodic function  $f_\lambda$  in the following way:  $f_\lambda(t) = f(t)$  for  $t \in [-\lambda\pi, \lambda\pi]$ , and  $f_\lambda(t - \lambda 2\pi) = f_\lambda(t)$ . Then the Fourier series of  $f_\lambda$  is given by  $f_\lambda(t) = \frac{1}{2\pi\lambda} \sum_n \hat{f}(\frac{n}{\lambda}) e^{i\frac{n}{\lambda}t}$ , where equality holds pointwise by our choice of  $f$ . Note that we also have

$$f(t) = f_\lambda(t) \chi_{[-\lambda\pi, \lambda\pi]}(t), \quad (8.27)$$

which yields

$$f(t) = \frac{1}{2\pi\lambda} \sum_n \hat{f}\left(\frac{n}{\lambda}\right) e^{i\frac{n}{\lambda}t}. \quad (8.28)$$

Now let us fix a function  $G$  such that  $G$  and  $G'$  are in  $L^1(\mathbb{R})$  and  $G(\xi)$  is  $O(1/\xi)$ . Let  $f$  be a continuous function which is compactly supported with  $\text{supp}(f) \in [-\pi, \pi]$  and whose Fourier transform  $\hat{f}$  satisfies

$$|\hat{f}(\xi)| \leq G(\xi). \quad (8.29)$$

Denote the collection of all such functions by  $\mathcal{B}_G$ . For  $f \in \mathcal{B}_G$ , we clearly have

$$\left| \frac{\hat{f}\left(\frac{n}{\lambda}\right)}{G\left(\frac{n}{\lambda}\right)} \right| < 1. \quad (8.30)$$

Thus we can use the standard first-order sigma-delta scheme, as in (1.10), to quantize the sequence  $\left(\frac{\hat{f}\left(\frac{n}{\lambda}\right)}{G\left(\frac{n}{\lambda}\right)}\right)$ : Consider the recursion relations

$$\begin{aligned} (\Delta v)_n &= \frac{\hat{f}\left(\frac{n}{\lambda}\right)}{G\left(\frac{n}{\lambda}\right)} - q_n^\lambda \\ q_n^\lambda &= \text{sign}(v_{n-1} + \frac{\hat{f}\left(\frac{n}{\lambda}\right)}{G\left(\frac{n}{\lambda}\right)}) \end{aligned} \quad (8.31)$$

Since we have (8.30), we have  $|v_n| \leq 1$  for all  $n$  by (1.11). Let us define  $T_G$  as the mapping that maps the sequence  $(f(\frac{n}{\lambda}))$  to the sequence  $\tilde{q}$ , where  $\tilde{q}_n := G(\frac{n}{\lambda})q_n$ . Note that

$$G\left(\frac{n}{\lambda}\right)(v_n - v_{n-1}) = \hat{f}\left(\frac{n}{\lambda}\right) - \tilde{q}_n. \quad (8.32)$$

Now define

$$\tilde{f}_{\lambda,1}(t) := \frac{1}{2\pi\lambda} \sum_n \tilde{q}_n e^{i\frac{n}{\lambda}t}. \quad (8.33)$$

Clearly,

$$\begin{aligned} |f(t) - \tilde{f}_{\lambda,1}(t)| &= \frac{1}{2\pi\lambda} \left| \sum_n (f(\frac{n}{\lambda}) - \tilde{q}_n) e^{i\frac{n}{\lambda}t} \right| \\ &= \frac{1}{2\pi\lambda} \left| \sum_n G(\frac{n}{\lambda})(v_n - v_{n-1}) e^{i\frac{n}{\lambda}t} \right| \\ &= \frac{1}{2\pi\lambda} \left| \sum_n v_n (G(\frac{n}{\lambda}) e^{i\frac{n}{\lambda}t} - G(\frac{n+1}{\lambda}) e^{i\frac{n+1}{\lambda}t}) \right| \\ &\leq \frac{1}{2\pi\lambda} \sum_n |G(\frac{n}{\lambda}) e^{i\frac{n}{\lambda}t} - G(\frac{n+1}{\lambda}) e^{i\frac{n+1}{\lambda}t}|; \end{aligned} \quad (8.34)$$

the second equality is the result of summation by parts (note that the boundary values vanish again since  $G(\frac{n}{\lambda})$  tends to zero as  $|n|$  approaches infinity); the final inequality follows because  $|v_n|$  is bounded by 1. Now set  $\Gamma(z, t) := G(z)e^{izt}$  and rewrite (8.34) as

$$\begin{aligned} |f(t) - \tilde{f}_{\lambda,1}(t)| &\leq \frac{1}{2\pi\lambda} \sum_n \left| \Gamma(\frac{n}{\lambda}, t) - \Gamma(\frac{n+1}{\lambda}, t) \right| \\ &\leq \frac{1}{2\pi\lambda} \sum_n \int_{\frac{n+1}{\lambda}}^{\frac{n}{\lambda}} |\partial_1 \Gamma(z, t)| dz \\ &\leq \frac{1}{2\pi\lambda} \|\partial_1 \Gamma(\cdot, t)\|_{L^1}. \end{aligned} \quad (8.35)$$

Note that the second inequality is due to the fact that  $\Gamma$  is smooth.

