JOURNAL OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 17, Number 1, Pages 229–242
S 0894-0347(03)00436-3
Article electronically published on August 1, 2003

APPROXIMATING A BANDLIMITED FUNCTION USING VERY COARSELY QUANTIZED DATA: IMPROVED ERROR ESTIMATES IN SIGMA-DELTA MODULATION

C. SİNAN GÜNTÜRK

1. Introduction

This paper concerns fine analytical error estimates in analog-to-digital conversion of bandlimited functions using the method of sigma-delta modulation (also called $\Sigma\Delta$ quantization). This method has found widespread usage in practice due to several advantages in its implementation compared to conventional methods (see [1, 11]). A recent mathematical treatment of this problem appears in [3]. Below we give a quick introduction as well as setting our notation.

We define the class $\mathcal{B}_\Omega$ of Ω-bandlimited functions to be the space of real-valued continuous functions in $L^\infty(\mathbb{R})$ whose Fourier transforms (as distributions) have supports contained in $[-\Omega, \Omega]$. We denote the Fourier transform of x by $\hat{x}$, which is defined by

$$\hat{x}(\xi) = \int_{-\infty}^{\infty} x(t) e^{-i\xi t}\, dt$$

for $x \in L^1(\mathbb{R})$ and extended to the space of tempered distributions in the usual way. It is known via the Paley-Wiener-Schwartz theorem that any function $x \in \mathcal{B}_\Omega$ is the restriction to $\mathbb{R}$ of an entire function of exponential type Ω. Finiteness of the range of frequencies and boundedness of the amplitude make $\mathcal{B}_\Omega$ a useful model space for audio signals. Throughout this paper, we normalize the bandwidth by setting $\Omega = \pi$.

Sampling is the first and most basic step of almost any analog-to-digital conversion algorithm. For any $\lambda > 0$, referred to as the sampling rate, it involves the operator $S_\lambda : C(\mathbb{R}) \to \mathbb{R}^{\mathbb{Z}}$ given by

$$(S_\lambda x)_n := x(\tfrac{n}{\lambda}).$$

The sampling operation can be inverted for bandlimited functions. Indeed, for a Schwartz function φ, let $T_{\lambda,\varphi} : \ell^\infty(\mathbb{Z}) \to C^\infty(\mathbb{R})$ denote the interpolation operator given by

$$(T_{\lambda,\varphi} s)(t) := \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} s_n \varphi(t - \tfrac{n}{\lambda}).$$

Received by the editors April 11, 2003.
2000 *Mathematics Subject Classification.* Primary 94A20, 11K06; Secondary 11L07, 41A25.
Key words and phrases. A/D conversion, sigma-delta modulation, sampling, quantization, uniform distribution, discrepancy, exponential sums.
The author's research was supported in part by the Francis Robbins Upton honorific fellowship from Princeton University, the NSF Grant 97-29992 at the Institute for Advanced Study, and the NSF Grant DMS-0219072.

If, in addition, $\hat{\varphi}$ satisfies

$$\hat{\varphi}(\xi) = \begin{cases} 1, & \text{if } |\xi| \le \pi, \\ 0, & \text{if } |\xi| \ge \lambda_0 \pi, \end{cases} \tag{1}$$

for some $\lambda_0 > 1$, then

$$T_{\lambda,\varphi} S_\lambda \, x = x, \quad \text{for all } x \in \mathcal{B}_\pi \text{ and } \lambda \ge \lambda_0.$$

This is the *sampling theorem*, and φ is called the reconstruction kernel. The proof follows easily via taking the Fourier transform of both sides and identifying the Fourier series expansion of $\hat{x}$ on $[-\lambda\pi, \lambda\pi]$. (See, e.g., [9] for a complete proof.)

At the heart of analog-to-digital conversion lies quantization, which is the reduction of the sample values from their continuous range $\mathbb{R}$ to a discrete set A. We denote this operation by the mapping $Q : \mathbb{R}^{\mathbb{Z}} \to A^{\mathbb{Z}}$, whose action is specific to each algorithm. We then set

$$\tilde{x}_\lambda := \tilde{x}_{\lambda,Q} := T_{\lambda,\varphi} Q S_\lambda x,$$

which represents an approximate reconstruction after quantization. Note that Q is necessarily a nonlinear operator. If $(Qs)_n$ depends only on s_n, then the operation is said to be memoryless. The operator Q_δ defined by

$$(Q_\delta s)_n := \delta \left\lfloor \frac{s_n}{\delta} \right\rfloor,$$

where $\lfloor w \rfloor$ denotes the greatest integer less than or equal to w, is the most basic example of a memoryless quantization operator. Note that $\|s - Q_\delta s\|_{\ell^\infty} \le \delta$. Due to the decay and regularity of φ, one has

$$\sup_{\lambda \ge \lambda_0} \sup_{t \in \mathbb{R}} \frac{1}{\lambda} \sum_n |\varphi(t - \tfrac{n}{\lambda})| =: C_\varphi < \infty,$$

which implies that the operator $T_{\lambda,\varphi}$ is bounded from $\ell^\infty(\mathbb{Z})$ to $L^\infty(\mathbb{R})$ with a λ-uniform bound C_φ on its norm:

$$\|T_{\lambda,\varphi} s\|_{L^\infty} \le C_\varphi \|s\|_{\ell^\infty}.$$

This implies, with φ as in (1), that for any $\lambda \ge \lambda_0$,

$$\|x - \tilde{x}_{\lambda,Q_\delta}\|_{\ell^\infty} = \|T_{\lambda,\varphi}(S_\lambda x - Q_\delta S_\lambda x)\|_{\ell^\infty} \le C_\varphi \delta,$$

and in particular,

$$\lim_{\delta \to 0} \tilde{x}_{\lambda,Q_\delta} = x.$$

This trivial algorithm hardly finds any usage in practice due to the hardware implementation cost of Q_δ when δ is small. It turns out that it can be much cheaper to use very coarse quantization algorithms in which the set A consists of as few as two elements. This is commonly referred to as one-bit quantization, due to the fact that a single binary digit (bit) is sufficient to label each quantization level. To compensate for the lack of resolution in amplitude, one then increases the sampling rate λ, which is still relatively inexpensive in its cost of implementation.

