INDUSTRIAL
MATHEMATICS
INSTITUTE

2001:24

Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order

R. DeVore and I. Daubechies

IMI
Preprint Series

Department of Mathematics
University of South Carolina

Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order

Ingrid Daubechies
Mathematics Department and
Program in Applied and Computational Mathematics
Princeton University

Ron DeVore
Mathematics Department and
Industrial Mathematics Institute
University of South Carolina

July 11, 2001

1 Introduction

Digital signal processing has revolutionized the storage and transmission of audio signals, images and video, in consumer electronics as well as in more scientific settings (such as medical imaging). The main advantage of digital signal processing is its robustness: although all the operations have to be implemented with necessarily not quite ideal hardware, the a priori knowledge that all ideal outcomes must lie in a very restricted set of well separated numbers makes it possible to recover the ideal outcomes by rounding off appropriately. When bursty errors can compromise this scenario (as is the case in many communication channels, as well as for storage in memory), making the "perfect" data unrecoverable by rounding off, knowledge of the type of expected contamination can be used to protect the data, prior to transmission or storage, by encoding them with error correcting codes; this is again done entirely in the digital domain. All these advantages have contributed to the present widespread use of digital signal processing. Many signals, however, are inherently "analog" rather than digital in nature; audio signals, for instance, correspond to functions $f(t)$, modeling rapid pressure oscillations, which depend on a "continuous" variable t (i.e. t ranges over $\mathbb{R}$ or an interval in $\mathbb{R}$, and not over a discrete set), and the range of f typically also fills an interval in $\mathbb{R}$. For this reason, the first step in any digital processing of such signals must consist in a conversion of the Analog signal to the Digital world, usually abbreviated as A/D conversion. For different types of signals, different A/D schemes are used; in this paper, we restrict our attention to audio signals, and a particular class of A/D conversion schemes adapted to audio signals. Note that at the other end of the chain, after the signal has been processed, stored, retrieved, transmitted, ..., all in digital form, it needs to be reconverted to an analog signal that can be understood by a human hearing system; we thus need a D/A conversion there.

The digitization of an audio signal rests on two pillars: *sampling* and *quantization.* It is standard to model audio signals by *bandlimited* functions, i.e. functions $f \in L^2(\mathbb{R})$ for which the Fourier transform

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\xi x} dx$$

vanishes outside an interval $|\xi| \leq \Omega$. Note that our Fourier transform is normalized so that it is equal to its inverse, up to a sign change,

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} d\xi \ .$$

The bandlimited model is justified by the observation that for the audio signals of interest to us, observed over realistic intervals $[-T,T]$, $\|\chi_{|\xi|>\Omega}(\chi_{|t|\leq T} f)^\wedge\|_2$ is negligible compared with $\|\chi_{|\xi|\leq\Omega}(\chi_{|t|\leq T} f)^\wedge\|_2$ for $\Omega \simeq 2\pi \cdot 20,000$ Hz. For band-limited functions one can use a well-known sampling theorem, the derivation of which is so simple that we include it here for completeness: since $\hat{f}$ is supposed on $[-\Omega, \Omega]$, it can be represented by a Fourier series converging in $L^2(-\Omega, \Omega)$, i.e.

$$\hat{f}(\xi) = \sum_{n \in \mathbb{Z}} c_n e^{-in\xi\pi/\Omega} \quad \text{for} \quad |\xi| \leq \Omega \ ,$$

where

$$c_n = \frac{1}{2\Omega} \int_{-\Omega}^{\Omega} \hat{f}(\xi) e^{in\xi\pi/\Omega} = \frac{1}{\Omega} \sqrt{\frac{\pi}{2}}\, f\left(\frac{n\pi}{\Omega}\right) \ .$$

We thus have

$$\hat{f}(\xi) = \frac{1}{\Omega} \sqrt{\frac{\pi}{2}} \sum_{n \in \mathbb{Z}} f\left(\frac{n\pi}{\Omega}\right) e^{-in\xi\pi\Omega} \quad \chi_{|\xi|\leq\Omega} \ ,$$

which by the inverse Fourier transform leads to

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

This formula reflects the well-known fact that an Ω-bandlimited function is completely characterized by sampling it at the corresponding Nyquist frequency $\frac{\Omega}{\pi}$. However, (1) is not useful in practice, because $sinc(x) = x^{-1} \sin x$ decays too slowly. If, as is to be expected, the samples $f\left(\frac{n\pi}{\Omega}\right)$ are not known perfectly, and have to be replaced, in the reconstruction formula (1) for $f(x)$, by $\tilde{f}_n = f\left(\frac{n\pi}{\Omega}\right) + \varepsilon_n$, then the corresponding reconstructed $\tilde{f}(x)$ may differ appreciably from $f(x)$. Indeed, the infinite sum $\sum_n \varepsilon_n sinc(\Omega x - n\pi)$ need not converge; even if we assume that we sum only over the finitely many n satisfying $\left|n\frac{\pi}{\Omega}\right| \leq T$ (using the tacit assumption that the $f\left(\frac{n\pi}{\Omega}\right)$ decay rapidly for n outside this interval), we will still not be able to ensure a better bound then

$$|f(x) - \tilde{f}(x)| \leq C\varepsilon \log T \ ;$$

since T may well be large, this is not satisfactory.

To circumvent this, it is useful to introduce oversampling. This amounts to viewing $\hat{f}$ as an element of $L^2(-\lambda\Omega, \lambda\Omega)$, with $\lambda > 1$; for $|\xi| \leq \lambda\Omega$ we can then represent $\hat{f}$ by a Fourier series in which the coefficients are proportional to $f\left(\frac{n\pi}{\lambda\Omega}\right)$,

$$\hat{f}(\xi) = \frac{1}{\lambda\Omega}\sqrt{\frac{\pi}{2}} \sum_{n\in\mathbb{Z}} f\left(\frac{n\pi}{\lambda\Omega}\right) e^{-in\xi\pi/\lambda\Omega} \ \text{ for } \ |\xi| \leq \lambda\pi \ .$$

Consider now a function g such that $\hat{g}$ is C^∞, and $\hat{g}(\xi) = \frac{1}{\sqrt{2\pi}}$ for $|\xi| \leq \pi$, $\hat{g}(\xi) = 0$ for $|\xi| > \pi$. Then

$$\hat{f}(\xi) = \frac{\pi}{\lambda\Omega} \sum_{n\in\mathbb{Z}} f\left(\frac{n\pi}{\lambda\Omega}\right) e^{-in\xi\pi/\lambda\Omega} \ \hat{g}\left(\frac{\pi\xi}{\Omega}\right) \ ,$$

resulting in

$$f(x) = \frac{1}{\lambda} \sum_{n\in\mathbb{Z}} f\left(\frac{n\pi}{\lambda\Omega}\right) g\left(\frac{\Omega}{\pi}x - \frac{n}{\lambda}\right) \ . \tag{2}$$

Because g is smooth with fast decay, this series now converges absolutely and uniformly; moreover if the $f\left(\frac{n\pi}{\lambda\Omega}\right)$ are replaced by $\tilde{f}_n = f\left(\frac{n\pi}{\lambda\Omega}\right) + \varepsilon_n$ in (2), with $|\varepsilon_n| < \varepsilon$, then the difference between the reconstructed $\tilde{f}(x)$ and $f(x)$ can be bounded uniformly:

$$|f(x) - \tilde{f}(x)| \leq \varepsilon \frac{1}{\lambda} \sum_{n\in\mathbb{Z}} \left| g\left(\frac{\Omega}{\pi}x - \frac{n}{\lambda}\right) \right| \leq \varepsilon C_g \tag{3}$$

where $C_g = \lambda^{-1}\|g'\|_{L^1} + \|g\|_{L^1}$ does not depend on T. Oversampling thus buys the freedom of using reconstruction formulas, like (2), that weigh the different samples in a much more localized way than (1) (only the $f\left(\frac{n\pi}{\lambda\Omega}\right)$ with $\left|x - \frac{n\pi}{\lambda\Omega}\right|$ "small" contribute significantly). In practice, it is customary to sample audio signals at a rate that is about 10 or 20% higher than the Nyquist rate. For high quality audio, a traditional sampling rate is 44,000 Hz.

The above discussion shows that moving from "analog time" to "discrete time" can be done without any problems or serious loss of information: for all practical purposes, f is completely represented by the sequence $\left(f\left(\frac{n\pi}{\lambda\Omega}\right)\right)_{n\in\mathbb{Z}}$. At this stage, each of these samples is still a real number. The transition to a discrete representation for each sample is called *quantization.*

The simplest way to "quantize" the samples $f\left(\frac{n\pi}{\lambda\Omega}\right)$ would be to replace each by a truncated binary expansion. If we know a priori that $|f(x)| \leq A < \infty$ for all x (a very realistic assumption), then we can write

$$f\left(\frac{n\pi}{\lambda\Omega}\right) = -A + A\sum_{k=0}^{\infty} b_k^n 2^{-k} \ ,$$

If we can "spend" κ bits per sample, then a natural solution is to just select the $(b_k^n)_{0\leq k\leq\kappa-1}$; reconstructing $\tilde{f}(x)$ from the approximate $\tilde{f}_n = -A + A\sum_{k=0}^{\kappa-1} b_k^n 2^{-n}$ then leads to $|f(x) - \tilde{f}(x)| \leq C2^{-\kappa+1}A$, where C is independent of κ or f (see above). Quantized representations of this type are used for the digital representations of audio signals, but they are, in fact, not the solution of choice for the A/D conversion step. (Instead, they are used after the A/D conversion, once one is firmly in the digital world.) The main reason for this is that it is very hard (and therefore very costly) to build analog devices that can divide the amplitude range $[-A, A]$ into $2^{-\kappa+1}$ precisely equal bins.

It turns out that it is much easier (= cheaper) to increase the oversampling rate, and to spend fewer bits on each approximate representation $\tilde{f}_n$ of $f\left(\frac{n\pi}{\Omega\lambda}\right)$. By appropriate choices of $\tilde{f}_n$ one can then hope that the error will decrease as the oversampling rate increases. Sigma-Delta (abbreviated by $\Sigma\Delta$) quantization schemes are a very popular way to do exactly this, oversampling significantly, and then spending very few bits per sample, achieving nevertheless a close approximation for the overall function f when the coarsely quantized $\tilde{f}_n$ are used instead of the true samples $f\left(\frac{n\pi}{\lambda\Omega}\right)$ in (3). In the most extreme case, every sample $f\left(\frac{n\pi}{\lambda\Omega}\right)$ is replaced by just one bit, i.e. by Aq_n with $q_n \in \{-1, 1\}$; in this paper we shall restrict our attention to such 1-bit $\Sigma\Delta$ quantization schemes. Although multi-bit $\Sigma\Delta$ schemes are becoming more popular in applications, there are many instances where 1-bit $\Sigma\Delta$ quantization is used. In the next section we explain the algorithm underlying $\Sigma\Delta$ quantization in its simplest version, we review the mathematical results that are known, and we formulate several questions, some of which we shall address in this paper in section 3. We conclude, in section 4, with many open problems and outlines for future research.

2 First order $\Sigma\Delta$-quantization

2.1 The simplest bound

For the sake of convenience, we shall set (by choosing appropriate units if necessary) $\Omega = \pi$ and $A = 1$. We are thus concerned with coarse quantization of functions $f \in \mathcal{C}_2 = \{h \in L^2;\ \|h\|_{L^\infty} \leq 1, \text{support } \hat{h} \subset [-\pi, \pi]\}$; for most of our results we also can consider the larger class

$$\mathcal{C}_1 = \{h;\ \hat{h} \text{ is a finite measure supported in } [-\pi,\pi],\ \|h\|_{L^\infty} \leq 1\}$$

.

