Annals of Mathematics, **158** (2003), 679–710

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

By Ingrid Daubechies and Ron DeVore

1. Introduction

Digital signal processing has revolutionized the storage and transmission of audio and video signals as well as still images, in consumer electronics and 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, of necessity, not quite ideal hardware, the *a priori* knowledge that all correct outcomes must lie in a very restricted set of well-separated numbers makes it possible to recover them by rounding off appropriately. Bursty errors can compromise this scenario (as is the case in many communication channels, as well as in memory storage devices), making the "perfect" data unrecoverable by rounding off. In this case, 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 done entirely in the digital domain. These advantages have contributed to the present widespread use of digital signal processing.

Many signals, however, are not digital but analog in nature; audio signals, for instance, correspond to functions $f(t)$, modeling rapid pressure oscillations, which depend on the "continuous" time 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 a particular class of A/D conversion schemes adapted to audio signals. Note that at the 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*, both of which we now briefly discuss.

We start with sampling. It is standard to model audio signals by *band-limited* functions, i.e. functions $f \in L^2(\mathbb{R})$ for which the Fourier transform

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

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,

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

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. Here and later in this paper, $\|\cdot\|_2$ denotes the $L_2(\mathbb{R})$ norm. For bandlimited 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 supported 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

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

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 $\mathrm{sinc}(x) = x^{-1}\sin x$ 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 $f(t)$, by $\widetilde{f}_n = f\left(\frac{n\pi}{\Omega}\right) + \varepsilon_n$, with all $|\varepsilon_n| \leq \varepsilon$, then the corresponding approximation $\widetilde{f}(t)$ may differ appreciably from $f(t)$. Indeed, the infinite sum $\sum_n \varepsilon_n \mathrm{sinc}(\Omega t - n\pi)$ need not converge. Even if we assume that we sum only over the finitely many n

satisfying $|n\frac{\pi}{\Omega}| \le T$ (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 than $|f(t)-\widetilde{f}(t)| \le C\varepsilon \log T$; 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 $|\xi| \le \lambda\Omega$ we can then represent $\hat{f}$ by a Fourier series in which the coefficients are proportional to $f\left(\frac{n\pi}{\lambda\Omega}\right)$,

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

Introducing a function g such that $\hat{g}$ is C^∞, and $\hat{g}(\xi) = \frac{1}{\sqrt{2\pi}}$ for $|\xi| \le \pi$, $\hat{g}(\xi) = 0$ for $|\xi| > \lambda\pi$, we can write

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

resulting in

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

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 $\widetilde{f}_n = f\left(\frac{n\pi}{\lambda\Omega}\right) + \varepsilon_n$ in (2), with $|\varepsilon_n| < \varepsilon$, then the difference between the approximation $\widetilde{f}(x)$ and $f(x)$ can be bounded uniformly:

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

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 $\left|t - \frac{n\pi}{\lambda\Omega}\right|$ "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 $|f(t)| \le A < \infty$ for all t (a very realistic assumption), then we can write

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

with $b_k^n \in \{0,1\}$ for all k,n. If we can "spend" κ bits per sample, then a natural solution is to just select the $(b_k^n)_{0\le k\le \kappa-1}$; constructing $\widetilde{f}(x)$ from the approximations $\widetilde{f}_n = -A + A\sum_{k=0}^{\kappa-1} b_k^n 2^{-n}$ then leads to $|f(t) - \widetilde{f}(t)| \le C2^{-\kappa+1}A$, where C is independent of κ or f. Quantized representations of this type are used for the digital representations of audio signals, but they are 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 $\widetilde{f}_n$ of $f\left(\frac{n\pi}{\Omega\lambda}\right)$. By appropriate choices of $\widetilde{f}_n$ 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. In the most extreme case, every sample $f\left(\frac{n\pi}{\lambda\Omega}\right)$ in (1) is replaced by just one bit, i.e. by a q_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.

The following is an outline of the content of the paper. In Section 2 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.

In Section 3, we generalize the simple first-order $\Sigma\Delta$ scheme of Section 2 to higher orders, leading to better bounds. In particular, we show, for any $k \in \mathbb{N}$, an explicit mathematical algorithm that defines, for every function f that is bandlimited (i.e. the inverse Fourier transform of a finite measure supported in $[-\Omega,\Omega]$) with absolute value bounded by $a < 1$, and for all $n \in \mathbb{Z}$, "bits" $q_n^{(k)} \in \{-1,1\}$ such that, uniformly in t,

$$\left| f(t) - \frac{1}{\lambda}\sum_n q_n^{(k)} g\left(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\right)\right| \le C_g^{(k)}\lambda^{-k} \quad . \tag{4}$$

Moreover, we prove that our algorithm is *robust* in the following sense. Since we have to make a transition from real-valued inputs $f\left(\frac{n\pi}{\lambda\Omega}\right)$ to the discrete-valued $q_n \in \{-1,1\}$, we have to use a discontinuous function as part of our algorithm. In our case, this will be the *sign* function, $\operatorname{sign}(A) = 1$ if $A \ge 0$, $\operatorname{sign}(A) = -1$ if $A < 0$. In practice, one cannot build, except at very high cost, an implementation of *sign* that "toggles" at exactly 0; we shall therefore allow every occurrence of $\operatorname{sign}(A)$ to be replaced by $Q(A)$, where Q can vary from one time step to the next, or from one component of the algorithm to another, with only the restrictions that $Q(A) = \operatorname{sign}(A)$ for $|A| \ge \tau$ and $|Q(A)| \le 1$ for $|A| \le \tau$, where $\tau > 0$ is known. (Note that this allows for both continuous and

discontinuous Q; if we impose *a priori* that $Q(t)$ can take the values 1 and -1 only, then the restrictions reduce to the first condition.) Moreover, whenever our algorithm uses multiplication by some real-valued parameter P, we also allow for the replacement of P by $P(1+\epsilon)$, where ϵ can again vary, subject only to $|\epsilon| \leq \mu < 1$, where the tolerance μ is again known *a prioiri*. We can now formulate what we mean by robustness: despite all this wriggle room, we prove that (4) holds independently of the (possibly time-varying) values of all the ϵ and Q, within the constraints.

We conclude, in Section 4, with 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 $f \in \mathcal{C}_2 = \{h \in L^2;\ \|h\|_{L^\infty} \leq 1,$ 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

$$f(t) = \frac{1}{\lambda} \sum_n f\left(\frac{n}{\lambda}\right) g\left(t - \frac{n}{\lambda}\right)\ , \tag{5}$$

with g as described before; i.e.,

$$\hat{g}(\xi) = \frac{1}{\sqrt{2\pi}} \text{ for } |\xi| \leq \pi,\ \hat{g}(\xi) = 0 \text{ for } |\xi| > \lambda\pi \text{ and } \hat{g} \in C^\infty\ . \tag{6}$$

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

$$\widetilde{f}_{\mathbf{q}^\lambda}(t) = \frac{1}{\lambda} \sum q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \tag{7}$$

provides a good approximation to f. Taking simply $q_n^\lambda =$ sign $\left(f\left(\frac{n}{\lambda}\right)\right)$ does not work because there exist infinitely many independent bandlimited functions φ that are everywhere positive (such as the lowest order prolate spheroidal wave functions [16], [14] for arbitrary time intervals and symmetric frequency intervals contained in $[-\pi, \pi]$); 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 $\widetilde{f}_{\mathbf{q}^\lambda}$. 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

$$\begin{cases} u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) - q_n^\lambda \\ q_n^\lambda = \ \text{sign}\left(u_{n-1} + f\left(\frac{n}{\lambda}\right)\right) \ , \end{cases} \tag{8}$$

and with an "initial condition" u_0 arbitrarily chosen in $(-1,1)$. In circuit implementation, the range of n in (8) is $n \geq 1$. However, for theoretical reasons, we view (8) as defining the u_n and q_n for all n. At first glance, this means the u_n are defined implicitly for $n < 0$. However, as we shall see below, it is possible to write u_n and q_n directly in terms of u_{n+1} and f_{n+1} when $n < 0$.

We shall now show by a simple inductive argument that the u_n of (8) are all bounded by 1. We prove this in two steps:

LEMMA 2.1. *For any $f \in \mathcal{C}_1$ and $|u_0| < 1$, the sequence $(u_n)_{n\in\mathbb{N}}$ defined by the recursion* (8) *is uniformly bounded, $|u_n| < 1$ for all $n \geq 0$.*

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

For negative n, we first have to transform the system (8) into a recursion in the other direction. To do this, observe that for $n \geq 1$,

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

In all cases we have, thus, $\text{sign}\left(u_n - f\left(\frac{n}{\lambda}\right)\right) = -\ \text{sign}(u_{n-1} + f\left(\frac{n}{\lambda}\right))$. The recursion (8) therefore implies, for $n \geq 1$,

$$u_{n-1} = u_n - f\left(\frac{n}{\lambda}\right) - \ \text{sign}(u_n - f\left(\frac{n}{\lambda}\right)) \ , \tag{9}$$

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 2.2. *The recursion* (8)*, 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 $|f(t) - \widetilde{f}_{\mathbf{q}^\lambda}(t)|$.

