DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

Number Theoretical Error Estimates in a Quantization Scheme for Bandlimited Signals

C. Sinan Güntürk

Abstract. Sigma-delta quantization is a way of representing bandlimited signals (functions with compactly supported Fourier transforms) by $\{0,1\}$ sequences for each sampling density such that convolving these sequences with appropriately chosen filters produces approximations of the original signals. Approximations are refined by increasing the sampling density; this is what makes such a scheme fundamentally different from more conventional quantization schemes, where the sampling density is not varied. We present various examples of how tools from analytic number theory are employed in sharpening the error estimates in sigma-delta systems.

1. Introduction

Consider the problem of representing real numbers in $[0,1]$ by binary sequences in the following translation invariant manner: Each $x \in [0,1]$ is mapped to a sequence $q \in \{0,1\}^{\mathbb{Z}}$ such that for some appropriate sequence $h \in l^1(\mathbb{Z})$, called the reconstruction *filter*, one has

$$q * h = x, \tag{1.1}$$

where $*$ denotes the additive convolution of two sequences, and the symbol x also denotes the constant sequence $(\dots, x, x, \dots)$. A natural normalization for h is that $\sum h(n) = 1$, so that the number 1 is necessarily represented by the sequence $q = (\dots, 1, 1, \dots)$. As we shall show in Section 2, this problem is too strict in terms of the reconstruction formula (1.1) to be solvable for all x: a solution exists if and only if x is rational, and the solution q is necessarily periodic. An alternative approach is to ask for a sequence of filters $(h_\lambda)_{\lambda>0}$ such that

$$q * h_\lambda \to x, \tag{1.2}$$

uniformly, or at least pointwise, as $\lambda \to \infty$. The normalization condition can also be relaxed to the weaker form $\sum h_\lambda(n) \to 1$ as $\lambda \to \infty$. Clearly, there would not be any gain in introducing this alternative for a choice of sequence (h_λ) that converges in l^1, since the problem would then be immediately reduced to the case (1.1). Indeed,

2000 *Mathematics Subject Classification.* Primary 94A20, 11K06; Secondary 11L07, 41A25.
The author's research was supported by a Francis Robbins Upton honorific graduate fellowship from Princeton University.

a typical choice is $h_\lambda(n) = \frac{1}{\lambda}\chi_{[0,1)}(\frac{n}{\lambda})$, which converges to 0 uniformly as $\lambda \to \infty$, but not in l^1. It turns out that in this new formulation, the problem has plenty of solutions for *all* x. Yet another possibility is to let the binary representation also vary with λ and ask for

$$q_\lambda * h_\lambda \to x. \tag{1.3}$$

The last two settings are quite flexible and the main question consists in finding efficient representations in the sense that (1.2) or (1.3) converges rapidly in λ, where the filters h_λ are scaled versions of an averaging window as in the example we have just given. It may also be desirable to determine the exact rate of convergence for particular schemes that have other features of interest.

In this more general context, it is also possible to formulate the problem for more general functions. Let $x(\cdot)$ be a function on $\mathbb{R}$, taking values in $[0, 1]$. Given a sequence u, we define the measure $\mu_\lambda(u)$ by

$$\mu_\lambda(u) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} u(n)\delta_{n/\lambda}, \tag{1.4}$$

where δ_a denotes the Dirac mass at the point a. Then, for each function x in some appropriate class $\mathcal{C}$, the problem is to find a family (q_λ) of binary representations such that a pre-chosen filter φ (or a sequence (φ_λ) of filters) in $L^1(\mathbb{R})$ can decode x in the sense that

$$(\mu_\lambda(q_\lambda) * \varphi)(\cdot) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} q_\lambda(n)\varphi(\cdot - \tfrac{n}{\lambda}) \quad \longrightarrow \quad x(\cdot) \tag{1.5}$$

in a given functional sense, as $\lambda \to \infty$. Note that (1.3) is already contained in (1.5) if we define the discrete filter h_λ by $h_\lambda(n) = \frac{1}{\lambda}\varphi(\frac{n}{\lambda})$, and restrict our attention to constant functions. However, the analogy between (1.3) and (1.5) is not exact due to the difference between the two settings.

Functions for which there exist solutions to the above problem include *bandlimited* functions, defined by

$$\mathcal{B}_\Omega = \{x : \mathbb{R} \to \mathbb{R} \mid \hat{x} \text{ is a finite Borel measure supported on } [-\Omega, \Omega]\}\,. \tag{1.6}$$

Here, $\hat{x}$ denotes the Fourier transform of x. Such a function is the restriction to $\mathbb{R}$ of an entire function of exponential type. Perhaps, the most important property of bandlimited functions is that they can be recovered from their samples taken at or above the critical density π/Ω, called the *Nyquist density*. This is called the *sampling theorem*. In our discussion, we assume $\Omega = \pi$ to ease the notation; otherwise, the analysis can be transposed by a rescaling of the argument. In this case, if the filter φ is such that $\hat{\varphi}$ is a continuous cut-off function satisfying

$$\hat{\varphi}(\xi) = \begin{cases} 1 & \text{if } |\xi| \le \pi, \text{ and} \\ 0 & \text{if } |\xi| > \lambda\pi, \end{cases} \tag{1.7}$$

then one has the reconstruction formula

$$x(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} x(\tfrac{n}{\lambda})\varphi(t - \tfrac{n}{\lambda}), \tag{1.8}$$

where we assume $\lambda > 1$. The equality holds pointwise, at least for all values of t for which the right hand side converges, and in general everywhere if the summation method uses Cesàro means. Typically, $\hat{\varphi}$ is chosen to be smooth so that the corresponding fast decay of φ enables an almost "local" reconstruction, which removes any concern about the method of summation. (The formula holds also in the L^2 sense when $x \in L^2(\mathbb{R})$, including the case $\lambda = 1$. However, in this critical sampling case, a smooth $\hat{\varphi}$ cannot be chosen; $\hat{\varphi} = \chi_{[-\pi,\pi]}$ is the only candidate.)