With these normalizations (3) simplifies to

$$f(x) = \frac{1}{\lambda}\sum_n f\left(\frac{n}{\lambda}\right) g\left(x - \frac{n}{\lambda}\right) \ , \tag{4}$$

with g as described before, i.e.

$$\hat{g}(\xi) = \frac{1}{\sqrt{2\pi}} \text{ for } |\xi| \leq \pi,\ \hat{g}(\xi) = 0 \text{ for } |\xi| > \lambda\pi \text{ and } \hat{g} \in C^\infty \ . \tag{5}$$

It is not immediately clear how to construct sequences $\mathbf{q}^\lambda = (q_n^\lambda)_{n\in\mathbb{Z}}$, with $q_n^\lambda \in \{-1, 1\}$ for each $n \in \mathbb{Z}$, such that

$$\tilde{f}_{\mathbf{q}^\lambda}(x) = \frac{1}{\lambda}\sum q_n^\lambda g\left(x - \frac{n}{\lambda}\right) \tag{6}$$

provides a good approximation to f. The most naive idea would be to take simply $q_n^\lambda =$ sign $\left(f\left(\frac{n}{\lambda}\right)\right)$. This doesn't work, for the simple reason that there exist infinitely many independent bandlimited functions φ that are everywhere positive (such as the lowest order prolate spheroidal wave functions for arbitrary time intervals and sufficiently small symmetric frequency intervals; see e.g. [14, 12]; picking the signs of samples as candidate q_n^λ would make it impossible to distinguish between any two functions in this class.

First order $\Sigma\Delta$-quantization circumvents this by providing a simple iterative algorithm in which the q_n^λ are constructed by taking into account not only $f\left(\frac{n}{\lambda}\right)$ but also past $f\left(\frac{m}{\lambda}\right)$; we shall see below how this leads to good approximate $\tilde{f}_{\mathbf{q}^\lambda}$. Concretely, one introduces an auxiliary sequence $(u_n)_{n\in\mathbb{Z}}$ (sometimes described as giving the "internal state" of the $\Sigma\Delta$ quantizer) iteratively defined by

$$\begin{cases} u_n = u_{n-1} + f\left(\dfrac{n}{\lambda}\right) - q_n^\lambda \\ q_n^\lambda = \text{ sign}\left(u_{n-1} + f\left(\dfrac{n}{\lambda}\right)\right) \ , \end{cases} \tag{7}$$

and with an "initial condition" u_0 arbitrarily chosen in $(-1, 1)$. These u_n are then all bounded by 1 by a simple inductive argument. We prove this in two steps:

Lemma 2.1 *For any $f \in \mathcal{C}_1$, the sequence $(u_n)_{n\in\mathbb{N}}$ defined by the recursion (7) is uniformly bounded, $|u_n| < 1$ for all $n \geq 0$, if $|u_0| < 1$.*

Proof
Suppose $|u_{n-1}| < 1$. Because $f \in \mathcal{C}_1$, $\left|f\left(\frac{n}{\lambda}\right)\right| \leq 1$, so that $\left|f\left(\frac{n}{\lambda}\right) + u_{n-1}\right| < 2$, hence $\left|f\left(\frac{n}{\lambda}\right) + u_{n-1} - \text{ sign}\left(f\left(\frac{n}{\lambda}\right) + u_{n-1}\right)\right| < 1$. ∎

For negative n, we first have to transform the system (7) into a recursion in the other direction. To do this, observe that for $n \geq 1$ we have

$$\begin{aligned} u_{n-1} + f\left(\frac{n}{\lambda}\right) &> 0 \Rightarrow u_n - f\left(\frac{n}{\lambda}\right) = u_{n-1} - 1 < 0 \\ u_{n-1} + f\left(\frac{n}{\lambda}\right) &< 0 \Rightarrow u_n - f\left(\frac{n}{\lambda}\right) = u_{n-1} + 1 > 0. \end{aligned}$$

In all cases we have thus sign $(u_n - f\left(\frac{n}{\lambda}\right)) = -$ sign$(u_{n-1} + f\left(\frac{n}{\lambda}\right))$. The recursion (7) therefore implies, for $n \geq 1$,

$$u_{n-1} = u_n - f\left(\frac{n}{\lambda}\right) - \text{ sign}(u_n - f\left(\frac{n}{\lambda}\right)) \ , \tag{8}$$

which we can now extend to all n, making it possible to compute u_n for $n < 0$ corresponding to the "initial" value $u_0 \in (-1, 1)$. The same inductive argument then proves that these u_n are also bounded by 1. We have thus

Proposition 2.1 *The recursion (7), with $|u_0| < 1$ and $f \in \mathcal{C}_1$, defines a sequence $(u_n)_{n\in\mathbb{Z}}$ for which $|u_n| < 1$ for all $n \in \mathbb{Z}$.*

¿From this we can immediately derive a bound for the approximation error $|f(x) - \tilde{f}_{\mathbf{q}^\lambda}(x)|$.

Proposition 2.2 *For $f \in \mathcal{C}_1,\ \lambda > 1$, we define the sequence $\mathbf{q}^\lambda$ through the recurrence (7), with u_0 chosen arbitrarily in $(-1,1)$. Let g be a function satisfying (5). Then*

$$\left| f(x) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(x - \frac{n}{\lambda}\right)\right| \le \frac{1}{\lambda} \|g'\|_{L^1} \ . \tag{9}$$

Proof

Using (4), summation by parts, and the bound $|u_n| < 1$, we derive

$$\begin{aligned}
\left| f(x) \right. & - \left. \frac{1}{\lambda} \sum_n q_n^\lambda g\left(x - \frac{n}{\lambda}\right)\right| = \frac{1}{\lambda}\left| \sum_n \left(f\left(\frac{n}{\lambda}\right) - q_n^\lambda \right) g\left(x - \frac{n}{\lambda}\right)\right| \\
& = \frac{1}{\lambda}\left| \sum_n u_n \left(g\left(x - \frac{n}{\lambda}\right) - g\left(x - \frac{n+1}{\lambda}\right)\right)\right| \\
& \le \frac{1}{\lambda} \sum_n \left| g\left(x - \frac{n}{\lambda}\right) - g\left(x - \frac{n+1}{\lambda}\right)\right| \\
& \le \frac{1}{\lambda} \sum_n \int_{x-\frac{n+1}{\lambda}}^{x-\frac{n}{\lambda}} |g'(y)| dy = \frac{1}{\lambda}\|g'\|_{L^1} \qquad \blacksquare
\end{aligned}$$

This extremely simple bound is rather remarkable in its generality. What makes it work is, of course, the special construction of the q_n^λ via (7); the q_n^λ are chosen so that, for any N, the sum $\sum_{n=1}^N q_n^\lambda$ closely tracks $\sum_{n=1}^N f\left(\frac{n}{\lambda}\right)$, since

$$\left| \sum_{n=1}^N f\left(\frac{n}{\lambda}\right) - \sum_{n=1}^N q_n^\lambda \right| = |u_N - u_0| < 2 \ .$$

If we choose $u_0 = 0$ (as is customary), then we even have

$$\left| \sum_{n=1}^N f\left(\frac{n}{\lambda}\right) - \sum_{n=1}^N q_n^\lambda \right| = |u_N| < 1 \ ; \tag{10}$$

this requirement (which can be extended to negative N) clearly fixes the q_n^λ unambiguously. The "Σ" in the name $\Sigma\Delta$-modulation or $\Sigma\Delta$-quantization stems from this feature of tracking "sums" in defining the q_n^λ; $\Sigma\Delta$-modulation can be viewed as a refinement of earlier Δ-modulation schemes, to which the sum-tracking was added. There exists a vast literature on $\Sigma\Delta$-modulation in the electrical engineering community; see e.g. the review books [2] and [13]. This literature is mostly concerned with the design of, and the study of good design criteria for, more complicated $\Sigma\Delta$-schemes. The one given by (7) is the oldest and simplest [2], but is not, as far as we know, used in practice. We shall see below how better bounds than (9), i.e. bounds that decay faster as $\lambda \to \infty$, can be obtained by replacing (7) by other recursions, in which higher order differences play a role. Before doing so, we spend the remainder of this section on further comments on the first-order scheme and its properties.

2.2 Finite filters

In practice, one cannot use filter functions g that satisfy the condition in (5) because they require the full sequence $(q_n^\lambda)_{n\in\mathbb{Z}}$ to reconstruct even one value $f(x)$. It would be closer to the common practice to use G that are compactly supported (and for which the support of $\hat{G}$ is therefore all of $\mathbb{R}$, in contrast with (5)). In this case, the reconstruction formula (4) no longer holds, and the approximation error has additional contributions. Suppose G is supported in $[-R,R]$, so that, for a given x, only the q_n^λ with $|x-\frac{n}{\lambda}|<R$ can contribute to the sum $\sum_n q_n^\lambda G(x-\frac{n}{\lambda})$. Define $I_{\lambda,R}|x| = \{n;\ |x-\frac{n}{\lambda}|<R\}$. Then we have

$$\begin{aligned}\left|f(x) \right.&- \left.\frac{1}{\lambda}\sum_{n\in I_{\lambda,R}(x)} q_n^\lambda G\left(x-\frac{n}{\lambda}\right)\right| \qquad (11)\\ &\le \left|f(x)-\frac{1}{\lambda}\sum_n f\left(\frac{n}{\lambda}\right)G\left(x-\frac{n}{\lambda}\right)\right| + \frac{1}{\lambda}\left|\sum_n\left(f\left(\frac{n}{\lambda}\right)-q_n^\lambda\right)G\left(x-\frac{n}{\lambda}\right)\right| .\end{aligned}$$

The second term can be bounded as before. We can bound the first term by introducing again an "ideal" reconstruction function g, satisfying supp $\hat{g}\subset[-\lambda\pi,\lambda\pi]$ and $\hat{g}|_{[-\pi,\pi]}\equiv(2\pi)^{-1/2}$. Then

$$\begin{aligned}\left|f(x)\right. &- \left.\frac{1}{\lambda}\sum_n f\left(\frac{n}{\lambda}\right)G\left(x-\frac{n}{\lambda}\right)\right| = \frac{1}{\lambda}\left|\sum_n f\left(\frac{n}{\lambda}\right)\left[g\left(x-\frac{n}{\lambda}\right)-G\left(x-\frac{n}{\lambda}\right)\right]\right|\\ &\le \frac{1}{\lambda}\sum_n\left|g\left(x-\frac{n}{\lambda}\right)-G\left(x-\frac{n}{\lambda}\right)\right| \le \|G-g\|_{L^1}+\lambda^{-1}\|G'-g'\|_{L^1} ;\end{aligned}$$

by imposing on G that the L^1 distance of G and G'/λ to g and g'/λ, respectively, must be less than C/λ for at least one suitable g, this term becomes comparable to the estimate for the first term. (This means that G depends on λ; the support of G typically increases with λ.) For the special case where $|\hat{f}|$ is integrable, we can also bound the first term in a simpler way, by Fourier transforming and applying the Poisson summation formula:

$$\begin{aligned}\left|f(x)\right. &- \left.\frac{1}{\lambda}\sum_n f\left(\frac{n}{\lambda}\right)G\left(x-f\left(\frac{n}{\lambda}\right)\right)\right| \le\\ &\le \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|\hat{f}(\xi)-\frac{1}{\sqrt{2\pi}}\lambda\sum_n\int_{-\infty}^{\infty}\hat{f}(\zeta)e^{-i\frac{n}{\lambda}(\zeta-\xi)}\hat{G}(\xi)d\zeta\right|d\xi\\ &\le \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|\hat{f}(\xi)-\sqrt{2\pi}\hat{G}(\xi)\sum_k\hat{f}(\xi+2\pi k\lambda)|d\xi\\ &\le \int_{-\pi}^{\pi}|\hat{f}(\xi)|\left|\frac{1}{\sqrt{2\pi}}-\hat{G}(\xi)\right|d\xi+\sum_{k\neq 0}\int_{-\pi}^{\pi}|\hat{f}(\xi)|\ |\hat{G}(\xi-2\pi k\lambda)|d\xi ,\end{aligned}$$

where we have exploited that $f\in\mathcal{C}_1$, and $\hat{f}$ is supported in $[-\pi,\pi]$. It follows that if $\hat{G}|_{[-\pi,\pi]}$ is closer to $\frac{1}{\sqrt{2\pi}}$ than $\frac{C_1}{\lambda}$, and if $\hat{G}$ has sufficient decay so that $\sum_{k\neq 0}|\hat{G}(\xi-2\pi k\lambda)|\hat{\chi}_{|\xi|\le\pi}\le\frac{C_2}{\lambda}$ ($|\hat{G}(\xi)|\le C_3(1+|\xi|)^{-2-\varepsilon}$ would be sufficient), then (11) is bounded by C/λ.