We fix the set A to be $\{0, 1\}$. To match this normalization with the averaging property of $T_{\lambda,\varphi}$, we also assume that the input functions x satisfy, after an appropriate scaling and shift, $0 \le x(t) \le 1$ for all t. Hence, for each such function $x \in \mathcal{B}_\pi$, we are interested in approximations of the form

$$\tilde{x}_\lambda(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda \varphi(t - \tfrac{n}{\lambda}), \tag{2}$$

where $q_n^\lambda = (QS_\lambda x)_n \in \{0,1\}$ for every n, such that

$$\tilde{x}_\lambda \to x, \qquad \text{as } \lambda \to \infty. \tag{3}$$

At first, it is not immediate that this objective can be achieved. For example, it is easily seen that the first (and natural) choice $Q = 2Q_{1/2}$ for the quantization operator (with $Q(1)$ redefined to equal 1) results in the constant approximation $\tilde{x}_\lambda \equiv 0$ or $\tilde{x}_\lambda \equiv 1$ for any function x whose range is contained within $[0, 1/2)$ or $[1/2, 1]$, respectively. Algorithms used in practice circumvent this problem by introducing regular oscillations in the output of quantization, even when the input is constant. Perhaps, the most basic example of these algorithms is the so-called "first order" sigma-delta modulation, the quantization rule of which can be described in one equation as

$$Q := Q_{\Sigma\Delta} := \Delta Q_1 \Sigma, \tag{4}$$

where Δ is the standard difference operator defined by $(\Delta u)_n = u_n - u_{n-1}$, $Q_1 : \mathbb{R} \to \mathbb{Z}$ is the quantization operator Q_δ for $\delta = 1$, and Σ is the integration operator defined to be the inverse of Δ with zero initial condition, i.e., $(\Sigma u)_n = w_n$ where $(\Delta w)_n = u_n$ with $w_0 = 0$.

Note that $a - a' \in [0, 1]$ implies $\lfloor a \rfloor - \lfloor a' \rfloor \in \{0, 1\}$. Let $s \in [0, 1]^{\mathbb{Z}}$ and $q = Q_{\Sigma\Delta} s$. Since $s_n = (\Sigma s)_n - (\Sigma s)_{n-1}$, and $q_n = (Qs)_n = \lfloor (\Sigma s)_n \rfloor - \lfloor (\Sigma s)_{n-1} \rfloor$, we get that $Q_{\Sigma\Delta}$ maps $[0, 1]^{\mathbb{Z}}$ to $\{0, 1\}^{\mathbb{Z}}$.

In practice, the operator $Q_{\Sigma\Delta}$ is not implemented in the form of $\Delta Q_1 \Sigma$ due to the fact that the integration operator Σ would in general produce unbounded sequences when applied to arbitrary sequences (in our case all the entries of which are nonnegative). Alternatively, the output sequence q can be generated using a simple recurrence relation that only involves bounded quantities. Indeed, setting $u = \Sigma s - Q_1 \Sigma s = \Sigma s \pmod 1$ results in $u_n \in [0, 1)$, and

$$u_n = u_{n-1} + s_n - q_n \tag{5}$$

for all n. Since $(\Sigma s)_0 = 0$ by definition, we obtain the initial condition $u_0 = 0$. We have $u_{n-1} + s_n \in [0, 2)$, and $q_n \in \{0, 1\}$, which imply together with (5) that q_n satisfies the formula

$$q_n = \begin{cases} 0 & \text{if} \quad u_{n-1} + s_n < 1, \\ 1 & \text{if} \quad u_{n-1} + s_n \geq 1. \end{cases} \tag{6}$$

This very simple and inexpensive recursive algorithm in fact makes the first order sigma-delta quantization extremely popular in practice.

Let us see how $\tilde{x}_\lambda = T_{\lambda,\varphi} q^\lambda = T_{\lambda,\varphi} \Delta Q_1 \Sigma S_\lambda x$ approximates x. For general λ, let us define the sequence

$$u^\lambda := \Sigma S_\lambda x - Q_1 \Sigma S_\lambda x, \tag{7}$$

so that the error function $e_\lambda := x - \tilde{x}_\lambda$ satisfies

$$e_\lambda = T_{\lambda,\varphi} \Delta (\Sigma S_\lambda x - Q_1 \Sigma S_\lambda x) = T_{\lambda,\varphi} \Delta u^\lambda.$$

To each sequence $s \in \ell^\infty(\mathbb{Z})$, associate the measure

$$\mu_\lambda(s) := \frac{1}{\lambda} \sum_n s_n \delta_{n/\lambda},$$

where δ_a denotes the Dirac mass at the point a. Then we have

$$T_{\lambda,\varphi} s = \mu_\lambda(s) * \varphi.$$

Define also Δ_η to be the operator whose action on a measure is given by $\Delta_\eta \nu(\cdot) := \nu(\cdot) - \nu(\cdot - \eta)$. Then clearly $\mu_\lambda(\Delta s) = \Delta_{1/\lambda}\mu_\lambda(s)$, so that by commutation of convolutional operators, we obtain the error formula

$$\begin{aligned} e_\lambda &= T_{\lambda,\varphi}\Delta u^\lambda \\ &= \Delta_{1/\lambda}\mu_\lambda(u^\lambda) * \varphi \\ &= \mu_\lambda(u^\lambda) * \Delta_{1/\lambda}\varphi \\ &= \frac{1}{\lambda}\sum_n u_n^\lambda \Delta_{1/\lambda}\varphi(\cdot - \tfrac{n}{\lambda}). \end{aligned} \tag{8}$$

Taking the sup norm of the resulting function yields the error bound

$$\big\|e_\lambda\big\|_{L^\infty} \le \Big\|\frac{1}{\lambda}\|u^\lambda\|_{\ell^\infty} \sum_n \big|\Delta_{1/\lambda}\varphi(\cdot - \tfrac{n}{\lambda})\big|\Big\|_{L^\infty} \le \frac{1}{\lambda}\mathrm{Var}(\varphi),$$

where $\mathrm{Var}(\varphi)$ denotes the total variation of φ which, in this case, is clearly equal to $\|\varphi'\|_1$. This is the error bound first given in [3] for the class $\mathcal{B}_\pi$. Assuming that the reconstruction kernel φ is fixed once for all, we can summarize the above "basic estimate" as

$$\big\|x - \tilde{x}_\lambda\big\|_{L^\infty} \ll \lambda^{-1}. \tag{9}$$

(In this paper, we shall frequently use the common notation $X \ll_{\alpha,\beta,\dots} Y$ to refer to the inequality $X \le CY$ where the constant C may depend on $\alpha, \beta, \dots$, but no other variable.)