We shall search solutions to (1.5) for the class $\mathcal{C} = \{x : \mathbb{R} \to [0,1] \mid x \in \mathcal{B}_\pi\}$. An important class of solutions are generated by the so-called $\Sigma\Delta$ *quantization* (or $\Sigma\Delta$ *modulation*), which transforms the sequence of samples $(x(\frac{n}{\lambda}))$ into the sequence of output bits $(q_\lambda(n))$. The *first order* $\Sigma\Delta$ quantizer operates according to a very simple principle: Given the sequence of samples $(x(\frac{n}{\lambda}))$ taking values in $[0,1]$, a binary sequence q_λ is constructed such that

$$\sum_{n_1}^{n_2} x(\tfrac{n}{\lambda}) \sim \sum_{n_1}^{n_2} q_\lambda(n), \tag{1.9}$$

for all n_1 and n_2. This is done by the following procedure: Define the sequences X_λ, Q_λ and q_λ (which we denote by X, Q, and q when x is constant) by

$$X_\lambda(n) = \sum_{m=1}^{n} x(\tfrac{m}{\lambda}), \tag{1.10}$$

$$Q_\lambda(n) = \lfloor X_\lambda(n) \rfloor, \quad \text{and} \tag{1.11}$$

$$q_\lambda(n) = Q_\lambda(n) - Q_\lambda(n-1). \tag{1.12}$$

Since x takes values in $[0,1]$, we have $q_\lambda(n) \in \{0,1\}$; and at the same time (1.9) is satisfied, up to an error less than 1. X_λ can be defined naturally for negative indices as well, by integrating backwards.

The property (1.9) implies an immediate estimate for constants; if the resulting sequence q is filtered using a rectangular averaging window of size λ, i.e. $h_\lambda(n) = \frac{1}{\lambda}\chi_{[0,1)}(\frac{n}{\lambda})$, then

$$\|q * h_\lambda - x\|_{l^\infty} \leq \frac{1}{\lambda}. \tag{1.13}$$

It turns out, as shown in [**5**], that the same estimate holds (in the sense of (1.5)) for the whole class of bandlimited functions, using a smooth window that satisfies (1.7). One of the main results we shall present in this paper is that the exponent of λ in this estimate can be improved using number theoretical tools, both for the special case of constants and the general case of bandlimited functions. After a discussion of general considerations in Section 2, we shall present these number theoretical estimates in Sections 3 and 4.

The summation, truncation and differencing steps in (1.10) to (1.12) explain the name "$\Sigma\Delta$ quantization". The reason for the name "first order" is that only first order sums and differences have been used; a k-th order scheme involves k-th order sums and differences. The first construction of stable $\Sigma\Delta$ quantization schemes for all orders is due to Daubechies and DeVore [**5**]. We shall return to this briefly in Section 5. Let us also mention the references [**1**] and [**2**] which cover the history as well as recent advances in the theory and applications of $\Sigma\Delta$ quantization in electrical engineering.

2. General Considerations for Constants

Let us return to the problem (1.1). We claimed that a solution exists if and only if x is rational. This claim will follow simply as a corollary to a theorem of Szegö. We first recall the definition of spectral set for bounded sequences.

Definition 2.1. Let a be a sequence in l^∞. Then the spectral set $\sigma(a)$ is the set of all $\xi \in \mathbb{T}$ such that $b * a = 0$ ($b \in l^1$) implies $\sum_n b(n)e^{-in\xi} = 0$.

The spectral set of a sequence a in l^1 is precisely the closed support of its Fourier transform $\hat{a}(\xi) = \sum_n a(n)e^{-in\xi}$. Definition 2.1 extends this notion for arbitrary sequences in l^∞, whose Fourier transforms in general are not functions. It is always true that $\sigma(a)$ is a closed subset of $\mathbb{T}$.

Szegö's theorem, as stated in the following form in [**7, 8**] by Helson, deals with spectral sets of sequences whose terms come from finite sets:

Theorem 2.2 (Szegö-Helson). *Let a be a sequence whose terms are all drawn from a finite set S of complex numbers. Unless the sequence is periodic, its spectral set fills $\mathbb{T}$.*

Let us apply this powerful theorem to the sequence $q - x$. Certainly, the terms of this sequence come from the finite set $\{-x, 1-x\}$. On the other hand, $h * x = x$, since $\sum h(n) = 1$. Thus, (1.1) is a restatement of $(q - x) * h = 0$. According to the conclusion of the theorem, a nonperiodic q (hence a nonperiodic $q - x$) would mean $\sigma(q - x) = \mathbb{T}$. This implies $\hat{h}(\xi) = \sum h(n)e^{-in\xi} = 0$ for all $\xi \in \mathbb{T}$, and hence $h \equiv 0$, a contradiction. Hence, only periodic solutions of (1.1) may exist.

Now, consider a periodic solution q, whose period is N. Then $q * h$ is also periodic and its period divides N. It is a simple calculation to show that

$$\sum_{n=1}^{N}(q * h)(n) = \sum_{n=1}^{N} q(n). \tag{2.1}$$

Since the left hand side is Nx and the right hand side is an integer $0 \leq M \leq N$, it follows that $x = M/N$, i.e., x is rational. This proves the "only if" part of the assertion. The "if" part follows from a trivial construction. Consider a rational number $x = M/N$ and let h be the rectangular averaging window of length N. If q is such that it has period N and exactly M of $\{q(k) : 1 \leq k \leq N\}$ are equal to 1, then it is clear that $q * h = x$. In summary:

Theorem 2.3. *There is a solution to (1.1) if and only if x is rational. The solution q is necessarily periodic and its period is a multiple of the denominator of x in its reduced form.*