Note that in practice, and except at the final D/A step mentioned in the introduction, bandlimited models for audio signals are always represented in *sampled* form. This means that once a digital sequence $(q_n^\lambda)_{n\in\mathbb{Z}}$ is determined, all the filtering and manipulations will be digital, and an estimate closer to the electrical engineering practice would seek to bound errors of the type

$$\left| f\left(\frac{m}{\lambda}\right) - \sum_{\lambda} q_n^\lambda G_{m-n}^\lambda \right| , \tag{12}$$

using discrete convolution with finite filters G^λ, rather than expressions of the type (9) or (11). If we were interested in optimizing constants relevant for practice, we should concentrate on (12) directly. For our present level of modeling however, in which we want to study the dominant behavior as a function of λ, working with (9) or (11), or their equivalent forms for higher order schemes, below, will suffice, since (12) will have the same asymptotic behavior as (11), for appropriately chosen G_m^λ. Unless specified otherwise, we shall assume, for the sake of convenience, that we work with reconstruction functions g satisfying (5).

2.3 More refined bounds

In practice, one observes better behavior for $|f(x) - \tilde{f}_{\mathbf{q}^\lambda}(x)|$ than that proved in Proposition 2.2. In particular, it is believed that, for arbitrary $f \in \mathcal{C}_1$, one has

$$\lim_{T\to\infty} \frac{1}{2T} \int_{|t|\le T} \left| f(t) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right|^2 dt \le \frac{C}{\lambda^3} , \tag{13}$$

with C independent of $f \in \mathcal{C}_1$ or of the initial condition u_0 for the recursion (7). Whether the conjecture (13) holds, either for each $f \in \mathcal{C}_1$, or in the mean (taking an average over a large class of functions in $\mathcal{C}_1$ or $\mathcal{C}_2$, with respect to an appropriate probability measure) is still an open problem.

A priori, it is not surprising that a better bound than (9) would hold. After all, we used very little in the derivation of (9). In particular, we never used explicitly that the $f\left(\frac{n}{\lambda}\right)$ were samples of the very smooth (because bandlimited) function f —we would have obtained the same C/λ bound if the samples $f\left(\frac{n}{\lambda}\right)$ had been replaced by any other bounded sequence $(a_n)_{n\in\mathbb{Z}}$ for which

$$f(x) = \frac{1}{\lambda} \sum_n a_n g\left(x - \frac{n}{\lambda}\right) . \tag{14}$$

(There exist many such sequences, because the $g\left(x - \frac{n}{\lambda}\right)$ are typically not independent. For instance, if $\hat{g}(\xi) = 0$ for $\xi \in I = [-\lambda\pi, -\lambda(1-\varepsilon)\pi] \cup [\lambda(1-\varepsilon)\pi, \pi]$, and if $\hat{b}$ is a finite measure supported in I, then $\hat{b}\hat{g} \equiv 0$, implying $\sum_n b\left(\frac{n}{\lambda}\right) g\left(x - \frac{n}{\lambda}\right) = 0$ for all x. We could thus take $a_n = f\left(\frac{n}{\lambda}\right) + b\left(\frac{n}{\lambda}\right)$ for any such b, and satisfy (14). If in addition we impose $\hat{b}$ to be real and even, and $|\hat{b}| \le \gamma$, then we would still have $a_n \in [-1, 1]$ for real $f \in C_1$ with $\|f\|_{L^\infty} \le 1 - \gamma$. An example would be $b_n = \gamma \cos n\pi \left(1 - \frac{\varepsilon}{2}\right)$.) We could use such much "wilder" sequences $(a_n)_{n\in\mathbb{Z}}$ in (7) and generate other possible 1-bit sequences $\mathbf{q}$, for which we do not expect $f_{\mathbf{q}}$ to approximate f as well as when the true samples are used. This argument leads us to suspect that the decay in λ^{-1} in (9) is not optimal. However, it turns out to be easier to derive the bound (9) than to improve on it.

For some special cases, i.e. for very restricted classes of functions f, (13) has been proved. In particular, it was proved by R. Gray [5] that if one restricts oneself to $f = f_a$, where $a \in [-1, 1]$ and $f_a(t) \equiv a$, then

$$\int_{-1}^{1} \left[\lim_{T\to\infty} \frac{1}{2T} \int_{|t|\leq T} \left| f_a(t) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right|^2 dt \right] da \leq \frac{C}{\lambda^3} \; ; \tag{15}$$

in Gray's analysis the integral over t is a sum over samples, and g is replaced by a discrete filter G^λ (see above), but his analysis applies equally well to our case. A different proof can be found also in [9]. Gray's result was later extended by Gray, Chou and Wong [6] to the case where the input function $f(t)$ is a sinusoid, $f(t) = a \sin bt$, with $|b| < \pi$.

For general bandlimited functions, there were no results, to our knowledge, until the work of S. Güntürk [7, 8], who proved, by a combination of tools from number theory and harmonic analysis, that, for all $f \in \mathcal{C}_1$ and all t for which $f'(t) \neq 0$,

$$\left| f(t) - \sum_n q_n^\lambda g^\lambda \left(t - \frac{n}{\lambda}\right) \right| \leq C\lambda^{-\frac{4}{3}+\varepsilon} \, . \tag{16}$$

In Güntürk's analysis the value of C depends on $|f'(t)|$; his g^λ (into which the $1/\lambda$ factor from (9) has been absorbed) is compactly supported, and has to satisfy various technical conditions. Although there is no mathematical proof for the moment, numerical simulations of intermediate results in Güntürk's work suggest that (16) may still hold, for general $f \in \mathcal{C}_1$, if the upper bound $C\lambda^{-\frac{4}{3}+\varepsilon}$ is replaced by $C\lambda^{-\frac{3}{2}+\varepsilon}$. For more details concerning the whole analysis and this discussion in particular, we refer the reader to [8].

2.4 Robustness

In practice, the recursive scheme (7) would be implemented by a simple feed-back loop circuit, with block diagram given in Figure 1.

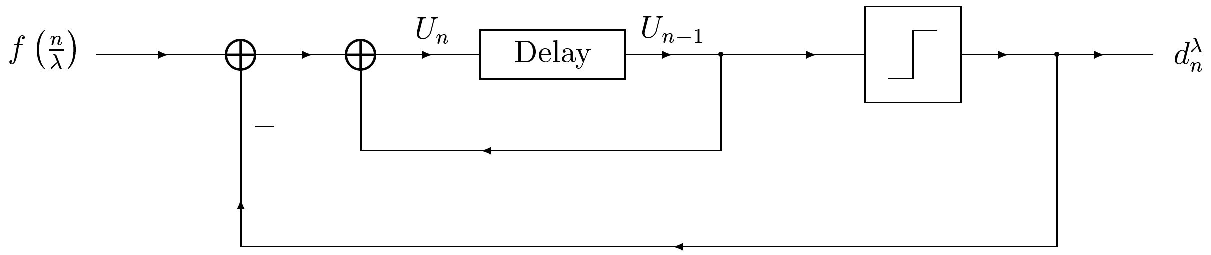

Figure 1: Block diagram implementing a first-order $\Sigma\Delta$ modulation. The symbol ⌠ stands for 1-bit quantization: the output of this block is simply the sign of the input.

For readers not accustomed to reading these diagrams, here is a simple road map. The whole diagram represents an algorithm, in which several quantities are computed from new inputs or previously computed quantities; all the computations are done instantaneously, but only at regularly occurring "clock times", represented here by values of $n \in \mathbb{Z}$ (in practice, nothing is ever instantaneous, of course, but we assume here that the computation time is significantly shorter than the clock time). Every arrow corresponds to a number that gets transmitted (sometimes in different directions) and that can be used in computations (adders or subtractors in our diagram), transformed (the quantizer box near the right of the diagram), or held back for one cycle (stored in memory and recovered, as in the "Delay" box). At every clocking time (i.e. at successive values of $n \in \mathbb{Z}$), a new sample value $f\left(\frac{n}{\lambda}\right)$ is read in, and a new d_n^λ can be read out. By giving the quantity entering the Delay box the label U_n, we see that Figure 1 is equivalent to

$$\begin{cases} U_n = U_{n-1} + f\left(\frac{n}{\lambda}\right) - d_n^\lambda \\ d_n^\lambda = \ \mathrm{sign}(U_{n-1})\ . \end{cases}$$

This is not quite the same as (7), but upon defining $u_n := U_{n+1} - f\left(\frac{n+1}{\lambda}\right), q_n^\lambda := d_{n+1}^\lambda$, it reduces exactly to (7).

Remarkably, one can give a very similar block diagram for a circuit or algorithm that computes, from the input sequence $x_0 = \alpha \in (-1,1)$, $x_n = 0$ for $n > 0$, the successive entries $(b_n)_{n\in\mathbb{N}}$ of the binary expansion of $\frac{1}{2}(\alpha+1)$; this block diagram is given in Figure 2.

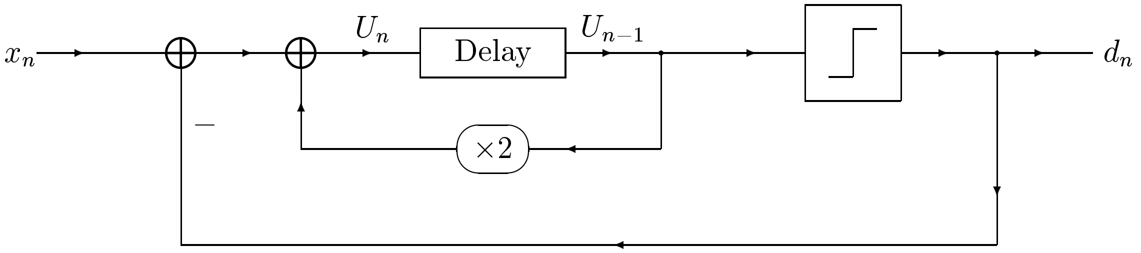

Figure 2: With input $x_0 = \alpha \in (-1,1), x_n = 0$ for $n > 0$, the output $(d_n)_n$ of this block diagram gives a binary representation for α; more precisely, $\frac{1}{2}(1+d_n)$ are the entries in the binary expansion of $\frac{1}{2}(1+\alpha)$.

The recursive algorithm reads now

$$\begin{aligned} U_n &= 2U_{n-1} + x_n - d_n \\ d_n &= \ \mathrm{sign}(U_{n-1})\ . \end{aligned}$$

After the transformation $\tilde{u}_n = U_{n+1} - \frac{1}{2}x_{n+1}$, $\tilde{b}_n = d_{n+1}$, this becomes

$$\begin{cases} \tilde{u}_n = 2\tilde{u}_{n-1} + x_n - \tilde{b}_n \\ \tilde{b}_n = \ \mathrm{sign}(2\tilde{u}_{n-1} + x_n) \end{cases} \tag{17}$$

One easily checks that if $\tilde{u}_{-1}$ is chosen to be 0, if $x_0 = \alpha$ with $\alpha \in (-1,1)$, and $x_n = 0$ for $n > 0$, then $|\tilde{u}_n| < 1$ for all n (the same induction argument as before works, since

$|2\tilde{u}_{n-1} + x_n|$ equals either $|x_0| < 2$ for $n = 0$, or $2|\tilde{u}_{n-1}| < 2$ if $n > 1$), so that

$$\left|\alpha - \sum_{n=0}^{N} 2^{-n}\tilde{b}_n\right| = \left|\sum_{n=0}^{N} 2^{-n}(x_n - \tilde{b}_n)\right|$$

$$= |\sum_{n=0}^{N} 2^{-n}(\tilde{u}_n - 2\tilde{u}_{n-1})| = |2^{-N}\tilde{u}_N| < 2^{-N} \to 0 \text{ as } N \to \infty$$

converging exponentially like a binary expansion. (Since the $b_n \in \{-1, 1\}$, this is not quite a binary expansion; however, the $b_n = \frac{1+\tilde{b}_n}{2} \in \{0, 1\}$ are the digits for the binary expansion of $\frac{1+\alpha}{2}$.)

The only difference between the two block diagrams lies in the presence of the multiplier by 2 in the feedback loop in Figure 2, absent in Figure 1. How can this be squared with our claim in the Introduction, i.e. how can $\Sigma\Delta$ quantization, which uses the circuit in Figure 1, be so much cheaper to implement than binary quantization of less frequent samples, which would use a circuit akin to Figure 2? The answer is that both circuits behave very differently when imperfections, in particular imperfect quantizers, are introduced. Quantizers are never perfect. Although we might desire to use $q(x) = \text{sign}(x)$ for our 1-bit quantizer, we must expect that in practice we may have, e.g., $q(x) = \text{sign}(x + \delta)$, where δ is unknown (and may vary from one chip to another), except for the specification $|\delta| < \tau$. A good algorithm or circuit is one that will perform well even without very stringent requirements on τ; if extremely tight specifications on τ are necessary to make everything work, then this will translate into an expensive circuit.