It has been observed via numerical simulation that the error decay in λ is in fact faster than λ^{-1}. The folklore in the electrical engineering literature is that for the class of bandlimited functions, the error decays "on the average" like $\lambda^{-3/2}$. There are few results on the rigorous side, however. In the particular case of *constant* functions $x(t) = c$, Gray [4] proved using spectral arguments that a particular "root mean square" norm that averages the error over $\mathbb{R}$ as well as over the value of the constant $c \in [0, 1]$ decays asymptotically like $\lambda^{-3/2}$. A similar estimate was given also for pure sinusoids.

We are interested in finding the true error behavior of the first order sigma-delta quantization. In this paper, we shall present improvements on the error bound (9) for arbitrary bandlimited functions. The following is our main theorem:

Theorem 1. *For all $\epsilon > 0$, there exists a family $\{\varphi_\lambda\}_{\lambda>1}$ of reconstruction kernels such that for all π-bandlimited functions x with range in $[0, 1]$, and for all t for which $x'(t) \neq 0$,*

$$|x(t) - \tilde{x}_\lambda(t)| \ \ll_{\epsilon, x'(t)} \ \lambda^{-4/3+\epsilon},$$

where $\tilde{x}_\lambda = T_{\lambda,\varphi_\lambda} q^\lambda$ is the approximate reconstruction of x from the first order sigma-delta quantized bit sequence $q^\lambda = Q_{\Sigma\Delta} S_\lambda x$.

This result relies heavily on the theory of uniform distribution for point sequences, and stationary phase methods for exponential sums. In Section 2, we summarize the basic definitions and theorems that we shall use in our analysis. Section 3 is of a technical nature; we provide upper bounds on the discrepancy of the iterates of a time-varying dynamical system on the circle, which then leads to the proof of Theorem 1 in Section 4. In the special case of constant functions, the error bound of Theorem 1 can be improved further using some classical results in

Diophantine approximation. Details of this improvement are given in Section 5. We conclude the paper with remarks on higher order sigma-delta modulation schemes.

2. Preliminaries

Uniform distribution. Let $u = (u_n)_{n=1}^{\infty}$ be a sequence of points in $[0,1)$ identified with the 1-torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. Recall that the sequence u is said to be *uniformly distributed* (in short, *u.d.*) if

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N \ : \ u_n \in I\}}{N} = |I| \tag{10}$$

for every arc I in $\mathbb{T}$. For a finite nonempty set $S \subset [0,1)$ of points (possibly with multiplicity), define the *discrepancy* of S to be

$$\mathrm{Discr}(S) := \sup_{I\subset\mathbb{T}} \left| \frac{\#(S\cap I)}{\#S} - |I| \right|. \tag{11}$$

Then the N-term discrepancy of the sequence u is defined as

$$D_N(u) := \mathrm{Discr}\big(\{u_n\}_{n=1}^N\big)\,. \tag{12}$$

It is an elementary result that u is u.d. if and only if $D_N(u) \to 0$ as $N \to \infty$. Two equivalent characterizations of uniform distribution are given by *Weyl's criterion*:

$$\begin{aligned}(u_n) \text{ is u.d.} \quad &\Longleftrightarrow \quad \frac{1}{N}\sum_{n=1}^{N} e^{2\pi i k u_n} \to 0 \ \text{ for each nonzero } k \in \mathbb{Z},\\ &\Longleftrightarrow \quad \frac{1}{N}\sum_{n=1}^{N} f(u_n) \to \int_{\mathbb{T}} f(u)\,du \ \text{ for every Riemann-integrable}\\ &\qquad\qquad \text{(or, equivalently, continuous) } f \text{ on } \mathbb{T}.\end{aligned}$$

These are "qualitative" statements. The quantitative theory aims to find out how quickly the convergence takes place in the above. We shall need the following two well-known results:

Theorem 2 (Koksma's inequality, [10])**.** *For any sequence of points $u_1,\dots,u_N$ in $[0,1)$, and any function $f : [0,1] \to \mathbb{R}$ of bounded variation,*

$$\left|\frac{1}{N}\sum_{n=1}^{N} f(u_n) - \int_0^1 f(t)dt\right| \le \mathrm{Var}(f)\,\mathrm{Discr}\big(\{u_n\}_{n=1}^N\big)\,, \tag{13}$$

where $\mathrm{Var}(f)$ *is the total variation of f.*

Theorem 3 (Erdős-Turán inequality, [10])**.** *For any sequence of points $u_1,\dots,u_N$ in $[0,1)$, and any positive integer K,*

$$\mathrm{Discr}\big(\{u_n\}_{n=1}^N\big) \ll \frac{1}{K} + \sum_{k=1}^{K} \frac{1}{k}\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i k u_n}\right|. \tag{14}$$

Exponential sums. Erdős-Turán inequality provides us with a tool to estimate the discrepancy of a sequence of points by turning it into the problem of estimating an associated family of exponential sums. At least two types of exponential sums are relevant to the study of sigma-delta modulation. The first type is the well-studied class of *Weyl sums*

$$S = \sum_{n=1}^{N} e^{2\pi i f(n)}, \tag{15}$$

where, by definition, f is a polynomial (with real coefficients). Weyl sums arise in sigma-delta modulation schemes with constant input, though the interesting cases appear only in "higher order" schemes [6] (see Section 6 for a short description of what this means). The main concentration of this paper is on the first order case, and therefore we will not be dealing with Weyl sums directly. The second type of sums are given by more general functions f in (15), that are not necessarily polynomials, yet still have a certain amount of smoothness. These sums, on the other hand, *will* arise in this paper, when the input is an arbitrary bandlimited function. For both types of sums, extremely sophisticated tools are available in the mathematical literature to estimate their sizes. We shall require here only outcomes of more "general purpose" tools, for they already lead to substantial improvements of the basic estimates. We shall make use of the *truncated Poisson formula* and *van der Corput's Lemma*, which we give below.