The rectangular filter of size N would work for all $x = M'/N'$, where N' divides N. (So, if the size of the window is chosen to be l.c.m.$(1, \ldots, N)$, then all numbers in the Farey sequence $\mathcal{F}_N$ can be decoded with it.[1]) Now, let us consider the implications of (2.1) on the filter h. If a rational number $x = r/s$ (in its reduced form) is represented by a sequence q of period N, then s must divide N and precisely $M = Nr/s$ of $\{q(k) : k = 1, \ldots, N\}$ are equal to 1. Let the corresponding indices be $l_1, \ldots, l_M$ and define $P(\xi) = \sum_{j=1}^{M} e^{-il_j\xi}$. Then, a straightforward calculation shows that $q * h = x$ implies $P(\xi)\hat{h}(\xi) = 0$ for all $\xi = 2\pi p/N$, $p = 1, \ldots, N-1$. If $M \neq N$ (i.e., $x \neq 1$), then $\hat{h}(\xi)$ must vanish at least at one of these points. Indeed, for almost any choice of P, almost all of these roots would belong to $\hat{h}$. This leads us to conjecture the following:

[1] $\mathcal{F}_N$ is the set of all rationals in $[0, 1]$ with denominators less than or equal to N.

CONJECTURE 2.4. *It is impossible to construct a "universal" filter that can decode all rationals simultaneously. Equivalently, there is no collection of sequences in* $\{0,1\}^{\mathbb{Z}}$ *such that for every* $x \in \mathbb{Q} \cap [0,1]$*, there is a sequence* q *in this collection for which* $q * h = x$*, where* h *is fixed.*

Let us consider the alternative formulation (1.2). We have already stated in the introduction that it is now possible to find solutions for all $x \in [0,1]$. Even stronger, *universal* filter sequences (h_λ) can be employed. Before looking at $\Sigma\Delta$ quantization more closely as a scheme that generates such solutions, let us state the following result which is merely an extension of Theorem 2.3.

THEOREM 2.5. *A solution* q *to (1.2) for an irrational* x *is necessarily non-periodic.*

PROOF. Suppose (1.2) is satisfied for some x and a periodic $q \in \{0,1\}^{\mathbb{Z}}$. Let N be the period of q. Then, as in (2.1), we have

$$\sum_{n=1}^{N} (q * h_\lambda)(n) \rightarrow \sum_{n=1}^{N} q(n). \tag{2.2}$$

Combined with (1.2), this implies $x = \frac{1}{N} \sum\limits_{n=1}^{N} q(n)$, i.e. $x \in \mathbb{Q}$. □

3. $\Sigma\Delta$ Quantization: Constants

Let us look at the very simple first order $\Sigma\Delta$ system (given in (1.10) to (1.12)) for constant input $x \in [0,1]$. Define the auxiliary variable $u(n)$ to be $X(n) - Q(n)$. It is equal to the fractional part of $X(n)$, which we denote by $\langle X(n) \rangle$. In practice, neither $X(n)$ nor $Q(n)$ are computed in an electronic circuit, since these variables are unbounded. However, the sequence u is bounded and satisfies the recursion relation

$$u(n) - u(n-1) = x - q(n), \qquad u(0) = 0. \tag{3.1}$$

In fact, this recursion is taken as the starting point in practice. One asks for a bounded solution u of (3.1) such that $q \in \{0,1\}^{\mathbb{Z}}$. The particular construction of q we have considered is just one of the solutions of (3.1).

We are now in the setting (1.2). For a given filter h_λ, let us derive an estimate for the error $e_\lambda = x - q * h_\lambda$, where h_λ obeys the scaling relation $h_\lambda(n) = \frac{1}{\lambda}\varphi(\frac{n}{\lambda})$ for some function φ. For $a \in l^1$, define $S(a) = \sum a(n)$. The error may be bounded by a sum of two contributions:

$$|e_\lambda(n)| \leq |x(1 - S(h_\lambda))| + \Big| \sum_k (x - q(k)) h_\lambda(n-k) \Big|. \tag{3.2}$$

Let us call the two terms e_λ^1 and e_λ^2. It is possible to choose φ such that the first error term is zero for all λ. For instance, $\varphi = \chi_{[0,1]}$ has this property for all integer λ; on the other hand, $\varphi \in BV \cap \mathcal{B}_\pi$ with $\hat{\varphi}(0) = 1$ has it for all real $\lambda > 1$ (which may easily be seen using Poisson's summation formula). Assume a choice of φ with this property, which leaves us with $e_\lambda^2 = (x - q) * h_\lambda$.

THEOREM 3.1. *For all* λ*,* $\|e_\lambda\|_{l^\infty} \leq \frac{1}{\lambda}\mathrm{Var}(\varphi)$*.*

PROOF. Let Δ denote the difference operator acting on sequences, defined by $\Delta u(n) = u(n) - u(n-1)$. Using the recursion relation (3.1), and that u is bounded by 1, it follows that

$$\begin{aligned} \|e_\lambda\|_{l^\infty} &= \|\Delta u * h_\lambda\|_{l^\infty}, \\ &= \|u * \Delta h_\lambda\|_{l^\infty}, \qquad (3.3)\\ &\leq \|u\|_{l^\infty}\|\Delta h_\lambda\|_{l^1}, \\ &\leq \frac{1}{\lambda}\mathrm{Var}(\varphi). \qquad (3.4) \end{aligned}$$

□

We call this the "basic estimate", in the sense that only boundedness of u was used in the derivation. We shall improve the exponent of λ by examining the expression (3.3) more closely.