Let us replace the "sign" function in (7) by such a non-ideal quantizer; the new recursion is then

$$\begin{cases} u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) - q_n \\ q_n = Q\left(u_{n-1} + f\left(\frac{n}{\lambda}\right)\right) \end{cases}, \tag{18}$$

and let us assume that $Q(t) = \text{sign}(t - \delta)$ for some δ with $|\delta| \le \tau < 1$. It turns out that the u_n are still bounded, uniformly in $\delta \in [-\tau, \tau]$:

Proposition 2.3 *Let f be in $\mathcal{C}_1$, let u_n, q_n be as defined in (18), and let $\delta \in [-\tau, \tau]$ be fixed. If $|u_0| \le 1 - \tau$, then $|u_n| \le \tau + 1$ for all n.*

Proof

For $n \ge 0$, we use induction again. Suppose $|u_{n-1}| \le \Delta + 1$. Because $f \in \mathcal{C}_1$, $\left|f\left(\frac{n}{\lambda}\right)\right| \le 1$. We now distinguish three cases. If $u_{n-1} + f\left(\frac{n}{\lambda}\right) > \Delta$, then $u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) - 1 \in (\Delta - 1, \Delta + 1)$. Likewise, if $u_{n-1} + f\left(\frac{n}{\lambda}\right) < -\Delta$, then $u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) + 1 \in (-\Delta - 1, -\Delta + 1)$. Finally, if $-\Delta \le u_{n-1} + f\left(\frac{n}{\lambda}\right) \le \Delta$, then $Q(u_{n-1} + f\left(\frac{n}{\lambda}\right))$ could be either $+1$ or -1, so $u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) - Q(u_{n-1} + f\left(\frac{n}{\lambda}\right)) \in (-\Delta - 1, \Delta + 1)$.

To discuss the case $n \le 0$, we need to transform (18) again. Because $Q(t) = \text{sign}(t - \delta)$ for some unknown δ, we have in fact $u_n \in (-1 + \delta, 1 + \delta)$ for all $n \ge 0$, if we start from u_0 in this range. (The choice $u_0 \in (-1 + \Delta, 1 - \Delta)$ ensures this.) It then follows that $u_n - f\left(\frac{n}{\lambda}\right) < \delta$ if $u_{n-1} + f\left(\frac{n}{\lambda}\right) > \delta$, and $u_n - f\left(\frac{n}{\lambda}\right) > \delta$ if $u_{n-1} + f\left(\frac{n}{\lambda}\right) < \delta$, so that (18) can also be rewritten as

$$\begin{aligned} u_{n-1} &= u_n - f\left(\tfrac{n}{\lambda}\right) - Q\left(u_n - f\left(\tfrac{n}{\lambda}\right)\right) \\ q_{n-1} &= Q\left(u_n - f\left(\tfrac{n}{\lambda}\right)\right) . \end{aligned}$$

This implies, again by induction, $u_n \in (-1+\delta, 1+\delta) \subset (-1-\Delta, 1+\Delta)$ for all $n \leq 0$. ∎

By the same argument as in the proof of Proposition 2.2, this has as an immediate consequence.

Corollary 2.1 *Let f be in $\mathcal{C}_1$, let λ be > 1, and suppose g satisfies (5). Suppose the sequence $(q_n^\lambda)_{n\in\mathbb{Z}}$ is generated by (18), with the imperfect quantizer $Q(t) =$ sign$(t+\delta)$, where $\delta \in [-\tau, \tau]$ is arbitrary, and $\tau < 1$. Then, for all $t \in \mathbb{R}$,*

$$\left| f(t) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| \leq \frac{1+\tau}{\lambda} \|g'\|_{L^1} \ . \tag{19}$$

We can also consider the case where $\delta \in [-\tau, \tau]$ need not be fixed, i.e. where we have

$$\begin{aligned} u_n &= u_{n-1} + f\left(\frac{n}{\lambda}\right) - q_n \\ q_n &= \text{sign}\left(u_{n-1} + f\left(\frac{n}{\lambda}\right) - \delta_n\right) \ , \end{aligned} \tag{20}$$

with $(\delta_n)_n$ a sequence such that $|\delta_n| \leq \tau$ for all n.

Proposition 2.4 *Let f be in $\mathcal{C}_1$, let u_n, q_n be defined as in* (20)*, with $|\delta_n| \leq \tau$ for all n. If $|u_\ell| \leq 1+\tau$, then $|u_n| \leq 1+\tau$ for all $n \geq \ell$.*

Proof
Same as in the first half of the proof of Proposition 2.3. ∎

Because we now have only a one-sided estimate for the $|u_n|$ (i.e. only for $n \geq \ell$), we have to replace (19) by a one-sided estimate in t as well. One has then

Corollary 2.2 *Let f be in $\mathcal{C}_1$, let λ be > 1, and suppose g satisfies* (5) *as well as $|g(t)| \leq\subset (|t|+1)^{-M-1}$, for some $M \geq 1$. Choose $N \geq \lambda^{|T|/M}$. Suppose the sequence $(q_n^\lambda)_{n\in\mathbb{Z}, n\geq -N}$ is generated by* (20)*, starting with $|u_{-N-1}| \leq 1+\tau$, and assuming $|\delta_n| \leq \tau$ for all $n \geq -N$. Then, for all $\tau \geq 0$,*

$$\left| f(t) - \frac{1}{\lambda} \sum_{n=-N}^{\infty} q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| \leq \frac{1}{\lambda} \left(\frac{C}{M} + (1+\tau)\|g'\|_{L^1} \right) \tag{21}$$

Proof

We have, for $t \geq 0$,

$$\begin{aligned}
\left| f(t) \right. &- \left. \frac{1}{\lambda} \sum_{n=-N}^{\infty} q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| \\
&\leq \frac{1}{\lambda} \left| \sum_{n=-\infty}^{-N-1} f\left(\frac{n}{\lambda}\right) g\left(t + \frac{n}{\lambda}\right) \right| + \frac{1}{\lambda} \left| \sum_{n=-N}^{\infty} (u_n - u_{n-1}) g\left(t - \frac{n}{\lambda}\right) \right| \\
&\leq \frac{1}{\lambda} \sum_{m=N+1}^{\infty} \left| g\left(t + \frac{m}{\lambda}\right) \right| + \frac{1}{\lambda} \left| \sum_{n=-N}^{\infty} u_n \left(g\left(t - \frac{n}{\lambda}\right) - g\left(t - \frac{n+1}{\lambda}\right) \right) \right| \\
&\leq \subset \int_{N/\lambda}^{\infty} (1 + |t + s|)^{-M-1} ds + \frac{1}{\lambda}(1 + \tau) \|g'\|_{L^1} \\
&\leq \frac{C}{M} \left(1 + \frac{N}{\lambda}\right)^{-M} + \frac{1}{\lambda}(1 + \tau) \|g'\|_{L^1} .
\end{aligned}$$

Since $N \geq \lambda^{1+1/M}$, the desired estimate follows. ∎

If one replaces the "perfect" reconstruction function g by a suitable compactly supported G^λ, as in §2.2, then one can also derive one-sided estimates similar to (21), exploiting the compactness of support(G^λ). Although we must pay some penalty for the imperfection of the quantizer in all these cases (the constants increase), the precision that can be attained is nevertheless not limited by the imperfection: by choosing λ sufficiently large, the approximation error can be made arbitrarily small.

The same is not true for the binary expansion schemes (17). Suppose we use (17) to generate bits $\tilde{b}_n \in \{-1, 1\}$, and consider the approximation $\alpha_N = \sum_{n=0}^{N} 2^{-n} \tilde{b}_n$ to the input α, as before; the quantizer has been changed to $Q(t) =$ sign$(t - \delta)$, however. Suppose now $\alpha = \frac{\delta}{2}$; for the sake of definiteness, assume $\delta > 0$. Then (17), with this imperfect quantizer, will give $\tilde{b}_0 = -1$, so that $\alpha_N = \tilde{b}_0 + \sum_{n=1}^{N} 2^{-n} \tilde{b}_n \leq -2^{-N}$ for all N, implying $|\alpha - \alpha_N| > \frac{\delta}{2}$ for all N. The mistake made by the imperfect quantizer cannot be recovered by working harder, in contrast to the self-correcting property of the $\Sigma\Delta$-scheme. In order to obtain good precision overall with the binary quantizer, one must therefore impose very strict requirements on τ, which would make such quantizers very expensive in practice (or even impossible if τ is too small). On the other hand [3], $\Sigma\Delta$-quantizers are robust under such imperfections of the quantizer, allowing for good precision even if cheap quantizers are used (corresponding to less stringent restrictions on τ). It is our understanding that it is this feature that makes $\Sigma\Delta$-schemes so successful in practice. It would be better, however, to see the approximation error decay faster with λ; faster even than the $\lambda^{-\frac{3}{2}}$ estimate conjectured to hold for first order $\Sigma\Delta$-quantization of bandlimited functions (see §2.3 above). For this faster decay we must turn to higher order schemes.

3 Higher order $\Sigma\Delta$-quantization

3.1 The general principle

The proof of Proposition 2.2 suggests a mechanism by which better decay for $|f(x) - \tilde{f}_{\mathbf{q}^\lambda}(x)|$ could be obtained. The argument relied completely on the fact that $f\left(\frac{n}{\lambda}\right) - q_n^\lambda$ was rewritten as the first difference of a bounded sequence; summation by parts then gave the estimate. The following proposition states that if we can work with k-th order (instead of first-order) differences of bounded sequences, then we shall obtain a λ^{-k} decay for $|f(x) - \tilde{f}_{\mathbf{q}^\lambda}(x)|$ instead of the λ^{-1} decay of (9).

Proposition 3.1 *Take $f \in \mathcal{C}_1$; take $\lambda > 1$, and suppose g satisfies (5). Suppose that the $q_n^\lambda \in \{-1, 1\}$ are such that there exists a bounded sequence $(v_n)_{n\in\mathbb{Z}}$ for which*

$$f\left(\frac{n}{\lambda}\right) - q_n^\lambda = \Delta_n^k(v) := \sum_{l=0}^{k} (-1)^l \binom{k}{l} v_{n-l}\ . \tag{22}$$

Then, for all $x \in \mathbb{R}$,

$$\left| f(x) - \frac{1}{\lambda} \sum_{n\in\mathbb{Z}} q_n^\lambda g\left(x - \frac{n}{\lambda}\right) \right| \leq \frac{1}{\lambda^k} \|v\|_{l^\infty} \left\| \frac{d^k g}{dx^k} \right\|_{L^1} \tag{23}$$

Proof
It follows from (22) that

$$\begin{aligned} |f(x) \quad - \quad & \frac{1}{\lambda} \sum_n q_n^\lambda g\left(x - \frac{n}{\lambda}\right)\Bigg| = \frac{1}{\lambda} \left| \sum_n \left(f\left(\frac{n}{\lambda}\right) - q_n^\lambda \right) g\left(x - \frac{n}{\lambda}\right) \right| \\ = \quad & \frac{1}{\lambda} \left| \sum_n \Delta_n^k(v) g\left(x - \frac{n}{\lambda}\right) \right| = \frac{1}{\lambda} \left| \sum_n v_n \overline{\Delta}_n^k \left(g\left(x - \frac{\cdot}{\lambda}\right)\right) \right|\ , \end{aligned}$$

where $\overline{\Delta}^k$ is the k-th order forward difference. Thus (see [4], p. 137)

$$\overline{\Delta}_n^k \left(g\left(x - \tfrac{\cdot}{\lambda}\right)\right) = \textstyle\sum_{l=0}^{k} (-1)^l \binom{k}{l} g\left(x - \frac{n+l}{\lambda}\right)$$

$$= \tfrac{1}{\lambda^{k-1}} \textstyle\int_0^{k/\lambda} g^{(k)}\left(x - \frac{n+k}{\lambda} + t\right) \varphi_k(\lambda t) dt\ ,$$

where φ_k is the k-th order B-spline, $\varphi_k = \chi_{[0,1]} * \cdots * \chi_{[0,1]}$ (k convolution factors). Note that φ_k is positive, and supported on $[0, k]$ (so that we can just as well replace the integration limits by $-\infty$ and ∞). Moreover, $\sum_{m\in\mathbb{Z}} \varphi_k(y + m) = 1$ for all $y \in \mathbb{R}$, and all $k \in \mathbb{N} \setminus \{0\}$.

