

# Stability analysis for several second-order sigma-delta methods of coarse quantization of bandlimited functions <sup>1</sup>

Özgür Yılmaz

Program in Applied and Computational Mathematics  
Princeton University

## Abstract

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 positively 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.

## 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,  $f\left(\frac{n\pi}{\Omega}\right)$ :

$$f(t) = \sum_{n \in \mathbb{Z}} f\left(\frac{n\pi}{\Omega}\right) \frac{\sin(\Omega t - n\pi)}{\Omega t - n\pi}. \quad (1)$$

Of course, the reconstruction is only perfect 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

$$|f(t) - \tilde{f}(t)| \leq C\epsilon \log N, \quad (2)$$

where  $\tilde{f}(t)$  is calculated by replacing the sample values of  $f$  in (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  $f_n = f\left(\frac{n\pi}{\Omega}\right)$ , let us sample more frequently and use  $f_n^\lambda = f\left(\frac{n\pi}{\lambda\Omega}\right)$ ,  $\lambda > 1$ , to reconstruct. In this case, one can prove that

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f\left(\frac{n\pi}{\lambda\Omega}\right) g\left(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\right), \quad (3)$$

if  $g$  satisfies:

$$\hat{g}(\xi) = \begin{cases} \frac{1}{\sqrt{2\pi}} & |\xi| \leq \Omega \\ 0 & |\xi| \geq \lambda\Omega \end{cases} \quad (4)$$

$$\hat{g} \in C^\infty. \quad (5)$$

---

<sup>1</sup>AMS Classification: 41A99, 42A99, 93C55

Keywords and phrases: redundant representations, sigma-delta quantization, coarse quantization

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

$$|f(t) - \tilde{f}(t)| \leq \epsilon C_g \frac{\Omega}{\pi}, \quad (6)$$

where

$$C_g \leq \frac{\Omega}{\pi} (\|g\|_{L^1} + \frac{1}{\lambda} \|g'\|_{L^1}), \quad (7)$$

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 [1]. In this paper 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 extreme case where we 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 (3) we know that

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f \left( \frac{n}{\lambda} \right) g \left( t - \frac{n}{\lambda} \right). \quad (8)$$

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

$$\tilde{f}(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda g \left( t - \frac{n}{\lambda} \right) \quad (9)$$

is a “good” approximation of  $f$ .

### 1.3.1 First-order $\Sigma\Delta$ -quantization

First-order  $\Sigma\Delta$ -quantizer produces  $(q_n^\lambda)_{n \in \mathbb{Z}}$  via the following scheme:

$$\begin{aligned} v_n - v_{n-1} &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(v_{n-1} + f_n^\lambda), \end{aligned} \quad (10)$$

where  $v$  is an internal state variable, with  $v_0 \in (-1, 1)$ . In this case, one can show that [2]

- $|v_n| < 1$  for all  $n$ ,
- $|f(t) - \tilde{f}(t)| \leq \frac{1}{\lambda} \|g'\|_{L^1}$ .

### 1.3.2 Higher order $\Sigma\Delta$ -quantization schemes

Define  $\Delta_n^k(v) = \sum_{l=0}^k (-1)^l \binom{k}{l} v_{n-l}$ . Note that  $\Delta_n^0(v) = v_n$  and

$\Delta_n^1(v) = v_n - v_{n-1}$ . A  $k^{\text{th}}$  order  $\Sigma\Delta$ -quantization scheme is defined by the following system of difference equations:

$$\begin{aligned} \Delta_n^k(v) &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(F(\Delta_n^0(v), \dots, \Delta_n^{k-1}(v), f_n^\lambda)), \end{aligned} \quad (11)$$

where  $F$  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  $F$ , that  $(v_n)_{n \in \mathbb{Z}}$ , produced by (11), is a bounded sequence. Then, for all  $t \in \mathbb{R}$ ,*

$$\left| f(t) - \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| \leq \frac{1}{\lambda^k} \|v\|_{l^\infty} \|g^{(k)}\|_{L^1} \quad (12)$$

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 [2]. In this paper we will mostly discuss properties of second-order  $\Sigma\Delta$ -schemes, for different rules  $F$ , 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 are 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  $F$  used in the quantizer (11) is never known exactly; for instance, if  $F$  is a linear function, then all its coefficients will be specified within a certain tolerance. In addition, the quantizer itself is not entirely precise, leading to the replacement of  $\text{sign}(F)$  in (11) by  $\text{sign}(F + \epsilon)$ , where  $\epsilon$  is again known within a certain tolerance. 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 this paper, 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 mapped *into* itself by the dynamical system underlying the  $\Sigma\Delta$ -quantizer (in this case we say that the set  $R$  is *positively invariant* under the corresponding map); moreover, this  $R$  can be constructed so that it is itself robust under changes in  $F$  and the quantizer.

In Section 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.

## 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:

$$\begin{aligned}\Delta_n^2(v) &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(F(\Delta_n^1(v), \Delta_n^0(v), f_n^\lambda)).\end{aligned}\tag{13}$$

Let us put  $u_n = v_n - v_{n-1}$ . Then (13) becomes:

$$\begin{aligned}u_n - u_{n-1} &= f_n^\lambda - q_n^\lambda \\ v_n - v_{n-1} &= u_n \\ q_n^\lambda &= \text{sign}(F(u_{n-1}, v_{n-1}, f_n^\lambda)).\end{aligned}\tag{14}$$

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

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

Some examples from the literature are [3, 2, 6]:

- $F(u, v, x) = u + \gamma v$  with  $\gamma > 0$ ,
- $F(u, v, x) = u + x + M \text{sign}(v)$  with  $M > 0$ ,
- $F(u, v, x) = \frac{6x-7}{3} + (u + \frac{x+3}{2})^2 + 2(1-x)v$ .

Note that in either region,  $\Lambda_+(x)$  or  $\Lambda_-(x)$ , the system described in (14), is affine. Indeed, we can write:

$$\begin{pmatrix} u_n \\ v_n \end{pmatrix} = \begin{cases} S_l^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in \Gamma_+ \\ S_r^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in \Gamma_- \end{cases}\tag{15}$$

$$:= S(u_{n-1}, v_{n-1}, f_n^\lambda),\tag{16}$$

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  $(q_n^\lambda)_{n \in \mathbb{Z}}$ , the output 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

$$\left| \sum_{n=1}^N f_n^\lambda - \sum_{n=1}^N q_n^\lambda \right| \leq |u_N - u_0| < 2C, \quad (17)$$

where  $u_n = \Delta_n^{(k-1)}(v)$ , and  $C$  is a constant bounding  $\Delta_n^{(k-1)}(v)$  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.