PROPOSITION 2.3. *For $f \in \mathcal{C}_1$, $\lambda > 1$, define the sequence $\mathbf{q}^\lambda$ through the recurrence* (8), *with u_0 chosen arbitrarily in $(-1, 1)$. Let g be a function satisfying* (6). *Then*

$$\left| f(t) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| \leq \frac{1}{\lambda} \|g'\|_{L^1} \,. \tag{10}$$

Proof Using (5), summation by parts, and the bound $|u_n| < 1$, we derive

$$\begin{aligned} \left| f(t) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| &= \frac{1}{\lambda} \left| \sum_n \left(f\left(\frac{n}{\lambda}\right) - q_n^\lambda \right) g\left(t - \frac{n}{\lambda}\right) \right| \\ &= \frac{1}{\lambda} \left| \sum_n u_n \left(g\left(t - \frac{n}{\lambda}\right) - g\left(t - \frac{n+1}{\lambda}\right) \right) \right| \\ &\leq \frac{1}{\lambda} \sum_n \left| g\left(t - \frac{n}{\lambda}\right) - g\left(t - \frac{n+1}{\lambda}\right) \right| \\ &\leq \frac{1}{\lambda} \sum_n \int_{t - \frac{n+1}{\lambda}}^{t - \frac{n}{\lambda}} |g'(y)| dy = \frac{1}{\lambda} \|g'\|_{L^1}. \end{aligned}$$

□

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 (8); 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

$$\left| \sum_{n=1}^N f\left(\frac{n}{\lambda}\right) - \sum_{n=1}^N q_n^\lambda \right| = |u_N| < 1 \,; \tag{11}$$

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 [15]. 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 (8) 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 (10), i.e. bounds that decay faster as

$\lambda \to \infty$, can be obtained by replacing (8) 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 (6) because they require the full sequence $(q_n^\lambda)_{n\in\mathbb{Z}}$ to approximate even one value $f(t)$. 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 (6)). In this case, the reconstruction formula (5) no longer holds, and the approximation error has additional contributions. Suppose G is supported in $[-R,R]$, so that, for a given t, only the q_n^λ with $|t-\frac{n}{\lambda}| < R$ can contribute to the sum $\sum_n q_n^\lambda G(t-\frac{n}{\lambda})$. Then we have

$$\left|f(t)-\frac{1}{\lambda}\sum_n q_n^\lambda G\left(t-\frac{n}{\lambda}\right)\right| \leq \left|f(t)-\frac{1}{\lambda}\sum_n f\left(\frac{n}{\lambda}\right)G\left(t-\frac{n}{\lambda}\right)\right| + \frac{1}{\lambda}\left|\sum_n\left(f\left(\frac{n}{\lambda}\right)-q_n^\lambda\right)G\left(t-\frac{n}{\lambda}\right)\right| . \tag{12}$$

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

$$\begin{aligned}\left|f(t)-\frac{1}{\lambda}\sum_n f\left(\frac{n}{\lambda}\right)G\left(t-\frac{n}{\lambda}\right)\right| &= \frac{1}{\lambda}\left|\sum_n f\left(\frac{n}{\lambda}\right)\left[g\left(t-\frac{n}{\lambda}\right)-G\left(t-\frac{n}{\lambda}\right)\right]\right| \\ &\leq \frac{1}{\lambda}\sum_n\left|g\left(t-\frac{n}{\lambda}\right)-G\left(t-\frac{n}{\lambda}\right)\right| \leq \|G-g\|_{L^1}+\lambda^{-1}\|G'-g'\|_{L^1} .\end{aligned}$$

By imposing on G that the L^1 distance of G and G'/λ to g and g'/λ, respectively, be less than C/λ for at least one suitable g, we see that 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.)

In practical applications, one is generally interested only in approximating $f(t)$ for t after some starting time t_0, $t > t_0$. If finite filters are used this means that one needs the q_n^λ only for n exceeding some corresponding n_0. There is then no need to consider the "backwards" recursion (9), introduced to extend Lemma 2.1 (bound on the $|u_n|$ uniform in $n \geq 0$) to Proposition 2.2 (bound on the $|u_n|$ uniform in n).

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

$$\left| f\left(\frac{m}{\lambda}\right) - \sum_n q_n^\lambda G_{m-n}^\lambda \right| , \tag{13}$$

using discrete convolution with finite filters G^λ, rather than expressions of the type (10) or (11). If we were interested in optimizing constants relevant for practice, we should concentrate on (13) directly. For our present level of modeling however, in which we want to study the dominant behavior as a function of λ, working with (10) or (11), or their equivalent forms for higher order schemes, below, will suffice, since (13) 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 (6). Since such g are supported on all of $\mathbb{R}$, we will always need to define q_n for all $n \in \mathbb{Z}$ (rather than $\mathbb{N}$). For first-order $\Sigma\Delta$, we could easily "invert" the recursion so as to reach $n < 0$. For the higher order $\Sigma\Delta$ considered from Section 3 onwards, such an inversion is not straightforward; instead we will simply give, for every algorithm that defines q_n for $n \geq 0$, a parallel prescription that defines q_n for $n < 0$.

2.3. *More refined bounds.* In practice, one observes better behavior for $|f(t) - \widetilde{f}_{\mathbf{q}^\lambda}(t)|$ than that proved in Proposition 2.3. In particular, it is believed that, for arbitrary $f \in \mathcal{C}_1$,

$$\lim_{T\to\infty} \frac{1}{2T} \int_{|t|\leq T} \left| f(t) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right|^2 dt \leq \frac{C}{\lambda^3} , \tag{14}$$

with C independent of $f \in \mathcal{C}_1$ or of the initial condition u_0 for the recursion (8). Whether the conjecture (14) 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 $\mathcal{C}_2$) is still an open problem.

It is not surprising that a better bound than (10) would hold, since we used very little in its derivation. In particular, we never used explicitly that the $f\left(\frac{n}{\lambda}\right)$ were samples of the *entire* (because bandlimited) function f.

For some special cases, i.e. for very restricted classes of functions f, (14) 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 in [10]. 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], [9], 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 $|f'(t)|$ as well as ϵ; his g^λ (into which the $1/\lambda$ factor from (10) 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], [9].

2.4. *Robustness.* Remarkably, an iterative procedure very similar to (8) can be used to compute the binary expansion of a number in $(0, 1)$. Consider the recursion

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

with initial condition $\widetilde{u}_{-1} = \alpha/2$, $\widetilde{b}_0 = \operatorname{sign}(\alpha)$, and with the sequence $(x_n)_n$ defined by $x_0 = \alpha$, $x_n = 0$ for $n > 0$; here α is any number in $(-1, 1)$. By induction one derives again that $|\widetilde{u}_n| < 1$ for all n, so that

$$\begin{aligned} \left| 2\alpha - \sum_{n=0}^{N} 2^{-n} \widetilde{b}_n \right| &= \left| \alpha + \sum_{n=0}^{N} 2^{-n} (x_n - \widetilde{b}_n) \right| \\ &= |2\widetilde{u}_{-1} + \sum_{n=0}^{N} 2^{-n} (\widetilde{u}_n - 2\widetilde{u}_{n-1})| \\ &= |2^{-N} \widetilde{u}_N| < 2^{-N} \to 0 \ \text{ as } N \to \infty, \end{aligned}$$

which converges exponentially like a binary expansion. (Since the $\widetilde{b}_n \in \{-1, 1\}$, $\sum_{n=0}^{\infty} 2^{-n} \widetilde{b}_n$ is not quite a binary expansion; however, for $n \geq 1$, the $b_n = (1 + \widetilde{b}_{n-1})/2 \in \{0, 1\}$ are the digits for the binary expansion of $\frac{1+\alpha}{2}$.)

The only difference between the two recursions is the presence of the multiplications by 2 in (17). When the recursive equations are converted into block diagrams for circuits that would implement these recursions in practice, the diagram for (17) would require only one item more (a multiplier by 2) than the diagram for (8). The similarity of the two algorithms or circuits

seems to contradict the claim in the introduction, that $\Sigma\Delta$ quantization is much cheaper to implement than binary quantization of less frequent samples. However, the two algorithms behave very differently when imperfections, in particular imperfect quantizers, are introduced. Quantizers are never perfect. Although we desire to use $q(x) = \text{sign}(x)$ for our 1-bit quantizer, in practice we may have, e.g., $q(x) = \text{sign}(x + \delta)$, where δ is unknown except for the specification $|\delta| < \tau$; the value of δ may vary from one circuit to another, and it may even, due to thermal fluctuations, vary from one time step n to the next. More generally, we may have $Q(x) = \text{sign}(x)$ for $|x| \geq \tau$, whereas for $|x| \leq \tau$, we have only the bound $|Q(x)| \leq 1$. (Note that if Q is restricted to take only the values 1 and -1, the second condition is automatically satisfied, implying that for $|t| < \tau$, the behavior of $Q(t)$ can be completely arbitrary.) 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 well, then this will translate into an expensive circuit.