It follows that we can estimate

$$\begin{aligned}\left|f(x)\right. &- \left.\frac{1}{\lambda}\sum_{n\in\mathbb{Z}} q_n^\lambda g\left(x-\frac{n}{\lambda}\right)\right| \leq \\ &\leq \frac{1}{\lambda^k}\|v\|_{l^\infty}\sum_n \int_{-\infty}^{\infty} |g^{(k)}(x-\frac{n+k}{\lambda}+t)|\varphi_k(\lambda t)dt \\ &= \frac{1}{\lambda^k}\|v\|_{l^\infty}\sum_n \int_{-\infty}^{\infty} |g^{(k)}(y)|\varphi_k(\lambda y-\lambda x+n+k)dy \\ &= \frac{1}{\lambda^k}\|v\|_{l^\infty}\|g^{(k)}\|_{L^1}\ . \qquad\qquad \blacksquare\end{aligned}$$

Remark
As in the first-order case, we can adapt this to the case where only finitely many consecutive samples $f\left(\frac{n}{\lambda}\right)$ are known, and only the corresponding finitely many q_n^λ are computed. If (22) holds for $n=0,\ldots,N$, then we can still conclude, by decomposing the error into two components as in section 2.2, that

$$\begin{aligned}\left|f(x)\right. &- \left.\frac{1}{\lambda}\sum_{n=0}^{N} q_n^\lambda g\left(x-\frac{n}{\lambda}\right)\right| \\ &\leq C_K\left[(1+x^2)^{-K}+(1+\left(x-\frac{N}{\lambda}\right)^2)^{-K}\right]+\frac{1}{\lambda^k}(\max_{-k\leq n\leq N}|v_n|)\left\|\frac{d^kg}{dx^k}\right\|_{L^1}\ ,\end{aligned}$$

where we assumed that $|g(x)| \leq\subset (1+x^2)^{-K-1/2}$. We shall therefore concentrate on the forward recursion only, and on proving bounds on $\max_{-k\leq n\leq N}|v_n|$ that are independent of N. The key to better decay in λ for the approximation rate is thus to construct algorithms of type (22) with $k>1$ and uniformly bounded v_n. A $\Sigma\Delta$ algorithm which has such uniform bounds on the "internal state variables" is called "stable" in the electrical engineering literature; see e.g. [11]. We are thus concerned here with establishing the existence of stable $\Sigma\Delta$ schemes of arbitrary order. We first discuss the case $k=2$ and 3, before proceeding to general k.

3.2 Second-order $\Sigma\Delta$ schemes

We shall consider the recursion

$$\begin{cases} u_n = u_{n-1}+x_n-q_n \\ v_n = v_{n-1}+u_n \\ q_n = \ \mathrm{sign}[F(u_{n-1},v_{n-1},x_n)]\ , \end{cases} \qquad (24)$$

where the function F still needs to be specified. We are interested in applying this to the case where the x_n are samples of a function $f\in\mathcal{C}_1$; however, our discussion of the boundedness of u_n, v_n is valid for arbitrary input sequences $(x_n)_{n\in\mathbb{Z}}$, provided $|x_n|\leq a<1$. (Note that this means we need to impose the additional requirement $\|f\|_{L^\infty}\leq a<1$ on $f\in\mathcal{C}_1$ if we want to derive a bound of type (23).

Several choices for F have been considered in the literature; see e.g. [2]. One family of choices described in [2] is

$$F(u,v,x) = u + \gamma v + x \ , \tag{25}$$

where γ is a fixed parameter. A detailed discussion of the mathematical properties of this family is given in [17]. Another very interesting choice, proposed by N. Thao [15], is

$$F(u,v,x) = \frac{6x - 7sign(x)}{3} + \left(u + \frac{x + 3sign(x)}{2}\right)^2 + 2(1-|x|)v \ . \tag{26}$$

In both cases, one can prove that there exists a bounded set $A_a \subset \mathbb{R}^2$ so that if $|x_n| \leq a$ for all n, and $(u_0, v_0) \in A_a$, then $(u_n, v_n) \in A_a$ for all n; see [17]. It follows that we have uniform boundedness for the v_n if $x_n = f\left(\frac{n}{\lambda}\right)$ for bandlimited f with $\|f\|_{L^\infty} \leq a$, implying a λ^{-2} bound according to (23). As in the first order case, it turns out that for (26) this λ^{-2} bound can be improved by a more detailed analysis; for constant input, one achieves a $\lambda^{-5/2}$ bound. For (25) an additional x-dependent offset has to be inserted in F to obtain such a $\lambda^{-5/2}$ bound, without the insert one finds only λ^{-2}. We refer to [10, 16, 17] for a detailed analysis and discussion of these schemes. Robustness is an issue for second-order (and higher-order) schemes, just as it was for the first-order case. In fact, the problem becomes trickier because the quantization scheme should be able to deal not only with imperfect quantizers, but also with imprecisions in the multiplicative factors defining F in (26) or (27) below. The analysis in [17] shows that we do indeed have such robustness, for a wide family of second-order sigma-delta schemes. Proving more refined bounds than (23) for higher order $\Sigma\Delta$ schemes, even for constant input, turns out to be much harder than for first order (where already the analysis leading to (16) is highly nontrivial – see [8]). This is mainly because even for $x_n \equiv x$ constant, the dynamical system (24) is much more complex that (7). In particular, the map

$$\begin{aligned} R_{1,x} : \ & \mathbb{R} \to \mathbb{R} \\ & u \mapsto u + x - \ \text{sign}(u + x) \end{aligned}$$

has $[-1, 1]$ as an invariant set, regardless of the value of $x \in [-1, 1]$. In contrast, the maps

$$\begin{aligned} R_{2,x} \ : \ & \mathbb{R}^2 \to \mathbb{R}^2 \\ & \begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u + x - \ \text{sign}(u + \frac{v}{2} + x) \\ v + u + x - \ \text{sign}(u + \frac{v}{2} + x) \end{pmatrix} \end{aligned} \tag{27}$$

have similar invariant sets Γ_x, which now depend on the value of $x \in (-1, 1)$, however. The sets Γ_x have fascinating properties which are still poorly understood; for instance, for each fixed x, Γ_x seems to be a tile for $\mathbb{R}^2$ under translations by $2\mathbb{Z}^2$. (This tiling property is observed for many F, although we know of no proof in general.) For $x \neq 0$, the Γ_x for (25) can have interesting fractal boundaries; for "large" x, these Γ_x are disconnected. (See Figure 3.)

On the other hand, the sets Γ_x for (26) are connected neighborhoods of $(0, 0)$ bounded by four parabolic arcs (see Figure 4); because of the explicit characterization of these sets, a proof that the $2\mathbb{Z}^2-$ translates of Γ_x tile $\mathbb{R}^2$ is straightforward in this case. The smoothness of the boundaries also makes it possible to refine (23) for this choice of F and for constant input (see [10]).

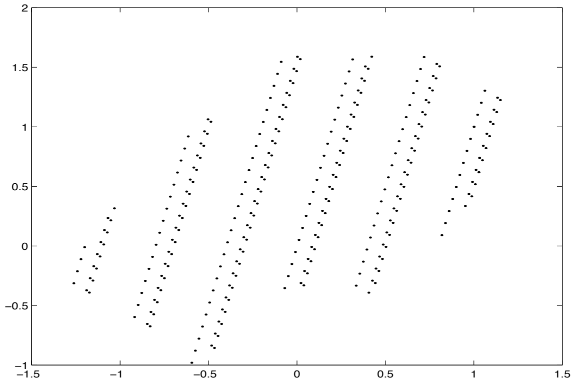

Figure 3: The attracting invariant sets Γ_x for two values of x (left: $x = .2$, right: $x = .8$) and for the choice (25) for F, with $\gamma = .5$. For $x = .2$, Γ_x is polygon, with sides that can be computed explicitly [10] ; for $x = .8$, Γ_x is disconnected and fractal.

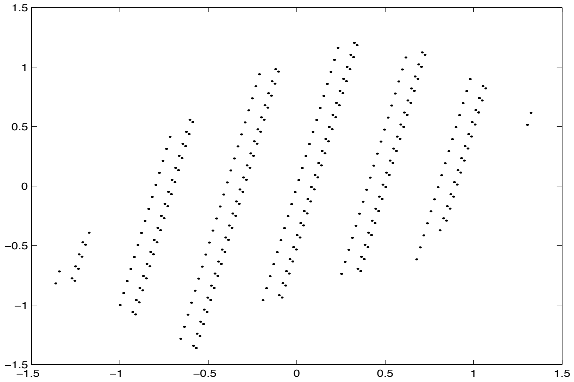

Figure 4: The attracting invariant sets Γ_x for two values of x (left: $x = .5$, right: $x = .8$) for the choice (26) for F.

Neither of the two schemes (25) or (26) are easy to generalize to higher order. We shall therefore concentrate our attention here on yet another choice for F,

$$F(u, v, x) = u + x + M \text{ sign}(v) \ , \tag{28}$$

with $M > 1$ to be fixed below. For this choice of F, we shall prove explicitly that the u_n, v_n remain bounded if $|x_n| \leq a < 1$ via an argument that we will be able to generalize to arbitrary order. With F as in (28), q_n^λ in (24) has two regimes:

- if $|u_{n-1} + x_n| > M$, then $q_n = \text{ sign}(u_{n-1} + x_n)$
- if $|u_{n-1} + x_n| \leq M$, then $q_n = \text{ sign}(v_{n-1})$.

We now have

Proposition 3.2 *Suppose $|x_n| \leq a < 1$ for all $n \in \mathbb{N}$. Choose F as in (28), with $M \geq 1+2a$, and let u_n, v_n, q_n be as defined by (24). Then, if $|u_0| \leq M + 1$, we have $|u_n| \leq M + 1$ for all $n \in \mathbb{N}$. Moreover, if $v_0 = 0$, then $|v_n| \leq \frac{(M+2-a)^2}{2(1-a)}$ for all $n \in \mathbb{N}$.*

We start by proving the following succession of Lemmas 3.1-3.4. In each of these lemmas, we make the same assumptions as in the statement of Proposition 3.2.

Lemma 3.1 *If $|u_0| \le M+1$, then $|u_n| \le M+1$ for all $n \in \mathbb{N}$.*

Proof
By induction. Suppose $|u_{n-1}| \le M+1$. If $|u_{n-1}+x_n| > M$, then $|u_n| = |u_{n-1}+x_n - \operatorname{sign}(u_{n-1}+x_n)| = |u_{n-1}+x_n| - 1 \le |u_{n-1}| + a - 1 < M+1$. If $|u_{n-1}+x_n| \le M$, then $|u_n| \le |u_{n-1}+x_n| + 1 \le M+1$. ■

Lemma 3.2 *Suppose $v_k \le 0$, and $v_{k+1}, v_{k+2}, \ldots, v_{k+L} > 0$. Define κ to be the smallest integer strictly larger than $\frac{2M}{1-a}+1$. If $L \ge \kappa$, then there exists at least one $l \in \{1,\ldots,\kappa\}$ such that $u_{k+l} + x_{k+l+1} < -M+1+a$.*

Proof
Suppose $u_{k+1}+x_{k+2}, \ldots, u_{k+\kappa-1}+x_{k+\kappa}$ are all $\ge -M+1+a$. Because $v_{k+1}, \ldots, v_{k+\kappa+1}$ are all > 0, we have $q_{k+2} = \ldots = q_{k+\kappa} = 1$, which implies

$$\begin{aligned} u_{k+\kappa} + x_{k+\kappa-1} &= u_k + \sum_{l=2}^{\kappa}(x_{k+l} - q_{k+l}) + x_{k+\kappa+1} \\ &\le M+1+(\kappa-1)(a-1)+a < M+1+a-(1-a)\frac{2M}{1-a} \\ &= -M+1+a\ . \qquad \blacksquare \end{aligned}$$

Lemma 3.3 *Let $v_k, v_{k+1}, \ldots, v_{k+L}$ be as in Lemma 3.2. If $u_{k+l}+x_{k+l+1} < -M+1+a$ for some $l \in \{1,\ldots,L\}$, then we have, for all l' satisfying $l \le l' \le L$,*

$$u_{k+l'} + x_{k+l'+1} < -M+1+a\ .$$

Proof
By induction. Suppose $u_{k+n}+x_{k+n+1} < -M+1+a$ with $n \in \{1,\ldots,L-1\}$; we prove that this implies $u_{k+n+1}+x_{k+n+2} < -M+1+a$. If $u_{k+n}+x_{k+n+1} \ge -M$, then $q_{k+n+1} = 1$ (since $v_{k+n} > 0$), hence $u_{k+n+1}+x_{k+n+2} < -M+1+a+x_{k+n+2}-1 < -M+1+a$. On the other hand, if $u_{k+n}+x_{k+n-1} < -M$, then $q_{k+n+1} = -1$, and $u_{k+n+1}+x_{k+n+2} < -M+1+x_{k+n+2} \le -M+1+a$. ■

Lemma 3.4 *Let $v_k, v_{k+1}, \ldots, v_{k+L}$ be as above. Then the u_{k+l} decrease monotonically in l, with $u_{k+l-1} - u_{k+l} \ge 1-a$, until $u_{k+l}+x_{k+l+1}$ drops below $-M+1+a$. All subsequent $u_{k+l'}$ with $l' \le L$ remain negative.*

Proof
As long as $u_{k+n}+x_{k+n+1} \ge -M+1+a$ with $n \le L$, we have $q_{k+n+1} = 1$, so $u_{k+n} - u_{k+n+1} = -x_{k+n+1}+1 \ge 1-a$. If $u_{k+l}+x_{k+l+1} < -M+1+a$, then $u_{k+l'}+x_{k+l'+1} < -M+1+a$ by Lemma 3.3 if $l' \in \{l,\ldots,L\}$, so that $u_{k+l'} < -M+1+2a \le 0$. ■

It is now easy to complete the proof of Proposition 3.2:

Proof
Lemma 3.1 already proved our assertion for the u_n. We shall prove here that $v_n \leq \frac{(M+2-a)^2}{2(1-a)}$. The lower bound follows analogously.