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

**Proposition 1.** *Let  $(x_n)$  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. 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 pattern automatically arise when the input  $f(\frac{q}{\lambda})$  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 (11) by

$$\begin{aligned} \Delta_n^k(v) &= f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \begin{cases} 0; & \text{if } |F(\cdot)| \leq 0.5/\eta \\ \text{sign}(F(\cdot)); & \text{otherwise} \end{cases} \quad := m(\eta F(\cdot)). \end{aligned} \quad (18)$$

for some fixed  $\eta > 0$ . Indeed, for the first-order case, the tri-level quantization scheme described in (18) with  $\eta = 1$  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 (18) with  $k = 1$ ,  $\eta = 1$  and  $F(v, x) = v + x$ , 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  $q_n = m(v_{n-1}) = 0$ . 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  $F$  be chosen such that the system, described in (18) 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  $u_n$  is as defined just after (13) (regardless of the value of  $\eta$ ).*

**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,  $q_n = m(\eta F(u_{n-1}, v_{n-1}, 0)) = m(\eta F(u_0, v_0, 0)) = q_1 = 0$ .

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 (18) 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  $m(\eta F(u, v)) = 0$  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:

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda \\ q_n^\lambda &= \text{sign}(\beta_\lambda u_{n-1} + f_n^\lambda) \end{aligned} \quad (19)$$

The system defined in (19) is equivalent to the first-order  $\Sigma\Delta$ -quantization scheme given in (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  $u_{n-1}$ , after one time unit, we have  $\beta_\lambda u_{n-1}$  in memory. Physically one always encounters some leakage and this is usually considered to be a problem (or an imperfection)[7, 8].

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

$$\beta_\lambda = e^{-\frac{c}{\lambda}}, \quad (20)$$

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  $u_{n-1}$  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 (20); when the scheme is implemented digitally, as in D/A converters, we always have the freedom to choose  $\beta_\lambda$  as in (20).)

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 (4) and (5) with  $\Omega = \pi$ . Let the leakage factor be  $\beta_\lambda = e^{-\frac{c}{\lambda}}$ . Assume that the sequence  $(v_n)$  generated by (19) is bounded. Then if  $(q_n^\lambda)$  is the output of the first-order leaky  $\Sigma\Delta$ -quantizer given in (19), then*

$$|f(t) - \tilde{f}(t)| \leq \frac{\|v\|_{l^\infty}}{\lambda} (\|g'\|_{L^1} + cC_g), \quad (21)$$

where  $C_g$  is as in (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

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

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

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda \\ v_n &= \beta_\lambda v_{n-1} + u_n \\ q_n^\lambda &= \text{sign}(F(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})). \end{aligned} \quad (22)$$

**Theorem 3.** *Let  $f, g$  and  $\beta_\lambda$  be as in Theorem 2. Assume that  $(v_n)$ , generated by (22) is bounded. Then if  $(q_n^\lambda)$  is the output of the second-order leaky  $\Sigma\Delta$ -quantizer given in (22), then*

$$|f(t) - \tilde{f}(t)| \leq \frac{\|v\|_{l^\infty}}{\lambda^2} (\|g''\|_{L^1} + 2c\|g'\|_{L^1} + 2c^2C_g), \quad (23)$$

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

$$\begin{aligned} f(t) - \tilde{f}(t) &= \frac{1}{\lambda} \left( \sum \Delta_n^2(v) g(t - \frac{n}{\lambda}) + 2(1 - \beta_\lambda) \sum v_{n-1} g(t - \frac{n}{\lambda}) \right. \\ &\quad \left. - (1 - \beta_\lambda^2) \sum v_{n-2} g(t - \frac{n}{\lambda}) \right) \\ &= \frac{1}{\lambda} \left( \sum \Delta_n^2(v) g(t - \frac{n}{\lambda}) + 2(1 - \beta_\lambda) \sum (v_{n-1} - v_{n-2}) g(t - \frac{n}{\lambda}) \right. \\ &\quad \left. + (1 - \beta_\lambda)^2 \sum v_{n-2} g(t - \frac{n}{\lambda}) \right). \end{aligned} \quad (24)$$

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

$$|f(t) - \tilde{f}(t)| \leq \frac{\|v\|_{L^\infty}}{\lambda} \left( \frac{1}{\lambda} \|g''\|_{L^1} + \frac{2c}{\lambda} \|g'\|_{L^1} + \frac{2c^2}{\lambda} C_g \right). \quad (25)$$

□

We define the tri-level finite memory second-order  $\Sigma\Delta$ -quantizer by replacing *sign* in (22) by  $m$ , as in (18), i.e.

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda \\ v_n &= \beta_\lambda v_{n-1} + u_n \\ q_n^\lambda &= m(\eta F(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})). \end{aligned} \quad (26)$$

where  $\eta > 0$  is fixed, and  $F$  and  $\eta$  are chosen in such a way that the sequences  $u$  and  $v$  stay bounded. Note that Theorem 3 still holds as long as 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 (26) with  $F(u, v) = u + \gamma v$ . Let the input sequence  $(x_n)$  be identically equal to zero for all  $n \geq N$  for some  $N$ ,  $q_N = 0$  and*

$$|u_{N-1}| < \frac{(1 - \beta_\lambda)}{2\eta\gamma\beta_\lambda^2}. \quad (27)$$

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

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

$$|\beta_\lambda(u_{N-1} + \gamma v_{N-1})| \leq 0.5/\eta, \quad (28)$$

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

$$\begin{aligned} u_N &= \beta_\lambda u_{N-1} \\ v_N &= \beta_\lambda(v_{N-1} + u_{N-1}) \end{aligned} \quad (29)$$

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

$$q_{N+1} = m(\eta(\beta_\lambda(\beta_\lambda(u_{N-1} + \gamma v_{N-1}) + \gamma\beta_\lambda u_{N-1}))). \quad (30)$$

But by (27) and (28) we have

$$\begin{aligned}
& |\beta_\lambda(\beta_\lambda(u_{N-1} + \gamma v_{N-1}) + \gamma\beta_\lambda u_{N-1})| \\
& \leq \beta_\lambda |\beta_\lambda(u_{N-1} + \gamma v_{N-1})| + \gamma\beta_\lambda^2 |u_{N-1}| \\
& \leq 0.5/\eta,
\end{aligned} \tag{31}$$

which implies that  $q_{N+1} = 0$ . Since  $0 < \beta_\lambda < 1$ , (30) implies  $|u_N| < |u_{N-1}|$  and by induction we are done.  $\square$