Finally we will calculate  $\|\partial_1 \Gamma(\cdot, t)\|_{L^1}$ . Clearly,

$$\partial_1 \Gamma(z, t) = G'(z)e^{izt} + itG(z)e^{izt}, \quad (8.36)$$

which yields

$$\|\partial_1 \Gamma(\cdot, t)\|_{L^1} \leq \frac{1}{2\pi\lambda} (\|G'\|_{L^1} + |t|\|G\|_{L^1}). \quad (8.37)$$

But since  $f$  is supported on  $[-\pi, \pi]$ , we can conclude that

$$\sup_t |f(t) - \tilde{f}_{\lambda,1}(t)| \leq \frac{C}{\lambda} \quad (8.38)$$

with  $C = \frac{1}{2\pi}\|G'\|_{L^1} + \frac{1}{2}\|G\|_{L^1}$ . Thus we proved

**Theorem 26.** *Let  $G$  be a fixed function such that  $G$  and  $G'$  are in  $L^1(\mathbb{R})$  and  $G(\xi)$  is  $O(1/|\xi|)$ . Suppose  $f$  be in  $\mathcal{B}_G$ . Then*

$$\sup_t |f(t) - \tilde{f}_{\lambda,1}(t)| \leq \frac{\frac{1}{2\pi}\|G'\|_{L^1} + \frac{1}{2}\|G\|_{L^1}}{\lambda}, \quad (8.39)$$

where  $\tilde{f}_{\lambda,1}$  is defined as in (8.33) with  $\tilde{q} = T_G(f(\frac{n}{\lambda}))$ .

**Remark:** It is straight-forward to define the higher-order versions of the above described scheme. The  $k^{\text{th}}$ -order scheme can be defined by replacing the first-order backward difference operator in (8.31) by a  $k^{\text{th}}$ -order backward difference operator, i.e. a  $k^{\text{th}}$ -order quantizer is defined by the following recursion relations:

$$\begin{aligned} (\Delta^k v)_n &= \frac{\hat{f}(\frac{n}{\lambda})}{G(\frac{n}{\lambda})} - q_n^\lambda \\ q_n^\lambda &= \text{sign}(M((\Delta^0 v)_{n-1}, \dots, (\Delta^{k-1} v)_{n-1} \frac{\hat{f}(\frac{n}{\lambda})}{G(\frac{n}{\lambda})})) \end{aligned} \quad (8.40)$$

where  $M$  is chosen such that  $v$  is uniformly bounded in  $l^\infty$ . The existence of such  $M$  due to [3] for arbitrary  $k$ .

In this case, the approximation error is  $O(\lambda^{-k})$ . The proof is similar to the proof of Theorem 20 and follows from Corollary 3.

## 8.2 Higher-order schemes

We will define the  $k^{th}$ -order TF $\Sigma\Delta$ -II scheme as follows: Let  $(\varphi, \tau_o, \xi_0)$  be a tight Weyl-Heisenberg frame with frame bound  $A$ . Suppose  $f$  is in  $\mathcal{B}_G^\varphi$ . Let  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$ , and denote by  $a_{n,m}$  and  $b_{n,m}$  the real and imaginary parts of  $c_{n,m}$  respectively. Define  $\tilde{a}_{n,m} := \frac{a_{n,m}}{G(n\tau_0, m\xi_0)}$  and  $\tilde{b}_{n,m} := \frac{b_{n,m}}{G(n\tau_0, m\xi_0)}$ , as before. Now consider the recursion relations

$$\begin{aligned} (\Delta_1^k u^R)_{n,m} &= \tilde{a}_{n,m} - p_{n,m}^R \\ p_{n,m}^R &= \text{sign}(M(\Delta_1^0 u_{n-1,m}^R, \dots, \Delta_1^{k-1} u_{n-1,m}^R, \tilde{a}_{n,m})) \end{aligned} \quad (8.41)$$