Suppose $v_{k+1}, \ldots, v_{k+L}$ is a stretch of strictly positive v_n, preceded by $v_k \leq 0$. We have then, for all $m \in \{1, \ldots, L\}$,

$$v_{k+m} = v_k + \sum_{l=1}^{m} u_{k+l} \leq \sum_{l=1}^{m} u_{k+l} \ .$$

By Lemma 3.4, these u_{k+l} decrease monotonically by at least $(1-a)$ at every step until they drop below a certain negative value, after which they stay negative. Consequently, $u_{k+l} \leq u_{k+1} - (1-a)(l-1) \leq (M+1) - (1-a)(l-1)$, at least until this last expression drops below zero. It follows that

$$\begin{aligned} v_{k+m} &\leq \max_{n\geq 1} \sum_{l=1}^{n} ((M+1) - (1-a)(l-1)) \\ &\leq \frac{1}{2}\left(\left\lfloor\frac{M+1}{1-a}\right\rfloor + 1\right)\left(2(M+1) - \left\lfloor\frac{M+1}{1-a}\right\rfloor(1-a)\right) \\ &\leq \frac{(M+2-a)^2}{2(1-a)} \ . \qquad \blacksquare \end{aligned}$$

Remarks

1. The bound on $|v_n|$ is significantly larger than that on $|u_n|$. For $a = .5$, for instance and $M = 1 + 2a = 2$, we have $|u_n| \leq 3$ and $|v_n| \leq 12.25$. Although we could certainly tighten up our estimates, the growth of the bounds on the interval state variables, as we go to higher order schemes, is unavoidable. We shall come back to this later.

2. It is not really necessary to suppose $v_0 = 0$. If $|v_0| \leq A$, then we have $|v_n| \leq A + \frac{1}{2}\frac{(M+2-a)^2}{1-a}$ for all n; moreover, once an index ℓ is reached for which v_ℓ and $v_{\ell+1}$ differ in sign, we have $|v_n| \leq \frac{1}{2}\frac{(M+2-a)^2}{1-a}$ for all subsequent n.

3.3 A third-order $\Sigma\Delta$ scheme

Let us consider the construction we discussed for second order, but take it one step further. We define the recursion

$$\left\{ \begin{aligned} u_n^{(1)} &= u_{n-1}^{(1)} + x_n - q_n \\ u_n^{(2)} &= u_{n-1}^{(2)} + u_n^{(1)} \\ u_n^{(3)} &= u_{n-1}^{(3)} + u_n^{(2)} \\ q_n &= \text{sign}\left[u_{n-1}^{(1)} + x_n + M_1\left(\text{sign}(u_{n-1}^{(2)} + M_2\ \text{sign}(u_{n-1}^{(3)}))\right)\right] \end{aligned} \right. \tag{29}$$

where M_1, M_2 will be fixed below in such a way as to ensure uniform boundedness of the $(|u_n^{(3)}|)_{n\in\mathbb{N}}$, provided we start from appropriate boundary conditions. We assume again that $|x_n| \leq a < 1$ for all $n \geq 0$.

There are now *three* regimes for the determination of the q_n:

- if $|u_{n-1}^{(1)} + x_n| > M_1$, then $q_n = \operatorname{sign}(u_{n-1}^{(1)} + x_n)$
- if $|u_{n-1}^{(1)} + x_n| \leq M_1$, then two possibilities exist:
  - if $|u_{n-1}^{(2)}| > M_2$, then $q_n = \operatorname{sign}(u_{n-1}^{(2)})$
  - otherwise, $q_n = \operatorname{sign}(u_{n-1}^{(3)})$.

Let us indicate here how the arguments of subsection 3.2 can be adapted to deal with this case. We shall keep this discussion to a sketch only; a formal proof of this third order case will be implied by the formal proof for arbitrary order in the next subsection. This preliminary discussion will help understand the more general construction, however.

First of all, exactly the same argument as in the proof of Lemma 3.1 establishes that $|u_n^{(1)}| \leq M_1 + 1 = M_1'$.

Next, imagine a long stretch of $u_{n+1}^{(2)}, u_{n+2}^{(2)}, \ldots$, all $> M_2$. Then the corresponding q_{n+l+1} are all 1, unless $u_{n+l}^{(1)} < -M_1$. Arguments similar to those in the proofs of Lemma 3.2-3.4 then show that if $u_{n+1}^{(1)} > -M_1 + 1 + 2a \geq 0$, the $u_{n+l}^{(1)}$ will decrease monotonically, by at least $(1-a)$ at each step, until $u_{n+l}^{(1)} + x_{n+l+1}$ drops below $-M_1 + 1 + a$ (in at most κ_1 steps), after which all the subsequent $u_{n+l'}^{(1)}$, in the stretch are negative. As before, this argument (and its mirror for the lower bound) leads to $|u_n^{(2)}| \leq M_2'$.

One could then imagine repeating the same argument again to prove the desired bound on the $|u_n^{(3)}|$: prove that if one has a long stretch of $u_{l+1}^{(3)}, \ldots, u_{l+L}^{(3)}$ that are all positive, then necessarily the corresponding $u_{l+m}^{(2)}$ must dip to negative values and remain negative, in such a way that the total possible growth of the $u_{l+m}^{(3)}$ must remain bounded. We will have to make up for a missing argument, however: when we followed this reasoning at the previous level, we were helped by the priori knowledge that consecutive $u_n^{(1)}$ just differ by some minimal amount, $|u_{n+1}^{(1)} - u_n^{(1)}| \geq 1 - a$. We used this to ensure the speediness of the dropping $u_{l+m}^{(1)}$, and thus to bound the $u_{l+m}^{(2)}$. In our present case, we have no such a priori bound on $|u_{n+1}^{(2)} - u_n^{(2)}|$, so we need to find another argument to ensure sufficiently fast decrease of the $u_{l+m}^{(2)}$. What follows sketches how this can be done.

Suppose $u_l^{(3)} \leq 0, u_{l+1}^{(3)}, \ldots, u_{l+L}^{(3)} > 0$. Then we must have, within the first κ_2 indices of this stretch (with κ_2, independent of L, to be determined below) that some $u_{l+m}^{(2)} \leq -M_2$. Indeed, if $u_{l+1}^{(2)}, \ldots, u_{l+\kappa_2-1}^{(2)} > -M_2$, then the corresponding q_{l+m} are 1, unless $u_{l+m-1}^{(1)} < -M_1$. As before, this forces the $u_{l+m}^{(1)}$ down, until they hit below $-M_1 + a$ in at most κ_1 steps, after which they remain below this negative value. This forces the $u_{l+m}^{(2)}$ to decrease, and one can determine κ_2 so that if $u_{l+1}^{(2)}, \ldots, u_{l+\kappa_2-1}^{(2)} > -M_2$, then $u_{l+\kappa_2} \leq -M_2$ must follow. Once $u_{l+l'}^{(2)}$ has dropped below $-M_2$, the picture changes. We can get $q_{l+l'+k} = -1$, and the argument that kept the $u_{l+m}^{(1)}$ down can then no longer be applied. In fact, some of the $u_{l+m}^{(1)}$ with

$m > l'$ may become positive again, causing the $u^{(2)}_{l+m}$ to increase. However, as soon as we have κ_1 consecutive $u^{(2)}_n > -M_2$, we must have for at least one of the corresponding indices, that $u^{(1)}_n < -M_1 + 1 + a$, which forces the subsequent $u^{(1)}_n$, below this value too, and we are back in our cycle forcing the $u^{(2)}_n$, down until they hit below $-M_2$. So if $-M_2 + \kappa_1 M'_1 \leq 0$, then the $u^{(2)}_n$ don't get a chance to grow to positive values within the first κ_1 indices after $u^{(2)}_{l+l'} < -M_2$. This forces *all* the $u^{(2)}_{l+m}$ to be negative for $m = l'+1, \ldots, L$; since $l' \leq \kappa_2$, this then leads, by the same argument as on the previous level, to a bound on $u^{(3)}_{l+m}$.

In the next subsection we present this argument formally, for schemes of arbitrary order; the proof consists essentially of careful repeats of the last paragraph at every level. This then also leads to estimates for the bounds M'_j.

3.4 Generalization to arbitrary order

We assume again that $|x_n| \leq a < 1$ for all $n \in \mathbb{N}$. To define the $\Sigma\Delta$ scheme of order J for which we shall prove uniform boundedness of all internal variables, we first introduce the following constants.

$$\begin{aligned} \kappa_1 &= \left\lfloor \frac{5(1+a)}{1-a} \right\rfloor + 2 \\ M_1 &= 2(1+a) \\ M_j &= (3\kappa_1)^{j-1} 4^{(j-1)(j-2)} M_1 \qquad j = 2, \ldots, J-1 \\ M_J &= 0 \end{aligned} \tag{30}$$

The scheme itself is then defined as follows

$$\left\{ \begin{aligned} u^{(1)}_n &= u^{(1)}_{n-1} + x_n - q_n \\ u^{(j)}_n &= u^{(j)}_{n-1} + u^{(j-1)}_n \qquad j = 2, \ldots, J \\ q_n &= \mathrm{sign}\Big\{ u^{(1)}_{n-1} + M_1 \,\mathrm{sign}\Big[u^{(2)}_{n-1} + M_2 \,\mathrm{sign}\Big(u^{(3)}_{n-1} + \cdots \\ &\quad + M_{J-2}\mathrm{sign}\Big(u^{(J-1)}_{n-1} + M_{J-1}\,\mathrm{sign}(u^{(J)}_{n-1})\Big)\Big)\Big]\Big\} \end{aligned} \right. \tag{31}$$

If we write $U_n = u^{(J)}_n$, then $u^{(j)}_n = \Delta^{J-j}_n(U) := \sum_{l=0}^{J-j} \binom{J-j}{l} (-1)^l U_{n-l}$. As initial conditions we take $u^{(j)}_{-1} = 0$ for $j = 1, \ldots, J$; we start the recursion at $n = 0$. We then have the following proposition:

Proposition 3.3 *Suppose $|x_n| \leq a < 1$ for all $n \in \mathbb{N}$. Let M_j for $j = 1, \ldots, J$, be defined as in (31), and let the sequences $(q_n)_{n\in\mathbb{N}}$ and $(u^{(j)}_n)_{n\in\mathbb{N}}, j = 1, \ldots, J$, be as defined by (22), with initial conditions $u^{(j)}_{-1} = 0$ for $j = 1, \ldots, J$. Then $U_n := u^{(J)}_n$ satisfies $|U_n| \leq \frac{1}{2}(3\kappa_1)^{J-1} 4^{(J-1)(J-2)} M_1$ for all $n \in \mathbb{N}$.*

Remark

Note that this scheme is slightly different from the ones considered so far, in that the formula for q_n includes $u_{n-1}^{(1)}$ only and not the combination $u_{n-1}^{(1)} + x_n$. This is done merely for convenience: it avoids having to single out the case $j = 1$ as a special case whenever we write general lemmas for the $u_n^{(j)}$, below. Similar bounds can be proved when x_n is included in the formula for q_n; we expect that the numerical constants might be slightly better (as they are in the first and second order case) but their general behavior will be similar.