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

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

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 positively 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  $l \leq L$  if  $(u_{n-1}, v_{n-1}) \in \Lambda$  and  $x_{n+l} = 0$  for all  $l \leq L$ . 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 set  $R$  that have periodic orbits outside  $\Lambda$ . For example, take  $\eta = 1$ ,  $\gamma = 0.2$ ,  $\beta_\lambda = 0.9$  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  $S_{LT}^2(u, v, 0) = (u, v)$ . Because  $m(\beta_\lambda(u + \gamma v)) = 1$  and  $m(\beta_\lambda(-u - \gamma v)) = -1$ , 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  $F$  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  $F$  used in the quantizer. Then we will extend this to “non-standard” schemes of interest to us, using the same  $F$ . In particular, we will consider the non-standard schemes with

- a tri-level quantizer, i.e. the scheme described in (18) with  $k = 2$ ,
- a finite memory quantizer, i.e. the scheme described in (22),
- a finite memory tri-level quantizer, i.e. the scheme described in (26).

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

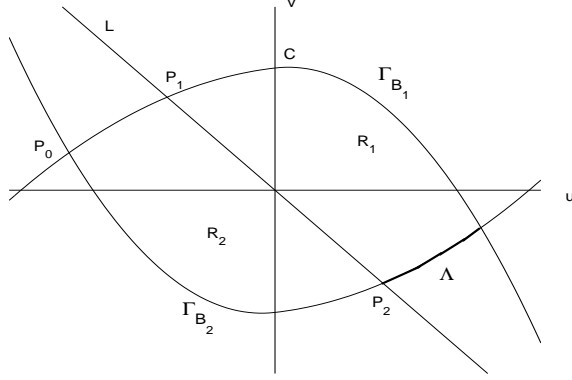


Figure 1:  $\Gamma_{B_1}$  and  $\Gamma_{B_2}$  are the graphs of the functions  $B_1$  and  $B_2$ , respectively.  $L$  is the line consisting of the points  $(u, v)$  for which  $F(u, v) = 0$ .

### 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 (14), with the quantization rule  $F(u, v, x) = u + \gamma v$  for a range of  $\gamma$  to be specified later. Since any  $F$  of this form does not depend on  $x$ , we will drop  $x$  from its argument, i.e.  $F = F(u, v)$ .

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 (14) can be rewritten as

$$\begin{aligned} (u_n, v_n) &= \begin{cases} S_t^{\delta_n}(u_{n-1}, v_{n-1}) = (u_{n-1} - \delta_n, u_{n-1} + v_{n-1} - \delta_n); & \text{if } q_n = 1 \\ S_r^{\delta_n}(u_{n-1}, v_{n-1}) = (u_{n-1} + \delta_n, u_{n-1} + v_{n-1} + \delta_n); & \text{if } q_n = -1 \end{cases} \\ q_n &= \text{sign}(F(u_{n-1}, v_{n-1})), \end{aligned} \quad (32)$$

In this case we will also write

$$(u_n, v_n) = S(u_{n-1}, v_{n-1}, \delta_n). \quad (33)$$

Let us define now the functions

$$B_1(u) = \begin{cases} -\frac{1}{2\delta_-}(u - \frac{\delta_-}{2})^2 + \frac{\delta_-}{8} + C; & \text{if } u \geq 0 \\ -\frac{1}{2\delta_+}(u - \frac{\delta_+}{2})^2 + \frac{\delta_+}{8} + C; & \text{if } u < 0 \end{cases}, \quad (34)$$

$$B_2(u) = \begin{cases} \frac{1}{2\delta_+}(u + \frac{\delta_+}{2})^2 - \frac{\delta_+}{8} - C; & \text{if } u \geq 0 \\ \frac{1}{2\delta_-}(u + \frac{\delta_-}{2})^2 - \frac{\delta_-}{8} - C; & \text{if } u < 0 \end{cases}, \quad (35)$$

where the constant  $C$  will have to be determined below. Note that the graphs of  $B_1$  and  $B_2$  are symmetric about the origin, i.e.  $B_2(u) = -B_1(-u)$ .

Figure 1 illustrates the graphs  $\Gamma_{B_1}$ , respectively  $\Gamma_{B_2}$ , of the function  $B_1$ , respectively  $B_2$ , 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  $F(u, v) = 0$ ), 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,

$$\begin{aligned} R_1 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), v \geq l(u)\} \\ R_2 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), v \leq l(u)\} \\ R &= R_1 \cup R_2. \end{aligned} \tag{36}$$

Note that any line  $L$  with a  $v$ -axis intercept between  $-C$  and  $C$  intersects  $\Gamma_{B_i}$ ,  $i = 1, 2$ , at two points. We will define  $(L \cap \Gamma_{B_i})_<$  to be the intersection point of  $L$  and  $\Gamma_{B_i}$  with the *smallest* first coordinate; similarly  $(L \cap \Gamma_{B_i})_>$  will refer to the intersection point of  $L$  and  $\Gamma_{B_i}$  with the *largest* 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 the graphs of  $B_1$  and  $B_2$  by  $P_0 = (u_0, v_0)$ ,  $(L \cap \Gamma_{B_1})_<$  by  $P_1 = (u_1, v_1)$ , and  $(L \cap \Gamma_{B_2})_>$  by  $P_2 = (u_2, v_2)$ , where  $L$  is the graph of the line consisting of points  $(u, v)$  for which  $F(u, v) = 0$ . Note that  $P_0 = (u_0, v_0)$ , shown in Figure 1, is given by

$$\begin{aligned} u_0 &= -[2C(1-a^2)]^{1/2}, \\ v_0 &= B_1(u_0), \end{aligned} \tag{37}$$

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

$$\delta_+ - \delta \leq B_1(u - \delta) - B_1(u - \delta_+). \tag{38}$$

But,

$$B_1(u - \delta) - B_1(u - \delta_+) = \frac{-1}{2\delta_+}(\delta_+ - \delta)(2u - \delta - 2\delta_+). \tag{39}$$