Theorem 4 (Truncated Poisson, [10])**.** *Let f be a real-valued function and suppose that f' is continuous and increasing on $[a,b]$. Put $\alpha = f'(a)$, $\beta = f'(b)$. Then*

$$\sum_{a\le m\le b} e^{2\pi i f(m)} = \sum_{\alpha-1\le\nu\le\beta+1} \int_a^b e^{2\pi i(f(\tau)-\nu\tau)}\,d\tau + O(\log(2+\beta-\alpha)). \tag{16}$$

(If f' is decreasing on $[a,b]$, taking the complex conjugate of the above expression applied to $-f$ leads to the same expression with α and β switched.)

Theorem 5 (van der Corput, [12])**.** *Suppose ϕ is real-valued and smooth in the interval (a,b), and that $|\phi^{(r)}(t)| \ge \mu$ for all $t \in (a,b)$ and for a positive integer r. If $r = 1$, suppose additionally that ϕ' is monotonic. Then*

$$\Big| \int_a^b e^{i\phi(t)}\,dt \Big| \ll_r \mu^{-1/r}. \tag{17}$$

Discrepancy of arithmetic progressions modulo 1. Perhaps the most important uniformly distributed sequences are arithmetic progressions modulo 1, defined by $u_n = n\alpha \pmod 1$, with $\alpha \in \mathbb{R}\backslash\mathbb{Q}$. These sequences arise in first order sigma-delta modulation with constant inputs, and the corresponding discrepancy estimates, as we will show, directly relate to the error estimates. For simplicity, we shall only make use of metric results which are valid for almost every α (with respect to the Lebesgue measure), and none of the results that depend on the finer Diophantine properties of α.

Denote by $\|u\|$ the distance between a real number u and the set of integers. Let $\psi : \mathbb{Z}^+ \to \mathbb{R}^+$ be a given nondecreasing function. An irrational number α is said to be *of type* $<\psi$ if the inequality $n\|n\alpha\| \ge 1/\psi(n)$ holds for all positive integers n.

An important metric result (due to Khinchine) is the following: Let $\epsilon > 0$ be given. Then, almost all α are of type $<C_\alpha \psi_\epsilon$ where $\psi_\epsilon(q) = \log^{1+\epsilon}(2q)$, and C_α is a constant that may depend on α. This result leads to the following theorem:

Theorem 6 ([8])**.** *For any $\epsilon > 0$, the N-term discrepancy of $u_n = n\alpha \pmod 1$ satisfies*

$$D_N(u) \ll_\alpha N^{-1} \log^{2+\epsilon} N \tag{18}$$

for almost all α.

It is true that the same estimate holds uniformly (with the same constant) for any translate (in n) of the sequence u. This strengthens the qualitative result that for irrational α, $(n\alpha)$ is not only u.d. (mod 1) but also *well distributed* [8].

If α is of type $<\psi$ for a constant function ψ, then one says α is *of constant type*. For instance, all quadratic irrationals are in this category. For these, the discrepancy satisfies $D_N(u) \ll_\alpha N^{-1} \log N$. This is the smallest possible order of discrepancy for *any* infinite sequence u due to the following lower bound: $D_N(u) \geq c\, N^{-1} \log N$ for infinitely many N, where c is an absolute constant.

Bernstein's inequality [9]**.** For any $1 \leq p \leq \infty$, if $x \in \mathcal{B}_\Omega \cap L^p$, then

$$\|x'\|_{L^p} \leq \Omega \|x\|_{L^p}. \tag{19}$$

3. A local discrepancy estimate for u^λ

Note that u_n^λ, defined by (7), is simply the fractional part of $(\Sigma S_\lambda x)_n$. One can also describe this sequence by saying that u_n^λ is the nth iterate of the time-varying dynamical system

$$u_n^\lambda = R_{x(\frac{n}{\lambda})}(u_{n-1}^\lambda),$$

where $R_\theta : [0,1) \to [0,1)$ denotes the rotation map $w \mapsto w + \theta \pmod 1$. Clearly, in the case when x is equal to a constant function with an irrational value, the sequence u^λ is uniformly distributed in $[0,1)$; in this case, we shall employ the discrepancy estimate given by Theorem 6.

For the general case, we define a local discrepancy quantity associated to the sequence u^λ by

$$d(t, I, \lambda) := \mathrm{Discr}\big(\big\{u_n^\lambda : \tfrac{n}{\lambda} - t \in I\big\}\big)\,.$$

Lemma 1. *There exist two absolute constants $C_1 > 0$ and $C_2 > 0$ such that for all t at which $x'(t) \neq 0$, and for all intervals I and numbers λ satisfying $I \subset [-C_1|x'(t)|, C_1|x'(t)|]$, and $\lambda > \max\big(|I|^{-1}, C_2|x'(t)|^{-1}\big)$, one has*

$$d(t, I, \lambda) \ll \frac{1}{\lambda^{1/3}} + \frac{1}{|I|\sqrt{|x'(t)|}} \frac{1}{\lambda^{1/2}}. \tag{20}$$

3.1. **Analytic interpolation of the sequences u^λ.** The proof of Lemma 1 will rely on the following proposition:

Proposition 1. *For each $\lambda > 1$, there exists an analytic function X_λ such that*

$$u_n^\lambda = X_\lambda(n) \pmod 1, \tag{21}$$

and

$$\Big\| X_\lambda' - x\Big(\frac{\cdot}{\lambda}\Big)\Big\|_{L^\infty} \ll \frac{1}{\lambda}. \tag{22}$$

Proof. Define $\hat{X}_\lambda$ to be the compactly supported distribution

$$\hat{X}_\lambda(\xi) = \frac{\lambda \hat{x}(\lambda\xi)}{1 - e^{-i\xi}} + c\,\delta_0(\xi)$$

where $c = c(\lambda)$ is chosen such that $X_\lambda(0) = u_0^\lambda = 0$. Then X_λ is an analytic function that satisfies

$$X_\lambda(t) - X_\lambda(t-1) = x\Big(\frac{t}{\lambda}\Big)$$

for all t. This shows (21). Let φ be a fixed smoothing kernel as defined in (1). Then it is clear by Fourier inversion that

$$X_\lambda' - x\Big(\frac{\cdot}{\lambda}\Big) = \phi_\lambda * x\Big(\frac{\cdot}{\lambda}\Big)$$

where

$$\hat{\phi}_\lambda(\xi) = \Big(\frac{i\xi}{1 - e^{-i\xi}} - 1\Big)\hat{\varphi}(\lambda\xi).$$

