On the Robustness of Single Loop Sigma-Delta Modulation

C. Sinan Güntürk
Program in Applied and Computational Mathematics
Princeton University
Princeton, NJ 08544

Jeffrey C. Lagarias
Information Sciences Research
AT&T Labs
180 Park Avenue, Florham Park, NJ 07932

Vinay Vaishampayan
Information Sciences Research
AT&T Labs
180 Park Avenue, Florham Park, NJ 07932

(July 24, 2000)

Abstract

Sigma-delta modulation is a widely used method of analog-to-digital signal conversion. One reason for the popularity of this method is its apparent robustness to hardware imperfections. That is, bitstreams that may have been generated by slightly imprecise hardware components can still be decoded comparably well. In this paper we formulate a model for robustness and give a rigorous analysis in the simplest case of a single-loop sigma-delta modulation system applied to constant signals (dc inputs) for N time cycles, with an arbitrary (small enough) initial condition u_0, and where the quantizer may contain an offset error. The mean square error (MSE) of any decoding scheme for this quantizer (with u_0 and the offset error known) is bounded below by $\frac{1}{96}N^{-3}$. We also determine asymptotically best possible MSE as $N \to \infty$ for perfect decoding for the cases $u_0 = 0$ and $\frac{1}{2}$. The robustness result is the upper bound that a triangular linear filter decoder (with both u_0 and the offset error unknown) achieves an MSE of $\frac{40}{3}N^{-3}$. These results establish the known result that the $O(1/N^3)$ decay of the MSE with N is optimal in the single-loop case, under weaker assumptions than previous analyses, and show that a suitable linear decoder is robust against offset

error. These results are obtained using methods from number theory and Fourier analysis.

1. Introduction

Modern techniques of high accuracy analog to digital conversion of bandlimited signals is based on using single-bit quantization together with oversampling, as a practical alternative to using a multi-bit quantizer on a sequence sampled at the Nyquist rate. This is true for reasons related to both the quantizer and to oversampling. Single-bit quantizers are preferable to multi-bit quantizers because they are easier and cheaper to build. Also, single-bit sigma-delta modulation is robust against circuit imperfection, owing to the feedback which compensates for deviations in the quantization thresholds. The deviations in two-level feedback, which occurs in the case of single-bit quantization, only amounts to an offset error combined with an amplification error, while deviations from a multi-level feedback would cause irreversible harmonics. We consider this robustness viewpoint further below. Oversampling facilitates implementations in various ways, including making the job of analog filtering easier, see Candy and Temes [3] for a general discussion.

Sigma-delta modulation is a widely used method for analog-to-digital conversion, see [3, 20, 22, 25] . It transforms a bandlimited signal by oversampling using a single bit quantizer with feedback, to produce a two-valued signal stream which we call the *coded signal.* This signal stream is then appropriately filtered- usually with a linear filter- to produce a (vector) quantized version of the original signal. This filtering step may be regarded as a form of decoding. The simplest version of sigma-delta modulation is the single loop version originally introduced in [23, 1, 14, 8], but many more complicated multi-loop systems have since been considered.

The sigma-delta modulator is a nonlinear system with feedback and is notoriously difficult to analyze. One of the first rigorous analyses of this system was performed by Gray for constant inputs [8], where he showed that filtering the quantization output sequence with a rectangular window of length N results in a reconstruction error bounded by $O(1/N)$, for all initial conditions and values of the constant input. Gray [9] later used a connection with ergodic theory to show that the mean-squared error (MSE) decays asymptotically as $O(1/N^3)$, as $N \to \infty$, where N is the number of taps of the filter used in decoding. The notion of MSE used here is taken over a uniform distribution of the value x to be estimated, but also requires a time average[1] of the error signal. Concerning lower bounds, Hein and Zakhor [18] and Hein, Ibrahim and Zakhor [17] showed that for any decoding scheme for dc input the quantization error must be at least as large as a constant times $1/N^3$, where the constant depends on the initial value u_0 of the integrator in the sigma-delta modulator at the beginning of the quantization interval. In actual practice sigma-delta modulators are used for A/D conversion of

[1] See Gray [9, equation (6.1 ff)]. This time averaged quantity $M\{(\hat{x}_n - \frac{x}{2b})^2\}$ is equivalent to an average over initial conditions that are the values of u_n occurring in an infinitely long string of samples with fixed input value x. In fact, Gray's paper is mainly concerned with the statistics of the "quantization noise" $\{q(u_n) - u_n \ : \ n \geq 1\}$.

bandlimited signals and not for dc signals. Constant signals are apparently the worst case for such modulators, and engineering practice recommends adding a high frequency dither signal to make the input vary, cf. Candy and Temes [4, p. 14]. Regarding the performance on more general classes of signals, and the use of nonlinear decoding methods, see [11], [21], [6], [27, 28, 29]; we discuss this further in the concluding section.

This paper studies the robustness of sigma-delta schemes against certain hardware imperfections. This seems to be one of the main reasons they are used in practice [4, p. 13]. Non-idealities for circuits using a single-bit quantizer can include offset quantizer threshold, offset quantization level, leakage in the integrator, non-unitary integrator gain, nonzero initial state and random noise. Because of their complexity, robustness of a given system has generally been studied by simulation, see [21] for an example. Feely and Chua [7] present a rigorous study of the effect of integrator leakage on quantization of dc inputs. In this case there can be constant size errors which do not go to zero as the number N of output bits increases, because there is a mode-locking effect.[2] In this paper we study the effects of offset quantizer thresholds and of nonzero initial state of the encoder, and rigorously establish that robust decoders do exist for constant signals.

The issue of robustness is one of information theory insofar as it concerns the design of both the encoder and decoder. From the robustness viewpoint an analog-to-digital conversion scheme is a (necessarily nonlinear) encoding operation that produces an output stream of bits describing a signal. One has an ideal system $E(\cdot)$ used for circuit design while the actual hardware produces a system $\tilde{E}(\cdot)$ that approximates the behavior of $E(\cdot)$ but which makes certain "systematic" errors. Whenever feedback is present the nonlinear nature of the system can potentially lead after some time to large discrepancies between the internal states of system $E(\cdot)$ and system $\tilde{E}(\cdot)$ given the same input, no matter how small the "systematic" errors are. Here the "errors" which relate the difference of the actual system $\tilde{E}(\cdot)$ to the ideal system are highly correlated, but can lead to extreme changes in the output bit stream (large Hamming distance). The robust encoding problem is to design an ideal encoder $E(.)$ for which there exists some decoder robust against certain types of hardware imperfections $\tilde{E}(\cdot)$. Given a robust encoder, the robust decoding problem is to design a decoder $D(\cdot)$ which produces an adequate reconstruction of the signal from the digital bit streams produced by any of the (not precisely known) systems $\tilde{E}(\cdot)$. The robust coding problem is quite different from the classical model in information theory of the binary symmetric channel, in which errors occur randomly. In the classical case error correcting coding is introduced in advance of transmission over the channel, but that is not available in this context. Here the errors are not random but are systematic, caused by the uncertainty in the encoder used.

[2] This provides a reason that constant signals are to be avoided in practice using sigma-delta modulators.