Let us replace the *sign* function in (8) by such a nonideal quantizer; the new recursion is then

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

and let us assume that, for all $n \in \mathbb{Z}$,

$$\begin{array}{rcll} Q_n(x) & = & \text{sign}(x) & \text{for } |t| \geq \tau \\ |Q_n(x)| & \leq & 1 & \text{for } |t| \leq \tau \ . \end{array} \tag{19}$$

It turns out that the u_n are then still bounded, uniformly, independently of the detailed behavior of Q_n, as long as (19) is satisfied:

LEMMA 2.4. *Let f be $\in \mathcal{C}_1$, let u_n, q_n be as defined in* (18)*, and let Q_n satisfy* (19) *for all n. If $|u_0| \leq 1 + \tau$, then $|u_n| \leq 1 + \tau$ for all $n \geq 0$.*

Proof. We use induction again. Suppose $|u_{n-1}| \leq \tau + 1$. Because $f \in \mathcal{C}_1$, $|f\left(\frac{n}{\lambda}\right)| \leq 1$. We now distinguish three cases. If $u_{n-1} + f\left(\frac{n}{\lambda}\right) > \tau$, then $u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) - 1 \in (\tau - 1, \tau + 1)$. Likewise, if $u_{n-1} + f\left(\frac{n}{\lambda}\right) < -\tau$, then $u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) + 1 \in (-\tau - 1, -\tau + 1)$. Finally, if $-\tau \leq u_{n-1} + f\left(\frac{n}{\lambda}\right) \leq \tau$, then $|Q_n(u_{n-1} + f\left(\frac{n}{\lambda}\right))| \leq 1$, so that $u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) - Q_n(u_{n-1} + f\left(\frac{n}{\lambda}\right)) \in (-\tau - 1, \tau + 1)$. □

Note that Lemma 2.4 holds regardless of how large τ is; even $\tau \gg 1$ is allowed. To discuss the case $n \leq 0$, we need to reconsider the recursion, because for generic Q_n, we can no longer "invert" the relationship between u_n and u_{n-1}. Therefore, we simply posit the following recursion for $n < 0$,

inspired by (9),

$$\left\{ \begin{array}{rcl} u_n & = & u_{n+1} - f\left(\frac{n+1}{\lambda}\right) + q_n \\ q_n & = & -Q_n\left(u_{n+1} - f(\frac{n+1}{\lambda})\right) \end{array} \right. . \tag{20}$$

An immediate generalization of Lemma 2.4 is then

LEMMA 2.5. *Let f be in $\mathcal{C}_1$, let u_n, q_n be as defined in* (18) *or* (20), *and let Q_n satisfy* (19) *for all $|n| > 1$. Assume also that $|u_0| \leq 1 + \tau$. Then $|u_n| \leq \tau + 1$ for all $n \in \mathbb{Z}$.*

By the same argument as in the proof of Proposition 2.3, Lemma 2.5 has as an immediate consequence the following:

COROLLARY 2.6. *Let f be in $\mathcal{C}_1$, let λ be > 1, and suppose g satisfies* (6). *Suppose, also, the sequence $(q_n^\lambda)_{n\in\mathbb{Z}}$ is generated by* (18), *with imperfect quantizers $Q_n(t)$ that satisfy* (19). *Then, for all $t \in \mathbb{R}$,*

$$\left| f(t) - \frac{1}{\lambda} \sum_n q_n^\lambda g\left(t - \frac{n}{\lambda}\right) \right| \leq \frac{1+\tau}{\lambda} \|g'\|_{L^1} . \tag{21}$$

If one replaces the "perfect" reconstruction function g by a suitable compactly supported G^λ, as in subsection 2.2, then one can also derive estimates similar to (21), exploiting the compactness of the support of 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-type schemes (17). Suppose we use (17) to generate bits $\widetilde{b}_n \in \{-1, 1\}$, and consider the approximation $\alpha_N = \sum_{n=0}^N 2^{-n}\widetilde{b}_n$ to the input α, as before; however, the quantizer has been changed to, say, $Q_n(t) = \text{sign}(t - \delta_n)$, with $|\delta_n| < \tau$. Suppose now $\alpha = \frac{\delta_0}{2}$; for the sake of definiteness, assume $\delta_0 > 0$. Then (17), with this imperfect quantizer, will give $\widetilde{b}_0 = -1$, so that $\alpha_N = \widetilde{b}_0 + \sum_{n=1}^N 2^{-n}\widetilde{b}_n \leq -2^{-N}$ for all N, implying $|\alpha - \alpha_N| > \frac{\delta_0}{2}$ for all N. The mistake made by the imperfect quantizer cannot be recovered by computing more bits, 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 2.3 suggests a mechanism by which better decay for $|f(t)-\widetilde{f}_{\mathbf{q}^\lambda}(t)|$ can 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. If we can work with k-th order (instead of first-order) differences of bounded sequences, then we obtain a λ^{-k} decay for $|f(t)-\widetilde{f}_{\mathbf{q}^\lambda}(t)|$ instead of the λ^{-1} decay of (10):

PROPOSITION 3.1. *Take $f\in\mathcal{C}_1$; take $\lambda>1$, and suppose g satisfies* (6). *Suppose that the $q_n^\lambda\in\{-1,1\}$ are such that there exists a bounded sequence $(u_n)_{n\in\mathbb{Z}}$ for which*

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

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

$$\left|f(t)-\frac{1}{\lambda}\sum_{n\in\mathbb{Z}}q_n^\lambda g\left(t-\frac{n}{\lambda}\right)\right|\le\frac{1}{\lambda^k}\|u\|_{l^\infty}\left\|\frac{d^kg}{dt^k}\right\|_{L^1}\ . \tag{23}$$

Proof. It follows from (22) that

$$\begin{aligned}\left|f(t)-\frac{1}{\lambda}\sum_n q_n^\lambda g\left(x-\frac{n}{\lambda}\right)\right| &= \frac{1}{\lambda}\left|\sum_n\Delta_n^k(u)g\left(t-\frac{n}{\lambda}\right)\right| \\ &= \frac{1}{\lambda}\left|\sum_n u_n\overline{\Delta}_n^k\left(g\left(t-\frac{\cdot}{\lambda}\right)\right)\right|\ ,\end{aligned} \tag{24}$$

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

$$\begin{aligned}\overline{\Delta}_n^k\left(g\left(t-\frac{\cdot}{\lambda}\right)\right) &= \sum_{l=0}^{k}(-1)^l\binom{k}{l}g\left(t-\frac{n+l}{\lambda}\right) \\ &= (-1)^k\frac{1}{\lambda^{k-1}}\int_0^{k/\lambda}g^{(k)}\left(t-\frac{n+k}{\lambda}+s\right)\varphi_k(\lambda s)ds\ ,\end{aligned} \tag{25}$$

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 $y \in \mathbb{R}$. It follows that we can estimate

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

□

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 u_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. [13]. We are thus concerned here with establishing the existence of stable $\Sigma\Delta$ schemes of arbitrary order. We first discuss the cases $k = 2$ and 3, before proceeding to general k.

3.2. *Second-order* $\Sigma\Delta$ *schemes.* We shall consider the recursion

$$\text{(26)}\qquad \begin{cases} v_n = v_{n-1} + x_n - q_n \\ u_n = u_{n-1} + v_n \\ q_n = \ \text{sign}[F(u_{n-1}, v_{n-1}, x_n)]\ , \end{cases}$$

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 $|x_n| \le a < 1$.

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

$$\text{(27)}\qquad F(u,v,x) = \gamma u + v + x\ ,$$

where γ is a fixed parameter. A detailed discussion of the mathematical properties of this family is given in [19]. Another very interesting choice, proposed by N. Thao [17], is

$$\text{(28)}\qquad F(u,v,x) = \frac{6x - 7\ \text{sign}(x)}{3} + \left(v + \frac{x + 3\ \text{sign}(x)}{2}\right)^2 + 2(1-|x|)u\ .$$

In both cases, one can prove that there exists a bounded set $A_a \subset \mathbb{R}^2$ so that if $|x_n| \le a$ for all n, and $(u_0, v_0) \in A_a$, then $(u_n, v_n) \in A_a$ for all $n \in \mathbb{N}$; see [19].

It follows that we have uniform boundedness for the u_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 (28) this λ^{-2} bound can be improved by a more detailed analysis; for constant input one achieves, in a root-mean-squared sense, a $\lambda^{-9/4+\epsilon}$ bound. Numerical observations suggest that this result can be improved to a $\lambda^{-5/2}$ decay rate for appropriately "balanced" F; they also suggest that this result can be extended to general band-limited functions (instead of constants). We refer to [11], [18], [19] 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 (28) or (30) (below). The analysis in [19] 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], [9]). This is mainly because even for $x_n \equiv x$ constant, the dynamical system (26) is much more complex than (8). In particular, the map

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

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

$$\begin{aligned} R_{2,x} \quad : \quad & \mathbb{R}^2 \to \mathbb{R}^2 \\ & \begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u + x - \ \operatorname{sign}(u + \frac{v}{2} + x) \\ v + u + x - \ \operatorname{sign}(u + \frac{v}{2} + x) \end{pmatrix} \end{aligned} \tag{29}$$

have invariant sets Γ_x that depend on the value of $x \in (-1, 1)$. 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, and we conjecture that it holds for a large family of F, even though we can prove only a few special cases – see below.) For $x \neq 0$, the Γ_x for (27) can have interesting fractal boundaries; for "large" x, these Γ_x are disconnected. (See Figure 1.)

On the other hand, the sets Γ_x for (28) are connected neighborhoods of $(0, 0)$ bounded by four parabolic arcs (see Figure 2); 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 [11]).

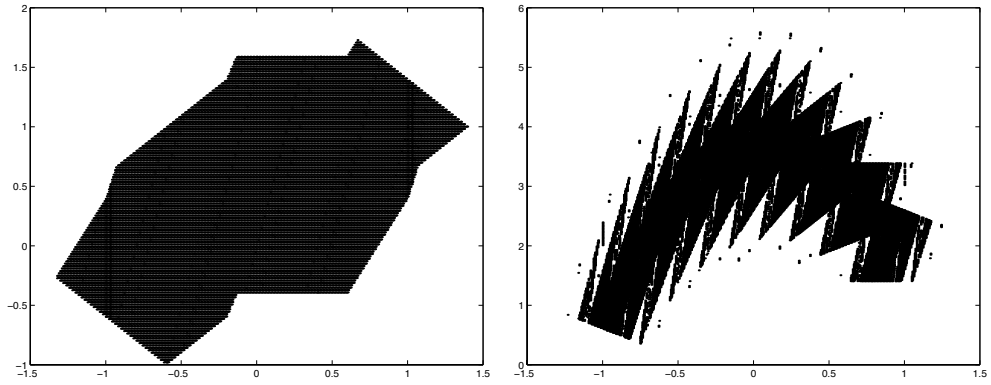

Figure 1. The attracting invariant sets Γ_x for two values of x (left: $x = .2$, right: $x = .8$) and for the choice (27) for F, with $\gamma = .5$. For $x = .2$, Γ_x is polygon, with sides that can be computed explicitly [11]; for $x = .8$, Γ_x is disconnected and fractal.