First we note that $u * \Delta h_\lambda = (u-c) * \Delta h_\lambda$ for any constant c. Next, we consider a particular filter, the triangle function $\varphi(t) = (1-|t|)\chi_{[-1,1]}(t)$. Then,

$$(u * \Delta h_\lambda)(n) = \frac{1}{\lambda^2}\sum_{k=0}^{\lambda-1}\big(u(n+k) - \frac{1}{2}\big) - \frac{1}{\lambda^2}\sum_{k=1}^{\lambda}\big(u(n-k) - \frac{1}{2}\big) \qquad (3.5)$$

In the above expression, the value of c was chosen to be $1/2$ in order to exploit the fact that the state variable $u(n) = \langle X(n)\rangle = \langle nx\rangle$ forms a uniformly distributed sequence in $[0,1]$ for all irrational x. Using well known results in the theory of uniform distribution, we now prove the following improved estimate:

THEOREM 3.2. *Let $\epsilon > 0$ be given. Then for almost every $x \in [0,1]$, one has the estimate $\|e_\lambda\|_{l^\infty} \leq C_x \lambda^{-2}\log^{2+\epsilon}\lambda$, using the triangular filter.*

PROOF. *Koksma's inequality*[2] reduces the problem to considering the discrepancy values for the two sequences $u(n-\lambda), \dots, u(n-1)$ and $u(n), \dots, u(n+\lambda-1)$. The discrepancy can be bounded using the *Erdős-Turán inequality*[3]. In our case, these two inequalities result in

$$\Big|\frac{1}{\lambda}\sum_{k=a+1}^{a+\lambda} u(k) - \frac{1}{2}\Big| \leq \inf_K\, C\left(\frac{1}{K} + \sum_{k=1}^{K}\frac{1}{k}\Big|\frac{1}{\lambda}\sum_{m=1}^{\lambda} e^{2\pi ikmx}\Big|\right), \qquad (3.8)$$

uniformly in a and for all λ. The precise behaviour of this quantity depends on the behaviour of the continued fraction expansion of x (see, e.g. [**11, 12, 13**]). We shall not pursue such a detailed analysis, but rather use a metric result due to Khinchine, which yields the bound $O(\lambda^{-1}\log^{2+\epsilon}\lambda)$ for almost every (Lebesgue) x (see [**11**, pp. 131]). Let us note that the involved constant in general depends on x. This, together with (3.5) gives us the desired estimate. □

[2] For any function $f \in BV([0,1])$ and a finite sequence of points $x_1, \dots, x_N$ in $[0,1]$,

$$\Big|\frac{1}{N}\sum_{n=1}^{N} f(x_n) - \int_0^1 f(t)dt\Big| \leq \mathrm{Var}(f) D_N, \qquad (3.6)$$

where D_N denotes the *discrepancy* of the sequence $x_1, \dots, x_N$ and $\mathrm{Var}(f)$ is the variation of f.

[3] The discrepancy D_N of any real numbers $x_1, \dots, x_N$ is bounded by

$$D_N \leq C\left(\frac{1}{K} + \sum_{k=1}^{K}\frac{1}{k}\Big|\frac{1}{N}\sum_{m=1}^{N} e^{2\pi ikx_m}\Big|\right) \qquad (3.7)$$

for any positive integer K, where C is an absolute constant.

Let us use the notation $e_{\lambda,x}$ to denote the dependence of the error e_λ on the input value x. We assume that the triangular filter is used, so that (3.5) and (3.8) are in effect. We have already mentioned above that the precise behaviour of $e_{\lambda,x}$ can only be described by means of the continued fraction expansion of x. However, the mean behaviour is simpler. Let us consider the mean squared error (MSE)

$$\text{MSE}(e_\lambda) = \int_0^1 \|e_{\lambda,x}\|_{l^\infty}^2 \, dx \tag{3.9}$$

A straightforward bound for $\text{MSE}(e_\lambda)$ can be found directly from (3.8). Let $P_{\lambda,k}(x)$ denote the trigonometric polynomial $\lambda^{-1}\sum_{m=1}^{\lambda} e^{2\pi i k m x}$. Note that $\|P_{\lambda,k}\|_{L^2([0,1])} = \lambda^{-1/2}$. Then,

$$\begin{aligned}
\text{MSE}(e_\lambda) &\le \frac{C}{\lambda^2}\int_0^1 \inf_K \Big(\frac{1}{K} + \sum_{k=1}^{K} k^{-1}|P_{\lambda,k}(x)|\Big)^2 dx \\
&\le \frac{C'}{\lambda^2}\inf_K \int_0^1 \left(\frac{1}{K^2} + \Big(\sum_{k=1}^{K} k^{-1}|P_{\lambda,k}(x)|\Big)^2\right) dx \\
&\le \frac{C'}{\lambda^2}\inf_K \left(\frac{1}{K^2} + \sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{kl}\int_0^1 |P_{\lambda,k}(x)||P_{\lambda,l}(x)|dx\right) \\
&\le \frac{C'}{\lambda^2}\inf_K \left(\frac{1}{K^2} + \frac{1}{\lambda}\log^2 K\right),
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the last step. Finally, by choosing $K \sim \lambda^{1/2}$, we arrive at the bound

$$\text{MSE}(e_\lambda) \le C\lambda^{-3}\log^2\lambda. \tag{3.10}$$

The exponent of λ in this estimate is optimal. Indeed, using number theoretical tools, it is shown in [**14**] that

$$C_1\lambda^{-3} \le \int_0^1 |e_{\lambda,x}(0)|^2 dx \le C_2\lambda^{-3}. \tag{3.11}$$

It is natural to conjecture that this is also true for $\|\int_0^1 |e_{\lambda,x}(\cdot)|^2 dx\|_{l^\infty}$, a quantity smaller than $\text{MSE}(e_\lambda)$. It is shown in [**3**] that the quantity

$$\int_0^1 \frac{1}{2N+1}\sum_{n=-N}^{N} |e_{\lambda,x}(n)|^2 dx \tag{3.12}$$

(which is yet even smaller) also behaves as $O(\lambda^{-3})$ as $N \to \infty$.