Besides formulating a framework for the robust encoding problem in the context of sigma-delta modulators, the object of this paper is to provide an analysis of robustness in the "simplest" situation. We give a rigorous performance analysis of single-loop sigma-delta modulation in the case of a constant signal, which includes the effect of nonzero initial state u_0 and of possible offset in the quantizer. As indicated above, the decay of the MSE for such signals is well known to be of order $O(N^{-3})$, under various hypotheses. Here we rigorously demonstrate robustness by showing that a simple linear decoder achieves MSE of order $O(N^{-3})$ with the initial condition and dc offset in the quantizer unknown. To describe the precise results, for constant signals the sigma-delta modulator can be viewed as a (scalar) quantizer, in which the quantization assigned to a constant signal x is the sequence of quantized bits produced over the N time periods, which will depend on x, the initial state u_0 of the sigma-delta modulator, and the quantizer used, allowing offset error δ. Our analysis for single loop sigma delta modulation with constant input signal x is valid for any fixed small enough initial state u_0, for N time periods, allowing offset error in the quantizer. More precisely, for offset error δ with $|\delta| < \frac{1}{8}$ we may allow $-\frac{3}{8} < u_0 < \frac{11}{8}$. We give upper and lower bounds for the MSE of this quantizer, assuming only that the dc signal x is uniformly distributed in $[0,\ 1]$. In particular, no assumption is imposed on the quantizer noise statistics within the N time periods. The lower bound of $\frac{1}{96}N^{-3}$ is valid for the optimal quantizer, which assumes that both u_0 and the offset δ are known to the decoder. The proof uses the same idea as [18] and [17] but sharpens it slightly in obtaining a uniform bound independent of u_0. We also obtain asymptotically exact bounds for the MSE of optimal quantization for the special cases $u_0 = 0$ and $u_0 = \frac{1}{2}$ as $N \to \infty$, using detailed facts about Farey fractions. The result for $u_0 = 0$ has the constant $\frac{2}{21}\frac{\zeta(2)}{\zeta(3)} - \frac{1}{16} = 0.06782$, which sets a limit on how much the lower bound can be improved. The robustness result is the upper bound, which is $\frac{40}{3}N^{-3}$, for the MSE using the triangular linear filter decoder, which treats both the initial integrator value u_0 and the offset δ as unknown to the decoder. The proof uses Fourier series and an estimate from elementary number theory. The choice of triangular filter decoder is explained at the end of §2. These MSE bounds improve on the analysis of Gray [9] in that they do not do any averaging of the signal over input values u_0, and are valid for each fixed initial value u_0 separately. The specific constants $\frac{1}{96}$ and $\frac{40}{3}$ obtained in the analysis can be further improved, with more detailed estimates, which we do not attempt to do. A small improvement related to the upper bound is indicated at the end of Appendix C.

Compared to a multi-bit quantizer which can achieve an exponentially small MSE of order $O(2^{-N})$, the sigma-delta quantization output sequence is not efficient. However, this does not mean that the quantizer is very nonuniform: It is well-known, e.g. [17], that the number of distinct codewords of length N produced by the first order sigma-delta modulator on constant

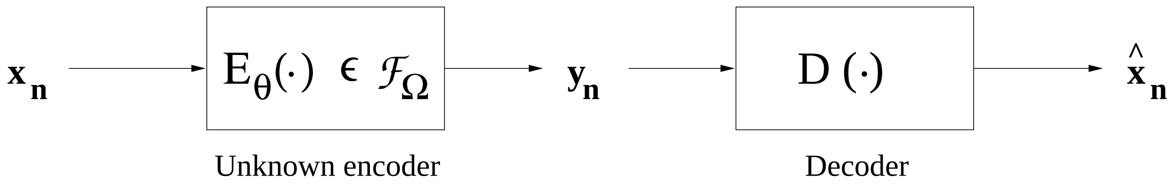

Figure 1: Robust encoding problem.

signals, for fixed u_0 and δ, is bounded by $O(N^2)$. Hence, an exponential rate-distortion function is still achievable with further coding of the output bit-stream. The particular set of $O(N^2)$ admissible output codewords depends on the parameters u_0 and δ. In our model with no random errors, a received codeword contains some information about u_0 and δ. It is this extra information that makes robust decoding possible in this case.

2. Problem Formulation

We first formulate the robust encoding problem as the simple block diagram given in Fig 1. We are given a family of encoders $\mathcal{F}_\Omega = \{E_\theta \ : \ \theta \in \Omega\}$ where θ is a (vector) parameter. We may think of these as representing an ideal encoder with the parameter θ measuring the deviation from ideality of a particular hardware implementation. The parameter θ is not under any control, except to lie in a fixed compact set Ω representing the manufacturing tolerances. The decoder's performance is to be measured in the worst case against all the allowable encoders. We use mean square error as a performance measure. This model scheme does not include any source of random errors, just systematic encoding errors embodied in the parameter θ. For the sigma-delta modulator we can take $\theta = (u_0, \delta)$, but as we show below, the behavior of the system really depends only on the single parameter $\tilde{\theta} = u_0 + \delta$. We study the asymptotic behavior as the number N of output bits becomes large. The possible existence of a robust decoder depends on the family of encoders $\mathcal{F}_\Omega$ considered. For example Feely and Chua [7] consider encoders $\mathcal{F}_\Omega$ that include leaky integrators, and their results imply that for constant inputs and optimal decoding, the mean square error does not go to zero with increasing N, so that a robust decoder does not exist in the asymptotic sense considered here.

We consider systems that use a single bit quantizer. An *ideal quantizer* has a threshold at 0.5 and reconstruction levels 1 and 0, whose *quantizing map* $q(\cdot)$ is given by

$$q(u) = \begin{cases} 1, & u \geq 1/2, \\ 0, & u < 1/2. \end{cases} \tag{2.1}$$

An *offset quantizer* has a threshold at $0.5 - \delta$, where we assume $-\delta_{max} \le \delta \le \delta_{max}$ for some δ_{max} (say $\delta_{max} = 0.125$), and reconstruction levels at 1 and 0, hence is the *quantizing map* $q_\delta(\cdot)$ given by

$$q_\delta(u) = \begin{cases} 1, & u \ge 1/2 - \delta, \\ 0, & u < 1/2 - \delta. \end{cases} \tag{2.2}$$

A single loop (or first-order) sigma-delta modulator is illustrated in Fig. 2. The sigma-delta modulator consists of a quantizer in a feedback loop. The behavior of the system with an ideal quantizer is described by

$$u_{n+1} = x_n + u_n - q(u_n), \ n = 0, 1, \ldots \tag{2.3}$$

while with the non-ideal quantizer it is

$$\tilde{u}_{n+1} = x_n + \tilde{u}_n - q_\delta(\tilde{u}_n), \ n = 0, 1, \ldots \tag{2.4}$$

The output vector at time n is denoted $Y_n := q_\delta(\tilde{u}_n)$. The following simple fact, observed in [19], simplifies the robust quantizing problem.

Lemma 2.1. *Let x_n be a fixed input sequence. The output bit sequence Y_n for the non-ideal first order sigma-delta modulator with initial value $\tilde{u}_0$ and offset δ is identical to the output bit sequence for the ideal first-order sigma-delta modulator with the modified input value $u_0 := \tilde{u}_0 + \delta$.*

Proof. Since $q_\delta(\cdot) = q(\cdot + \delta)$, on setting $u_n := \tilde{u}_n + \delta$, the system (2.4) becomes equivalent to the system (2.3) with the initial condition $u_0 = \tilde{u}_0 + \delta$. ■

Lemma 2.1 shows that studying robustness of a first-order sigma delta modulator against arbitrary initial value $\tilde{u_0}$ and offset error reduces to the special case of studying the ideal system (2.3) with arbitrary (unknown) initial condition u_0. This reduction is special to first-order sigma-delta modulation. In higher-order schemes the initial value u_0 and offset parameter δ are independent sources of error.

Let us suppose the offset error δ satisfies $|\delta| \le \delta_{max} < \frac{1}{2}$. The system (2.3) with the ideal quantizer maps the interval $[-1/2, 3/2)$ into itself, so if the original initial condition $\tilde{u}_0$ for $q_\delta(.)$ satisifes

$$-\frac{1}{2} - \delta \le \tilde{u}_0 < \frac{3}{2} - \delta, \tag{2.5}$$

then this condition is preserved under iteration (2.4). As long as $-1/2 + \delta_{max} \le \tilde{u}_0 < 3/2 - \delta_{max}$, then (2.5) is satisfied for all allowable values of $|\delta| \le \delta_{max}$, and the subsequent analysis applies.