Then (38) 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

$$-(\delta - \delta_-) \leq B_1(u - \delta) - B_1(u - \delta_-), \quad (40)$$

which reduces to showing that

$$B_1(u - \delta) - B_1(u - \delta_-) = \frac{1}{2\delta_-}(\delta_- - \delta)(2u - \delta - 2\delta_-) \geq -(\delta - \delta_-). \quad (41)$$

But (41) 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,

$$B_1(u) = -\frac{1}{2\delta_-}(u - \frac{\delta_-}{2})^2 + \frac{\delta_-}{8} + C, \quad (42)$$

and

$$B_1(u - \delta) = -\frac{1}{2\delta_+}(u - \frac{\delta_+}{2})^2 + \frac{\delta_+}{8} + C. \quad (43)$$

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

$$\frac{1}{2}(\frac{1}{\delta_+} - \frac{1}{\delta_-}) + u(1 - \frac{\delta}{\delta_+}) + \frac{\delta}{2}(\frac{\delta}{\delta_+} - 1) \leq 0. \quad (44)$$

But the left hand side equals

$$(1 - \frac{\delta}{\delta_+})(-\frac{1}{2\delta}u^2 + u - \frac{\delta}{2}) + \frac{1}{2}u^2(\frac{1}{\delta} - \frac{1}{\delta_-}). \quad (45)$$

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 (44). □

**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  $F(u, v) = u + \gamma v$  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  $F(u, v) = u + \gamma v = 0$ , and  $\Gamma_{B_1}$  as shown in Figure 1, i.e.  $P_1 = (L \cap \Gamma_{B_1})_{<}$ . Suppose*

$$u_0 + \delta_+ \leq u_1 \leq -\delta_+. \quad (46)$$

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

**Proof:** By Lemma 1, we know that  $S_i^\delta(R_1)$  lies under  $\Gamma_{B_1}$ . Therefore, we need to prove only that  $S_i^\delta(R_1)$  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 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  $S_i^\delta$  is affine 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  $S_i^\delta$  is affine 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,  $S_i^{\delta_+}$  maps  $P_1$  to a point on  $\Gamma_{B_1}$  by construction. Moreover,  $u(S_i^{\delta_+}(P_1)) = u_1 - \delta_+ \geq u_0$ . Therefore,  $S_i^{\delta_+}(P_1)$  is above  $\Gamma_{B_2}$ .  
 $S_i^{\delta_-}(P_1)$  is on the line through  $S_i^{\delta_+}(P_1)$  with slope 1 in the increasing direction. Since  $B_2(u) \leq 1$  for  $u \leq 0$ , and  $u(S_i^{\delta_-}(P_1)) \leq 0$ , it follows that  $S_i^{\delta_-}(P_1)$  is above  $\Gamma_{B_2}$ .
2. We know that  $u_2 = u(P_2) = -u_1$ . Then,  $u(S_i^{\delta_+}(P_2)) = -u_1 - \delta_+ \geq 0$  by our condition on  $u_1$ . But  $B_2$  is increasing in  $u$  for  $u \geq 0$ , thus  $B_2(u(S_i^{\delta_+}(P_2))) < B_2(u_2)$ . We also know that  $v(S_i^{\delta_+}(P_2)) > v_2 = v(P_2)$ . Hence we have that  $S_i^{\delta_+}(P_2)$  is above  $\Gamma_{B_2}$ . Because  $0 < \delta_- < \delta_+$ , we also conclude that  $u(S_i^{\delta_-}(P_2)) \geq 0$ , and hence  $S_i^{\delta_-}(P_2)$  is above  $\Gamma_{B_2}$ .
3. Finally, we want to show that  $S_i^\delta(\Lambda)$  lies above  $\Gamma_{B_2}$ . But by our condition this is obvious:  $u(S_i^\delta(P))$  will be positive for any  $P$  on  $\Lambda$  for  $\delta \in [\delta_-, \delta_+]$ . Therefore, since  $v(S_i^\delta(P)) \geq v(P)$  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  $S_i^\delta(\Lambda)$  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

$$C \geq 2 \frac{1 + \alpha}{1 - \alpha}. \quad (47)$$

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

$$\frac{1}{\gamma} \geq \frac{[2C(1 - a^2)]^{1/2} + 2\alpha C}{2\{[2C(1 - a^2)]^{1/2} - (1 + \alpha)\}}, \quad (48)$$

and

$$\frac{1}{\gamma} \leq \frac{C - (1 + \alpha)}{1 + \alpha}. \quad (49)$$

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

3. 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 (33) with the rule  $F(u, v) = u + \gamma v$ . Suppose  $C$  and  $\gamma$  satisfy (47), (48) and (49) 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 (13) with  $F(u, v) = u + \gamma v$ . If  $C$  and  $\gamma$  satisfy (47), (48) and (49),  $(u_n, v_n) \in R$  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 positively invariant region  $R$  and the trapping region  $R_c$  described by Pinault and Lopresti in [4]. One difference is that we did not impose any conditions on the input sequence  $(x_n)$ , except that it remains bounded,  $|x_n| < \alpha$ . 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 positively invariant set  $R$ , because it is completely determined by  $C$ ,  $\gamma$  and  $\alpha$ , the upper bound on the input sequence. As long as  $\|f\|_{L^\infty} < \alpha$ , and  $C$  and  $\gamma$  satisfy (47), (48), and (49), the set  $R$ , defined by (36), is positively invariant for any input sequence  $(f(\frac{n}{\lambda}))$ , so that unlike the trapping region  $R_c$  in [4], our positively invariant region  $R$  stays fixed when we change the oversampling ratio.

The idea of finding a positively invariant set to prove stability of second (and higher) order sigma-delta quantizers is also used by Schreier et al. in [5]; their construction however is numerical and applies to a more restricted class of inputs.

**Remark:**

Clearly, for any  $(u_0, v_0) \in \mathbb{R}^2$ , there exists a  $C$  such that  $(u_0, v_0) \in R_C$  (note that by construction the set  $R$  depends on  $C$ ); thus Corollary 1 holds for any  $\gamma$  satisfying (48) and (49) with this  $C$ . However, for a given quantizer with a fixed quantization rule, i.e. fixed  $\gamma$ , there will be an upper bound,  $C_{max}$ , on the permissible values of  $C$  (imposed by (48) after fixing  $\gamma$ ) if  $R_C$  will be positively invariant. Thus, if we choose  $(u_0, v_0) \notin R_{C_{max}}$  there will be no guarantee that  $(u_n, v_n)$  will be trapped in a bounded region.