Since

$$\Big\| X_\lambda' - x\Big(\frac{\cdot}{\lambda}\Big)\Big\|_{L^\infty} \le \|\phi_\lambda\|_{L^1},$$

it suffices to show that $\|\phi_\lambda\|_{L^1} \ll \lambda^{-1}$. To see this, first note that

$$\frac{i\xi}{1 - e^{-i\xi}} = 1 + \frac{i}{2}\xi + O(|\xi|^2).$$

Hence, for $|\xi| \le \pi$, one has

$$\Big|\frac{i\xi}{1 - e^{-i\xi}} - 1\Big| \ll |\xi|$$

and

$$\Big|\frac{d}{d\xi}\Big(\frac{i\xi}{1 - e^{-i\xi}} - 1\Big)\Big| \ll 1.$$

Since $|\xi| \le \lambda_0/\lambda$ in the support of $\hat{\phi}_\lambda$, we obtain

$$|\hat{\phi}_\lambda(\xi)| \ll \frac{1}{\lambda}$$

and

$$\Big|\frac{d\hat{\phi}_\lambda}{d\xi}\Big| \ll |\hat{\varphi}(\lambda\xi)| + |\xi|\,|\lambda(\hat{\varphi})'(\lambda\xi)| \ll 1.$$

This implies that

$$\|\phi_\lambda\|_{L^\infty} \ll \int |\hat{\phi}_\lambda(\xi)| d\xi \ll \int_{|\xi| \le \lambda_0/\lambda} \frac{1}{\lambda}\, d\xi \ll \frac{1}{\lambda^2},$$

and

$$\Big\|\frac{d\hat{\phi}_\lambda}{d\xi}\Big\|_{L^2} \ll \Big(\int_{|\xi| \le \lambda_0/\lambda} d\xi\Big)^{1/2} \ll \frac{1}{\sqrt{\lambda}}\,.$$

Combining these two estimates, we get, for any $A > 0$,

$$\begin{aligned}\|\phi_\lambda\|_{L^1} &\leq \int_{|t|\leq A} \|\phi_\lambda\|_{L^\infty} dt + \int_{|t|>A} \frac{1}{|t|}\,|t\phi_\lambda(t)|\,dt \\ &\ll \frac{A}{\lambda^2} + \Big(\int_{|t|>A} \frac{1}{|t|^2}\,dt\Big)^{1/2}\Big(\int |t\phi_\lambda(t)|^2\,dt\Big)^{1/2} \\ &\ll \frac{A}{\lambda^2} + \frac{1}{\sqrt{A}}\Big\|\frac{d\hat{\phi}_\lambda}{d\xi}\Big\|_{L^2} \\ &\ll \frac{A}{\lambda^2} + \frac{1}{\sqrt{A\lambda}}.\end{aligned}$$

By choosing $A = \lambda$, we obtain $\|\phi_\lambda\|_{L^1} \ll \lambda^{-1}$. This completes the proof. □

Corollary 1. *There exist two absolute constants $C_1 > 0$ and $C_2 > 0$ such that for all t at which $x'(t) \neq 0$, and for all τ and λ satisfying $|\frac{\tau}{\lambda} - t| \leq C_1|x'(t)|$ and $\lambda > C_2|x'(t)|^{-1}$, one has*

$$\frac{1}{2}\frac{|x'(t)|}{\lambda} \leq |X''_\lambda(\tau)| \leq \frac{3}{2}\frac{|x'(t)|}{\lambda}. \tag{23}$$

Proof. Since $X'_\lambda - x(\frac{\cdot}{\lambda})$ is in $\mathcal{B}_{\pi/\lambda}$, Bernstein's inequality with Proposition 1 implies that

$$\Big\|X''_\lambda - \frac{1}{\lambda}x'\Big(\frac{\cdot}{\lambda}\Big)\Big\|_{L^\infty} \leq \frac{\pi}{\lambda}\Big\|X'_\lambda - x\Big(\frac{\cdot}{\lambda}\Big)\Big\|_{L^\infty} \leq C_0\frac{1}{\lambda^2}$$

for some absolute constant C_0. Let C_1 and C_2 be constants satisfying

$$\frac{C_0}{C_2} + \pi^2 C_1 \leq \frac{1}{2}.$$

Then for any τ and λ satisfying $|\frac{\tau}{\lambda} - t| \leq C_1|x'(t)|$ and $\lambda > C_2|x'(t)|^{-1}$, one has

$$\begin{aligned}\Big|X''_\lambda(\tau) - \frac{1}{\lambda}x'(t)\Big| &\leq \Big|X''_\lambda(\tau) - \frac{1}{\lambda}x'\Big(\frac{\tau}{\lambda}\Big)\Big| + \frac{1}{\lambda}\Big|x'\Big(\frac{\tau}{\lambda}\Big) - x'(t)\Big| \\ &\leq \frac{C_0}{\lambda^2} + \frac{1}{\lambda}\|x''\|_{L^\infty}\Big|\frac{\tau}{\lambda} - t\Big| \\ &\leq \frac{1}{\lambda}\left(\frac{C_0}{C_2}|x'(t)| + \pi^2 C_1|x'(t)|\right) \\ &\leq \frac{1}{2}\frac{|x'(t)|}{\lambda},\end{aligned}$$

hence the result of the corollary. □

3.2. **Proof of Lemma 1.** Since $u_n^\lambda = X_\lambda(n) \pmod 1$, Erdős-Turán inequality gives

$$d(t, I, \lambda) \ll \frac{1}{K} + \sum_{k=1}^{K}\frac{1}{k}\left|\frac{1}{\lfloor\lambda|I|\rfloor}\sum_{n\in\mathbb{Z}\cap\lambda(I+t)} e^{2\pi i k X_\lambda(n)}\right| \tag{24}$$

for any positive integer K, where we have used the lower bound $\lfloor\lambda|I|\rfloor$ for the number of integers n such that $\frac{n}{\lambda} - t \in I$. Let

$$S_k(t, I, \lambda) := \sum_{n\in\mathbb{Z}\cap\lambda(I+t)} e^{2\pi i k X_\lambda(n)}$$

for $k \geq 1$, and consider the phase function $f = kX_\lambda$. Corollary 1 implies that f' is monotonic on the interval $\lambda(I+t)$ (since $f'' = kX_\lambda''$ is continuous and bounded away from zero) and that

$$\frac{k}{2}\frac{|x'(t)|}{\lambda} \leq |f''(\tau)| \leq \frac{3k}{2}\frac{|x'(t)|}{\lambda}.$$