In view of the Lemma 2.1, in analyzing robustness against offset error, it suffices to treat the case of a first-order sigma-delta modulator with an ideal quantizer, and consider robustness against the choice of initial value

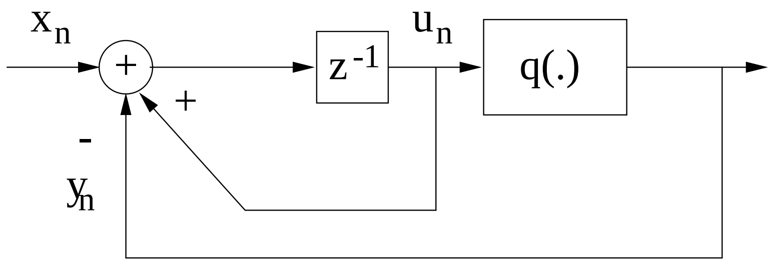

Figure 2: Encoder of a single stage sigma-delta modulator.

u_0, and this we do in the remainder of the paper. We treat u_0 as given, and average over the input value x, assumed uniformly distributed on $[0,1]$ and independent of the value u_0.

The ideal quantizer output at time n is denoted Y_n and is related to input u_n by $Y_n := q(u_n)$. From (2.3), it follows that the quantizer output is given by

$$Y_n := q(u_n) = q(u_0 + \sum_{i=0}^{n-1} x_i - \sum_{i=0}^{n-1} q(u_i)), \quad n = 0, 1, \ldots \tag{2.6}$$

Equation (2.6) defines a map $f_{u_0}^N(\cdot) : \mathbf{R}^N \to \{0,1\}^N$, with input $(x_0, x_1, \ldots, x_{N-1})$ and output $(Y_1, Y_2, \ldots, Y_N)$. As the x_i vary, this map changes value at points where

$$u_0 + \sum_{i=0}^{n-1} x_i - \sum_{i=0}^{n-1} q(u_i) = 1/2, \quad n = 1, \ldots, N. \tag{2.7}$$

These points constitute the boundaries of the level sets of $f_{u_0}^N$, in other words, the n-bit quantization bins. Thus, a partition of $[0,1]^N$ is created. Note that implicit in (2.7), u_n is a nonlinear function of the input $x_0, \ldots, x_{n-1}$. The resulting bins are very irregularly shaped.

Now suppose that we have constant input signal $x_i = x$, for $1 \le i \le N$. This corresponds to looking at the intersections of the sets in the partition with the principal diagonal of $[0,1]^N$ given by $x_0 = x_1 = \ldots = x_{N-1}$. This naturally induces a partition of $[0,1]$, which we refer to as the *effective quantizer*, in order to distinguish it from the binary quantizer in the loop. The thresholds of the effective quantizer are obtained by determining values of x in $[0,1]$ that solve

$$u_0 + nx - \{\sum_{i=0}^{n-1} q(u_i)\} = 1/2, \quad n = 1, 2, \ldots, N. \tag{2.8}$$

Lemma 2.2. *For a general u_0 in the range $-\frac{1}{2} \le u_0 < \frac{3}{2}$, the quantization thresholds are given by*

$$\mathcal{S}_N(u_0) := \{(j - \beta_0)/n \ : \ \ j = 1, \ldots, n; \ n = 1, 2, \ldots, N\}. \tag{2.9}$$

where $\beta_0 = \langle u_0 + 1/2 \rangle$, and $\langle \alpha \rangle := \alpha - \lfloor \alpha \rfloor$ denotes the fractional part of α.

We give the derivation in Appendix A. For the special cases $u_0 = 0$ and $u_0 = 1$, the set of thresholds is the set

$$\mathcal{S}_N := \{(2j-1)/2n,\ j = 1, \ldots, n;\ n = 1, 2, \ldots, N\}.$$

The set $\mathcal{S}_N$ is related to the *Farey series* $\mathcal{F}_N$ of order N, which is the set of fractions k/n in reduced form with $0 < k < n$, $1 \le n \le N$, with $GCD(k, n) = 1$, on the interval $(0, 1)$, arranged in ascending order [13, Chapter 3]. To be more specific, let $\mathcal{G}_{2N}$ be the subset of the Farey series of order $2N$ that has *even* denominators. Then it follows that $S_N = \mathcal{G}_{2N}$. On the other hand, the special cases $u_0 = -1/2$ and $u_0 = 1/2$ have $\mathcal{S}_N(-1/2) = \mathcal{S}_N(1/2) = \mathcal{F}_N \backslash \{0\}$, the Farey series itself. The connection of breakpoints for $u_0 = 0$ to the Farey series was observed in Hein and Zakhor [20, p. 24].

Our problem is to give lower and upper bounds for the mean squared quantization error for fixed u_0 with the constant dc input x assumed drawn from the uniformly distribution on [0, 1]. For lower bounds we assume optimal decoding, where u_0 is known to the decoder. The optimal MSE quantizer is described using the map $Q_{opt}(x; u_0)$, which maps x to the midpoint of the interval J that x lies in, the endpoints of J being successive elements of the thresholds of the effective quantizer with initial value u_0. The map $Q_{opt}(x; u_0)$ is the optimal quantizer under our assumption that the quantity x being quantized is uniformly distributed in $[0, 1]$ and independent of u_0. Our objective is to lower bound the mean squared-error, given by the integral

$$MSE_{u_0}(Q_{opt}) := \int_0^1 (x - Q_{opt}(x; u_0))^2 dx. \tag{2.10}$$

In the upper bound case, we suppose u_0 is fixed but unknown to the decoder, and consider a decoding algorithm which uses a particular linear filter of length N, the triangular filter. Let $Q_h(x; u_0)$ denote the triangular filtered estimate for x, which depends on x and u_0. Our objective is to upper bound

$$MSE_{u_0}(Q_h) := \int_0^1 (x - Q_h(x; u_0))^2 dx. \tag{2.11}$$

for all u_0. The choice of triangular filter for analysis is explained at the end of §4.

3. Lower Bound for Mean Square Error

In this section we suppose that the initial value u_0 is fixed and known, with $\frac{1}{2} \le u_0 < \frac{3}{2}$. The unit interval [0, 1] is partitioned into subintervals $J = J(\mathbf{Y})$, where $\mathbf{Y} := (Y_1, Y_2, \ldots, Y_N)$ is the data available to the decoder. The *optimal decoding algorithm*[3] $Q_{opt}(x; u_0)$ assigns to the quantization data

[3]The optimality of this algorithm is a consequence of the assumption that the quantity x being quantized is uniformly distributed in $[0, 1]$. When conditioned on the data $\mathbf{Y}$ the distribution of x is uniform on the quantization interval $J(\mathbf{Y})$.

$\mathbf{Y}$ associated to x the midpoint of the interval $J(\mathbf{Y})$. There are at most $\frac{(N+1)(N+2)}{2}$ quantization intervals determined by the values given in (2.9). For $u_0 = 0$ some of the values are repeated[4], and the number of distinct values is asymptotic to $\frac{3}{\pi^2}N^2$ as $N \to \infty$, using [13, Theorem 330], since the points of $\mathcal{G}_{2N}$ can be put in one-to-one correspondence with the Farey sequence $\mathcal{F}_N$. It is well known that the intervals produced by the Farey sequence $\mathcal{F}_{2N}$ range in size from $\frac{1}{2N}$ down to size $\frac{1}{4N^2}$, see [13, Theorem 35]. The interval $[0, \frac{1}{2N}]$ contributes $\frac{1}{96}N^{-3}$ all by itself to the MSE of the optimal decoding algorithm. We now show that the same bound holds for an arbitrary u_0.

