Optimal Trade-off Between Sampling Rate and Quantization Precision in Sigma-Delta A/D Conversion

Alon Kipnis*, Andrea J. Goldsmith* and Yonina C. Eldar†

Abstract—The optimal sampling frequency in a Sigma-Delta analog-to-digital converter with a fixed bitrate at the output is studied. We consider the mean squared error performance metric where the input signal statistics are known. Fixing the output bitrate introduces a trade-off between the sampling rate and the number of bits used to quantize each sample. That is, while increasing the sampling rate reduces the in-band quantization noise, it also reduces the number of bits available to quantize each sample and therefore increases the magnitude of the quantization noise. The optimal sampling rate is the result of the interplay between these two phenomena. In this work we analyze the sampling rate of a Sigma-Delta modulator of arbitrary order under the approximation that the quantization error behaves like additive white noise that is uncorrelated with the signal. We show that for a signal with a spectrum that is constant over its bandwidth, the optimal sampling rate is either the Nyquist rate or the maximal sampling rate corresponding to the output bitrate. The choice between the two is approximately a function of the Sigma-Delta system order and the bitrate per unit bandwidth.

I. INTRODUCTION

A. Background

In analog to digital conversion (ADC) an analog signal is converted into a sequence of bits. Shannon’s distortion-rate function [1] gives the theoretical minimal error as a function of the bitrate of the digital sequence, however it does not provide concrete methods for the A/D conversion. Practical ADC schemes involve operations of sampling and quantization. The overall bitrate in the resulting digital representation is the product of the sampling rate with the average number of bits used to store each sample.

In this work we are interested in the trade-off between these two quantities in A/D conversion using Sigma-Delta modulation (Σ∆M). In this ADC scheme, the input process is oversampled (sampled above its Nyquist rate) and quantized using a low-resolution quantizer (usually 1-bit). Σ∆M also employs a negative feedback loop and an integrator so that quantization error of previous samples will be considered in quantizing consecutive samples.

While oversampled modulation does not provide any theoretical improvement over sampling at the Nyquist rate or at the minimal rate that achieves the rate-distortion function [2], Σ∆M is commonly used in applications due to its relatively cheap and simple hardware implementation. However, its high sampling rates may be hard to implement in some applications [3]. This makes a performance analysis of Σ∆M relevant for all sampling frequencies, and not only in the high over-sampling rate regime.

In this work we analyze the Σ∆M as a source coding scheme, that is, we are interested in the minimal error as a function of the bitrate. For that purpose we assume a statistical model on the input process and mean squared error (MSE) as our performance metric. We use the additive white noise assumption [4] for quantization error, where the variance of the quantization noise decreases exponentially with the number of bits per sample q.

If the analog source is sampled at frequency $f_s$, the memory rate at the output of the quantizer is $R = q f_s$ bits per time unit. Since Σ∆M uses oversampling to reduce the amount of in-band quantization noise, increasing $f_s$ decreases the error and effectively improves the resolution of the quantizer. This implies that fixing the memory rate $R$ introduces an interplay between $f_s$ and $q$ that induces a trade-off between the amount of in-band quantization noise and the magnitude of this noise.

B. Related Work

A Σ∆M is based on the principle of oversampling and a negative feedback loop that includes an integrator. The paper [5] provides an extended tutorial of the theoretical and practical aspects in Σ∆M. As in other systems which involve quantization, in Σ∆M analysis it is common to approximate the difference between the quantizer input and output by a white additive noise, see e.g. [6]. While the conditions under which this assumption yields a good approximation are not usually met in Σ∆M, it has been shown in several cases that the white noise assumption does not significantly change the performance results obtained through a rigorous analysis which does not make this assumption. We will discuss this approximation more in Section II.

The feedback loop in the Σ∆M is sometimes referred to as a quantization noise shaping system. The white quantization noise assumption implies that in the absence of the feedback loop (zero order Σ∆M), the power of the quantization noise within the signal band decreases linearly with $f_s$. With a simple noise shaping system [6] the quantization noise is attenuated even more and the in-band noise power decreases by a factor of $f_s^{2L+1}$, where $L$ is the number of consecutive feedback loops or the modulator order. This implies a mean squared error (MSE) reduction of $R^{2L+1}$ in the bitrate. This error reduction is still much slower than the exponential reduction of the optimal distortion-rate trade-off in Shannon’s distortion-rate function [1]. Oversampling schemes which try to bridge this gap and

---

*Department of Electrical Engineering, Stanford University, CA
†Department of EE Technion, Israel Institute of Technology, Haifa, Israel
achieve exponential MSE reduction in bitrate using a single bit quantizer were suggested in [7] and [8].

In this work we do not restrict ourselves to single bit measurements. Instead, we study the error in \( \Sigma \Delta \) for all sampling rates under a fixed bitrate \( R \) where the quantizer resolution \( q \) and the sampling rate \( f_s \) may vary.

C. Contribution

We derive an expression for the MSE under optimal linear estimation in an \( L \)th order \( \Sigma \Delta \) of Fig. 1 and a fixed bitrate \( R \) at the output of the modulator, under the assumption that the quantization error behaves like an additive white noise that is uncorrelated with the input signal. In addition, we analyze the sampling rate \( f_s^* \) that minimizes the MSE for this bitrate.

For input signals whose spectrum is constant over their entire bandwidth \( 2f_B \), the optimal sampling rate is found to be either the Nyquist rate or the maximal sampling rate allowed by the system. Which of the two is optimal is a function of \( L, R \) and \( f_B \), that can be well-approximated by \( R/(2f_B) \), which is the number of bits per unit bandwidth.