4. ΣΔ Quantization: Arbitrary Bandlimited Inputs

In this section, we generalize the results of the previous section to arbitrary bandlimited functions. The difference equation now reads as

$$u_\lambda(n) - u_\lambda(n-1) = x_\lambda(n) - q_\lambda(n), \qquad u(0) = 0; \tag{4.1}$$

where $x_\lambda(n)$ and $u_\lambda(n)$ are defined to be $x(\frac{n}{\lambda})$ and $X_\lambda(n) - Q_\lambda(n) = \langle X_\lambda(n)\rangle$, respectively. Clearly, u_λ takes its values in $[0,1]$. Our setting is (1.5), and in general we allow φ to depend on λ, which will be denoted by φ_λ.

The Basic Estimate. We start with the corresponding "basic estimate" given in [**5**], where it is also possible to use a fixed filter φ for all λ.

THEOREM 4.1 ([**5**]). *Let $x \in \mathcal{B}_\pi$ with $0 \leq x(t) \leq 1$ for all t, and $\varphi \in BV(\mathbb{R})$ satisfying (1.7) for some fixed $\lambda_0 > 1$. Then, $\|x - \mu_\lambda(q_\lambda) * \varphi\|_{L^\infty} \leq \frac{1}{\lambda}\mathrm{Var}(\varphi)$ for all $\lambda \geq \lambda_0$.*

PROOF. The sampling theorem states that $x = \mu_\lambda(x_\lambda) * \varphi$ for all $\lambda \geq \lambda_0$. Let Δ_η be the difference operator whose action on a measure is given by $\Delta_\eta \mu(\cdot) = \mu(\cdot) - \mu(\cdot - \eta)$ and let $\mathbf{1}$ denote the constant sequence of 1's. Then,

$$\begin{aligned} x - \mu_\lambda(q_\lambda) * \varphi &= \mu_\lambda(x_\lambda - q_\lambda) * \varphi, \\ &= \mu_\lambda(\Delta u_\lambda) * \varphi, \\ &= \Delta_{1/\lambda}\mu_\lambda(u_\lambda) * \varphi, \\ &= \mu_\lambda(u_\lambda) * \Delta_{1/\lambda}\varphi, \end{aligned} \tag{4.2}$$

so that

$$\begin{aligned} \|x - \mu_\lambda(q_\lambda) * \varphi\|_{L^\infty} &\leq \|\mu_\lambda(\mathbf{1}) * |\Delta_{1/\lambda}\varphi|\,\|_{L^\infty}, \\ &\leq \frac{1}{\lambda}\mathrm{Var}(\varphi). \end{aligned} \tag{4.3}$$

In the last step, we made use of the identity $\mu_\lambda(\mathbf{1}) * f = \frac{1}{\lambda}\sum f(\cdot - \frac{n}{\lambda})$. □

An Improved Estimate. We shall apply the ideas of the previous section to prove the following theorem, which is an improvement of the above basic estimate:

THEOREM 4.2. *For all $\eta > 0$, there exists a family $\{\varphi_\lambda\}_{\lambda \geq 1}$ of filters such that, for all x in Theorem 4.1, and all t for which $x'(t)$ does not vanish, we have*

$$|x(t) - (\mu_\lambda(q_\lambda) * \varphi_\lambda)(t)| \leq C\lambda^{-4/3+\eta} \tag{4.4}$$

for some constant $C = C(\eta, x'(t))$.

Note that bandlimited functions are analytic, so that the derivative x' of a non-constant bandlimited function x has at most countably many zeros, with no accumulation point. It is also possible to carry out a higher order analysis at the zeros of x', but we shall not touch upon this here.

PROOF. We divide the proof into a number of steps.
1. Fix t (for which $x'(t) \neq 0$). For each λ, let $N_\lambda = \lfloor \lambda t \rfloor$, and define the sequence U_λ by

$$U_\lambda(n) - U_\lambda(n-1) = u_\lambda(n) - \frac{1}{2}, \qquad U_\lambda(N_\lambda) = 0. \tag{4.5}$$

Let also $t_\lambda = N_\lambda/\lambda$ and $\delta_\lambda = t - t_\lambda$. Note that $|\delta_\lambda| \leq 1/\lambda$. Now, (4.2) can be written as

$$\begin{aligned} x(t) - (\mu_\lambda(q_\lambda) * \varphi_\lambda))(t) &= \frac{1}{\lambda}\sum_n \Delta U_\lambda(n)\Delta_{1/\lambda}\varphi(t - \tfrac{n}{\lambda}), \\ &= \frac{1}{\lambda}\sum_n U_\lambda(n)\Delta^2_{1/\lambda}\varphi(t - \tfrac{n}{\lambda}), \\ &= \frac{1}{\lambda}\sum_n U_\lambda(N_\lambda + n)\Delta^2_{1/\lambda}\varphi(-\tfrac{n}{\lambda} + \delta_\lambda), \end{aligned} \tag{4.6}$$

for any φ that decays sufficiently fast. Denote the error expression by $e_\lambda(t)$.

2. Our purpose is to find non-trivial bounds for $U_\lambda(N_\lambda+n)$ by accounting for the cancellations in

$$U_\lambda(N_\lambda+n)=\sum_{m=1}^{n}\Big(u_\lambda(N_\lambda+m)-\frac{1}{2}\Big), \tag{4.7}$$

where we have assumed $n>0$, the other case being essentially the same. Note that the trivial bound is $|n|/2$. We shall prove the following estimate:

$$|U_\lambda(N_\lambda+n)|\le C_1\Big(\lambda^{2/3}+\frac{\lambda^{1/2}}{|x'(t)|^{1/2}}\Big), \tag{4.8}$$

for all $n\le C_2|x'(t)|\lambda$ and $\lambda>C_3|x'(t)|^{-1}$. [4]

The inequalities of Koksma and Erdős-Turán result in the bound

$$|U_\lambda(N_\lambda+n)|\le\inf_K\, C\left(\frac{n}{K}+\sum_{k=1}^{K}\frac{1}{k}\Big|\sum_{m=1}^{n}e^{2\pi iku_\lambda(N_\lambda+m)}\Big|\right), \tag{4.9}$$

which reduces our task to analyzing the behaviour of the exponential sums

$$S_{\lambda,k}(n):=\sum_{m=1}^{n}e^{2\pi ikX_\lambda(N_\lambda+m)}, \tag{4.10}$$

since $u_\lambda(n)=\langle X_\lambda(n)\rangle$.