Figure 2 illustrates an example where for  $\gamma = 0.2$  the initial condition is set to  $(u_0, v_0) = (20, 20)$ . For  $\gamma = 0.2$ ,  $C_{max} = 88.9711$  and  $(20, 20)$  lies outside of  $R_{C_{max}}$ . This example shows that the sets  $R_C$  with  $C \leq C_{max}$  need not to be attractors for the full  $(u, v)$ -plane; for  $F(u, v) = u + 0.2v$  none of them are.

### 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 (12), for all  $\gamma$  within the allowed range specified in (48) and (49). Moreover, our analysis still holds even if  $\gamma$  does not remain fixed, but varies with  $n$ , i.e. if in (32) we replace  $\gamma$  in  $F$  by  $\gamma_n$  during the iteration, where  $\gamma_n$  all satisfy (48) and (49), 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 positively invariant as long as  $\gamma_n$  at each step satisfies (48) and (49), leading to the same uniform bound on  $|v_n|$ .

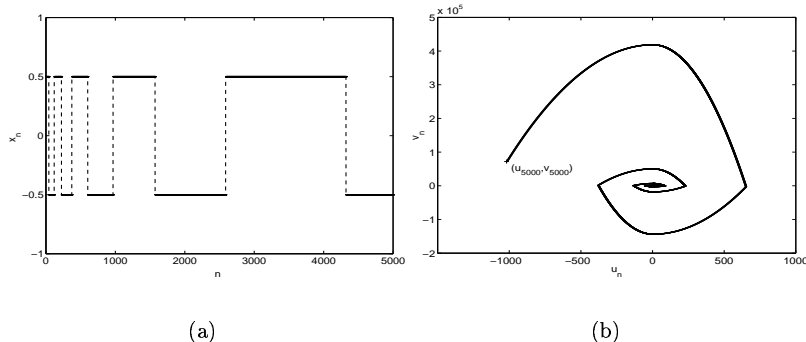


Figure 2: The input sequence  $x_n$  shown in Figure 2a produces the  $(u_n, v_n)$  shown in Figure 2b via the recursion described in (13) with the initial condition  $(u_0, v_0) = (20, 20)$ . The quantization rule is  $F(u, v) = u + 0.2v$ .

In this section we will show that the second-order  $\Sigma\Delta$ -scheme in (32) with  $F(u, v) = u + \gamma v$  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  $F$ , 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 positively 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 (32) with  $F(u, v) = u + \gamma v$ , with the positively invariant set  $R$  corresponding to some  $C$  and  $\gamma$ , where  $C$  satisfies (47) with strict inequality, and  $\gamma$  satisfies (48) and (49). In this case, we will show that there exists  $\epsilon_0 > 0$  such that  $R$  is positively invariant also with respect to the second-order  $\Sigma\Delta$ -scheme with  $F^\epsilon(u, v) = u + \gamma v + \epsilon$ , as long as  $|\epsilon| < \epsilon_0$ .

**Proposition 5.** *Let  $F^\epsilon(u, v) = u + \gamma v + \epsilon$  with  $|\epsilon| < \gamma C$  be the quantization rule used in the second-order  $\Sigma\Delta$ -scheme, described in (32). Let  $u_0$  be as in (38). Let  $L^\epsilon$  be the line consisting of points  $(u, v)$  that satisfy  $F^\epsilon(u, v) = 0$ , 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 (36), is positively invariant if both*

$$u_0 + 1 + \alpha < u(P_1^\epsilon) < -(1 + \alpha), \quad (50)$$

$$1 + \alpha < u(P_2^\epsilon) < -u_0 - (1 + \alpha), \quad (51)$$

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  $S_l^\delta(\tilde{R}_1) \subset R$ , for any  $\delta \in [\delta_-, \delta_+]$ . Note that  $\tilde{R}_1$  is convex, and  $S_l^\delta$  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  $S_l^\delta(P_1^\epsilon)$ ,  $S_l^\delta(P_2^\epsilon)$ , and  $S_l^\delta(\Lambda(P_2^\epsilon))$ , where we define

$$\Lambda(P) = \{(u, v) : v = B_2(u), u(P) < u < -u_0\}, \quad (52)$$

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

Suppose that  $P_1^\epsilon$  and  $P_2^\epsilon$  satisfy (50) and (51), respectively. Then, clearly, both  $P_1^\epsilon$  and  $-P_2^\epsilon$  satisfy (46); by Theorem 5  $S_l^{\delta_+}(P_i^\epsilon)$ , and  $S_l^{\delta_-}(P_i^\epsilon)$ ,  $i = 1, 2$ , are in  $R$ . Moreover, again by Theorem 5, if any point  $P$  satisfies (46), then both  $S_l^{\delta_+}$  and  $S_l^{\delta_-}$  map  $\Lambda(-P)$  into  $R$ . Since  $-P_2^\epsilon$  satisfies (46), we conclude that both  $S_l^{\delta_+}(\Lambda(P_2^\epsilon))$  and  $S_l^{\delta_-}(\Lambda(P_2^\epsilon))$  are in  $R$ .

This shows that  $S_l^\delta(\tilde{R}_1) \subset R$  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 (50), and  $-P_1^\epsilon$  satisfies (51)), we conclude that  $S_l^\delta(R \setminus \tilde{R}_1) \subset R$ , and hence  $S_l^\delta(R) \subseteq R$  for any  $\delta \in [\delta_-, \delta_+]$ .  $\square$

We now investigate under what condition on  $\epsilon$ , for fixed  $\gamma$  and  $C$ , (50) and (51) will be satisfied.

**Theorem 6.** *Let  $F(u, v) = u + \gamma v$  be a given quantization rule with  $\gamma$  satisfying (48) and (49) with strict inequalities for some  $C > 2\frac{1+\alpha}{1-\alpha}$ . Let  $u_0$  be as in (38) and let  $(u_1, v_1)$  be the intersection of  $L$  with  $\Gamma_{B_1}$ , as in Theorem 4. Take  $\epsilon$  such that*

$$|\epsilon| < \min\{u_1(\alpha, \gamma, C) - u_0(\alpha, C) - (1 + \alpha); -u_1(\alpha, \gamma, C) - (1 + \alpha)\}, \quad (53)$$

with

$$\begin{aligned}u_0(\alpha, C) &= -[2C(1 - a^2)]^{1/2}, \quad \text{as in (38),} \\ u_1(\alpha, \gamma, C) &= (1 + \alpha) \left( \frac{\gamma+2}{2\gamma} - \left[ \left( \frac{\gamma+2}{2\gamma} \right)^2 + \frac{2C}{1+\alpha} \right]^{1/2} \right).\end{aligned} \quad (54)$$

Then the second-order  $\Sigma\Delta$ -scheme with the quantization rule  $F^\epsilon(u, v) = u + \gamma v + \epsilon$  is stable with the positively invariant set  $R$ , where  $R$  is as in (36).