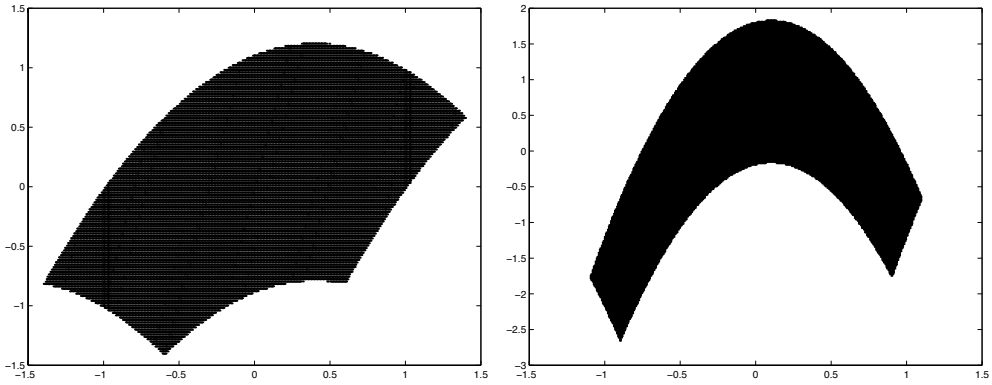

Figure 2. The attracting invariant sets Γ_x for two values of x (left: $x = .5$, right: $x = .8$) for the choice (28) for F.

Neither of the two schemes (27) or (28) is easy to generalize to higher order. We shall therefore concentrate our attention here on yet another choice for F,

$$F(u, v, x) = v + x + M \text{ sign}(u) \ , \tag{30}$$

with $M > 1$ to be fixed below. In addition, we shall also allow the *sign*-functions in (26) and (30) to be imperfect quantizers, and the multiplication by M to be imperfect as well. Our recursion thus reads, for $n > 0$,

$$\begin{cases} v_n = v_{n-1} + x_n - q_n \\ u_n = u_{n-1} + v_n \\ q_n = Q_n^1[v_{n-1} + x_n + M(1 + \epsilon_n)Q_n^2(u_{n-1})] \ , \end{cases} \tag{31}$$

where we assume that $Q_n^1,\ Q_n^2$ satisfy (19), and $|\epsilon_n| \le \mu < 1$.

The approach in [19] can be used to show that this second-order recursion does produce uniformly bounded u_n, v_n. We shall provide a different argument here, that, unlike the analysis in [19], generalizes to arbitrary order.

Prescribing initial values u_0, v_0 (or equivalently u_0, u_{-1}) the recursion (31) determines q_n, u_n, v_n, $n \geq 1$. In addition, we also need to give a prescription for $n \leq 0$. Observe that the equations for u_n, v_n can be rewritten as $u_n = 2u_{n-1} - u_{n-2} + x_n - q_n$; this suggests a symmetry between u_n and u_{n-2}. We use this to define the following recursion for u_n, q_n with $n < 0$,

$$\begin{cases} u_n &= 2u_{n+1} - u_{n+2} + x_{n+2} - q_{n+2} \\ q_{n+2} &= Q_n^1 \left[u_{n+1} - u_{n+2} + x_{n+2} + M(1+\varepsilon_n) Q_n^2(u_{n+1}) \right] \ , \end{cases}$$

to be used for $n \leq -2$. If we introduce also $v_n = u_n - u_{n+1}$ for $n < 0$, this becomes

$$\begin{cases} v_n &= v_{n+1} + x_{n+2} - q_{n+2} \\ u_n &= u_{n+1} + v_n \\ q_{n+2} &= Q_n^1 \left[v_{n+1} + x_{n+2} + M(1+\varepsilon_n) Q_n^2.(u_{n+1}) \right] \ , \end{cases} \tag{32}$$

We define $v_{-1} = -v_0$ and use this together with the already prescribed values u_0, u_{-1} in (32). This recursion will then serve to determine the values of q_j, u_j, v_j for $j \leq 0$. The sequences (u_n), (q_n) will then satisfy, for all n,

$$\Delta^2 u_n = x_n - q_n.$$

As pointed out at the end of Section 2.2, we introduce an algorithm to generate q_n for $n < 0$ because our approximation formula (5), using g supported on all of $\mathbb{R}$, requires them; in practice one uses only compactly supported G, and q_n with $n \leq 0$ are not needed. Since the negatively-indexed q_n are kept for only theoretical reasons, we would be justified in keeping the sign function "clean" in their recursion, i.e. without the $Q_n^1, Q_n^2, \varepsilon_n$ "imperfections"; we left them in for the sake of generality. It is clear, by comparing (32) with (31), that if we can prove that (31) implies uniform bounds on $|u_n|, |v_n|$ for $n > 0$, starting from some initial condition $|u_0| \leq U_0, |v_0| \leq V_0$ (with U_0, V_0 to be determined), then the same uniform bounds on $|u_n|, |v_n|$ for $n < 0$ will follow, provided $|u_{-1}| \leq U_0, |v_{-1}| \leq V_0$. Since $v_{-1} = -v_0$, we need to impose only the additional constraint $|u_0 + v_{-1}| = |u_0 - v_0| \leq U_0$ for this to hold. This will allow us to restrict our arguments to the $n > 0$ case. We then have:

PROPOSITION 3.2. *Suppose $|x_n| \leq a < 1$ for all $n \in \mathbb{Z}$. Let u_n, v_n, and q_n be defined as in* (31) *and* (32), *with $M \geq \frac{2a+\tau+1}{1-\mu}$. Then, if $|v_0| \leq M(1+\mu) + 1 + \tau$, there exists $|v_n| \leq M(1+\mu) + 1 + \tau$ for all $n \in \mathbb{Z}$. Moreover, if $|u_0|, |v_0| \leq \tau/2$, then $|u_n| \leq \tau + \frac{[M(1+\mu)+\tau+3/2-a/2]^2}{2(1-a)}$ for all $n \in \mathbb{Z}$.*

We start by proving a succession of lemmas, in each of which we make the same assumptions as in the statement of Proposition 3.2. The lemmas deal only with the case $n \in \mathbb{N}$.

LEMMA 3.3. *If $|v_0| \le M(1+\mu)+1+\tau$, then $|v_n| \le M(1+\mu)+1+\tau$ for all $n \in \mathbb{N}$.*

Proof. By induction. Suppose $|v_{n-1}| \le M(1+\mu)+1+\tau$. If $|v_{n-1}+x_n| > M(1+\epsilon_n)+\tau$, then

$$\begin{aligned} |v_n| = |v_{n-1}+x_n - Q_n^1(v_{n-1}+x_n)| &= |v_{n-1}+x_n| - 1 \\ &\le |v_{n-1}| + a - 1 < M(1+\mu)+1+\tau, \end{aligned}$$

where we have used that $|v_{n-1}+x_n| > \tau$. If $|v_{n-1}+x_n| \le M(1+\epsilon_n)+\tau$, then

$$|v_n| \le |v_{n-1}+x_n| + 1 \le M(1+\epsilon_n)+\tau+1 \le M(1+\mu)+1+\tau. \qquad \square$$

LEMMA 3.4. *Suppose $u_k \le \tau$, and $u_{k+1}, u_{k+2}, \dots, u_{k+L} > \tau$. 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 $l \in \{1, \dots, \kappa\}$ such that $v_{k+l} + x_{k+l+1} < -M(1-\mu)+1+a+\tau$.*

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

$$\begin{aligned} v_{k+\kappa} + x_{k+\kappa+1} &= v_{k+1} + \sum_{l=2}^{\kappa} (x_{k+l} - q_{k+l}) + x_{k+\kappa+1} \\ &\le M(1+\mu)+1+\tau+(\kappa-1)(a-1)+a \\ &< M(1+\mu)+1+\tau+a-(1-a)\frac{2M}{1-a} \\ &= -M(1-\mu)+1+a+\tau\,. \end{aligned}$$

$\square$

LEMMA 3.5. *Let $u_k, u_{k+1}, \dots, u_{k+L}$ be as in Lemma 3.4. If*

$$v_{k+l} + x_{k+l+1} < -M(1-\mu)+1+a+\tau$$

for some $l \in \{1, \dots, L\}$, then for all l' satisfying $l \le l' \le L$,

$$v_{k+l'} + x_{k+l'+1} < -M(1-\mu)+1+a+\tau\,.$$

Proof. By induction. Suppose $v_{k+n} + x_{k+n+1} < -M(1-\mu)+1+a+\tau$ with $n \in \{1, \dots, L-1\}$; we prove that this implies

$$v_{k+n+1} + x_{k+n+2} < -M(1-\mu)+1+a+\tau.$$

If $v_{k+n} + x_{k+n+1} \ge -M(1+\varepsilon_{n+k+1})+\tau$, then $q_{k+n+1} = 1$ (since $u_{k+n} > \tau$), hence

$$\begin{aligned} v_{k+n+1} + x_{k+n+2} &< -M(1-\mu)+1+a+\tau-1+x_{k+n+2} \\ &< -M(1-\mu)+1+a+\tau. \end{aligned}$$

On the other hand, if

$$v_{k+n} + x_{k+n+1} < -M(1+\varepsilon_{n+k+1})+\tau,$$

then

$$\begin{aligned} v_{k+n+1} + x_{k+n+2} &< -M(1+\varepsilon_{n+k+1}) + \tau + 1 + x_{k+n+2} \\ &\leq -M(1-\mu) + 1 + a + \tau. \end{aligned}$$