3. We will use the stationary phase methods of van der Corput to estimate the exponential sums given in (4.10). The following well-known theorems serve well for this purpose:

Theorem 4.3 (Truncated Poisson, [**9**]). *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*

$$\sum_{a\le m\le b}e^{2\pi if(m)}=\sum_{\alpha-1\le\nu\le\beta+1}\int_a^b e^{2\pi i(f(\tau)-\nu\tau)}d\tau+\ O(\log(2+\beta-\alpha)). \tag{4.11}$$

(If f' is decreasing, then by taking the complex conjugate of the above expression applied to $-f$, one finds the same expression with α and β switched.)

Theorem 4.4 (van der Corput, [**10**]). *Suppose ϕ is real-valued and smooth in (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 there exists an absolute constant c_r such that*

$$\Big|\int_a^b e^{i\phi(t)}\,dt\Big|\le c_r\mu^{-1/r}. \tag{4.12}$$

In our case, X_λ is initially only defined on the integers; however, (1.10) immediately yields an (analytic) interpolation of X_λ. We call this new function X_λ as

[4] As is customary, we shall use the notations $C,C',C_1,C_2,...$ for generic constants that may change value from one proof to another; constants of different values occurring in the same argument will be distinguished by different indices.

well, and show in the Appendix that for $\lambda \geq C|x'(t)|^{-1}$, and all real τ in the range $0 \leq \tau \leq C_3|x'(t)|\lambda$, one has

$$C_1\frac{|x'(t)|}{\lambda} \leq |X''_\lambda(N_\lambda+\tau)| \leq C_2\frac{|x'(t)|}{\lambda}, \tag{4.13}$$

where C, C_1, C_2, and C_3 are absolute numerical constants.

In Theorem 4.3, we set $f = kX_\lambda$, $a = N_\lambda + 1$, and $b = N_\lambda + n$, and assume $n \leq C_3|x'(t)|\lambda$. It follows from (4.13) that the number of integral terms in the right hand side of (4.11) is bounded by

$$\begin{aligned} |\beta-\alpha+3| &\leq 3 + k(n-1)\sup_{1\leq\tau\leq n}|X''_\lambda(N_\lambda+\tau)| \\ &\leq 3 + k(n-1)C_2\frac{|x'(t)|}{\lambda}. \end{aligned} \tag{4.14}$$

On the other hand, using Theorem 4.4 (for $r = 2$) with (4.13), each exponential integral term in (4.11) is bounded by $C(k|x'(t)|/\lambda)^{-1/2}$. Combining this with the bound on the number of terms that we have just found, we get

$$|S_{\lambda,k}(n)| \leq C_1 n\Big(\frac{k}{\lambda}\Big)^{1/2} + C_2\Big(\frac{k}{\lambda}\Big)^{-1/2}|x'(t)|^{-1/2} + \ O(\log(2+k)), \tag{4.15}$$

where in the first term we have made use of the fact that $\|x'\|_{L^\infty} \leq \pi$ (which follows from Bernstein's inequality[5]), and in the logarithmic term that $|\beta-\alpha| \leq k$ for the given range of n. Note that, for small k, this bound significantly improves the trivial bound n. Now, if in (4.9), one chooses $K \sim \lambda^{1/3}$, then (4.15) yields our desired estimate (4.8).

4. We finish the proof of Theorem 4.4 by bounding (4.6) for a particular family of filters which we construct next. For this, we fix a filter φ such that $\hat\varphi$ is C^∞, $\operatorname{supp}(\hat\varphi) \subset [-c_0\pi, c_0\pi]$ for some small fixed $c_0 > 1$, and $\hat\varphi(\xi) = 1$ on $[-\pi,\pi]$. Then φ is a Schwartz function, i.e., φ has rapidly decreasing derivatives: there are constants $C_N^{(l)}$ for all $N \geq 0$ and $l \geq 0$ such that

$$|\varphi^{(l)}(t)| \leq \frac{C_N^{(l)}}{(1+|t|)^N}. \tag{4.16}$$

For a small $\eta > 0$, we set $\Omega_\lambda = \lambda^{\eta/2}$ and define φ_λ by

$$\varphi_\lambda(t) = \Omega_\lambda\varphi(\Omega_\lambda t) \tag{4.17}$$

for $\lambda \geq 1$. Then $\hat\varphi_\lambda(\xi) = \hat\varphi(\xi/\Omega_\lambda)$ and hence $\{\varphi_\lambda\}$ is an admissible family of reconstruction filters. We turn back to the expression (4.6). For small n (i.e., for $|n| \leq c|x'(t)|\lambda$ for a sufficiently small constant c), we will use the estimate (4.8) in the form $O(|x'(t)|^{-1/2}\lambda^{2/3})$, and for large n, the trivial estimate $|n|/2$. Thus,

$$\begin{aligned} |e_\lambda(t)| &\leq \frac{1}{\lambda}\Big(O(|x'(t)|^{-1/2}\lambda^{2/3})\sum_{|n|\leq c|x'(t)|\lambda}|\Delta^2_{1/\lambda}\varphi_\lambda(-\tfrac{n}{\lambda}+\delta_\lambda)| \\ &\qquad + \sum_{|n|>|x'(t)|c\lambda}\frac{|n|}{2}|\Delta^2_{1/\lambda}\varphi_\lambda(-\tfrac{n}{\lambda}+\delta_\lambda)|\Big) \end{aligned} \tag{4.18}$$