Theorem 3.1. *Suppose that u_0 is fixed with $-1/2 \le u_0 < 3/2$ and let x be drawn uniformly from $[0, 1]$. Then single-loop oversampled sigma-delta modulation*

$$u_{n+1} = x + u_n - q(u_n) \quad 0 \le n \le N-1$$

with oversampling rate N, using the optimal quantizer Q_{opt} with u_0 known to the quantizer, has expected mean square error

$$MSE_{u_0}(Q_{opt}) = \int_0^1 (Q_{opt}(x; u_0) - x)^2 dx \ge \frac{1}{96}N^{-3}\ . \tag{3.1}$$

Proof. The value u_0 completely determines the quantization bins. The quantization bin endpoints consist of the points

$$x = \frac{j - \beta_0}{n}, \quad 1 \le j \le n, \quad 1 \le n \le N, \tag{3.2}$$

where $\beta_0 = \langle u_0 + 1/2 \rangle$. We will show that at least one of the open intervals $(0, \frac{1}{2N})$ or $(1 - \frac{1}{2N}, 1)$ contains no quantization threshold. This interval is of length $\frac{1}{2N}$, and since an interval of length $|I|$ contributes $\frac{1}{12}|I|^3$ to the MSE, the contribution of this interval is $\frac{1}{96}N^{-3}$.

Case 1. $0 \le \beta_0 < 1/2$.
For $1 \le j \le n$, and $1 \le n \le N$,

$$\frac{j - \beta_0}{n} \ge \frac{j - \beta_0}{N} \ge \frac{1 - \beta_0}{N} \ge \frac{1}{2N},$$

hence $(0, \frac{1}{2N})$ contains no quantization threshold.

Case 2. $1/2 \le \beta_0 < 1$.
For $1 \le j \le n$, and $1 \le n \le N$,

$$\frac{j - \beta_0}{n} \le \frac{n - \beta_0}{n} \le 1 - \frac{\beta_0}{N} \le 1 - \frac{1}{2N},$$

hence $(1 - \frac{1}{2N}, 1)$ contains no quantization threshold. ■

[4] That is, the values $\frac{j}{n}$ and $\frac{jk}{nk}$ for any $k \ge 2$ determine the same quantization threshold.

The optimal lower bound in Theorem 3.1 appears to have a constant on the order of five times larger than $\frac{1}{96}$ but seems hard to determine. However we can show the following exact result.

Theorem 3.2. *Suppose that $u_0 = 0$ or $u_0 = \frac{1}{2}$ and let x be drawn uniformly from $[0, 1]$. Then single-loop oversampled sigma-delta modulation*

$$u_{n+1} = x + u_n - q(u_n) \quad 0 \le n \le N - 1$$

with oversampling rate N, using the optimal quantizer Q_{opt} with u_0 known to the quantizer, has expected mean square error

$$MSE_{u_0}(Q_{opt}) = \alpha_{u_0} N^{-3} + O(N^{-4}\log N) \quad as \quad N \to \infty, \tag{3.3}$$

where

$$\alpha_0 := \frac{2}{21}\frac{\zeta(2)}{\zeta(3)} - \frac{1}{16} = 0.06782.$$

and

$$\alpha_{1/2} := \frac{1}{6}\frac{\zeta(2)}{\zeta(3)} = 0.22807.$$

We prove this result in Appendix B. The proof is based on the explicit relation of the set of quantization thresholds in these two cases with the Farey series. Theorem 3.2 sets a limit on how much improvement is possible in the constant $\frac{1}{96}$ appearing in Theorem 3.1, showing that the best constant can be no larger than $\frac{2}{21}\frac{\zeta(2)}{\zeta(3)} - \frac{1}{16}$. Numerical simulations suggest that this bound for $u_0 = 0$ is actually close to the minimum over all initial conditions u_0, and conceivably it might give the best constant.

4. Upper Bound for Mean Square Error

In this section we suppose that u_0 is viewed as fixed with $\frac{1}{2} \le u_0 < \frac{3}{2}$, but is unknown. The quantization values $(Y_1, Y_2, \ldots, Y_{N-1})$ are known to the decoder. For simplicity we assume that $N = 2M$ is even.

We consider a triangular filter decoder

$$Q_h(x, u_0) := \sum_{n=1}^{2M-1} h_n Y_n \tag{4.1}$$

in which $\{h_n : 1 \le n \le 2M - 1\}$ is the triangular filter of mass 1, given by

$$h_n = \begin{cases} \frac{1}{M} - \frac{(M-n)}{M^2} & 1 \le n \le M\ , \\ \\ \frac{1}{M} - \frac{(n-M)}{M^2} & M \le n \le 2M - 1\ . \end{cases} \tag{4.2}$$

We give a detailed analysis for the case $N = 2M$ only; for the case $N = 2M+1$ we may discard the value Y_N and use the above filter on the remaining values.

Theorem 4.1. *Suppose that u_0 is fixed with $-1/2 \leq u_0 < 3/2$, and let x be drawn uniformly in $[0,1]$. Then single-loop oversampled sigma-delta modulation*

$$u_{n+1} = x + u_n - q(u_n), \quad 0 \leq n \leq N-1 \ ,$$

at oversampling rate $N = 2M$, using quantizer Q_h, has expected mean square error

$$MSE_{u_0}(Q_h) = \int_0^1 (Q_h(x;u_0) - x)^2 dx \leq \frac{40}{3} N^{-3} \ . \qquad (4.3)$$

The proof uses two number-theoretic lemmas, whose proofs are given in Appendix C. In the following, (m,n) denotes the greatest common divisor of m and n.

Lemma 4.1. *For fixed constant β_0 and all positive integers n and m,*

$$\left| \int_0^1 \langle nx + \beta_0 \rangle \langle mx + \beta_0 \rangle \, dx - \frac{1}{4} \right| \leq \frac{1}{12} \frac{(n,m)^2}{nm} \ . \qquad (4.4)$$

Lemma 4.2. *For all positive integers L,*

$$\sum_{n=1}^{L} \sum_{m=1}^{L} \frac{(n,m)^2}{nm} \leq 5L. \qquad (4.5)$$

Proof of Theorem 4.1. Suppose $N = 2M$ is even. We set

$$\epsilon_N(x) := Q_h(x;u_0) - x \ .$$

where $Q_h(x;u_0)$ is the triangular filter decoder

$$Q_h(x;u_0) := \sum_{n=1}^{2M-1} h_n Y_n \ , \qquad (4.6)$$

has filter weights (4.2), and $Y_n = q(u_n)$. We have

$$\begin{aligned} \epsilon_N(x) &= \sum_{n=1}^{2M-1} (Y_n - x) h_n \\ &= \sum_{n=1}^{2M-1} (u_n - u_{n+1}) h_n. \end{aligned}$$

Summing this by parts and substituting (A.2) from Appendix A yields

$$\begin{aligned} \epsilon_N(x) &= \frac{1}{M^2}(u_1 + \ldots + u_M) - \frac{1}{M^2}(u_{M+1} + \ldots + u_{2M}) \\ &= \frac{1}{M^2}(w_1 + \ldots + w_M) - \frac{1}{M^2}(w_{M+1} + \ldots + w_{2M}). \end{aligned} \qquad (4.7)$$

Then we have

$$|\epsilon_N(x)| \leq \frac{1}{M^2}\left|\sum_{n=1}^{M} w_n - \frac{M}{2}\right| + \frac{1}{M^2}\left|\sum_{n=M+1}^{2M} w_n - \frac{M}{2}\right|, \tag{4.8}$$

which, upon taking the square yields

$$|\epsilon_N(x)|^2 \leq \frac{2}{M^4}\Big(\sum_{n=1}^{M} w_n - \frac{M}{2}\Big)^2 + \frac{2}{M^4}\Big(\sum_{n=M+1}^{2M} w_n - \frac{M}{2}\Big)^2. \tag{4.9}$$