□

LEMMA 3.6. *Let $u_k, u_{k+1}, \ldots, u_{k+L}$ be as above. Then the v_{k+l} decrease monotonically in l, with $v_{k+l-1} - v_{k+l} \geq 1 - a$, until $v_{k+l} + x_{k+l+1}$ drops below $-M(1-\mu) + 1 + a + \tau$. All subsequent $v_{k+l'}$ with $l' \leq L$ remain negative.*

Proof. As long as $v_{k+n} + x_{k+n+1} \geq -M(1-\mu) + 1 + a + \tau$ with $n \leq L$, we have $q_{k+n+1} = 1$, so $v_{k+n} - v_{k+n+1} = -x_{k+n+1} + 1 \geq 1 - a$. If $v_{k+l} + x_{k+l+1} < -M(1-\mu) + 1 + a + \tau$, then $v_{k+l'} + x_{k+l'+1} < -M(1-\mu) + 1 + a + \tau$ by Lemma 3.5 if $l \leq l' \leq L$, so that $v_{k+l'} < -M(1-\mu) + 1 + 2a + \tau \leq 0$. □

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

Proof. We first discuss the case $n > 0$. The bound on v_n is proved in Lemma 3.3; we now turn to u_n. Suppose $u_{k+1}, \ldots, u_{k+L}$ is a stretch of $u_n > \tau$, preceded by $u_k \leq \tau$. We have then, for all $m \in \{1, \ldots, L\}$,

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

By Lemma 3.6, these v_{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, $u_{k+l} \leq u_{k+1} - (1-a)(l-1) \leq M(1+\mu) + 1 + \tau - (1-a)(l-1)$, at least until this last expression drops below zero. It follows that

$$\begin{aligned} u_{k+m} &\leq \tau + \max_{n \geq 1} \sum_{l=1}^{n} [M(1+\mu) + 1 + \tau - (1-a)(l-1)] \\ &\leq \tau + \frac{[M(1+\mu) + 3/2 - 1/2 + \tau]^2}{2(1-a)} \end{aligned} \tag{33}$$

The initial condition $|u_0| \leq \tau/2$ ensures that the upper bound (33) holds for all $u_n, n \geq 0$. The lower bound, $u_n \geq -\tau - \frac{[M(1+\mu)+3/2-a/2+\tau]^2}{2(1-a)}$ for $n \geq 0$, is proved entirely analogously.

To treat $n < 0$, note that the "initial conditions" for the recursion (32) satisfy $|v_{-1}| = |v_0| \leq \tau/2$, and $|u_{-1}| = |u_0 - v_0| \leq \tau$. It follows that we can repeat the same arguments to derive an identical bound on $|u_n|$ for $n \leq -1$. □

Remarks. 1. The bound on $|u_n|$ is significantly larger than that on $|v_n|$. For $a = .5$ and $\tau = \mu = .1$, for instance, and $M = (2a + \tau + 1)/(1-\mu) = 7/3$, we have $|v_n| \leq 10/3$ and $|u_n| \leq 12.6$. 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 $|v_0|, |u_0| \leq \tau/2$. If $|v_0| \leq M(1+\mu)+1+\tau$, and $|u_0| \leq A$, then $|u_0 - v_0| \leq A' = A + M(1+\mu)+1+\tau$, and we have $|u_n| \leq A' + [M(1+\mu)+\tau+3/2-a/2]^2/[2(1-a)]$ for all $n \in \mathbb{Z}$; moreover, once an index ℓ is reached for which u_ℓ and $u_{\ell+1}$ differ in sign, we have $|u_n| \leq \tau + [M(1+\mu)+\tau+3/2-a/2]^2/[2(1-a)]$ for all $n > \ell$ if ℓ is positive, or all $n < \ell$ if ℓ is negative.

3.3. *A third-order $\Sigma\Delta$ scheme.* Let us consider the construction we discussed for second order, but take it one step further. For $n > 0$ define the recursion

(34)

$$\begin{cases} u_n^{(1)} &= u_{n-1}^{(1)} + x_n - q_n \\ u_n^{(2)} &= u_{n-1}^{(2)} + u_n^{(1)} \\ u_n^{(3)} &= u_{n-1}^{(3)} + u_n^{(2)} \\ q_n &= Q_n^1 \left[u_{n-1}^{(1)} + x_n + M_1(1+\varepsilon_n^1) Q_n^2 \left(u_{n-1}^{(2)} + M_2(1+\varepsilon_n^2) Q_n^3(u_{n-1}^{(3)}) \right) \right] \end{cases}$$

where Q_n^1, Q_n^2, Q_n^3 satisfy (19), $|\varepsilon_n^1|, |\varepsilon_n^2| \leq \mu$, and 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 initial conditions $u_0^{(1)}, u_0^{(2)}, u_0^{(3)}$. We assume again that $|x_n| \leq a < 1$ for all $n \geq 0$.

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 us understand the more general construction, however.

First of all, exactly the same argument as in the proof of Lemma 3.3 establishes that $|u_n^{(1)}| \leq M_1(1+\mu)+1+\tau =: M_1'$.

Next, imagine a long stretch of $u_{n+1}^{(2)}, u_{n+2}^{(2)}, \ldots$, all $> M_2(1+\mu)+1+\tau$. Then the corresponding q_{n+l+1} are all automatically equal to 1, unless $u_{n+l}^{(1)} + x_{n+l} < -M_1(1+\epsilon_{n+l}^1)+\tau$. Arguments similar to those in the proofs of Lemmas 3.4–3.6 then show that if $u_{n+1}^{(1)} > -M_1(1-\mu)+1+a+\tau \geq 0$, 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 $-M_1(1-\mu)+1+a+\tau$ (in at most $\kappa_1 = \left\lfloor \frac{2M_1}{1-a} \right\rfloor + 2$ steps), after which all the subsequent $u_{n+l'}^{(1)}$ in the stretch are negative, provided we chose $M_1 \geq \frac{1+2a+\tau}{1-\mu}$. As before, this argument leads to $|u_n^{(2)}| \leq M_2' := M_2(1+\mu)+\tau+\frac{M_1'+(1-a)/2}{2(1-a)}$.

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 *a 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 a minimum speed for 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 that 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 $u_l^{(3)} \leq \tau, u_{l+1}^{(3)}, \ldots, u_{l+L}^{(3)} > \tau$. 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 $u_{l+m}^{(2)} \leq -M_2(1-\mu)+\tau$. Indeed, if $u_{l+1}^{(2)}, \ldots, u_{l+\kappa_2-1}^{(2)} > -M_2(1-\mu)+\tau$, then the corresponding q_{l+m} are 1, unless $u_{l+m-1}^{(1)} < -M_1(1-\mu)+a+\tau$. As before, this forces the $u_{l+m}^{(1)}$ down, until they hit below $-M_1(1-\mu)+a+\tau$ 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 $u_{l+1}^{(2)}, \ldots, u_{l+\kappa_2-1}^{(2)} > -M_2(1-\mu)+\tau$, then $u_{l+\kappa_2}^{(2)} \leq -M_2(1-\mu)+\tau$ must follow. Once $u_{l+l'}^{(2)}$ has dropped below $-M_2(1-\mu)+\tau$, 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 exceed τ again, causing the $u_{l+m}^{(2)}$ to increase. However, as soon as we have κ_1 consecutive $u_n^{(2)} > -M_2(1-\mu)+\tau$, we must have, for at least one of the corresponding indices, that $u_n^{(1)} < -M_1(1-\mu)+1+a+\tau$, which forces the subsequent $u_n^{(1)}$ below this value too, and we are back in our cycle forcing the $u_n^{(2)}$ down, until they hit below $-M_2(1-\mu)+\tau$. So if $-M_2(1-\mu)+\tau+\kappa_1 M_1' \leq 0$, then the $u_n^{(2)}$ do not get a chance to grow to positive values within the first κ_1 indices after $u_{l+l'}^{(2)} < -M_2(1-\mu)+\tau$. This forces *all* the $u_{l+m}^{(2)}$ 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 $u_{l+m}^{(3)}$.

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', and corresponding conditions on the M_j.

3.4. *Generalization to arbitrary order.* We assume again that $|x_n| \le a < 1$ 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 need to introduce a number of constants. As before, the $\Sigma\Delta$-scheme will use nonideal quantizers with an inherent imprecision limited by τ, and all the multipliers in the algorithm will be known only up to a factor $(1+\epsilon)$, where $|\epsilon| \le \mu < 1$. We pick α so that $2\alpha < 1 - \mu$, and we define

$$
\begin{aligned}
M_1 &:= 2\,\frac{1+a+\tau}{1-\mu} \qquad\qquad \kappa_1 := \left\lfloor \frac{2M_1+1+a}{1-a} \right\rfloor + 2 \\
B &:= \frac{4}{1-\mu-\alpha} \qquad\qquad M_j := M_1 B^{j-1} \nu^{(j-1)^2} \\
\nu &:= \left\lfloor \max\left(\frac{4B}{\kappa_1(1-\mu)} + \frac{\kappa_1^2}{B}, 1 + \frac{B(3-\alpha-\mu)}{\alpha\kappa_1}, \kappa_1 \right) \right\rfloor + 1
\end{aligned} \tag{35}
$$

where j ranges from 1 to J. For $n \ge 0$, the scheme itself is then defined as follows

(36)

$$
\left\{
\begin{aligned}
u_n^{(1)} &= u_{n-1}^{(1)} + x_n - q_n \\
u_n^{(j)} &= u_{n-1}^{(j)} + u_n^{(j-1)}, \qquad j = 2, \dots, J \\
q_n &= Q_n^1 \Big\{ u_{n-1}^{(1)} + M_1(1+\epsilon_n^1) Q_n^2 \Big[u_{n-1}^{(2)} + M_2(1+\epsilon_n^2) Q_n^3 \Big(u_{n-1}^{(3)} + \cdots \\
&\qquad + M_{J-2}(1+\epsilon_n^{J-2}) Q_n^{J-1} \Big(u_{n-1}^{(J-1)} + \cdots \\
&\qquad + M_{J-1}(1+\epsilon_n^{J-1}) Q_n^J (u_{n-1}^{(J)}) \Big) \cdots \Big) \Big] \Big\} \quad ,
\end{aligned}
\right.
$$

where $|\varepsilon_n^1|, |\varepsilon_n^2|, \dots, |\varepsilon_n^{J-1}| \le \varepsilon$ and $Q_n^1, \dots Q_n^J$ satisfy (19) for all n. We start with initial conditions $u_0^{(1)}, \dots, u_0^{(J)}$, and we apply (36) recursively to determine $q_j, u_j^{(1)}, \dots, u_j^{(J)}$ for $j = 1, 2, \dots$. Prescribing these initial conditions is equivalent to prescribing $u_0^{(J)}, \dots, u_{-J+1}^{(J)}$.