Therefore Theorem 5 with $r = 2$ implies, for all ν,

$$\Big| \int_{\lambda(I+t)} e^{2\pi i(f(\tau)-\nu\tau)}\, d\tau \Big| \ll \sqrt{\frac{\lambda}{k|x'(t)|}}.$$

Since f' is monotonic and continuous, the number of integer values ν that are attained by f' on $\lambda(I+t)$ is bounded by $1+|f'(\lambda(I+t))|$, which is further bounded by

$$1 + \lambda|I| \sup_{\tau\in\lambda(I+t)} |f''(\tau)| \leq 1 + (\lambda|I|)\Big(\frac{3k}{2}\frac{|x'(t)|}{\lambda}\Big) \ll 1 + k|I||x'(t)|.$$

When coupled with Theorem 4, this yields the estimate

$$\begin{aligned} |S_k(t,I,\lambda)| &\ll \Big(k|I||x'(t)|+3\Big)\sqrt{\frac{\lambda}{k|x'(t)|}} + \log(k+2) \\ &\ll |I|\sqrt{|x'(t)|\lambda k} + \sqrt{\frac{\lambda}{k|x'(t)|}} + \log(k+2). \end{aligned}$$

We see that the above bound is better than the trivial bound $1+\lambda|I|$ for $k \ll \lambda$ except for very small values of k which we are not interested in. Note that the Erdős-Turán inequality can be exploited most for large K.

Plugging this into (24), we obtain

$$\begin{aligned} d(t,I,\lambda) &\ll \frac{1}{K} + \sqrt{\frac{|x'(t)|}{\lambda}}K^{1/2} + \frac{1}{|I|\sqrt{\lambda|x'(t)|}} + \frac{1}{|I|\lambda}\log^2(K) \\ &\ll \left(\frac{|x'(t)|}{\lambda}\right)^{1/3} + \frac{1}{|I|\sqrt{\lambda|x'(t)|}} + \frac{1}{|I|\lambda}\log^2\Big(\frac{\lambda}{|x'(t)|}\Big) \end{aligned}$$

where at the last step we chose the optimal value $K \sim \big(\frac{\lambda}{|x'(t)|}\big)^{1/3}$. Clearly, the third term can be absorbed in the second term, and also $|x'(t)|$ can be dropped from the first term. Therefore we get

$$d(t,I,\lambda) \ll \frac{1}{\lambda^{1/3}} + \frac{1}{|I|\sqrt{|x'(t)|}}\frac{1}{\lambda^{1/2}},$$

hence the proof of the lemma. □

4. Proof of Theorem 1

We have now gathered the necessary tools and results for the proof of the main theorem of this paper. The rest of the analysis consists of a number of gluing steps:

For each t and λ, let $\eta = \eta(\lambda,t)$ be the integer such that $-\frac{1}{\lambda} < \frac{\eta}{\lambda} - t \leq 0$, and define an auxiliary sequence $U := U^{\lambda,t}$ by

$$U_m - U_{m-1} = u^\lambda_{\eta+m} - \frac{1}{2}$$

with the initial condition $U_0 = 0$. Then we have the expression

$$U_m = \sum_{k=1}^{m} \left(u^{\lambda}_{\eta+k} - \frac{1}{2} \right), \qquad m \geq 1,$$

with a similar expression for $m \leq -1$. Using Koksma's inequality for the function $f(u) = u$, we obtain the bound

$$|U_m| \ll m \cdot \mathrm{Discr}\big(\{u^{\lambda}_{\eta+k} : \ 1 \leq k \leq m\}\big) = m \cdot d\left(t, \left[\frac{\eta+1}{\lambda} - t, \frac{\eta+m}{\lambda} - t\right], \lambda\right),$$

where we have used that

$$\{u^{\lambda}_{\eta+k} : \ 1 \leq k \leq m\} = \left\{ u^{\lambda}_n : \ \frac{n}{\lambda} - t \in \left[\frac{\eta+1}{\lambda} - t, \frac{\eta+m}{\lambda} - t\right] \right\}.$$

Therefore Lemma 1 yields the estimate

$$|U_m| \ll m \left(\frac{1}{\lambda^{1/3}} + \frac{\lambda}{m\sqrt{|x'(t)|}} \frac{1}{\lambda^{1/2}} \right) \ll \lambda^{2/3} + \frac{\lambda^{1/2}}{|x'(t)|^{1/2}} \tag{25}$$

for all m such that $1 \leq m \leq C_1 |x'(t)|\lambda$. A similar argument for $m \leq -1$ provides us with the same estimate for $|m| \leq C_1|x'(t)|\lambda$.

Let $\epsilon < 1$ be an arbitrary small positive number. Choose a Schwartz function φ_1 satisfying (1) with an arbitrary $\lambda_0 > 1$, and define

$$\varphi_\lambda(t) := \lambda^{\frac{\epsilon}{2}} \varphi_1(\lambda^{\frac{\epsilon}{2}} t).$$

Since $\hat{\varphi}_\lambda(\xi) = \hat{\varphi}_1(\lambda^{-\frac{\epsilon}{2}}\xi)$, and $\lambda \geq \lambda_0 \lambda^{\frac{\epsilon}{2}}$ for all $\lambda \geq \lambda_0^2$, we can employ φ_λ in the reconstruction process.