We now consider the expected mean square error,

$$MSE_{u_0}(Q_h) = \int_0^1 |\epsilon_N(x)|^2 \mathrm{d}x.$$

Substituting (A.4) from Appendix A into (4.9), and integrating, we get

$$MSE_{u_0}(Q_h) \leq \frac{2}{M^4}\int_0^1 \left\{\Big(\sum_{n=0}^{M-1}\langle nx+\beta_0\rangle - \frac{M}{2}\Big)^2 + \Big(\sum_{n=M}^{2M-1}\langle nx+\beta_0\rangle - \frac{M}{2}\Big)^2\right\} \mathrm{d}x. \tag{4.10}$$

We expand this expression, substitute $\int_0^1 \langle nx+\beta_0\rangle \ \mathrm{d}x = 1/2$ for positive n, and rearrange to get

$$\begin{aligned} MSE_{u_0}(Q_h) \ \leq \ & \frac{2}{M^4}\left\{(\beta_0 - \frac{1}{2})^2 + \sum_{n=1}^{M-1}\sum_{m=1}^{M-1}\Big(\int_0^1 \langle nx+\beta_0\rangle\langle mx+\beta_0\rangle \ \mathrm{d}x - \frac{1}{4}\Big)\right. \\ & \left. + \sum_{n=M}^{2M-1}\sum_{m=M}^{2M-1}\Big(\int_0^1 \langle nx+\beta_0\rangle\langle mx+\beta_0\rangle \ \mathrm{d}x - \frac{1}{4}\Big)\right\}. \end{aligned} \tag{4.11}$$

Next we apply Lemma 4.1 to (4.11) and replace the term $(\beta_0 - \frac{1}{2})^2$ by its maximum value $1/4$, to obtain

$$MSE_{u_0}(Q_h) \leq \frac{2}{M^4}\left(\frac{1}{4} + \frac{1}{12}\sum_{n=1}^{M-1}\sum_{m=1}^{M-1}\frac{(n,m)^2}{nm} + \frac{1}{12}\sum_{n=M}^{2M-1}\sum_{m=M}^{2M-1}\frac{(n,m)^2}{nm}\right). \tag{4.12}$$

Finally, we conclude our estimate of $MSE_{u_0}(Q_h)$ by applying Lemma 4.2 with $L = 2M-1 = N-1$ to (4.12) (combining the double sums) to obtain

$$MSE_{u_0}(Q_h) \leq \frac{40}{3}N^{-3} - \frac{16}{3}N^{-4}, \tag{4.13}$$

which yields the desired bound. ∎

Remarks. (1) The triangular filter was used in the analysis because of the identity (4.7) that it yields for the error expression: The "first order" terms of size $O(1/N)$ get cancelled out due to the subtraction, and this was exploited in the estimate (4.8). This is not the case for the rectangular filter; indeed, a telescoping argument gives that the error expression for this filter is equal to $(w_{N+1} - w_1)/N$, which is in general not smaller than $O(1/N)$. Other reasons based on the Fourier transform can be given, see He, Kuhlmann and Buzo [15, Sect IV.C].

(2) Gray [9] determines the optimal linear filter (in the context of [9]), whose general shape is similar to the triangular filter, but differs from it slightly. Hein and Zakhor [19] later constructed an "optimal" nonlinear decoding method.

(3) The proof of Theorem 4.1 did not determine the best constant for MSE using the triangular filter, and some improvements are possible on the constant $\frac{40}{3}$ by more careful argument. The constant in Lemma 4.2 can be improved slightly.

5. Conclusions

This paper gave rigorous upper and lower bounds on the mean square error for single loop sigma-delta modulation applied to constant inputs, where the quantizer may have offset error and an arbitrary fixed initial value u_0. It showed that a particular linear decoder is robust against such errors, and attains the optimal MSE within a multiplicative constant. In these special circumstances a nonlinear decoder can save at most a multiplicative constant in MSE over a linear decoder, and cannot achieve further asymptotic improvement in MSE as $N \to \infty$. These results show that the redundancy built into oversampled sigma-delta modulation schemes is serving the useful purpose of permitting robust decoding by a linear decoder. It seems likely that for the first-order scheme robust decoders should exist for a general class of non-constant signals, but that is a more difficult question which we have not addressed.

The methods of this paper exploited certain features specific to first-order sigma-delta modulation (e.g. Lemma 2.1), which do not hold for higher-order sigma-delta schemes. However the general approach of viewing higher-order schemes as discrete dynamical systems is a useful one, to which Fourier-analytic methods can be successfully applied, as in Daubechies and DeVore [6] and Güntürk [10], and for these number-theoretic ideas of a more sophisticated type may also be relevant.

It would be of great interest to extend robustness analyses to higher-order sigma-delta systems and to obtain bounds valid for general bandlimited signals rather than constant signals. For constant signals it is believed that a k-th order sigma-delta modulation scheme can achieve a mean square error that decays like $O(1/N^{2k+1})$ for signals of length N, and that this should be best possible. An upper bound $O(1/N^{2k+1})$ is demonstrated for certain k-th-order sigma-delta schemes in He, Kuhlmann and Buzo [15, Sects.3, 4], [16], but their analysis treats the input x as *fixed,* and then lets $N \to \infty$; the error estimates obtained are not uniform in x (and require that x be irrational,) hence their MSE bounds do not apply in the framework[5] of this paper. Does there exist a similar upper bound $O(1/N^{2k+1})$ for some k-th order sigma-delta modulation scheme, using the MSE-criterion of this paper, and do there exist such schemes that are robust against offset error in the quantizer, assuming perfect integrators are used? For signals drawn from a wider class of bandlimited signals, it is believed that the achievable [6] mean square error should be $O(1/N^{2k+2})$, see Thao[27]. Demonstrating this rigorously, with or without robustness, is apparently an open problem. A rigorous lower bound of $O(1/N^{2k+2})$ for bandlimited signals was obtained by Thao[27]. In the case $k = 1$ nonlinear coding schemes that experimentally achieve $O(N^{-4})$ for a class of sinusoidal signals are given in [21], [28], [29].

For general bandlimited functions, it is an open problem to rigorously establish whether non-linear decoding schemes for sigma-delta modulation schemes can offer an asymptotic improvement over linear decoding. If so, another issue would be whether there exists a nonlinear decoding achieving this improvement which is robust. It seems an important general problem to quantify the tradeoff between efficiency and robustness in such schemes, both theoretically and in practice.

Acknowledgment. We are indebted to Z. Cvetković and especially to N. T. Thao for helpful comments and references.

Appendix A. Proof of Lemma 2.1

Set $J := [-\frac{1}{2}, \frac{3}{2})$. Then, for $u_n \in J$, we have $q(u_n) = \lfloor u_n + \frac{1}{2} \rfloor$. This way, the recursion (2.3) can be rewritten as

$$u_{n+1} = x_n - \frac{1}{2} + \left\langle u_n + \frac{1}{2} \right\rangle \ . \tag{A.1}$$

Suppose $x_n \in [0, 1]$. Then from (A.1), $u_n \in J$ implies $u_{n+1} \in J$.

If $x_n = x$ is constant, then in fact, $u_n \in [x - \frac{1}{2},\ x + \frac{1}{2})$, for all n. If we now define

$$w_n := u_n - x + \frac{1}{2}, \tag{A.2}$$

then the iteration for w_n is just rotation on the unit circle

$$w_{n+1} := \langle w_n + x \rangle, \tag{A.3}$$

[5]The framework of this paper requires integrating their bounds over x, and the resulting integral diverges.