**Proof:** If  $C > 2\frac{1+\alpha}{1-\alpha}$  and if  $\gamma$  satisfies (48) and (49) with strict inequalities, both  $u_1 - u_0 - (1 + \alpha)$ , and  $-u_1 - (1 + \alpha)$  are positive; therefore (53) makes sense.

Note that (53) can be rewritten as

$$|\epsilon| + u_0 + \delta_+ < u_1 < -|\epsilon| - \delta_+. \quad (55)$$

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

$$|\epsilon| + \delta_+ < u_2 < -|\epsilon| - u_0 - \delta_+. \quad (56)$$

One checks by explicit calculation that  $u_1$  is as in (54). Let us denote the line consisting of the points  $(u, v)$  such that  $F^\epsilon(u, v) = 0$  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 (53) 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 (50) and (51), respectively, if they satisfy (55) and (56), respectively, since then by Proposition 5 we will be done.

Assume (55) and (56) are true. Then clearly we know that  $u_1 < 0$  and  $u_2 > 0$ . Also,

$$|u(P_i^\epsilon) - u_i| \leq |\epsilon|, \quad (57)$$

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

$$u_i - |\epsilon| < u(P_i^\epsilon) < u_i + |\epsilon|, \quad i = 1, 2. \quad (58)$$

The combination of (58) and (55) implies that  $P_1^\epsilon$  satisfies (50); similarly combining (58) and (56) we have that  $P_2^\epsilon$  satisfies (51). Hence we conclude that  $R$  is positively invariant under the second-order  $\Sigma\Delta$ -scheme with the rule  $F^\epsilon$ .  $\square$

#### 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. Because the positively invariant set is independent of the exact values of  $\epsilon$  and  $\gamma$  within certain tolerances - given by (53) for  $\epsilon$  and by (48) and (49) for  $\gamma$  -, the scheme can be safely used in practical A/D conversion where changes in temperature and other external factors can cause drifts in both  $\epsilon$  and  $\gamma$ , so that the  $n$ -th step in (32) would use  $F_n(u, v) = u + \gamma_n v + \epsilon_n$  rather than  $F(u, v) = u + \gamma v$ .
2. Theorem 6 also implies that the second-order sigma-delta quantizer with the rule  $F(u, v, x) = u + \gamma v + \mu x$  is also stable with the positively invariant set  $R$  if for the input sequence  $x = (x_n)_{n \in \mathbb{Z}}$   $\mu \|x\|_{l^\infty}$  satisfies (53), i.e.

$$\mu \|x\|_{l^\infty} < \min\{u_1(\alpha, \gamma, C) - u_0(\alpha, C) - (1 + \alpha); -u_1(\alpha, \gamma, C) - (1 + \alpha)\}. \quad (59)$$

3. We do not have to partition the plane by a line. Define  $L_r$  to be the line 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  $\tilde{F}$  as long as the set of points  $(u, v) \in R$  such that  $\tilde{F}(u, v) = 0$  constitutes a continuous curve that stays between the line segments  $L_r$  and  $L_l$ . An example is illustrated in Figure 3.
4. By above remark we observe that the set  $R$  corresponding to a sufficiently large  $C$  is also positively invariant under the second-order  $\Sigma\Delta$ -quantization scheme introduced by [2].

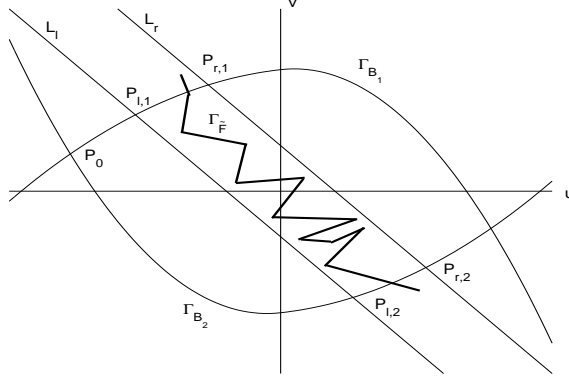


Figure 3:  $P_{r,i}$  and  $P_{l,i}$  are the points defined in the second remark above.  $\Gamma_{\tilde{F}}$  the curve consisting of the points  $(u, v)$  for which  $\tilde{F}(u, v) = 0$ ;  $\tilde{F}$  is such that the conditions described in the second remark above are satisfied.

## 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 (18) with  $k = 2$ , i.e.

$$\begin{aligned} u_n - u_{n-1} &= f_n^\lambda - q_n^\lambda \\ v_n - v_{n-1} &= u_n \\ q_n^\lambda &= \begin{cases} 1; & \text{if } F(u_{n-1}, v_{n-1}, f_n^\lambda) > 0.5/\eta \\ 0; & \text{if } |F(u_{n-1}, v_{n-1}, f_n^\lambda)| \leq 0.5/\eta \\ -1; & \text{if } F(u_{n-1}, v_{n-1}, f_n^\lambda) < -0.5/\eta \end{cases}, \end{aligned} \quad (60)$$

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

$$F(u, v) = u + \gamma v \quad (61)$$

for some range of  $\gamma$  and a fixed positive  $\eta$  whose range will be specified later. We will prove that, under some additional constraints this system is stable with the same positively invariant set  $R$  as in Theorem 5. Let  $L : F(u, v) = 0$ ,  $L_1 : F(u, v) = 0.5/\eta$  and  $L_2 : F(u, v) = -0.5/\eta$  be the lines whose graphs are shown in Figure 4. Define

$$\begin{aligned} R'_1 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), F(u, v) > 0.5/\eta\} \\ R'_2 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), F(u, v) < -0.5/\eta\} \\ R'_0 &= \{(u, v) : v \leq B_1(u), v \geq B_2(u), |F(u, v)| \leq 0.5/\eta\}, \end{aligned} \quad (62)$$