For $n < 0$, we mirror this system, obtaining

(37)

$$
\left\{
\begin{aligned}
u_n^{(1)} &= u_{n+1}^{(1)} + (-1)^J (x_{n+J} - q_{n+J}) \\
u_n^{(j)} &= u_{n+1}^{(j)} + u_n^{(j-1)}, \qquad j = 2, \dots, J \\
q_{n+J} &= (-1)^J Q_n^1 \Big\{ u_{n+1}^{(1)} + M_1(1+\epsilon_n^1) Q_n^2 \Big[u_{n+1}^{(2)} + M_2(1+\epsilon_n^2) Q_n^3 \Big(u_{n+1}^{(3)} + \cdots \\
&\qquad + M_{J-2}(1+\epsilon_n^{J-2}) Q_n^{J-1} \Big(u_{n+1}^{(J-1)} + M_{J-1}(1+\epsilon_n^{J-1}) Q_n^J (u_{n+1}^{(J)}) \Big) \cdots \Big) \Big] \Big\} .
\end{aligned}
\right.
$$

To set the recursion running for $n < 0$, we prescribe the mirrored initial conditions $u^{(j)}_{-J+1} = \sum_{l=1}^{j} (-1)^{j-l} u^{(l)}_0 \begin{pmatrix} j-1 \\ l-1 \end{pmatrix}$. These conditions are chosen to guarantee that $u^{(J)}_0, \ldots, u^{(J)}_{-J+1}$ are given the same values as in the prescription for the forward recurrence. We now use (37) recursively to generate the q_n, $n \leq 0$. If we take, for simplicity, $u^{(j)}_0 = 0$ for $j = 1, \ldots J$, then the "initial conditions" for the $n < 0$ recursion have likewise $u^{(j)}_{-J+1} = 0$ for $j = 1, \ldots J$. If we relax our constraints on the initial conditions somewhat, imposing $\left| u^{(j)}_0 \right| \leq A_j$ for appropriate A_j, then we also impose that $\left| \sum_{l=1}^{j} (-1)^{j-l} u^{(l)}_0 \begin{pmatrix} j-1 \\ l-1 \end{pmatrix} \right| \leq A_j$. In both cases, one readily sees, as before, that the proof of a uniform bound for the $|u^{(J)}_n|$ in the $n > 0$ recursion simultaneously provides the same uniform bound for the $|u^{(J)}_n|$ in the $n < 0$ recursion.

We then have the following proposition:

PROPOSITION 3.7. *Suppose $|x_n| \leq a < 1$ for all $n \in \mathbb{Z}$. Let M_j for $j = 1, \ldots, J$, be defined as in* (35), *let the imperfect quantizers $Q^1_n, \ldots Q^J_n$ satisfy* (19) *for all $n \in \mathbb{Z}$, 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* (36) *or* (37), *with initial conditions $u^{(j)}_0 = 0$ for $j = 1, \ldots, J$. Then $|u^{(J)}_n| \leq (2-\alpha) M_1 B^{J-1} \nu^{(J-1)^2}$ for all $n \in \mathbb{Z}$.*

Remarks. 1. Note that this scheme is slightly different from the ones considered so far, in that the formula for q_n includes $u^{(1)}_{n-1}$ only and not the combination $u^{(1)}_{n-1} + x_n$. 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 involving the $u^{(j)}_n$, 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.

2. In all the lemmas below, we treat the case $n \geq 0$ only. The case $n < 0$ is similar.

3. As in the second order case, it is not necessary (and in practice it would not be possible) to have initial conditions exactly zero. The bounds on the $|u^{(J)}_n|$ might increase somewhat in the initial regime if the $u^{(l)}_0$ are bounded but not zero, but essentially the estimates are the same.

The proof of Proposition 3.7 is essentially along the lines sketched for the third-order case, albeit more technical in order to deal with general J. 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 3.8. $|u_n^{(1)}| \le M_1(1+\mu) + 1 + a + \tau$ *for all* $n \in \mathbb{N}$.

Proof. The argument is very similar to that used in the proof of Lemma 3.3, except that x_n does not appear in the definition of q_n. We work by induction. Suppose $|u_{n-1}^{(1)}| \le M_1(1+\mu) + 1 + a + \tau$. If $|u_{n-1}^{(1)}| > M_1(1+\epsilon_n^1) + \tau$, then q_n and $u_{n-1}^{(1)}$ have the same sign, so that $|u_n^{(1)}| \le |u_{n-1}^{(1)}| - 1 + |x_n| \le |u_{n-1}^{(1)}| - 1 + a \le |u_{n-1}^{(1)}| \le M_1(1+\mu) + 1 + a + \tau$. If $|u_{n-1}^{(1)}| \le M_1(1+\epsilon_n^1) + \tau$, then $|u_n^{(1)}| \le |u_{n-1}^{(1)}| + 1 + a \le M_1(1+\mu) + 1 + a + \tau$. □

LEMMA 3.9. *If* $u_{n+1}^{(2)}, \dots, u_{n+N}^{(2)} > M_2(1+\mu) + \tau$, *with* $N \ge \kappa_1$, *then there must exist* $l \in \{1, \dots, \kappa_1\}$ *such that* $u_{n+l}^{(1)} < -M_1(1-\mu) + \tau$. *Moreover, for all* $l' \in \{l, \dots, N\}$, $u_{n+l'}^{(1)} < -M_1(1-\mu) + \tau + 1 + a$. *A similar statement holds if* $u_{n+1}^{(2)}, \dots, u_{n+N}^{(2)} < -M_2(1+\mu) - \tau$, *and other signs are reversed accordingly.*

Proof. The argument is again similar to the proofs of Lemmas 3.4–3.5. Suppose $u_{n+1}^{(1)}, \dots, u_{n+\kappa_1-1}^{(1)}$ are all $\ge -M_1(1-\mu) + \tau$. Then we have $q_{n+2} = \cdots = q_{n+\kappa_1} = 1$. Hence

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

This establishes that $u_{n+l}^{(1)} < -M_1(1-\mu) + \tau$ for some $l \in \{1, \dots, \kappa_1\}$. Next, suppose that $u_{n+r}^{(1)} < -M_1(1-\mu) + \tau + 1 + a$, for some r with $l \le r \le N-1$. If $u_{n+r}^{(1)} \ge -M_1(1-\mu) + \tau$, then $q_{n+r+1} = 1$, hence

$$u_{n+r+1}^{(1)} = u_{n+r}^{(1)} + x_{n+r+1} - 1 < u_{n+r}^{(1)} < -M_1(1-\mu) + \tau + 1 + a;$$

if $u_{n+r}^{(1)} < -M_1(1-\mu) + \tau$, then

$$u_{n+r+1}^{(1)} < -M_1(1-\mu) + \tau + 1 + |x_{n+r+1}| \le -M_1(1-\mu) + \tau + 1 + a.$$

In both cases, $u_{n+r+1}^{(1)} < -M_1(1-\mu) + \tau + 1 + a$, and we continue by induction. □

Next we introduce auxiliary constants, for $j = 1, \dots, J$:

$$\kappa_j := \nu^{2(j-1)}\kappa_1 \tag{38}$$

$$M_1' := (1+\mu)M_1 + \tau + 1 + a \qquad M_j' := (1+\mu)M_j + \tau + \kappa_{j-1}M_{j-1}' \text{ for } j \ge 2$$

$$M_1'' := (1-\mu)M_1 - \tau - 1 - a \qquad M_j'' := (1-\mu)M_j - \tau - \kappa_{j-1}M_{j-1}' \text{ for } j \ge 2$$

$$\widetilde{M}_j := M_j(1+\mu) + \tau$$

$$\widetilde{m}_j := M_j(1-\mu) - \tau \quad .$$

These have been tailored so that

LEMMA 3.10. *The constants defined above by* (37) *satisfy, for* $j = 2, \ldots, J$,

$$(1-\mu)M_j > \tau + \kappa_{j-1}(2-\alpha)M_{j-1}, \tag{39}$$

$$M_j' \leq (2-\alpha)M_j, \tag{40}$$

$$\kappa_j - \kappa_{j-1} \geq \frac{\widetilde{m}_j + M_j'}{M_{j-1}''}\,. \tag{41}$$

Proof. The first equation is proved by straight substitution:

$$\begin{aligned}
&(1-\mu)M_j - \tau - \kappa_{j-1}(2-\alpha)M_{j-1} \\
&\quad = B^{j-1}\nu^{(j-1)^2}M_1\left[1-\mu-\frac{\tau}{\nu^{(j-1)^2}B^{j-1}M_1} - \frac{(2-\alpha)\kappa_1}{B\nu}\right] \\
&\quad \geq B^{j-1}\nu^{(j-1)^2}M_1\left[1-\mu-\frac{\tau/M_1 + (2-\alpha)\kappa_1}{B\nu}\right] \\
&\quad \geq B^{j-1}\nu^{(j-1)^2}M_1\left[1-\mu-\frac{2(2-\alpha)(1-\alpha-\mu)}{4}\right] \;\geq\; \alpha M_j\ .
\end{aligned} \tag{42}$$