[6]This MSE averages over a larger class of (bounded) signals, so the contribution of constant signals to the MSE is reduced.

hence

$$w_{n+1} = \langle nx + \beta_0 \rangle \ , \tag{A.4}$$

as was originally observed by Gray [8]. Thus, substituting $\beta_0 = \langle u_0 + 1/2 \rangle$, one arrives at the formula

$$u_{n+1} = x - \frac{1}{2} + \langle nx + \beta_0 \rangle. \tag{A.5}$$

On the other hand, one also has

$$u_{n+1} = u_0 + (n+1)x - (Y_0 + \ldots + Y_n). \tag{A.6}$$

Combining the two and using $Y_0 = \lfloor u_0 + 1/2 \rfloor$, it follows that

$$Z_n := Y_1 + \ldots + Y_n = \lfloor nx + \beta_0 \rfloor. \tag{A.7}$$

Clearly, the sequence (Y_n) is uniquely determined by the sequence (Z_n), and vice versa. From (A.7), it follows that Z_n increments by one at the points $\{x = (j - \beta_0)/n; \quad j = 1, \ldots, n\}$, starting with 0 at $x = 0$ and ending with n at $x = 1$. So, the N-tuple $(Z_1, \ldots, Z_N)$ and consequently the quantization codeword $(Y_1, \ldots, Y_N)$ attains single and distinct values on each of the subintervals defined by the threshold points $\mathcal{S}_N(u_0)$ given in (2.9), which completes the proof. ■

Appendix B. Proof of Theorem 3.2

Proof of Theorem 3.2. Lemma 2.2 shows that for $u_0 = 0$ the set of thresholds is

$$\mathcal{G}_{2N} = \{\frac{2j-1}{2n} : 1 \le j \le n, 1 \le n \le N\},$$

while for $u_0 = \frac{1}{2}$ the set of thresholds is

$$\mathcal{F}_N = \{\frac{j}{n} : 1 \le j \le n, 1 \le n \le N\}$$

the Farey series of order N.

We first treat the case $u_0 = \frac{1}{2}$ and estimate

$$MSE_{1/2}(Q_{opt}) = MSE(\mathcal{F}_N) = \sum_{I \in \mathcal{F}_N} \frac{1}{12} |I|^3 \ ,$$

where $I \in \mathcal{F}_N$ means I is an interval determined by $\mathcal{F}_N$, and $|I|$ denotes its length. We will show that

$$MSE(\mathcal{F}_N) = \left(\sum_{q=1}^{\infty} \frac{\phi(q)}{q^3} \right) \frac{1}{6N^3} + O\left(\frac{\log N}{N^4} \right) \ , \tag{B.1}$$

in which $\phi(q)$ is Euler's ϕ-function, which counts the number of integers k with $1 \le k < q$ which are relatively prime to q. The intervals $[0, \frac{1}{N}]$ and $[\frac{N-1}{N}, 1]$ each contribute $\frac{1}{12}\frac{1}{N^3}$. For each $2 \le q \le N$ there are $\phi(q)$ fractions $s = \frac{p}{q}$ in lowest terms. Any two adjacent Farey fractions $\frac{p}{q}, \frac{p'}{q'} \in \mathcal{F}_N$ have $q + q' > N$, for otherwise their mediant $\frac{p+p'}{q+q'} \in \mathcal{F}_N$ and falls between $\frac{p}{q}$ and $\frac{p'}{q'}$ contradicting their being adjacent. Thus if $\frac{p^-}{q^-} < \frac{p}{q} < \frac{p^+}{q^+}$ are the two neighboring fractions in $\mathcal{F}_N$, $q \le \frac{N}{2}$ implies that $q^- > N/2$ and $q^+ > N/2$. A well-known property of Farey fractions is that $|pq' - p'q| = 1$ for adjacent fractions, hence the interval $I^- = [\frac{p^-}{q^-}, \frac{p}{q}]$ has length $l^-(s) = \frac{1}{qq^-}$ and $I^+ = [\frac{p}{q}, \frac{p^+}{q^+}]$ has length $l^+(s) = \frac{1}{qq^+}$. All these intervals are disjoint and they contribute

$$S_N' = \sum_{s \in \mathcal{F}_N \top q \le N/2} \frac{1}{12}\left(\frac{1}{(qq^-)^3} + \frac{1}{(qq^+)^3}\right)$$

to $MSE(\mathcal{F}_N)$. Since $N - q < q^+, q^- \le N$ we obtain the bounds

$$\frac{1}{6N^3} + \frac{1}{12}\left(\sum_{q=2}^{N/2} \frac{\phi(q)}{q^3}\frac{2}{N^3}\right) \le S_N' \le \frac{1}{6N^3} + \frac{1}{12}\sum_{q=2}^{N/2}\left(\frac{\phi(q)}{q^3}\frac{2}{(N-q)^3}\right) .$$

Now,

$$\frac{1}{(N-q)^3} - \frac{1}{N^3} = \frac{3N^2 q - 3Nq^2 + q^3}{N^3(N-q)^3}$$

so that

$$\sum_{q=2}^{N/2} \frac{\phi(q)}{q^3}\left(\frac{1}{(N-q)^3} - \frac{1}{N^3}\right) = O\left(\frac{\log N}{N^4}\right),$$

on using $(N-q)^3 \ge \frac{1}{8}N^3$, and $\frac{\phi(q)}{q^3} \le \frac{1}{q^2}$. These bounds imply

$$\frac{1}{6N^3} + \frac{1}{12}\left(\sum_{q=2}^{N/2}\frac{\phi(q)}{q^3}\right)\frac{2}{N^3} \le S_N' \le \frac{1}{6N^3} + \frac{1}{12}\left(\sum_{q=2}^{N/2}\frac{\phi(q)}{q^3}\right)\frac{2}{N^3} + O\left(\frac{\log N}{N^4}\right).$$

Since

$$\sum_{q=N/2}^{\infty} \frac{\phi(q)}{q^3} = O\left(\frac{1}{N}\right)$$

we obtain

$$S_N' = \left(\sum_{q=1}^{\infty}\frac{\phi(q)}{q^3}\right)\frac{1}{6N^3} + O\left(\frac{\log N}{N^4}\right) . \tag{B.2}$$

Let S_N'' denote the contribution to $MSE(\mathcal{F}_N)$ coming from all the remaining intervals. Each of them has endpoints $\frac{p}{q}, \frac{p'}{q'}$ with $q, q' \ge N/2$, hence their length $|I| = \frac{1}{qq'} \le \frac{4}{N^2}$. There are most N^2 such intervals, hence

$$S_N'' \le N^2\left(\frac{4}{N^2}\right)^3 \le \frac{64}{N^4} .$$

Combining this with (B.2) establishes (B.1).

We next consider the case $u_0 = 0$ and estimate

$$MSE_0(Q_{opt}) = \sum_{I \in \mathcal{G}_{2N}} \frac{1}{12}|I|^3 \ , \tag{B.3}$$

We will show that

$$MSE_0(Q_{opt}) = \left(\sum_{q=1}^{\infty} \frac{\phi(q)}{q^3}\right) \frac{1}{24N^3} + \left(\sum_{m=1}^{\infty} \frac{\phi(2m+1)}{(2m+1)^3}\right) \frac{1}{16N^3} + O\left(\frac{\log N}{N^4}\right) . \tag{B.4}$$

To begin, for $\mathcal{G}_{2N}$ we note that $\frac{1}{2}$ is a threshold and the set of thresholds is symmetric about $1/2$, so that

$$MSE_0(Q_{opt}) = 2 \sum_{\substack{I \subseteq [0,1/2] \\ I \in \mathcal{G}_{2N}}} \frac{1}{12}|I|^3 \ . \tag{B.5}$$

Next, we rescale the thresholds in $\mathcal{G}_{2N} \cap [0, 1/2]$ by a factor of 2 to obtain the *modified Farey series of order N*

$$\mathcal{F}_N^* := \left\{ \frac{p}{q} : \frac{p}{q} \in \mathcal{F}_N \quad \text{and} \quad p \equiv 1 (\text{mod } 2) \right\} \ .$$

Thus

$$MSE_0(Q_{opt}) = \frac{1}{4} MSE(\mathcal{F}_N^*) \ , \tag{B.6}$$

where

$$MSE(\mathcal{F}_N^*) := \sum_{I \in \mathcal{F}_N^*} \frac{1}{12}|I|^3 \ . \tag{B.7}$$