such that

$$R'_0 \cup R'_1 \cup R'_2 = R, \quad (63)$$

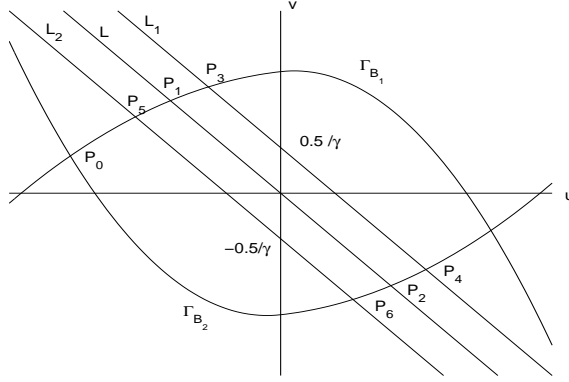


Figure 4:  $\Gamma_{B_1}$  and  $\Gamma_{B_2}$  are the graphs of  $B_1$  and  $B_2$ .  $L$ ,  $L_1$  and  $L_2$  are lines consisting of the points  $(u, v)$  such that  $F(u, v) = 0$ ,  $F(u, v) = 0.5/\eta$  and  $F(u, v) = -0.5/\eta$ , respectively.

where  $R$  is identical to the positively invariant set in Theorem 5. Note that, for sufficiently large  $\eta$  the sets  $R'_0$ ,  $R'_1$  and  $R'_2$  are all non-empty. Note that the scheme described in (60) is equivalent to

$$(u_n, v_n) = \begin{cases} S_1^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in R'_1 \\ S_r^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in R'_2 \\ S_0^{f_n^\lambda}(u_{n-1}, v_{n-1}); & \text{if } (u_{n-1}, v_{n-1}) \in R'_0 \end{cases}, \quad (64)$$

$$:= S_T(u_{n-1}, v_{n-1}, f_n^\lambda), \quad (65)$$

where

$$S_0^\delta : (u, v) \rightarrow (u + \delta, u + v + \delta). \quad (66)$$

To prove stability, we will show that  $S_0^\delta$  maps  $R'_0$  into  $R$ . We already know from Theorem 5 that  $S_1^\delta(R'_1) \subset S_1^\delta(R_1) \subset R$  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 therefore need to show only that  $S_0^\delta$  maps  $R'_0$  into  $R$  to conclude that the tri-level quantizer described in (60) 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  $F(u, v) = 0$ . Suppose  $C$ ,  $\alpha < 1$  and  $\eta > 0$  are such that*

$$u_0 + \delta_+ < u_1 < -\delta, \quad (67)$$

with  $\delta = 1 + \delta_-/2 + 0.5/\eta$  and  $\delta_- = 1 - \alpha$ , for some  $\gamma$  satisfying (48) and (49). Then  $S_0^\delta(R'_0) \subset R$ , and the system defined in (60) is stable with the positively invariant set  $R$ , where  $R$  is as in (36).

**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

$$\begin{aligned}\Lambda_1 &= \{(u, v) : v = B_1(u), u_5 < u < u_3\} \\ \Lambda_1 &= \{(u, v) : v = B_1(u), u_5 < u < u_3\}\end{aligned}\quad (68)$$

Clearly,  $u_3 \leq u_1 + 0.5/\eta < -1 - \delta_-/2$  by construction. This implies that  $P'_3 = (u_3 + 1, v_3 + 1)$  is in  $R$  (It is above  $B_2$  because  $u'_3 < -\delta_-/2$ ,  $v'_3 > v_3$ , and  $B_2(u)$  is decreasing for  $u < -\delta_-/2$ ; 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 (67) makes sense only if  $u_0 \leq -\delta_+ - \delta$ , which is equivalent to the condition

$$C \geq \frac{(5 + \alpha + \eta^{-1})^2}{8(1 - \alpha^2)}.\quad (69)$$

Note that (69) also specifies a range for  $\eta$ , i.e.

$$\eta \geq \frac{1}{2\sqrt{2C(1 - \alpha^2)} - 5 - \alpha}\quad (70)$$

for a given  $C$  which satisfies  $2\sqrt{2C(1 - \alpha^2)} - 5 - \alpha > 0$ .

2. For  $C$  satisfying (69) the set  $R$ , as in (36), is positively invariant for the tri-level scheme if

$$\frac{1}{\gamma} \geq \frac{B_1(u_0 + \delta_+)}{|u_0 + \delta_+|},\quad (71)$$

which is the same as (48), and

$$\frac{1}{\gamma} \leq \frac{B_1(-\delta)}{\delta},\quad (72)$$

with  $\delta$  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 (64) as follows.

$$(u_n, v_n) = \begin{cases} S_l^{\lambda}(u_{n-1}, v_{n-1}); & \text{if } u_{n-1} + \gamma v_{n-1} > 0.5/\eta \\ S_r^{\lambda}(u_{n-1}, v_{n-1}); & \text{if } u_{n-1} + \gamma v_{n-1} < -0.5/\eta \\ S_0^{\lambda}(u_{n-1}, v_{n-1}); & \text{if } |u_{n-1} + \gamma v_{n-1}| < 0.5/\eta \end{cases}, \quad (73)$$

Like the case with the standard scheme, the positively invariant set  $R$  of the tri-level scheme remains fixed for all  $\gamma$  that satisfy (71) and (72) by Theorem 7.

Now let us replace  $\gamma$  in (73) by  $\gamma_n$ , i.e. at every step of the iteration  $\gamma$  changes. Theorem 7 shows that as long as  $C$ , in definition of the functions  $B_1$  and  $B_2$ , satisfies (69),  $\gamma_n$  satisfies (71) and (72) for all  $n$ , the set  $R$ , as in (36), is positively invariant under the scheme in (73).