The first sum term can easily be bounded by

$$2\lambda^{-1}\|\varphi''_\lambda\|_{L^1} = 2\lambda^{-1+\eta}\|\varphi''\|_{L^1}, \tag{4.19}$$

[5] If $f \in \mathcal{B}_\Omega$, then $\|f^{(s)}\|_{L^\infty} \leq \Omega^s\|f\|_{L^\infty}$.

and the second sum term by

$$\sum_{|n|>c|x'(t)|\lambda} \frac{|n|}{2} \cdot \frac{2}{\lambda^2} \cdot \frac{\Omega_\lambda^3 C_N^{(2)}}{(1+\Omega_\lambda |n| \lambda^{-1})^N} \leq C(N, |x'(t)|)\Omega_\lambda^{-N+3}, \tag{4.20}$$

for all N. We choose N such that $(N-3)\eta/2 > 1/3$. Putting together, this results in the estimate

$$\epsilon_\lambda(t) \leq C(\eta, |x'(t)|)\lambda^{-4/3+\eta}, \tag{4.21}$$

concluding the proof. □

5. Other Results

In the introduction, we mentioned higher order $\Sigma\Delta$ schemes. In a k-th order $\Sigma\Delta$ scheme, one is interested in solutions $q \in \{0,1\}^{\mathbb{Z}}$ of the difference equation

$$\Delta^k u = x - q, \tag{5.1}$$

with $u \in l^\infty$ (such a scheme is called *stable*). Again, we use the notations u_λ, x_λ and q_λ when x is an arbitrary bandlimited function sampled at the rate λ.

The second order case is best understood among all higher order schemes, and there are many construction strategies for q that are stable. We shall not get into any specific details of these schemes, but state only the error estimates. The basic estimate is now $O(\lambda^{-2})$ both for constants and general bandlimited functions. However, in this case, the exponent can be improved only for constants [**15**]. A notable difficulty arises in analyzing a particular family of piecewise affine dynamical systems in the plane, whose attracting invariant sets turn out to tile the plane with the action of the integer lattice. Using similar number theoretical techniques (such as mean values of Gauss sums and higher dimensional versions of discrepancy and the inequalities of Koksma and Erdős-Turán), the MSE(e_λ) is shown to be bounded by $O(\lambda^{-4.5})$. A similar improvement of the exponent for the uniform error is also found using the corresponding metric results.

The basic estimate for a stable k-th order scheme is $O(\lambda^{-k})$, which makes higher order schemes more interesting in terms of their approximation properties. However in practice, the first order scheme is widely used for its many attractive features, such as its robustness to hardware imperfections. There are various (mostly ad hoc) constructions in the electrical engineering literature of schemes of higher (single digit) orders ([**1, 2**]), without proof of stability. A recent mathematical achievement in this direction has been the construction of a family of arbitrary order stable $\Sigma\Delta$ schemes in [**5**].

For a given order k, let C_k be the constant hidden in the $O(\lambda^{-k})$ estimate. Then, it is natural to let the order depend on λ and look for the best possible decay of the error in λ. For the scheme in [**5**], the authors show that $C_k \sim c^{k^2}$, which results in an $O(c^{-\log^2 \lambda})$ type decay using the optimal choice of order for each λ. On the other hand, a lower bound can easily shown to be $2^{-\lambda}$ using Kolmogorov entropy. It is an unsolved problem to determine whether exponential decay of error can be achieved.[6] Another step towards this lower bound is taken in [**17**] for the

[6] **Note added in October 2003:** This problem is now solved. In [**16**], we construct $\Sigma\Delta$ families which collectively achieve the error bound $\|e_\lambda\|_{L^\infty} = O(2^{-0.07\lambda})$ for arbitrary π-bandlimited functions. The best achievable constant in the exponent is still unknown.

constant input case, in which a purely number theoretical construction yields an $O(c^{-\sqrt{\lambda}})$ type of error decay.

6. Appendix

In this section, we present the construction of the analytic interpolation of the sequence X_λ, which was stated in the proof of Theorem 4.4. Using the Taylor expansion of x about the point t_λ, we proceed as follows:

$$\begin{aligned} X_\lambda(N_\lambda+m) &= X_\lambda(N_\lambda)+\sum_{l=1}^{m} x(t_\lambda+\tfrac{l}{\lambda}) \\ &= X_\lambda(N_\lambda)+\sum_{s=0}^{\infty}\frac{x^{(s)}(t_\lambda)}{\lambda^s s!}\sum_{l=1}^{m} l^s \\ &= X_\lambda(N_\lambda)+\sum_{s=0}^{\infty}\frac{x^{(s)}(t_\lambda)}{\lambda^s s!}\sum_{j=1}^{s+1} P_{s,j}m^j \qquad (6.1) \\ &= X_\lambda(N_\lambda)+\sum_{j=1}^{\infty} m^j \sum_{s=j-1}^{\infty}\frac{x^{(s)}(t_\lambda)}{\lambda^s s!}P_{s,j}, \qquad (6.2) \end{aligned}$$

where $P_{s,j}$ are related to Bernoulli numbers. We use this last expression to define

$$X_\lambda(N_\lambda+\tau) = X_\lambda(N_\lambda)+\sum_{j=1}^{\infty} a_j\tau^j \qquad (6.3)$$

for all $\tau \geq 0$, where a_j is the sum term appearing in (6.2). The simple bound $P_{s,j} \leq s!/j!$ and Bernstein's inequality easily yields

$$|a_j| < \frac{2}{j!}\Big(\frac{\pi}{\lambda}\Big)^{j-1} \qquad (6.4)$$

for $\lambda > 2\pi$. Let us show that

$$X'_\lambda(N_\lambda+\tau) = x(t_\lambda+\tfrac{\tau}{\lambda})+R_\lambda(\tau), \qquad (6.5)$$