The modified Farey series $\mathcal{F}_N^*$ is obtained from the Farey series $\mathcal{F}_N$ by removing all points $s = \frac{p}{q}$ with $2|p$, which we call "even" Farey points. Let the neighboring Farey points to the left and right of such an s be $\frac{p^-}{q^-} < \frac{p}{q} < \frac{p^+}{q^+}$, and note that neither of the neighboring points is "even". The associated Farey intervals $[\frac{p^-}{q^-}, \frac{p}{q}]$ and $[\frac{p}{q}, \frac{p^+}{q^+}]$ have lengths $l^+(s) = \frac{1}{qq^+}$ and $l^-(s) = \frac{1}{qq^-}$. In $\mathcal{F}_N^*$ these intervals are combined into a single interval of length $l^+(s) + l^-(s)$. The contribution to $MSE(\mathcal{F}_N^*)$ for this interval is $\frac{1}{12}(l^+(s) + l^-(s))^3$, rather than the two contributions $\frac{1}{12}l^+(s)^3 + \frac{1}{12}l^-(s)^3$ in $MSE(\mathcal{F}_N)$. Using the identity $(x+y)^3 - (x^3 + y^3) = 3xy(x+y)$, we have

$$MSE(\mathcal{F}_N^*) = MSE(\mathcal{F}_N) + T_N, \tag{B.8}$$

where

$$T_N = \frac{1}{12} \sum_{\substack{s \in \mathcal{F}_N \\ s \text{ "even"}}} 3l^+(s)l^-(s)(l^+(s) + l^-(s)) \, . \tag{B.9}$$

To estimate T_N, we split it into two subsums T_N' and T_N'', where T_N' sums over all "even" $s = \frac{p}{q}$ with $1 < q < \frac{N}{2}$, and T_N'' sums over those "even" s with $\frac{N}{2} \le q \le N$.

In the first subsum, we have $N - q \le q^+, q^- \le N$, hence

$$\frac{1}{4}\left(\sum_{\substack{q=3\\ q \text{ odd}}}^{N/2} \frac{\frac{1}{2}\phi(q)}{q^3}\frac{2}{N^3}\right) \le T_N' \le \frac{1}{4}\left(\sum_{\substack{q=3\\ q \text{ odd}}}^{N/2} \frac{\frac{1}{2}\phi(q)}{q^3}\frac{2}{(N-q)^3}\right). \tag{B.10}$$

Here s is "even", so its denominator q is odd, and there are exactly $\frac{1}{2}\phi(q)$ "even" numerators p. We obtain

$$T_N' = \frac{1}{4}\left(\sum_{m=1}^{\infty} \frac{\phi(2m+1)}{(2m+1)^3}\right)\frac{1}{N^3} + O\left(\frac{\log N}{N^4}\right).$$

in a similar fashion to the estimate (B.2).

To bound the second sum T_N'', we note that the mediant $\frac{p^+ + p^-}{q^+ + q^-}$ lies inside $[\frac{p^-}{q^-}, \frac{p^+}{q^+}]$. The only point of $\mathcal{F}_N$ inside this interval is s, so we conclude that

$$q^+ + q^- \ge q \ge \frac{N}{2}.$$

Thus at least one of q^+ and q^- exceeds $\frac{N}{4}$, hence

$$l^+(s)l^-(s)(l^+(s) + l^-(s)) \le \frac{1}{(N/2)^3}\frac{1}{N/4}\frac{2}{q_0^2} \le \frac{16}{N^4}\frac{1}{q_0^2}.$$

where $q_0 = \min(q^+, q^-)$. Each value q_0 can occur at most $2\phi(q_0)$ times, hence

$$T_N'' \le \frac{1}{6}\sum_{q_0=1}^{N} \frac{\phi(q_0)}{q_0^2}\frac{16}{N^4} \le O\left(\frac{\log N}{N^4}\right).$$

Combining these estimates gives[7]

$$T_N = T_n' + T_N'' = \left(\sum_{m=1}^{\infty} \frac{\phi(2m+1)}{(2m+1)^3}\right)\frac{1}{4N^3} + O\left(\frac{\log N}{N^4}\right). \tag{B.11}$$

Then combining (B.1), (B.6), (B.8) and (B.11) yields the desired formula (B.4).

It remains to obtain explicit formulas for the coefficients in the formulas (B.1) and (B.4) above, i.e. to determine the constants α_{u_0}. We use the fact ([30, p. 6]) that

$$\frac{\zeta(s-1)}{\zeta(s)} = \sum_{q=1}^{\infty} \frac{\phi(q)}{q^s}.$$

[7]The possible "even" point $s = \frac{N-1}{N}$ contributes only $O(\frac{1}{N^4})$ to T_N and goes in the error term.

One easily calculates that

$$\sum_{m=1}^{\infty} \frac{\phi(2m+1)}{(2m+1)^s} = \left(\frac{1-2^{1-s}}{1-2^{-s}}\right) \frac{\zeta(s-1)}{\zeta(s)} - 1 ,$$

since $\phi(2m) = \phi(m)$ if m is odd and $\phi(2m) = 2\phi(m)$ if m is even. We take $s = 3$ and obtain

$$\begin{aligned} MSE_{1/2}(Q_{opt}) &= \left(\frac{\zeta(2)}{\zeta(3)}\right) \frac{1}{6N^3} + O\left(\frac{\log N}{N^4}\right) \\ &= \frac{\alpha_{1/2}}{N^3} + O\left(\frac{\log N}{N^4}\right), \end{aligned}$$

where $\alpha_{1/2} = \frac{1}{6}\frac{\zeta(2)}{\zeta(3)}$, and

$$\begin{aligned} MSE_0(Q_{opt}) &= \left(\frac{\zeta(2)}{\zeta(3)}\right) \frac{1}{24N^3} + \left(\frac{6}{7}\frac{\zeta(2)}{\zeta(3)} - 1\right) \frac{1}{16N^3} + O\left(\frac{\log N}{N^4}\right) \\ &= \frac{\alpha_0}{N^3} + O\left(\frac{\log N}{N^4}\right) \end{aligned}$$

where $\alpha_0 = \frac{2}{21}\frac{\zeta(2)}{\zeta(3)} - \frac{1}{16}$. ■

Appendix C. Proof of Lemmas 4.1 and 4.2

Proof of Lemma 4.1. We shall prove the lemma by establishing the formula

$$\int_0^1 \langle nx + \beta_0 \rangle \langle mx + \beta_0 \rangle \mathrm{d}x = \frac{1}{4} + \frac{1}{12} \frac{(n,m)^2}{nm} \phi_{n,m} , \tag{C.1}$$

where $|\phi_{n,m}| \leq 1$. Denote the expression on the left hand side by $c_{n,m}$. We substitute $nx + \beta_0$ and $mx + \beta_0$ for x in the Fourier series expansion

$$\langle x \rangle = \sum_{k=-\infty}^{\infty} a_k e^{2\pi i k x},$$

where $a_k = (-2\pi i k)^{-1}$ for $k \neq 0$ and $a_0 = 1/2$. This Fourier series is only conditionally convergent, and is to be interpreted as the limit as $N \to \infty$ of the sum taken from $-N$ to N. However, its partial sums are uniformly bounded,

$$\left| \sum_{|k| \leq N} a_k e^{2\pi i k x} \right| \leq C, \qquad \text{for all } x \text{ and all } N. \tag{C.2}$$

In fact, one can take $C = \frac{1}{2} + \frac{5}{4\pi}$ [24, Ex. 4, p. 22]. Hence, using the bounded convergence theorem, one can change the order of integration and double sum to obtain

$$c_{n,m} = \sum_{k \in \mathbf{Z}} \sum_{l \in \mathbf{Z}} a_k \overline{a}_l \, e^{2\pi i (k-l)\beta_0} \int_0^1 e^{2\pi i (kn - lm)x} \mathrm{d}x. \tag{C.3}$$