The tri-level scheme is also robust with respect to small shifts of the offset of the line defined by  $L = \{(u, v) : F(u, v) = 0\}$ , i.e. there exists  $\epsilon_0 > 0$  such that the scheme obtained by replacing  $F$  in (60) by  $F^\epsilon(u, v) = u + \gamma v + \epsilon$  is stable with the same positively 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 (67). Let us replace the rule  $F$  in Theorem 7 by  $F^\epsilon(u, v) = u + \gamma v + \epsilon$  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 (67),  $R$ , as in (36), will be positively invariant for the tri-level quantizer with the quantization rule  $F^\epsilon$ . 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 (67). Hence we have proven:

**Theorem 8.** *Let  $F(u, v) = u + \gamma v$  be a given quantization rule with  $\gamma$  satisfying (71) and (72) for some  $C$  satisfying (69). Let  $u_0$  be as in (38) and let  $P_1 = (u_1, v_1)$  be  $(L \cap \Gamma_{B_1})_<$ , as in Theorem 4. Take  $\epsilon$  such that*

$$|\epsilon| < \min\{u_1(\alpha, \gamma, C) - u_0(\alpha, C) - (1 + \alpha); -u_1(\alpha, \gamma, C) - \delta\}, \quad (74)$$

*with  $\delta$  as in Theorem 7,  $u_0(\alpha, C)$  and  $u_1(\alpha, \gamma, C)$  as in (54). Then the tri-level second-order  $\Sigma\Delta$ -scheme obtained by replacing  $F$  in (60) by  $F^\epsilon(u, v) = u + \gamma v + \epsilon$  is stable with the positively invariant set  $R$ , where  $R$  is as in (36).*

## 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

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\ v_n &= \beta_\lambda v_{n-1} + \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\ q_n &= \text{sign}(F(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})), \end{aligned} \quad (75)$$

and the tri-level second-order finite-memory (leaky)  $\Sigma\Delta$ -scheme which is given by

$$\begin{aligned} u_n &= \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\ v_n &= \beta_\lambda v_{n-1} + \beta_\lambda u_{n-1} + f_n^\lambda - q_n^\lambda, \\ q_n &= m(\eta F(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1})). \end{aligned} \quad (76)$$

$F$  in the above equations will be specified when necessary. We will write

$$(u_n, v_n) = S_{LS}(u_{n-1}, v_{n-1}, \mathbf{f}_n^\lambda), \quad (77)$$

for the scheme in (75), and

$$(u_n, v_n) = S_{LT}(u_{n-1}, v_{n-1}, \mathbf{f}_n^\lambda), \quad (78)$$

for the scheme in (76). Note that if  $S$  and  $S_T$  are as in (33) and (65), respectively, then

$$S_{LS}(u_{n-1}, v_{n-1}, \mathbf{f}_n^\lambda) = S(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1}, \mathbf{f}_n^\lambda), \quad (79)$$

$$S_{LT}(u_{n-1}, v_{n-1}, \mathbf{f}_n^\lambda) = S_T(\beta_\lambda u_{n-1}, \beta_\lambda v_{n-1}, \mathbf{f}_n^\lambda). \quad (80)$$

Then we will have:

**Theorem 9.** *Fix  $0 < \alpha < 1$ . Let  $F(u, v) = u + \gamma v + \epsilon$  be such that the standard second-order  $\Sigma\Delta$ -scheme, as in (32), is stable with the positively invariant set  $R$  as in (36) for some  $C > 0$ . Then the standard second-order finite-memory  $\Sigma\Delta$ -scheme defined by (75) 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 (80),  $S_{LS}(u, v, \delta) = S(\beta_\lambda u, \beta_\lambda v, \delta)$ . 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 positively 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  $F(u, v) = u + \gamma v + \epsilon$  be such that the tri-level second-order  $\Sigma\Delta$ -scheme, as in (60), is stable with the positively invariant set  $R$  as in (36) for some  $\eta$  and  $C$  satisfying (69). Then the tri-level second-order finite-memory  $\Sigma\Delta$ -scheme defined by (76) 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.

## 6 Conclusions and future work

In this paper we have discussed stability and robustness for several second-order  $\Sigma\Delta$ -quantizers. In particular, for the two-dimensional dynamical system associated with the “standard second-order  $\Sigma\Delta$ -quantizer” defined in (32) we constructed a positively invariant bounded set  $R \subset \mathbb{R}^2$  provided the input sequence  $(x_n)$  satisfies  $\|x\|_{l^\infty} \leq \alpha < 1$  for some  $\alpha$ . Moreover, we proved that this set is also robust with respect to small changes of various parameters in the definition of the scheme. Then in Section 4 and Section 5 we showed that the same set  $R$  is also positively invariant and robust under the dynamical systems associated with two families of modified schemes, namely, the tri-level second-order quantizer and leaky versions of both the standard and the tri-level second-order  $\Sigma\Delta$ -quantizers. We introduced these modified schemes to overcome the infinite memory of the standard second-order scheme, as discussed in Section 2.

In this paper, we have defined the set  $R$  in such a way that  $(u_n, v_n)$  remain in  $R$  for any input sequence  $(x_n)$  in  $(-\alpha, \alpha)$ . In practice, we are interested in much more constrained sequences, which are given by frequent sampling of bandlimited functions; for these more restricted class of sequences there may well exist a correspondingly much smaller invariant set  $\tilde{R}$ , which might lead to tighter bounds. Alternatively, the problem could be approached in a probabilistic setting; it would be interesting to investigate the existence of a smaller invariant set for ‘almost all’ input sequences in  $(-\alpha, \alpha)$  with respect to a “reasonable” probability distribution on this sequence space. Finally, it is a completely open problem to prove results similar to the second-order results in this paper for schemes of order higher than two.

## 7 Acknowledgements

The author would like to thank I. Daubechies for her support, helpful comments and advice. The author would also like to thank Sinan Güntürk for helpful discussions. This work was partially funded by NSF grant DMS-0070689, NSF KDI grant DMS-9872890 and AFOSR grant F49620-98-1-0044. Finally, the author is grateful to the reviewer for a thoughtful report that led to improvements in the paper.

## References

- [1] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.
- [2] I. Daubechies and R. DeVore, “Reconstructing a bandlimited function from very coarsely quantized data: I. A family of stable sigma-delta modulators of arbitrary order”, preprint.
- [3] S.R. Norsworthy, R. Schreier and G.C. Themes, eds, *Delta-Sigma Data Converters*, IEEE Press, 1997.
- [4] S.C. Pinault and P.V. Lopresti, “On the behavior of the double-loop sigma-delta modulator”, *IEEE Transactions on Circuits and Systems-II*, vol.40, Aug. 1993.
- [5] R. Schreier, M.V. Goodson and B. Zhang, “An Algorithm for Computing Convex Positively Invariant Sets for Delta-Sigma Modulators”, *IEEE Transactions on Circuits and Systems-I*, vol.44, Jan. 1997.
- [6] N. Thao, “MSE behavior and centroid function of  $m$ th order asymptotic  $\Sigma\Delta$  modulators”, *IEEE Transactions on Circuits and Systems: Part II*, 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.

Ozgur Yilmaz  
 Program in Applied and Computational Mathematics,  
 Princeton University  
 Washington Road, Fine Hall  
 Princeton, NJ 08540  
 oyilmaz@math.princeton.edu