The second equation is proved by induction. First we consider the case $j = 2$:

$$M_2' - (2-\alpha)M_2 = (\mu+\alpha-1)M_2 - \tau - \kappa_1 M_1' < -\alpha M_2 - \tau - \kappa_1 M_1' < 0\ .$$

Now suppose that $M_j' \leq (2-\alpha)M_j$ holds for some $j \geq 2$. Then (42) immediately implies that

$$M_{j+1}'' > (1-\mu)M_{j+1} - \tau - \kappa_j(2-\alpha)M_j \geq \alpha M_{j+1}\ ,$$

leading to

$$M_{j+1}' = 2M_{j+1} - M_{j+1}'' \leq (2-\alpha)M_{j+1}\ .$$

It remains to prove the third inequality. Because the definition of M_{j-1}'' is slightly different for $j = 2$ than for $j > 2$, we handle the case $j = 2$ separately. Now

$$\begin{aligned}
M_1''(\kappa_2-\kappa_1) - \widetilde{m}_2 - M_2' &= M_1''\kappa_2 - 2M_1\kappa_1 - 2M_2 \\
&= (a+1+\tau)\nu^2\kappa_1 - 2M_1\kappa_1 - 2\nu B M_1 \\
&= (a+1+\tau)\left[\nu\left(\nu\kappa_1 - \frac{4B}{1-\mu}\right) - \frac{4\kappa_1}{1-\mu}\right] > 0\ ,
\end{aligned}$$

where we have used $\nu > \frac{4B}{\kappa_1(1-\mu)} + \frac{\kappa_1^2}{B}$.

For $j > 2$ we use $M'_j \le (2-\alpha)M_j$ and $M''_{j-1} \ge \alpha M_{j-1}$ to upper bound the right-hand side of (40), and we replace the various κ_s and M_s by their definitions; then we see that the equation holds if $\nu\kappa_1(1-\nu^{-2}) \ge B(3-\alpha-\mu)\alpha^{-1}$, or, equivalently, if $\nu^2 \ge B(3-\alpha-\mu)\nu(\alpha\kappa_1)^{-1}+1$. From the definition of ν one easily checks that this is indeed the case, completing the proof. □

We are now ready to state and prove our general lemmas, used in the induction argument of the proof of the proposition.

LEMMA 3.11 (j). *For all $n \in \mathbb{N}$, $|u_n^{(j)}| \le M'_j$.*

LEMMA 3.12 (j). *If $u_{n+1}^{(j+1)}, \dots, u_{n+N}^{(j+1)} > \widetilde{M}_{j+1}$, with $N \ge \kappa_j$, then there must be $l \in \{1, \dots, \kappa_j\}$ so that $u_{n+l}^{(j)} < -\widetilde{m}_j$. For all $l' \in \{l, \dots, N\}$, moreover, $u_{n+l'}^{(j)} < -M''_j$. A similar statement holds if $u_{n+1}^{(j+1)}, \dots, u_{n+N}^{(j+1)} < -\widetilde{M}_{j+1}$, and other signs are reversed appropriately.*

Our induction argument then alternates two steps:

Step a. Lemma 3.11(j) + Lemma 3.12(j) imply Lemma 3.11$(j+1)$.

Step b. Lemmas 3.11(k) + 3.12(k) for $k \le j$, together with Lemma 3.11$(j+1)$, imply Lemma 3.12$(j+1)$.

Since the case $j = 1$ is established (see Lemmas 3.8, 3.9), induction will ultimately get us to a proof of Lemma 3.11(J), establishing $|u_n^{(J)}| \le M'_J$. By (39) this then completes the proof of Proposition 3.7. It remains to prove Steps a and b.

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

Assume $u_n^{(j+1)} \le \widetilde{M}_{j+1}$, and $u_{n+1}^{(j+1)}, \dots, u_{n+N}^{(j+1)} > \widetilde{M}_{j+1}$. We need to show that none of these $u_{n+l}^{(j+1)}, l = 1, \dots, N$, 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 3.12$(j)$, at most the first κ_j terms in this sum can be positive, and each of these is bounded by M'_j by Lemma 3.11(j). Therefore, for each $l \in \{1, \dots, N\}$,

$$u_{n+l}^{(j+1)} \le \widetilde{M}_{j+1} + \kappa_j M'_j = M'_{j+1}\,. \qquad \square$$

Proof of Step b. This step is the most complicated. In order to prove it, we invoke a third technical lemma, that will itself be proved by induction. We put ourselves in the framework where Lemmas 3.11(k) are proved for $k \le j+1$, as well as Lemmas 3.12(k) for $k \le j$.

LEMMA 3.13 $(j+1)$. *Let $j \in \{1,\dots,J-2\}$ be fixed, and assume $k \in \{1,\dots,j\}$. Suppose $u_{n+1}^{(j+2)},\dots,u_{n+N}^{(j+2)} > \widetilde{M}_{j+2}$ 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 $M \geq \kappa_k$;*
- *$u_r^{(l)} \geq -\widetilde{m}_l$ 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 $u_r^{(k)} < -\widetilde{m}_k$. Moreover, once $u_r^{(k)} < -\widetilde{m}_k$, for an $r \in S$, then we have $u_{r'}^{(k)} \leq -M_k''$ for all $r' \in S, r' \geq r$.

Proof. By induction on k. We assume Lemmas 3.11(j') and 3.12(j') hold for $j' \leq j+1$ and $j' \leq j$ respectively.

1. The case $k = 1$.

   • We have $u_s^{(j')} \geq -\widetilde{m}_{j'}$ for all $s \in S$, and all $j' \in \{2,\dots,j\}$. We must prove that if there are $\kappa_1 - 1$ consecutive elements in S, numbered $r+1,\dots,r+\kappa_1-1$, for which $u_{r+1}^{(1)},\dots,u_{r+\kappa_1-1}^{(1)} \geq -\widetilde{m}_1$, then necessarily $u_{r+\kappa_1}^{(1)} < -\widetilde{m}_1$.

   If $u_{r+1}^{(1)},\dots,u_{r+\kappa_1-1}^{(1)} \geq -\widetilde{m}_1$, then $q_{r+2} = \cdots = q_{r+\kappa_1} = 1$ (because all the indices are in S, so that for each s, $u_s^{(j')} \geq -\widetilde{m}_{j'}$ if $j' \in \{2,\dots,j+1\}$, and $u_s^{(j+2)} > \widetilde{M}_{j+2}$). It follows that

   $$u_{r+\kappa_1}^{(1)} = u_{r+1}^{(1)} + \sum_{m=r+2}^{r+\kappa_1} (x_m - q_m) \leq M_1' + (\kappa_1 - 1)(a-1) < -\widetilde{m}_1 \tag{43}$$

   • Next we must show that if $u_r^{(1)} < -\widetilde{m}_1$ for some $r \in S$, then $u_{r'}^{(1)} < -M_1''$ for $r' \geq r, r' \in S$.

   This is again done as in the proof of Lemma 3.8, by induction on r':

   – assume $u_{r'-1}^{(1)} < -M_1''$,

   – if $u_{r'-1}^{(1)} < -\widetilde{m}_1$, then $u_{r'}^{(1)} < -\widetilde{m}_1 + a + 1 = -M_1''$, if $u_{r'-1}^{(1)} \geq -\widetilde{m}_1$, then $q_{r'} = 1$ and $u_{r'}^{(1)} = u_{r'-1}^{(1)} + a - 1 \leq -M_1'' + a - 1 < -M_1''$.

   This completes the proof of the case $k = 1$ of Lemma 3.13$(j+1)$.

2. Suppose the lemma holds for $k = 1,\dots,k_0-1$, with $2 \leq k_0 \leq j$. Let us 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 $u_r^{(k_0)} < -\widetilde{m}_{k_0}$. That is, we must prove that if there exist $u_{s+1}^{(k_0)}, \dots, u_{s+\kappa_{k_0}-1}^{(k_0)}$ that are all $\geq -\widetilde{m}_{k_0}$, then $u_{s+\kappa_{k_0}}^{(k_0)}$ must be $< -\widetilde{m}_{k_0}$.

Define $\widetilde{S} = \{s+1, \dots, s+\kappa_{k_0}-1\} \subset S$. Then $\widetilde{S}$ satisfies all the requirements in Lemma 3.13$(j+1)$ 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 $\widetilde{S}$ such that $u_t^{(k_0-1)} < -\widetilde{m}_{k_0-1}$. Moreover, for all $t' \in \widetilde{S}$ exceeding this t, $u_{t'}^{(k_0-1)} < -M''_{k_0-1}$. It follows that

$$\begin{aligned} u_{s+\kappa_{k_0}}^{(k_0)} &= u_{t-1}^{(k_0)} + \sum_{t'=t+1}^{s+\kappa_{k_0}-1} u_{t'}^{(k_0-1)} + u_t^{(k_0-1)} + u_{s+\kappa_{k_0}}^{(k_0-1)} \\ &< M'_{k_0} - (\kappa_{k_0} - 1 - \kappa_{k_0-1}) M''_{k_0-1} - \widetilde{m}_{k_0-1} + (-M''_{k_0-1} + \widetilde{m}_{k_0-1}) \\ &= M'_{k_0} - (\kappa_{k_0} - \kappa_{k_0-1}) M''_{k_0-1} \\ &\leq M'_{k_0} - \frac{\widetilde{m}_{k_0} + M'_{k_0}}{M''_{k_0-1}} M''_{k_0-1} = -\widetilde{m}_{k_0} \ , \end{aligned}$$

where in the first inequality, we used Lemma 3.12 $(k_0 - 1)$ to bound each of the entries in the sum and we bounded the last term by writing $u_{s+\kappa_{k_0}}^{(k_0-1)} = u_{s+\kappa_{k_0}-1}^{(k_0-1)} + u_{s+\kappa_{k_0}-1}^{(k_0-2)} \leq -M''_{k_0-1} + M'_{k_0-2}$, and using $M'_{k_0-2} < \widetilde{m}_{k_0-1}$ if $k_0 > 2$; if $k_0 = 2$, we use instead $u_{s+\kappa_2}^{(1)} \leq u_{s+\kappa_2-1}^{(1)} + 1 + a \leq -M''_1 + 1 + a < -M''_1 + \widetilde{m}_1$. In the second inequality of the derivation, we used Lemma 3.10.