The proof of Proposition 3.3 is essentially along the lines sketched for the third-order case, albeit more technical in order to deal with the general case. The whole argument is one big induction on j. We start by stating two lemmas for the lowest value of j, to start off the induction argument.

Lemma 3.5 $|u_n^{(1)}| \leq M_1 + 1 + a$ *for all* $n \in \mathbb{N}$.

Proof
The argument is very similar to that used in the proof of Lemma 3.1, except that x_n does not appear in the definition of q_n. We work by induction. Suppose $|u_{n-1}^{(1)}| \leq M_1 + 1 + a$. If $|u_{n-1}^{(1)}| > M_1$, then q_n and $u_{n-1}^{(1)}$ have the same sign, so $|u_n^{(1)}| \leq |u_{n-1}^{(1)}| - 1 + |x_n| \leq |u_{n-1}^{(1)}| - 1 + a \leq |u_{n-1}^{(1)}| \leq M_1 + 1 + a$.

If $|u_{n-1}^{(1)}| \leq M_1$, then $|u_n^{(1)}| \leq |u_{n-1}^{(1)}| + 1 + a \leq M_1 + 1 + a$. ∎

Lemma 3.6 *If* $u_{n+1}^{(2)}, \ldots, u_{n+N}^{(2)} > M_2$*, with* $N \geq \kappa_1$*, then there must exist* $l \in \{1, \ldots, \kappa_1\}$ *such that* $u_{n+l}^{(1)} < -M_1$*. Moreover, for all* $l' \in \{l, \ldots, N\}$ *we have then* $u_{n+l'}^{(1)} < -M_1 + 1 + a$*. A similar statement holds if we start with* $u_{n+1}^{(2)}, \ldots, u_{n+N}^{(2)} < -M_2$*, and other signs are reversed accordingly.*

Proof
The argument is again similar to the proofs of Lemmas 3.2-3.3. Suppose $u_{n+1}^{(1)}, \ldots, u_{n+\kappa_1-1}^{(1)}$ are all $\geq -M_1$. Then we have $q_{n+2} = \cdots = q_{n+\kappa_1} = 1$. Hence

$$\begin{aligned} u_{n+\kappa_1}^{(1)} &= u_{n+1}^{(1)} + \sum_{l=2}^{\kappa_1} (x_{n+l} - q_{n+l}) \leq M_1 + 1 + a - (\kappa_1 - 1)(1 - a) \\ &< 3(1a) - 5(+a) = -M_1 \, . \end{aligned}$$

This establishes that $u_{n+l}^{(1)} < -M_1$ for some $l \in \{1, \ldots, \kappa_1\}$. Next, suppose that $u_{n+r}^{(1)} < -M_1 + 1 + a$, for some r with $l \leq r \leq N - 1$. If $u_{n+r}^{(1)} \geq -M_1$, then $q_{n+r+1} = 1$, hence $u_{n+r+1}^{(1)} = u_{n+r}^{(1)} + x_{n+r+1} - 1 < u_{n+r}^{(1)} < -M_1 + 1 + a$; if $u_{n+r}^{(1)} < -M_1$, then $u_{n+r+1}^{(1)} < -M_1 + 1 + |x_{n+r+1}| \leq -M_1 + 1 + a$. In both cases, $u_{n+r+1}^{(1)} < -M_1 + 1 + a$, and we continue by induction. ∎

Next we introduce auxiliary constants, for $j = 1, \ldots, J - 1$,

$$\begin{aligned} \kappa_j &= 4^{2(j-1)} \kappa_1 \\ M_j' &= \frac{3}{2} M_j \\ M_j'' &= \frac{1}{2} M_j \end{aligned}$$

as well as $M'_J = \frac{1}{2}(3\kappa_1)^{J-1}4^{(J-1)(J-2)}M_1$. These have been tailored so as to satisfy, for $j = 2, \ldots, J-1$,

$$\begin{aligned} M_j &= 2M'_{j-1}\kappa_{j-1} \\ \kappa_j &= \kappa_{j-1} + (M_j + M'_j)/M''_{j-1} \\ M'_j &= M_j + M'_{j-1}\kappa_{j-1} \\ M''_j &= M_j - M'_{j-1}\kappa_{j-1} \ ; \end{aligned}$$

in addition, the third formula also holds for $j = J$.

We now state our general lemmas, used in the induction argument.

Lemma 3.7 (j)
$|u_n^{(j)}| \leq M'_j$ for $n \in \mathbb{N}$.

Lemma 3.8 (j)
If $u_{n+1}^{(j+1)}, \ldots, u_{n+N}^{(j+1)} > M_{j+1}$, with $N \geq \kappa_j$, then there must be $l \in \{1, \ldots, \kappa_j\}$ so that $u_{n+l}^{(j)} < -M_j$. Moreover, for all $l' \in \{l, \ldots, N\}$, one then has $u_{n+l'}^{(j)} < -M''_j$. A similar statement holds if we start with $u_{n+1}^{(j+1)}, \ldots, u_{n+N}^{(j+1)} < -M_{j+1}$, and other signs are reversed appropriately.

Our induction argument then alternates two steps:

a. Lemmas 3.7(j) + Lemma 3.8(j) implies Lemma 3.7$(j+1)$

b. Lemmas 3.7(k) + 3.8(k) for $k \leq j$, together with Lemma 3.7$(j+1)$, implies Lemma 3.8$(j+1)$.

Since the case $j = 1$ is established (see Lemmas 3.5, 3.6), induction will ultimately get us to a proof of Lemma 3.7(J), which will complete the proof of Proposition 3.3. It remains to prove steps **a** and **b**.

Proof of step **a**
We prove only that $u_n^{(j+1)} \leq M'_{j+1}$; the inequality $u_n^{(j+1)} \geq -M'_{j+1}$ is analogous.

Assume $u_n^{(j+1)} \leq M_{j+1}, u_{n+1}^{(j+1)}, \ldots, u_{n+N}^{(j+1)} > M_{j+1}$. We need to show that none of these $u_{n+l}^{(j+1)}, l = 1, \ldots, N$, can exceed M'_{j+1}. We have $u_{n+l}^{(j+1)} = u_n^{(j+1)} + \sum_{k=1}^{l} u_{n+k}^{(j)}$.

By Lemma 3.8(j), at most the first κ_j terms in the sum can be positive, and each of those is bounded by M'_j by Lemma 3.7(j). Therefore, for each $l \in \{1, \ldots, N\}$,

$$u_{n+l}^{(j+1)} \leq M_{j+1} + \kappa_j M'_j = M'_{j+1} \ . \qquad \blacksquare$$

Proof of step **b**
This step is the most complicated. In order to prove this induction step, we invoke a third technical Lemma, that will itself be proved by induction. We put ourselves in the framework where Lemmas 3.7(k) are proved for $k \leq j+1$, as well as Lemmas 3.8(k) for $k \leq j$.

Lemma 3.9 $(j+1)$

Let $j \in \{1,\dots,J-2\}$ be fixed. Take any $k \in \{1,\dots,j\}$. Suppose $u^{(j+2)}_{n+1},\dots,u^{(j+2)}_{n+N} > M_{j+2}$ with $N \geq \kappa_{j+1}$. Suppose that the set $S \subset \{n+1,\dots,n+N\}$ satisfies the following requirements:

- *S consists of consecutive indices only, and contains at least κ_k elements, i.e. $S = \{n+m+1,\dots,n+M+m\}$ for some $m \geq 0$ and $M \geq \kappa_k$*
- *$u^{(l)}_r \geq -M_l$ for all $r \in S$, all $l \in \{k+1,\dots,j+1\}$*

Then any κ_k consecutive elements in S must contain at least one r such that $u^{(k)}_r < -M_k$. Moreover, once $u^{(k)}_r < -M_k$, for an $r \in S$, then we have $u^{(k)}_{r'} \leq -M''_k$ for all $r' \in S, r' \geq r$.

Proof

By induction on k. We assume Lemmas 3.7(j') and 3.8(j') hold for $j' \leq j+1$ and $j' \leq j$ respectively.

1. The case $k=1$.

   - We have $u^{(j')}_s \geq -M_{j'}$ for all $s \in S$, and all $j' \in \{2,\dots,j\}$. We must prove that if we have $\kappa_1 - 1$ consecutive elements in S, numbered $r+1,\dots,r+\kappa_1-1$, for which $u^{(1)}_{r+1},\dots,u^{(1)}_{r+\kappa_1-1} \geq -M_1$, then necessarily $u^{(1)}_{r+\kappa_1} < -M_1$.

     But if $u^{(1)}_{r+1},\dots,u^{(1)}_{r+\kappa_1-1} \geq -M_1$, then $q_{r+2} = \cdots = q_{r+\kappa_1} = 1$ (because all the indices are in S, so that for each s, $u^{(j')}_s \geq -M_{j'}$ if $j' \in \{2,\dots,j\}$, and $u^{(j+1)}_s > M_{j+1}$.) It follows that

     $$\begin{aligned} u^{(1)}_{r+\kappa_1} &= u^{(1)}_{r+1} + \sum_{m=r+2}^{r+\kappa_1} (x_m - q_m) \\ &\leq M'_1 + (\kappa_1 - 1)(a-1) < -M_1 \end{aligned}$$

   - Next we must show that if $u^{(1)}_r < -M_1$ for some $r \in S$, then $u^{(1)}_{r'} < -M''_1$ for $r' \geq r, r' \in S$.

     This is again done as in the proof of Lemma 3.5, by induction on r':

     - assume $u^{(1)}_{r'-1} < -M''_1$
     - if $u^{(1)}_{r'-1} < -M_1$, then $u^{(1)}_{r'} < -M_1 + a + 1 = -M''_1$

       if $u^{(1)}_{r'-1} \geq -M_1$, then $q_{r'} = 1$ and $u^{(1)}_{r'} = u^{(1)}_{r'-1} + a - 1 \leq -M''_1 + a - 1 < -M''_1$.

   This completes the proof of the case $k=1$ of Lemma 3.9(j).

2. Suppose the lemma holds for $k = 1,\dots,k_0-1$, with $k_0 \leq j$. Let's then prove it for $k = k_0$.

   Take a set S that satisfies all the requirements for $k = k_0$.

- In a first part, we must prove that among any κ_{k_0} consecutive elements in S there is at least one r such that $u^{(k_0)}_r < -M_{k_0}$. That is, we must prove that if we can find $u^{(k_0)}_{s+1}, \ldots, u^{(k_0)}_{s+\kappa_{k_0}-1}$ that are all $\geq -M_{k_0}$, then $u^{(k_0)}_{s+\kappa_{k_0}}$ must be $< -M_{k_0}$.

  Define $\tilde{S} = \{s+1, \ldots, s+\kappa_{k_0}-1\} \subset S$. Then $\tilde{S}$ satisfies all the requirements in Lemma 3.9$(j+1)$ for $k = k_0 - 1$. By the induction hypothesis, it follows that there is a t among the first κ_{k_0-1} elements of $\tilde{S}$ such that $u^{(k_0-1)}_t < -M_{k_0-1}$. Moreover, for all $t' \in \tilde{S}$ exceeding this t, $u^{(k_0-1)}_{t'} < -M''_{k_0-1}$. It follows that

$$\begin{aligned} u^{(k_0)}_{s+\kappa_{k_0}} &= u^{(k_0)}_{t-1} + \sum_{t'=t+1}^{s+\kappa_{k_0}-1} u^{(k_0-1)}_{t'} + u^{(k_0-1)}_t + u^{(k_0-1)}_{s+\kappa_{k_0}} \\ &< M'_{k_0} - (\kappa_{k_0} - 1 - \kappa_{k_0-1})M''_{k_0-1} - M_{k_0-1} + (-M''_{k_0-1} + M'_{k_0-2}) \\ &= M'_{k_0} - (\kappa_{k_0} - \kappa_{k_0-1})M''_{k_0-1} - M_{k_0-1} + M'_{k_0-2} \\ &= M'_{k_0} - \frac{M_{k_0} + M'_{k_0}}{M''_{k_0-1}} M''_{k_0-1} - M_{k_0-1} + M'_{k_0-2} \\ &= -M_{k_0} - M_{k_0-1} + M'_{k_0-2} \leq -M_{k_0} \ , \end{aligned}$$

  where in the first inequality, we used Lemmar 3.7 $(k_0 - 1)$ to bound each of the entries in the sum and we bounded the last term by writing $u^{(k_0-1)}_{s+\kappa_{k_0}} = u^{(k_0-1)}_{s+\kappa_{k_0-1}} + u^{(k_0-2)}_{s+\kappa_{k_0-1}}$.