The error expression we derived earlier in (8) can now be written as

$$\begin{aligned} e_\lambda(t) &= \frac{1}{\lambda} \sum_m \big(u^{\lambda}_{\eta+m} - \frac{1}{2}\big) \Delta_{1/\lambda} \varphi_\lambda\Big(t - \frac{\eta+m}{\lambda}\Big) \\ &= \frac{1}{\lambda} \sum_m \Delta U_m \Delta_{1/\lambda} \varphi_\lambda\Big(t - \frac{\eta+m}{\lambda}\Big) \\ &= \frac{1}{\lambda} \sum_m U_m \Delta^2_{1/\lambda} \varphi_\lambda\Big(t - \frac{\eta+m}{\lambda}\Big). \end{aligned} \tag{26}$$

We split this sum into two pieces given by $I_1 := \{m \in \mathbb{Z} \ : \ |m| \leq C_1|x'(t)|\lambda\}$, and $I_2 = \mathbb{Z} \backslash I_1$. For the first piece, we use the estimate (25) in the form $|U_m| \ll_{x'(t)} \lambda^{2/3}$ and obtain

$$\begin{aligned} \left| \frac{1}{\lambda} \sum_{m \in I_1} U_m \Delta^2_{1/\lambda} \varphi_\lambda\Big(t - \frac{\eta+m}{\lambda}\Big) \right| &\ll_{x'(t)} \lambda^{-1/3} \sum_m \left| \Delta^2_{1/\lambda} \varphi_\lambda\Big(t - \frac{\eta+m}{\lambda}\Big) \right| \\ &\ll_{x'(t)} \lambda^{-1/3} \mathrm{Var}(\Delta_{1/\lambda}\varphi_\lambda) \\ &\ll_{x'(t)} \lambda^{-4/3+\epsilon}, \end{aligned} \tag{27}$$

where in the last step we have used the estimate

$$\mathrm{Var}(\Delta_{1/\lambda}\varphi_\lambda) = \big\|\Delta_{1/\lambda}\varphi'_\lambda\big\|_{L^1} \ll \lambda^{-1}\big\|\varphi''_\lambda\big\|_{L^1} = \lambda^{-1+\epsilon}\big\|\varphi''_1\big\|_{L^1} \ll \lambda^{-1+\epsilon}.$$

For the second piece, we use the trivial estimate $|U_m| \ll m$ and exploit the rapid decay of φ_λ and its derivatives. Note that for all positive integers N and l, we have

$$|\varphi_1^{(l)}(s)| \ll_{N,l} \ (1 + |s|)^{-N}$$

so that for $s > 4/\lambda$,

$$|\Delta^2_{1/\lambda}\varphi_\lambda(s)| \ll \lambda^{-2} \sup_{s-\frac{2}{\lambda}\le r\le s} |\varphi''_\lambda(r)| \ll_N \lambda^{-2-(N-3)\frac{\epsilon}{2}}|s|^{-N}.$$

Therefore

$$\begin{aligned}\left|\frac{1}{\lambda}\sum_{m\in I_2} U_m \Delta^2_{1/\lambda}\varphi_\lambda\Big(t-\frac{\eta+m}{\lambda}\Big)\right| &\ll_N \left|\frac{1}{\lambda}\sum_{m\in I_2}|m|\lambda^{-2-(N-3)\frac{\epsilon}{2}}\Big|\frac{m}{\lambda}\Big|^{-N}\right| \\ &\ll_{N,x'(t)} \lambda^{-1-(N-3)\frac{\epsilon}{2}}, \\ &\ll_{\epsilon,x'(t)} \lambda^{-2}, \end{aligned} \tag{28}$$

where at the last step we choose $N = N(\epsilon)$ such that $(N-3)\frac{\epsilon}{2} \ge 1$. Combining (27) and (28), we obtain the desired bound

$$|e_\lambda(t)| \ll_{\epsilon,x'(t)} \lambda^{-4/3+\epsilon}, \tag{29}$$

hence the proof of Theorem 1. □

5. Improvements for constant functions

In the case of constant functions, the uniform error bound can be improved significantly. The reason is that Khinchine's theorem provides us with a much better estimate for $|U_m|$ for almost every x, and moreover, which holds uniformly for all shifts of the sequence u^λ. Let $\epsilon > 0$ be given and assume x is such that the result of Theorem 6 holds. Then we have

$$|U_m| \ll_x \log^{2+\epsilon}|m| \qquad \text{for all } m.$$

Note that there is now a lot more freedom to choose φ, since the bandwidth of a constant function is zero. For simplicity, let us assume again that φ is a fixed Schwartz function satisfying (1). We again split the error expression given by (26) into two pieces, this time the center block being $|m| \le \lambda^2$. Then we obtain

$$\begin{aligned}\left|\frac{1}{\lambda}\sum_{|m|\le\lambda^2} U_m \Delta^2_{1/\lambda}\varphi\Big(t-\frac{\eta+m}{\lambda}\Big)\right| &\ll_x \frac{1}{\lambda}\sum_m \log^{2+\epsilon}(\lambda^2)\left|\Delta^2_{1/\lambda}\varphi\Big(t-\frac{\eta+m}{\lambda}\Big)\right| \\ &\ll_x \lambda^{-1}\left(\log^{2+\epsilon}\lambda\right)\operatorname{Var}(\Delta_{1/\lambda}\varphi) \\ &\ll_x \lambda^{-2}\log^{2+\epsilon}\lambda. \end{aligned} \tag{30}$$

For the second piece, we use the bound

$$|\Delta^2_{1/\lambda}\varphi(s)| \ll \lambda^{-2}|s|^{-2},$$

which now yields

$$\begin{aligned}\left|\frac{1}{\lambda}\sum_{|m|>\lambda^2} U_m \Delta^2_{1/\lambda}\varphi\Big(t-\frac{\eta+m}{\lambda}\Big)\right| &\ll_x \left|\frac{1}{\lambda}\sum_{|m|>\lambda^2}\left(\log^{2+\epsilon}|m|\right)\lambda^{-2}\Big|\frac{m}{\lambda}\Big|^{-2}\right| \\ &\ll_x \lambda^{-2}. \end{aligned} \tag{31}$$

Combining (30) and (31) gives

$$\big\|e_\lambda\big\|_{L^\infty} \ll_x \lambda^{-2}(\log\lambda)^{2+\epsilon}. \tag{32}$$

Remark. The error bound (32) can easily be reproduced for the nonbandlimited reconstruction kernel

$$\varphi(t) = \begin{cases} 1-|t|, & \text{if } |t| \le 1, \\ 0, & \text{otherwise,} \end{cases}$$

as well. This result appears (independently) in [2] and [5] in slightly different but essentially equivalent forms.