$$\begin{aligned} (\Delta_2^k v^R)_{n,m} &= \bar{u}_{n,m}^R - r_{n,m}^R \\ r_{n,m}^R &= \text{sign}(\Delta_2^0 v_{n,m-1}^R, \dots, \Delta_2^{k-1} v_{n,m-1}^R, \bar{u}_{n,m}^R) \end{aligned} \quad (8.42)$$

and

$$\begin{aligned} (\Delta_1^k u^I)_{n,m} &= \tilde{a}_{n,m} - p_{n,m}^I \\ p_{n,m}^I &= \text{sign}(M(\Delta_1^0 u_{n-1,m}^I, \dots, \Delta_1^{k-1} u_{n-1,m}^I, \tilde{b}_{n,m})) \end{aligned} \quad (8.43)$$

$$\begin{aligned} (\Delta_2^k v^I)_{n,m} &= \bar{u}_{n,m}^I - r_{n,m}^I \\ r_{n,m}^I &= \text{sign}(\Delta_2^0 v_{n,m-1}^I, \dots, \Delta_2^{k-1} v_{n,m-1}^I, \bar{u}_{n,m}^I), \end{aligned} \quad (8.44)$$

where  $\bar{u}^R := u^R/C_{k,M}$ ,  $\bar{u}^I := u^I/C_{k,M}$  and  $M$  is a function which guarantees that  $u^R$ ,  $v^R$ ,  $u^I$  and  $v^I$  are uniformly bounded in  $l^\infty$  by  $C_{k,M}$ . Note that the recursion relations (8.41), (8.42), (8.43) and (8.44) all correspond to  $k^{th}$ -order standard sigma-delta quantizers, as in (1.13), with  $a_{n,k}$ ,  $\bar{u}_{n,k}^R$ ,  $b_{n,k}$  and  $\bar{u}_{n,k}^I$  respectively as their input. Thus, since all these sequences are bounded in  $l^\infty$  by 1, such an  $M$  exists due to [3].

Now set  $v = v^R + iv^I$ ,  $p = p^R + ip^I$ , and  $r = r^R + ir^I$ . Note that

$$C_{k,M}G(n\tau_0, m\xi_0)(\Delta_1^k \Delta_2^k v)_{n,m} = c_{n,m} - G(n\tau_0, m\xi_0)(p_{n,m} + C_{k,M}\Delta_1^k r_{n,m}), \quad (8.45)$$

which suggests us the following definition of  $T_{TFII_k}$ :  $T_{TFII_k}$  will denote the mapping that maps  $c$  to  $\tilde{q}$  with  $\tilde{q}_{n,m} := G(n\tau_0, m\xi_0)(p_{n,m} + C_{k,M}\Delta_1^k r_{n,m})$ .

**Theorem 27.** *Let  $(\varphi, \tau_0, \xi_0)$  be a tight Weyl-Heisenberg frame with frame bound  $A$ . Suppose  $f$  is in  $\mathcal{B}_G^\varphi$ . Let  $c_{n,m} = \langle f, \varphi_{n,m} \rangle$  and put  $\tilde{q} = T_{TFII_k}(c)$ . Suppose  $G$  and  $\varphi$  are chosen such that for  $l, l'$  in  $\{0, \dots, k\}$ ,*

$$t^{l'} (|\varphi^{(l)}| \star G_{l,l'})(t) \in L^2(\mathbb{R}) \quad (8.46)$$

where we define  $G_{l,l'}$  by

$$G_{l,l'}(s) := \int \binom{k}{l'} |\partial_2^{(k-l')} \partial_1^{(k-l)} G(s, z)| dz. \quad (8.47)$$

In this case, the approximation error satisfies

$$\|f - \tilde{f}_{A,k}\|_{L^2} \leq \frac{C}{A^k} \quad (8.48)$$

where  $\tilde{f}_{A,k}$  is defined as before, i.e.  $\tilde{f}_{A,k} = \frac{1}{A} \sum_{n,m} \tilde{q}_{n,m} \varphi_{n,m}$ , and the constant  $C$  is given by

$$C := \sum_{l=0}^k \binom{k}{l} \sum_{l'=0}^k \|t^{l'} (|\varphi^{(l)}| \star G_{l,l'})\|_{L^2}. \quad (8.49)$$