where $R_\lambda(\tau)$ is small compared to $x(t_\lambda+\frac{\tau}{\lambda})$ for $\tau = O(\lambda)$. We start with noting that $P_{s,s+1} = 1/(s+1)$ for all s. Then, starting from (6.1),

$$\begin{aligned} X'_\lambda(N_\lambda+\tau) &= \sum_{s=0}^{\infty}\frac{x^{(s)}(t_\lambda)}{\lambda^s s!}\sum_{j=1}^{s+1} P_{s,j}j\tau^{j-1} \\ &= \sum_{s=0}^{\infty}\frac{x^{(s)}(t_\lambda)}{\lambda^s s!}\Big(\tau^s+\sum_{j=1}^{s} P_{s,j}j\tau^{j-1}\Big) \\ &= x(t_\lambda+\tfrac{\tau}{\lambda})+R_\lambda(\tau), \qquad (6.6) \end{aligned}$$

where

$$\begin{aligned} R_\lambda(\tau) &= \sum_{s=0}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} \sum_{j=1}^{s} P_{s,j} j\tau^{j-1} \\ &= \sum_{j=1}^{\infty} j\tau^{j-1} \sum_{s=j}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} P_{s,j} \\ &=: \sum_{j=1}^{\infty} j\tau^{j-1} b_j. \end{aligned} \tag{6.7}$$

A similar estimate for b_j is

$$|b_j| < \frac{2}{j!}\Big(\frac{\pi}{\lambda}\Big)^j, \tag{6.8}$$

which, through (6.6) and (6.7), provides us the estimate

$$|X''_\lambda(N_\lambda + \tau) - \tfrac{1}{\lambda}x'(t_\lambda + \tfrac{\tau}{\lambda})| \le 2\big(\tfrac{\pi}{\lambda}\big)^2 e^{\tau\pi/\lambda}. \tag{6.9}$$

Now,

$$|x'(t_\lambda + \tfrac{\tau}{\lambda}) - x'(t)| \le \frac{(\tau+1)}{\lambda}\pi^2, \tag{6.10}$$

so that

$$\frac{1}{2}|x'(t)| \le |x'(t_\lambda + \tfrac{\tau}{\lambda})| \le \frac{3}{2}|x'(t)|, \tag{6.11}$$

for all $0 \le \tau \le C\lambda|x'(t)|$, where C is a sufficiently small absolute constant. Hence, from (6.9) and (6.11), it follows that

$$\begin{aligned} |X''_\lambda(N_\lambda + \tau)| &\ge \frac{1}{\lambda}|x'(t_\lambda + \tfrac{\tau}{\lambda})| - 2e^{\pi\tau/\lambda}\big(\tfrac{\pi}{\lambda}\big)^2 \\ &\ge C_1 \frac{|x'(t)|}{\lambda} \end{aligned} \tag{6.12}$$

if $\lambda \ge C'|x'(t)|^{-1}$. It follows similarly that

$$|X''_\lambda(N_\lambda + \tau)| \quad \le \quad C_2 \frac{|x'(t)|}{\lambda} \tag{6.13}$$

for the same range of λ and τ.

7. Acknowledgments

The author would like to thank I. Daubechies for her support and helpful advice during this work, and P. Sarnak for suggesting [**9**] for number-theoretic results of possible relevance to this problem. The author would also like to thank W. Schlag for a remark that simplified an earlier version of the proof of Theorem 4.4 and which led to a generalization of the class of filters that can be used.

References

[1] S.R. Norsworthy, R. Schreier, G.C. Temes, eds., *Delta-Sigma Data Converters*, IEEE Press, 1997.

[2] J.C. Candy, G.C. Temes, eds., *Oversampling Delta-Sigma Data Converters*, IEEE Press, 1992.

[3] R.M. Gray, "Spectral Analysis of Quantization Noise in a Single-Loop Sigma-Delta Modulator with dc Input", *IEEE Transactions on Communications, Vol. 37, June 1989.*

[4] R.M. Gray, W. Chou, P.W. Wong, "Quantization Noise in Single-Loop Sigma-Delta Modulation with Sinusoidal Inputs", *IEEE Transactions on Communications, Vol. 37, Sept. 1989.*

[5] I. Daubechies, R. DeVore, "Reconstructing a Bandlimited Function From Very Coarsely Quantized Data: A Family of Stable Sigma-Delta Modulators of Arbitrary Order", *Ann. of Math.*, vol. 158, no. 2, pp. 679–710, Sept. 2003.

[6] C.S. Güntürk, "Approximating a Bandlimited Function Using Very Coarsely Quantized Data: Improved Error Estimates in Sigma-Delta Modulation", *J. Amer. Math. Soc.*, posted on August 1, 2003, PII S 0894-0347(03)00436-3 (to appear in print).

[7] H. Helson, *Harmonic Analysis*, Addison-Wesley, 1983.

[8] H. Helson, "On a theorem of Szegö", *Proc. Amer. Math. Soc.* 6 (1955), 235-242.

[9] H.L. Montgomery, *Ten Lectures on the Interface Between Analytic Number Theory and Harmonic Analysis*, AMS, 1994.

[10] E.M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, 1993.

[11] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, Wiley, 1974

[12] S. Lang, *Introduction to Diophantine Approximation*, Springer-Verlag, 1995.

[13] M. Drmota, R.F. Tichy, *Sequences, Discrepancies and Applications*, Springer, 1997.

[14] C.S. Güntürk, J.C. Lagarias, and V.A. Vaishampayan, "On the robustness of single loop sigma-delta modulation," *IEEE Transactions on Information Theory.*, 47 (5):1735–1744, July 2001.

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

[16] C.S. Güntürk, "One-Bit Sigma-Delta Quantization with Exponential Accuracy," *Comm. Pure Appl. Math.*, vol. 56, pp. 1608–1630, no. 11, 2003.

[17] S. Konyagin, private communication, 1999.

Program in Applied and Computational Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ, 08544.

Current address: Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY, 10012.

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