Summing up (C.3) over the nonzero indices k, l given by $kn = lm$ and straightforward manipulations result in

$$c_{n,m} = \frac{1}{4} + \frac{1}{4\pi^2}\frac{(n,m)^2}{nm}\sum_{d\neq 0}\frac{1}{d^2}e^{2\pi i d\beta_0(m-n)/(n,m)}, \tag{C.4}$$

which in turn implies using $\sum_{d\neq 0}\frac{1}{d^2} = \frac{\pi^2}{3}$,

$$c_{n,m} = \frac{1}{4} + \frac{1}{12}\frac{(n,m)^2}{nm}\phi_{n,m} \tag{C.5}$$

for some $|\phi_{n,m}| \leq 1$. ■

Remark. Using the formula

$$\sum_{d=1}^{\infty}\frac{1}{d^2}\cos d\theta = \frac{1}{4}(\theta-\pi)^2 - \frac{1}{12}\pi^2, \qquad 0 \leq \theta \leq 2\pi, \tag{C.6}$$

the exact value of $\phi_{n,m}$ is easily found to be

$$\phi_{n,m} = \frac{3}{2}(2\langle\beta_0\frac{m-n}{(n,m)}\rangle - 1)^2 - \frac{1}{2}\ .$$

Proof of Lemma 4.2. We have

$$\begin{aligned}
\sum_{n=1}^{L}\sum_{m=1}^{L}\frac{(n,m)^2}{nm} &= \sum_{d=1}^{L}\sum_{\substack{1\leq n,m\leq L\\(n,m)=d}}\frac{(n,m)^2}{nm} \\
&\leq \sum_{d=1}^{L}\sum_{j=1}^{\lfloor L/d\rfloor}\sum_{k=1}^{\lfloor L/d\rfloor}\frac{1}{jk} \\
&= \sum_{d=1}^{L}\left(\sum_{j=1}^{\lfloor L/d\rfloor}\frac{1}{j}\right)^2 \\
&\leq \sum_{d=1}^{L}(1+\log\frac{L}{d})^2
\end{aligned} \tag{C.7}$$

However, this last expression is bounded by

$$(1+\log L)^2 + \int_1^L (1+\log(L/y))^2 \mathrm{d}y = 5L - 4 - 2\log L,$$

which proves (4.5). ■

Remark. The constant 5 appearing in (4.5) can be improved to $\gamma^2 + 5\gamma/2 + 7/3 \cong 4.1$ by using the inequality

$$\sum_{j=1}^{L} \frac{1}{j} \leq \gamma + \log L + \frac{1}{2L} \ , \tag{C.8}$$

where γ is the Euler-Mascheroni constant defined by

$$\gamma = \lim_{N\to\infty} \left(\sum_{j=1}^{N} \frac{1}{j} - \log N \right) = 0.5772\ldots \tag{C.9}$$

Numerical experiments suggest the optimal constant to be ~ 3.

References

[1] J. C. Candy, "A use of double integration in sigma delta modulation," *IEEE Transactions on Communications*, vol. COM-33, pp. 249–258, Mar. 1985.

[2] J. C. Candy and O. J. Benjamin, "The structure of quantization noise from Sigma-Delta modulation," *IEEE Transactions on Communications.* vol. COM-34, pp. 1316–1323, Sept . 1981.

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

[4] J. C. Candy and G. C. Temes, "Oversampling Methods for A/D and D/A Conversion," in [3], pp. 1–29.

[5] W. Chou, P. W. Wong and R. M. Gray, "Multistage Sigma-Delta modulation," *IEEE Transactions on Information Theory* vol. 35, pp. 784–796, July 1989.

[6] I. Daubechies, R. DeVore, "Reconstructing a Bandlimited Function From Very Coarsely Quantized Data: I. A Family of Stable Sigma-Delta Modulators of Arbitrary Order", in preparation.

[7] O. Feely and L. O. Chua, "The Effect of Integrator Leak in $\Sigma-\Delta$ Modulation," *IEEE Transactions on Circuits and Systems* vol. 38, pp. 1293–1305, November 1991.

[8] R. M. Gray, "Oversampled sigma-delta modulation," *IEEE Transactions on Communications*, vol. COM-35, pp. 481–489, May 1987.

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

[10] S. Güntürk, "Reconstructing a Bandlimited Function From Very Coarsely Quantized Data: II. Improving the Error Estimate for First Order Sigma-Delta Modulators," in preparation.

[11] S. Güntürk, "Improved error estimates for first order sigma-delta modulation," *Sampling Theory and Applications, SampTA'99*, Loen, Norway, August 1999.

[12] S. Güntürk, "Number theoretical error estimates in a quantization scheme for bandlimited signals," In: *Unusual Applications of Number Theory*, DIMACS Volume, Amer. Math. Soc.: Providence, RI, 2000, to appear.

[13] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers.* Oxford University Press, 1979.

[14] T. Hayashi, Y. Inabe, K. Uchimure, and T. Kimura, "A multistage delta-sigma modulator without double integration loop," in *ISSCC Dig. Technical Papers*, pp. 182–183, February 1986.

[15] N. He, F. Kuhlmann and A. Buzo, "Double-loop sigma-delta modulation with dc input," *IEEE Trans. Comm.*, vol. 38, pp. 487–495, 1990.

[16] N. He, F. Kuhlmann and A. Buzo, "Multiloop sigma-delta quantization," *IEEE Trans. on Information Theory,* vol. 38, pp. 1015–1027, May 1992.

[17] S. Hein, K. Ibrahim and A. Zakhor, "New properties of sigma-delta modulators with dc inputs." *IEEE Transactions on Communications*, vol. COM-40, pp. 1375–1387, Aug. 1992.

[18] S. Hein and A. Zakhor, "Lower Bounds on the MSE of the single and double loop sigma delta modulators," Proc. IEEE Intl. Symp. on Circuits and Systems, 1990, Vol. 3, 1751–1755.

[19] S. Hein and A. Zakhor, "Optimal decoding for data acquision applications of Sigma Delta modulators," *IEEE Trans. Signal Processing*, vol. 41, pp. 602–616, Feb. 1993.

[20] S. Hein and A. Zakhor, *Sigma Delta Modulators: Nonlinear Decoding Algorithms and Stability Analysis*, Kluwer Academic Publ.: Dordrecht 1993.

[21] S. Hein and A. Zakhor, "Reconstruction of oversampled band-limited signals from $\Sigma\Delta$ encoded binary sequences," IEEE Trans. Signal Processing, vol. 42, pp. 799–811, April 1994.

[22] David F. Hoschele, Jr., *Analog-to-Digital and Digital-to-Analog Conversion Techniques*, John Wiley & Sons: New York 1994.

[23] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proceedings of the IEEE*, vol. 51, pp. 1524–1535, Nov. 1963.

[24] Y. Katznelson, *An Introduction to Harmonic Analysis*, John Wiley & Sons, 1968 (reprint: Dover Publs. Inc.).

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

[26] S. K. Tewksbury and R. W. Hallock, "Oversampled, linear predictive and noise shaping coders of order $4N > 1$," IEEE Trans. Circuits Syst., vol CAS-25, pp. 436–447, July 1978.

[27] N. T. Thao, Vector quantization analysis of Sigma-Delta modulation," IEEE Trans. Signal Processing, vol. 44, pp. 808–817, April 1996.

[28] N. T. Thao, "Deterministic analysis of Sigma-Delta modulation for linear and nonlinear signal reconstruction," International J. Circuit Theory Appl., Special Issue on Delta-Sigma Modulators and Noise Shaping, pp. 369–392, Sep.-Oct. 1997.

[29] N. T. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," IEEE Trans. Signal Processing, vol. 42, pp. 519–531, Mar. 1994.

[30] E. C. Titchmarsh and D. R. Heath-Brown, *The Theory of the Riemann Zeta-Function*, Second Edition, Oxford Univ. Press 1986.

email: `cgunturk@math.princeton.edu`
`jcl@research.att.com`
`vinay@research.att.com`