**Proof:** Let us start by writing the error term:

$$\begin{aligned}
f(t) - \tilde{f}_{A,k}(t) &= \frac{1}{A} \sum_{n,m} (c_{n,m} - \tilde{q}_{nm}) \varphi_{n,m} \\
&= \frac{1}{A} \sum_{n,m} C_{k,M} G(n\tau_0, m\xi_0) (\Delta_1^k \Delta_2^k v)_{n,m} \varphi_{n,m} \\
&= \frac{C_{k,M}}{A} \sum_{n,m} v_{n,m} \bar{\Delta}_1^k \bar{\Delta}_2^k G(n\tau_0, m\xi_0) \varphi_{n,m}, \tag{8.50}
\end{aligned}$$

where the second equality is due to (8.45); the third equality is obtained by partial summation. Now set  $\Gamma(s, z, t) = G(s, z) \varphi(t-s) e^{izt}$ , and rewrite (8.50) as

$$f(t) - \tilde{f}_{A,k}(t) = \frac{C_{k,M}}{A} \sum_{n,m} v_{n,m} \bar{\Delta}_1^k \bar{\Delta}_2^k \Gamma(n\tau_0, m\xi_0, t). \tag{8.51}$$

By techniques identical to those used in the proof of Theorem 20 we have

$$|f(t) - \tilde{f}_{A,k}(t)| \leq \frac{C_{k,M} \|v\|_{l^\infty}}{A^k} \|\partial_2^k \partial_1^k \Gamma(\cdot, \cdot, t)\|_{L^1}. \tag{8.52}$$

We complete the proof by estimating  $\|\partial_2^k \partial_1^k \Gamma(\cdot, \cdot, t)\|_{L^1}$ . Note that

$$\partial_2^k \partial_1^k \Gamma(s, z, t) = \sum_{l=0}^k (-1)^l \binom{k}{l} \varphi^{(l)}(t-s) \sum_{l'=0}^k \binom{k}{l'} (it)^{l'} \partial_2^{(k-l')} \partial_1^{(k-l)} G(s, z), \tag{8.53}$$

which yields

$$|\partial_2^k \partial_1^k \Gamma(s, z, t)| \leq \sum_{l=0}^k \binom{k}{l} |\varphi^{(l)}(t-s)| \sum_{l'=0}^k \binom{k}{l'} |t|^{l'} |\partial_2^{(k-l')} \partial_1^{(k-l)} G(s, z)|. \tag{8.54}$$

By integrating both sides, we get

$$\|\partial_2^k \partial_1^k \Gamma(s, z, t)\|_{L^1(\mathbb{R}^2)} \leq \sum_{l=0}^k \sum_{l'=0}^k |t|^{l'} \binom{k}{l} \int \left( |\varphi^{(l)}(t-s)| \int \binom{k}{l'} |\partial_2^{(k-l')} \partial_1^{(k-l)} G(s, z)| dz \right) ds. \tag{8.55}$$

Finally by setting  $G_{l,l'}(s) := \int \binom{k}{l'} |\partial_2^{(k-l')} \partial_1^{(k-l)} G(s, z)| dz$ , we obtain

$$|f(t) - \tilde{f}_{A,k}(t)| \leq \frac{C_{k,M} \|v\|_{l^\infty}}{A^k} \sum_{l=0}^k \sum_{l'=0}^k \binom{k}{l} |t|^{l'} (|\varphi| \star G_{l,l'})(t). \quad (8.56)$$

which, by taking the  $L^2$ -norm of both sides, yields

$$\|f - \tilde{f}_{A,k}\|_{L^2} \leq \frac{C}{A^k} \quad (8.57)$$

where  $C = \frac{C_{k,M} \|v\|_{l^\infty}}{A^k} \sum_{l=0}^k \sum_{l'=0}^k \binom{k}{l} \| |t|^{l'} (|\varphi| \star G_{l,l'})(t) \|_{L^2}$ .  $\square$

### 8.3 Numerical Experiment

We consider again the function

$$f(t) = 0.4e^{-i0.9t^3} e^{-0.05t^2}. \quad (8.58)$$

Let  $\varphi(t) = \pi^{1/4} e^{-\frac{t^2}{2}}$ . Then, as discussed before in Section 7.3,  $(\varphi, \tau_0, \xi_0)$  is an almost tight Weyl-Heisenberg frame if  $\tau_0$  and  $\xi_0$  are equal and sufficiently small. We use the same algorithm we used in Section 7.3 to compute the frame coefficients  $\langle f, \varphi_{n,m} \rangle$ . For simplicity, we use  $G(\tau, \xi) = 2|\langle f, \varphi_{\tau,\xi} \rangle|$  and quantize the frame coefficients according to the algorithm given in (8.3) for the first-order scheme, and according to the algorithm given in (8.41)-(8.44) for the second-order scheme. We apply the algorithms to different frame expansions of  $f$  with different frame bounds. The frames we use are  $(\varphi, \tau_0, \xi_0)$  with  $\tau_0 = \xi_0$  ranging from 0.2 to 1.5. Figure 8.1 shows the  $L^2$ -approximation error estimates depending on the frame coefficient of the expansion we used to obtain the approximation using the first-order quantizer. The graphs of  $\tilde{f}_A$  which is obtained from the original function  $f$  via the first-order TF $\Sigma\Delta$ -II is given in Figure 8.2. In Figure 8.3 we show the  $L^2$ -approximation error estimates that are obtained using the