- In this second part, we must prove that if, for $r \in S$, $u^{(k_0)}_r < -M_{k_0}$, then all $r' \in S$ with $r' \geq r$ must satisfy $u^{(k_0)}_{r'} < -M''_{k_0}$.

  For $r' > r$, let $r'' = \max\{t \leq r'; u^{(k_0)}_t < -M_{k_0}\}$. Then $u^{(k_0)}_{r''+1}, \ldots, u^{(k_0)}_{r'-1} \geq -M_{k_0}$. By the induction hypothesis, we must have, among the first κ_{k_0-1} of these (if the stretch is that long) an index t so that $u^{(k_0-1)}_t < -M_{k_0-1}$, and all later t' in the stretch will have $u^{(k_0-1)}_{t'} \leq -M''_{k_0-1}$. It follows that the $u^{(k_0)}_{r''+1}, \ldots, u^{(k_0)}_{r'-1}$ cannot increase after the first $\kappa_{k_0-1} - 1$ entries:

$$\begin{aligned} &\max\left[u^{(k_0)}_{r''+1}, \ldots, u^{(k_0)}_{r'-1}\right] \leq \max\left[u^{(k_0)}_{r''+1}, \ldots, u^{(k_0)}_{r''+\kappa_{k_0-1}-1}\right] \\ &\leq u^{k_0}_{r''} + \max_{l\in\{1,\ldots,\kappa_{k_0-1}-1\}} \textstyle\sum_{l'=1}^{l} u^{(k_0-1)}_{r''+l'} < -M_{k_0} + (\kappa_{k_0-1} - 1)M'_{k_0-1} \quad . \end{aligned}$$

  Hence $u^{(k_0)}_{r'} \leq u^{(k_0)}_{r'-1} + M'_{k_0-1} \leq -M_{k_0} + \kappa_{k_0-1}M'_{k_0-1} = -M''_{k_0}$. This completes the proof of Lemma 3.9(j). ∎

We can now use this to complete the

Proof of step **b**:
Assume Lemmas 3.7(j') and 3.8(j') hold for $j' \leq j$, as well as Lemma 3.7$(j+1)$. This also allows us to use Lemma 3.9(j') for $j' \leq j$.

- Suppose now $u^{(j+2)}_{n+1}, \ldots, u^{(j+2)}_{n+N} > M_{j+2}$ with $N \geq \kappa_{j+1}$. We have to prove that among the first κ_{j+1} elements of this stretch, we have one for which $u^{(j+1)}_{n+l} < -M_{j+1}$. As usual, we assume $u^{(j+1)}_{n+1}, \ldots, u^{(j+1)}_{n+\kappa_{j+1}-1} \geq -M_{j+1}$ (and we need to establish $u^{(j+1)}_{n+\kappa_{j+1}} < -M_{j+1}$). Define S by $S = \{n+1, \ldots, n+\kappa_{j+1}-1\}$, and fix $k = j$. Then, S, k satisfy all the conditions in Lemma 3.9 $(j+1)$. It follows that at most the first $\kappa_j - 1$ elements of S can correspond to $u^{(j)}_r \geq -M_j$. Therefore the max of $\{u^{(j+1)}_t; t \in S\}$ must be achieved among the first $\kappa_j - 1$ elements, and

$$\begin{aligned} u^{(j+1)}_{n+\kappa_{j+1}} &< \max\{u^{(j+1)}_t; t \in \{n+1, \ldots, n+\kappa_j+1\}\} - (\kappa_{j+1} - \kappa_j - 1)M''_j \\ &\leq M'_{j+1} - (\kappa_{j+1} - \kappa_j)M''_j = -M_{j+1} \end{aligned}$$

- Next, we need to prove that if $u^{(j+1)}_{n+l} < -M_{j+1}$ for some $l \in \{1, \ldots, N\}$, then $u^{(j+1)}_{n+l'} \leq -M''_{j+1}$ for $l' \in \{l, \ldots, N\}$. Define $l'' := \max\{t \leq l' : u^{(j+1)}_{n+t} < -M_{j+1}\}$. Then $u^{(j+1)}_{n+l''+1}, \ldots, u^{(j+1)}_{n+l'-1} \geq -M_{j+1}$. We have again that the max of these must be obtained among the first $\kappa_j - 1$ entries (since after that, the $u^{(j+1)}_s$ must decrease monotonously), so that

$$\begin{aligned} \max[u^{(j+1)}_{n+l''+1}, \ldots, u^{(j+1)}_{n+l'-1}] &\leq u^{(j+1)}_{n+l''} + \textstyle\sum_{s=1}^{\kappa_j-1} |u^{(j)}_{n+l''+s}| \\ &< -M_{j+1} + (\kappa_j - 1)M'_j \\ \Rightarrow u^{(j+1)}_{n+l'} \leq u^{(j+1)}_{n+l'-1} + M'_j &\leq -M_{j+1} + \kappa_j M'_j = -M''_{j+1} \ . \end{aligned}$$

- We have thus proved Lemma 3.8$(j+1)$, completing the proof of step **b** in our induction process. ∎

Remarks

- There is clearly a lot of room for obtaining tighter bounds. By being more careful, one can replace the factor $4^{(J-1)(J-2)}$ in M'_J by $\gamma^{(J-1)(J-2)}$ with $\gamma < 4$. We have not been able to reduce the growth in J of the exponent below a quadratic, however. We shall come back to this, and its implications, in the next section.

- As in the lower order special cases, it is not really crucial to start with $u^{(j)}_{-1} = 0$; other initial conditions can also be chosen, with minimal impact on the bounds.

4 Conclusions and open problems

Our construction in section 3 showed that it is possible to construct stable $\Sigma\Delta$-quantizers of arbitrary order. The quantizers (31) are, however, very far from schemes built in practice for 1-bit $\Sigma\Delta$-quantization. Often, such practical schemes involve not only higher order differences (as in our family), but also additional convolutional filters; it is not clear to us at this point what mathematical role is played by these filters. It may well be that they allow the bounds on the internal state variables to be smaller numerically than in our construction. This is certainly one topic for future work.

In addition, other notions of "stability" are often desirable in practice. For instance, audio signals often have stretches in time where they are uniformly small in amplitude. It would be of interest to ensure that the internal state variables of the system then also fall back (after a transition time) into a bounded range much smaller than their full dynamic range. At present, we know of no construction to ensure this mathematically.

The fast growth of our bounds M_j in subsection 3.4 is also unsatisfactory from the purely theoretical point of view. The combination of Propositions 3.1 and 3.3 leads, for $f \in \mathcal{C}_1$ with $\|f\|_{L^\infty} \leq a < 1$, to the estimate

$$\left| f(x) - \frac{1}{\lambda} \sum_n q_n^{(k),\lambda} g(x - \frac{n}{\lambda} \right| \leq C \frac{1}{\lambda^k} \alpha^k \beta^{k^2} ,$$

where we have absorbed the bound on $\|\frac{d^k g}{dx^k}\|_{L^1}$ into α^k (which is possible for appropriately chosen g, within the constraints of (5)), and where we write $q_n^{(k),\lambda}$ for the output of the k-th order $\Sigma\Delta$-quantizer (31), given input $\left(f(\frac{n}{\lambda})\right)_{n\in\mathbb{Z}}$. Given λ, we can then select the optimal k_λ, which leads to the estimate

$$\left| f(x) - \frac{1}{\lambda} \sum_n q_n^{(\lambda)} g(x - \frac{n}{\lambda} \right| \leq C' \lambda^{-\gamma \log \lambda} ,$$

where $q_n^{(\lambda)} = q_n^{(k_\lambda),\lambda}$. By spending λ bits per Nyquist interval, we thus obtain a precision with an asymptotic behavior that is better than any inverse polynomial in λ, but that is still far from the exponential decay in λ that one would get from spending the bits on binary approximations to samples taken at a frequency slightly above the Nyquist frequency. We don't know how much of this huge discrepancy is due to our method of proof, to our stable family itself, or to the limitation of $\Sigma\Delta$-quantization schemes in general. In [1] it is proved that $\Sigma\Delta$-schemes can never obtain the optimal accuracy of binary expansions; sub-optimal but still exponential decay in λ is not excluded by [1], however. It would be of great interest to see what the information-theoretic constraints are on $\Sigma\Delta$-schemes or other practical quantization schemes for redundant information; a first discussion (including other robust quantizers) will be given in [3], but there are still many open problems.

Acknowledgments

We thank Sinan Güntürk and Thao Nguyen for helpful discussions concerning the topic of this paper. One of us (I.D.) would also like to thank the Air Force Office for Scientific Research for partial support, as well as the Institute for Advanced Study in Princeton for its hospitality during the writing of this paper. The other author wishes to thank Princeton University for sabbatical support and the Office of Naval Research for support of his research. Both authors gratefully acknowledge the support of a National Science Foundation KDI grant supporting their work.

References

[1] A.R. Calderbank and I. Daubechies, Shortcomings of Democracy, *IEEE Trans. Information Theory*, submitted.

[2] J.C. Candy and G.C. Temes (Editors), Oversampling Delta-Sigma Data Converters Theory, Design and Simulation, *IEEE Circuits and Systems Society*

[3] I. Daubechies, R. DeVore, C. Güntürk and V. Vaishampayan, Quantization Error Correction in A/D Conversion, in preparation.

[4] R.L. DeVore and G.G. Lorentz, *Constructive Approximation*, Springer Grundlehien, vol. 303, 1993.

[5] R. M. Gray, Spectral Analysis of Quantization Noise in Single-Loop Sigma-Delta Modulator with dc Input, *IEEE Transactions on Communications*, vol. COM-35, pp.481-489, 1987.

[6] R. M. Gray, W. Chou and P. W. Wong, Quantization Noise in Single-Loop Sigma-Delta Modulation with Sinusoidal Inputs, *IEEE Transactions on Communications*, vol. COM-37, pp. 956–968, 1989.

[7] C. Güntürk, Improved error estimates for first order Sigma-Delta modulation, *Sampling Theory and Applications, SampTA'99*, Leon, Norway, August 1999.

[8] C. Güntürk, Reconstructing a Bandlimited Function from Very Coarsely Quantized Data: II. Improving the Error Estimate for First Order Sigma-Delta Modulators, in preparation. A more extended version appears in C. Güntürk, Harmonic Analysis of Two Problems in Signal Quantization and Compression, Ph.D. thesis, Program in Applied and Computational Mathematics, Princeton University.

[9] S. Güntürk, J.C. Lagarias and V. Vaishampayan, Robustness of Single Loop Sigma-Delta Modulation for Constant Inputs, *IEEE Transactions on Information Theory*, submitted.

[10] C. Güntürk and N. Thao, Refined analysis of MSE in one-bit second order Sigma-Delta modulation, in preparation

[11] S. Hein and A. Zakhor, On the Stability of Sigma Delta Modulators, *IEEE transactions on signal processing*, vol. 41, no. 7, July 1993.

[12] H. Landau and H.O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty, II, Bell Systems Technical Journal, **40** pp. 65–84, 1961.

[13] S.R. Norsworthy, R. Schreier and G.C. Temes (Editors), Delta-Sigma Data Converters Theory, Design and Simulation, *sponsored by IEEE Circuits and Systems Society.*

[14] D. Slepian and H.O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty, I, Bell Systems Technical Journal, **40** pp. 43–64, 1961.

[15] N. Thao, Quadratic one-bit second order sigma-delta modulators, *IEEE Trans. on Circuits and Systems*, submitted.

[16] N. Thao, C. Güntürk,I. Daubechies, and R. DeVore, A new approach to one-bit nth order $\Sigma\Delta$-modulation, in preparation.

[17] O. Yilmaz, Stability analysis for several Sigma-Delta methods of coarse quantization of bandlimited functions, submitted to *Constructive Approximation*.