6. Higher order schemes

There is a whole class of *higher order* schemes, which provide improved approximations by employing even smarter quantization algorithms. For a given positive integer m, suppose now that the quantization operator Q has the decomposition

$$Q = \Delta^m \tilde{Q} \Sigma^m, \tag{33}$$

where the superscript m refers to an m-fold composition. Here, $\tilde{Q}$ is to be designed such that, again, Q maps $[0,1]^{\mathbb{Z}}$ to $\{0,1\}^{\mathbb{Z}}$. We call such a map $\tilde{Q}$ (or, equivalently, Q) admissible. Similar to the first order case, let $u^\lambda := \Sigma^m S_\lambda x - \tilde{Q}\Sigma^m S_\lambda x$. If for a class of input functions x, an admissible quantization operator $\tilde{Q}$ also satisfies

$$\|u^\lambda\|_{\ell^\infty} = \|\Sigma^m S_\lambda x - \tilde{Q}\Sigma^m S_\lambda x\|_{\ell^\infty} \le C_m := C_m(\tilde{Q}) < \infty$$

for all λ, then we say that the sigma-delta modulator (or, equivalently, Q) is stable. Then, similar to the analysis for $m = 1$, one would have

$$\begin{aligned} \|T_{\lambda,\varphi}\Delta^m s\|_{L^\infty} &= \|\mu_\lambda(s) * \Delta^m_{1/\lambda}\varphi\|_{L^\infty} \\ &\le \frac{1}{\lambda}\|s\|_{\ell^\infty}\mathrm{Var}(\Delta^{m-1}_{1/\lambda}\varphi) \\ &\le \frac{1}{\lambda^m}\|s\|_{\ell^\infty}|\varphi|_{W_1^m}, \end{aligned}$$

where in the last step we have made use of the bound

$$\mathrm{Var}(\Delta^{m-1}_{1/\lambda}\varphi) = \|\Delta^{m-1}_{1/\lambda}\varphi'\|_{L^1} \le \frac{1}{\lambda^{m-1}}\|\varphi^{(m)}\|_{L^1}.$$

Here $|\varphi|_{W_1^m} = \|\varphi^{(m)}\|_{L^1}$ is the Sobolev W_1^m semi-norm of φ. Applying this result to $s = u^\lambda$, one obtains

$$\|x - \tilde{x}_\lambda\|_{L^\infty} \ll_{m,\tilde{Q}} \lambda^{-m},$$

where we now emphasize the dependence of the constant on m as well as on the algorithm $\tilde{Q}$, which needs to be specified. This error bound in this generality was given first in [3].

The naive operator Q_1 is unfortunately not admissible for $m > 1$. Indeed, consider $m = 2$ and a small constant $x > 0$. Then $Q_1\Sigma^2 S_\lambda x$ will always contain a substring of the form 0 1 1; therefore the second order difference of this substring will contain a -1. However, admissible and stable operators do exist. For certain such rules, and for the particular case of constant functions, we provide in [6] improved error estimates for an input-averaged square norm, using techniques analogous to the ones presented in this paper. These results, however, not only depend on the stability properties of u^λ, but also on further algebraic and analytic properties of the associated two dimensional sequence $\mathbf{u}^\lambda_n = (u^\lambda_n,\ u^\lambda_{n-1})$. For details we refer to [6].

The problem of finding stable operators with small $C_m(\tilde{Q})$ is an ongoing research problem. In [3], the first example of an infinite family of stable sigma-delta modulators is given. We construct other families in [7], and moreover which collectively yield the error bound $\|e_\lambda\|_{L^\infty} = O(2^{-0.07\lambda})$ for arbitrary π-bandlimited functions.

Acknowledgments

The author would like to thank Ingrid Daubechies for many valuable discussions on the results presented in this paper. The author would also like to thank Peter Sarnak for various suggestions and Wilhelm Schlag and Sergei Konyagin for two remarks that simplified an earlier version of the proof of Theorem 1.

References

[1] J. C. Candy and G. C. Temes, Eds., *Oversampling Delta-Sigma Data Converters: Theory, Design and Simulation*, IEEE Press, 1992.

[2] W. Chou, T. H. Meng, and R. M. Gray, "Time Domain Analysis of Sigma Delta Modulation," *Proceedings ICASSP-90*, Int. Conf. on Acoustics, Speech and Signal Processing, vol. 3, pp. 1751–1754, Albuquerque, NM, April 1990.

[3] I. Daubechies, R. DeVore, "Approximating a Bandlimited Function Using Very Coarsely Quantized Data: A Family of Stable Sigma-Delta Modulators of Arbitrary Order", to appear in Annals of Mathematics.

[4] R. M. Gray, "Spectral Analysis of Quantization Noise in a Single-Loop Sigma-Delta Modulator with dc Input," *IEEE Trans. on Comm.*, vol. COM-37, pp. 588–599, June 1989.

[5] C. S. Güntürk, "Improved Error Estimates for First Order Sigma-Delta Systems," *Proceedings SampTA-99*, Int. Workshop on Sampling Theory and Applications, Loen, Norway, August 1999.

[6] C. S. Güntürk and N. T. Thao, "Refined Analysis of MSE in Second Order Sigma-Delta Modulation with DC Inputs," submitted to IEEE Transactions on Information Theory, in revision.

[7] C. S. Güntürk, "One-Bit Sigma-Delta Quantization with Exponential Accuracy," to appear in Communications on Pure and Applied Mathematics.

[8] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, Wiley, 1974. MR **54:**7415

[9] Y. Meyer, *Wavelets and Operators*, Cambridge University Press, 1992. MR **92k:**42001

[10] H. L. Montgomery, *Ten Lectures on the Interface Between Analytic Number Theory and Harmonic Analysis*, AMS, 1994. MR **96i:**11002

[11] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta-Sigma Data Converters: Theory, Design and Simulation*, IEEE Press, 1996.

[12] E. M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, 1993. MR **95c:**42002

Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, New York 10012-1185

E-mail address: gunturk@cims.nyu.edu