second-order TF $\Sigma\Delta$ -II. Finally, in Figure 8.4 we present the graphs of  $\tilde{f}_{A,2}$  that are obtained from  $f$  via the second-order TF $\Sigma\Delta$ -II.

Figure 8.1: The  $L^2$ -approximation error vs. the frame bound  $A$  for the first-order case. The line is the graph of  $\{(A, 6/A)\}$

Figure 8.2: The graph of  $\tilde{f}_A$ , the  $L^2$ -approximation of  $f$  obtained via the first-order TF $\Sigma\Delta$ -II algorithm, for three different values of the frame bound  $A$ , along with the graph of  $f$ . In each figure, the top graph is the real part of  $f$  and  $\tilde{f}_A$ —the solid graph belongs to  $f$  and the dashed to  $\tilde{f}_A$ ; the bottom graph is the imaginary part of  $f$  and  $\tilde{f}_A$ —the solid belongs to  $f$  and the dashed to  $\tilde{f}_A$ . The left-most figure is the graph of  $\tilde{f}_{6.29}$  along with  $f$ ; the middle figure is the graph of  $\tilde{f}_{25.16}$  and the right-most figure is the graph of  $\tilde{f}_{157.25}$ .

Figure 8.3: The  $L^2$ -approximation error vs. the frame bound  $A$  for the second order case. The line is the graph of  $\{(A, 6/A^2)\}$ .

Figure 8.4: The graph of  $\tilde{f}_{A,2}$ , the  $L^2$ -approximation of  $f$  obtained via the second-order TF $\Sigma\Delta$ -II algorithm, for three different values of the frame bound  $A$ , along with the graph of  $f$ . In each figure, the top graph is the real part of  $f$  and  $\tilde{f}_{A,2}$ —the solid graph belongs to  $f$  and the dashed to  $\tilde{f}_{A,2}$ ; the bottom graph is the imaginary part of  $f$  and  $\tilde{f}_{A,2}$ —the solid belongs to  $f$  and the dashed to  $\tilde{f}_{A,2}$ . The left-most figure is the graph of  $\tilde{f}_{6.29,2}$  along with  $f$ ; the middle figure is the graph of  $\tilde{f}_{25.16,2}$  and the right-most figure is the graph of  $\tilde{f}_{157.25,2}$ .

# Bibliography

- [1] Özgür Yılmaz, “Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions”, *Constructive Approximation*, submitted.
- [2] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.
- [3] I. Daubechies and R. DeVore, “Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order”, submitted.
- [4] S.R. Norsworthy, R.Schreier and G.C. Themes, eds, *Delta-Sigma Data Converters*, IEEE Press, 1997.
- [5] S.C. Pinault and P.V. Lopresti, “On the behavior of the double-loop sigma-delta modulator”, *IEEE Transactions on Circuits and Systems*, vol.40, Aug. 1993.
- [6] N. Thao, “Quadratic one-bit second-order sigma-delta modulators”, *IEEE Transactions on Circuits and Systems*, submitted.
- [7] S.J. Park and R.M. Gray, “Sigma-delta modulation with leaky integration and constant input”, *IEEE Transactions on Information Theory*, vol.38, Sep.1992.
- [8] O. Feely and L.O. Chua, “The effect of integrator leak in  $\Sigma$ - $\Delta$  Modulation”, *IEEE Transactions on Circuits and Systems*, vol.38, Nov.1991.
- [9] I. Daubechies *Ten Lectures on Wavelets*, SIAM, 1992.

- [10] S. Mallat *A Wavelet Tour of Signal Processing*, Academic Press, 1999.
- [11] H.G. Feichtinger, Thomas Strohmer, *Gabor Analysis and Algorithms*, Birkhäuser, 1998.
- [12] C.S. Güntürk, “Reconstructing a Bandlimited Function from Very Coarsely Quantized Data: Improving the Error Estimate for First Order Sigma-Delta Modulators”, in preparation.
- [13] C.S. Güntürk, Harmonic Analysis of Two Problems in Signal Quantization and Compression, Ph.D. thesis, Program in Applied and Computational Mathematics, Princeton University, November 2000.