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

For $r' > r$, let $r'' = \max\{t \leq r'; u_t^{(k_0)} < -\widetilde{m}_{k_0}\}$. Then $u_{r''+1}^{(k_0)}, \dots, u_{r'-1}^{(k_0)} \geq -\widetilde{m}_{k_0}$. 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 $u_t^{(k_0-1)} < -\widetilde{m}_{k_0-1}$, and all later t' in the stretch will have $u_{t'}^{(k_0-1)} \leq -M''_{k_0-1}$. It follows that the $u_{r''+1}^{(k_0)}, \dots, u_{r'-1}^{(k_0)}$ cannot increase after the first $\kappa_{k_0-1} - 1$ entries:

$$\begin{aligned} \max\left[u_{r''+1}^{(k_0)}, \dots, u_{r'-1}^{(k_0)}\right] &\leq \max\left[u_{r''+1}^{(k_0)}, \dots, u_{r''+\kappa_{k_0-1}-1}^{(k_0)}\right] \\ &\leq u_{r''}^{k_0} + \max_{l \in \{1,\dots,\kappa_{k_0-1}-1\}} \sum_{l'=1}^{l} u_{r''+l'}^{(k_0-1)} \\ &< -\widetilde{m}_{k_0} + (\kappa_{k_0-1} - 1) M'_{k_0-1} \ . \end{aligned}$$

Hence $u^{(k_0)}_{r'} \le u^{(k_0)}_{r'-1} + M'_{k_0-1} < -\widetilde{m}_{k_0} + \kappa_{k_0-1}M'_{k_0-1} = -M''_{k_0}$. This completes the proof of Lemma 3.13$(j+1)$. □

We can now use this to complete the

Proof of Step b. Assume Lemmas 3.11(j') and 3.12(j') hold for $j' \le j$, as well as Lemma 3.11$(j+1)$. This also allows us to use Lemma 3.13(j') for $j' \le j+1$.

- Suppose now $u^{(j+2)}_{n+1}, \dots, u^{(j+2)}_{n+N} > \widetilde{M}_{j+2}$ with $N \ge \kappa_{j+1}$. We have to prove that among the first κ_{j+1} elements of this stretch, we have one for which $u^{(j+1)}_{n+l} < -\widetilde{m}_{j+1}$. As usual, we assume $u^{(j+1)}_{n+1}, \dots, u^{(j+1)}_{n+\kappa_{j+1}-1} \ge -\widetilde{m}_{j+1}$ (and we need to establish $u^{(j+1)}_{n+\kappa_{j+1}} < -\widetilde{m}_{j+1}$). Define S by $S = \{n+1, \dots, n+\kappa_{j+1}-1\}$, and fix $k = j$. Then S, k satisfy all the conditions in Lemma 3.13 $(j+1)$. It follows that at most the first $\kappa_j - 1$ elements of S can correspond to $u^{(j)}_r \ge -M''_j$. Therefore the max of $\{u^{(j+1)}_t; t \in S\}$ must be achieved among the first $\kappa_j - 1$ elements, and

$$\begin{aligned} u^{(j+1)}_{n+\kappa_{j+1}} &< \max\{u^{(j+1)}_t; t \in \{n+1, \dots, n+\kappa_j+1\}\} \\ &\quad -(\kappa_{j+1} - \kappa_j - 1)M''_j \le M'_{j+1} - (\kappa_{j+1} - \kappa_j)M''_j \le \widetilde{m}_{j+1} \end{aligned}$$

where we have used Lemma 3.10.

- Next, we need to prove that if $u^{(j+1)}_{n+l} < -\widetilde{m}_{j+1}$ for some $l \in \{1, \dots, N\}$, then $u^{(j+1)}_{n+l'} \le -M''_{j+1}$ for $l' \in \{l, \dots, N\}$. Define $l'' := \max\{t \le l' : u^{(j+1)}_{n+t} < -\widetilde{m}_{j+1}\}$. Then $u^{(j+1)}_{n+l''+1}, \dots, u^{(j+1)}_{n+l'-1} \ge -\widetilde{m}_{j+1}$. Again, the max of these must be obtained among the first $\kappa_j - 1$ entries (since after that, the $u^{(j+1)}_s$ must decrease monotonely), so that

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

- We have thus proved Lemma 3.12$(j+1)$, completing the proof of Step b in our induction process. □

Remarks. 1. There is clearly a lot of room for obtaining tighter bounds. We have not been able to reduce the growth in J of the exponent of ν below a quadratic, however, even in the "perfect" case, when $\tau = \mu = 0$. We shall come back to this, and its implications, in the next section.

2. As in the lower order special cases, it is not really crucial to start with $u_{-1}^{(j)} = 0$; 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 (36) 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. (Note added in revision: in very recent work [12], Güntürk has constructed a $\Sigma\Delta$ scheme with filters that achieves better bounds; see below.)

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 3.7 leads, for $f \in \mathcal{C}_1$ with $\|f\|_{L^\infty} \le a < 1$, to the estimate

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

where we have absorbed the bound on $\|\frac{d^k g}{dx^k}\|_{L^1}$ into γ^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 (36), 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

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

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 do not 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 (without filters)

in general. In [1] it is proved that 1-bit quantization schemes that allow convolutional approximation formulas can never obtain the optimal accuracy of binary expansions. On the other hand, sub-optimal but still exponential decay in λ is not excluded. In fact, the filter-$\Sigma\Delta$ scheme in [12] achieves such exponential decay (although it is no longer robust in the sense of this paper). It would be interesting 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) is given in [3], but there are still many open problems.

Acknowledgments. We thank Sinan Güntürk and Nguyen Thao for many 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 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.

Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ

E-mail address: ingrid@math.princeton.edu

Industrial Mathematics Institute and Mathematics Department, University of South Carolina, Columbia, SC

E-mail address: devore@math.sc.edu

References

[1] A. R. Calderbank and I. Daubechies, The pros and cons of democracy, *IEEE Trans. Inform. Theory* **48** (2002), 1721–1725.

[2] J. C. Candy and G. C. Temes (Editors), *Oversampling Delta-Sigma Data Converters' Theory, Design, and Simulation*, IEEE Press, New York, 1992.

[3] I. Daubechies, R. DeVore, C. Güntürk, and V. Vaishampayan, Exponential precision in A/D conversion with an imperfect quantizer, preprint.

[4] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, *Grundlehren Math. Wiss.* **303**, Springer-Verlag, New York, 1993.

[5] R. M. Gray, Spectral analysis of quantization noise in single-loop sigma-delta modulator with dc input, *IEEE Trans. on Commun.* **COM-37** (1989), 588–599.

[6] R. M. Gray, W. Chou, and P. W. Wong, Quantization noise in single-loop sigma-delta modulation with sinusoidal inputs, *IEEE Trans. on Commun.* **COM-37** (1989), 956–968.

[7] C. S. Güntürk, Improved error estimates for first order sigma-delta systems, *Internat. Workshop on Sampling Theory and Applications* (SampTA′99), Loen, Norway, August 1999.

[8] ———, Harmonic analysis of two problems in signal quantization and compression, Ph.D. thesis, Program in Applied and Computational Mathematics, Princeton University, 2000.

[9] ———, Approximating a bandlimited function using very coarsely quantized data: Improved error estimates in sigma-delta modulation, *J. Amer. Math. Soc.*, to appear.
[10] C. S. Güntürk, J. C. Lagarias, and V. Vaishampayan, On the robustness of single loop sigma-delta modulation, *IEEE Trans. Inform. Theory* **47** (2001), 1735–1744.
[11] C. S. Güntürk and N. T. Thao, Refined analysis of MSE in second order sigma delta modulation with DC inputs, preprint.
[12] C. S. Güntürk, One-bit sigma-delta quantization with exponential accuracy, *Commun. Pure Appl. Math.* **56**, no. 11 (2003), 1608–1630
[13] S. Hein and A. Zakhor, On the stability of sigma delta modulators, *IEEE Transactions on Signal Processing* **41** (1993), 2322–2348.
[14] H. Landau and H. O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty, II, *Bell System Tech. J.* **40** (1961), 65–84.
[15] S. R. Norsworthy, R. Schreier, and G. C. Temes (Editors), *Delta-Sigma Data Converters Theory, Design and Simulation*, IEEE Press, New York, 1997.
[16] D. Slepian and H. O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty. I, *Bell System Tech. J.* **40** (1961) 43–64.
[17] N. T. Thao, Quadratic one-bit second order sigma-delta modulators, preprint.
[18] N. T. Thao, C. Güntürk, I. Daubechies, and R. DeVore, A new approach to one-bit nth order $\Sigma\Delta$-modulation, in preparation.
[19] O. Yilmaz, Stability analysis for several second-order sigma-delta methods of coarse quantization of bandlimited functions, *Constr. Approx.* **18** (2002), 599–623.

(Received October 29, 2001)
(Revised December 2, 2002)