The rest of this paper is organized as follows: relevant background on \( \Sigma \Delta \) is given in Section II. In Section III we solve an MSE estimation problem from discrete-time measurements. In Section IV we describe our main results. Concluding remarks are provided in Section V.

II. PRELIMINARIES

We consider the \( L \)th order \( \Sigma \Delta \) of Fig. 1. The input process is an analog wide-sense stationary (WSS) process \( X(\cdot) = \{X(t), t \in \mathbb{R}\} \) with PSD

\[
S_X(f) \triangleq \int_{-\infty}^{\infty} E[X(t+\tau)X^*(\tau)]e^{-2\pi i f \tau} d\tau.
\]

The discrete-time process \( \bar{X}[\cdot] = \{\bar{X}[n], n \in \mathbb{Z}\} \) is obtained by uniformly sampling \( X(\cdot) \) at frequency \( f_s \), namely

\[
\bar{X}[n] \triangleq X(n/f_s), \quad n \in \mathbb{Z}.
\]

The signal \( Y[\cdot] \) at the output of the quantizer is given by

\[
Y[n] = X[n] + \eta[n], \quad n \in \mathbb{Z},
\]

where \( X[n] \) is the process at the input to the quantizer, and \( \eta[\cdot] \) is referred to as the quantization error process. In order to linearize the system in Fig. 1 we make the following assumption:

Assumption 1. The quantization error \( \eta[\cdot] \) is an i.i.d process independent of \( X[\cdot] \) with variance \( \sigma^2_\eta \) that decreases exponentially with the quantizer bit-resolution \( q \).

Assumption 1 says that quantization error is a white noise process independent of \( X[\cdot] \) with PSD

\[
S_\eta(e^{2\pi i \phi}) = \sigma^2_\eta = \frac{c_0}{(2^q - 1)^2} \approx \frac{c_0}{2^q}.
\]

Conditions under which Assumption 1 provides a good approximation for the true behavior of the quantization error were derived in [9]. While these conditions are not met in general in \( \Sigma \Delta \), we motivate our use of Assumption 1 by the following two facts: 1) error analysis which do not use the approximation of Assumption 1 as in [10], [11], [12] predict MSE reduction of no more than 3db per octave faster than \( R \) as compared to a simplified analysis using Assumption 1, e.g. [6]. 2) The optimal linear MSE derived under Assumption 1 is always higher than the optimal linear MSE in the case where the noise is correlated with \( X[\cdot] \) or with itself, under the same marginal distribution of the input signal and the noise. This last point is explained in more detail in [13].

Assumption 1 also implies that digital-to-analog converter (DAC) that reverses the quantizer operation preserves the linear input-output relation. It follows [6] that the relation between the input and the output of the \( \Sigma \Delta \) can be represented in the \( z \) domain by:

\[
Y(z) = STF(z) \bar{X}(z) + NTF(z) \eta(z),
\]

where the functions \( STF(z) \) and \( NTF(z) \) satisfy

\[
STF(z) = z^{-1},
\]

Fig. 1. \( L \)th order \( \Sigma \Delta \) modulator.
and \[ \text{NTF}(z) = (1 - z^{-1})^L. \]

Equation (3) leads to the following relation between the corresponding PSDs:

\[
S_Y(e^{2\pi j \phi}) = |\text{STF}(e^{2\pi j \phi})|^2 S_X(e^{2\pi j \phi}) + |\text{NTF}(e^{2\pi j \phi})|^2 S_N(e^{2\pi j \phi})
\]

\[
= S_X(e^{2\pi j \phi}) + |1 - e^{2\pi j \phi}|^2 L S_N(e^{2\pi j \phi})
\]

\[
= S_X(e^{2\pi j \phi}) + 2\sin(\pi f_s / f_c) L S_N(e^{2\pi j \phi})
\]

Equation (4) implies that only the quantization noise is statistically affected by the feedback of the system, a phenomena known as noise shaping.

III. MINIMAL MSE IN DISCRETE-TIME TO CONTINUOUS-TIME LINEAR ESTIMATION

In this section we are interested in the minimal MSE (MMSE) in linear estimation of the analog process \( X(\cdot) \) from the modulator output \( Y(\cdot) \), i.e. the process \( \hat{X}(\cdot) \) that minimizes

\[
\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} E[(X(t) - \hat{X}(t))^2] \, dt
\]

over all processes reconstructing \( X(\cdot) \) from \( Y(\cdot) \) of the form

\[
\hat{X}(t) = \sum_{n \in \mathbb{Z}} w(t, n) Y[n].
\]

This can be seen as a general estimation problem of a WSS continuous-time signal from its noisy samples described by (4). The solution to this estimation problem is given in the following theorem.

**Theorem 1.** Consider the input-output relation (4) where \( X(\cdot) \) is a WSS process bandlimited to \( f_B \), \( \hat{X}(\cdot) \) are its uniform samples at rate \( f_s \) and \( \eta(\cdot) \) is an independent WSS noise process. If \( f_s \geq 2f_B \), then the minimal time-averaged MSE in (5) in linear estimation of \( X(\cdot) \) from \( Y(\cdot) \) is given by

\[
\text{mmse} = \int_{-f_B}^{f_B} \frac{S_X(f)}{1 + \text{SNR}(f)} df,
\]

where \( \text{SNR}(f) \triangleq f_s S_X(f) / \left| \text{STF}(e^{2\pi j f_s / f_c}) \right|^2 \left| \text{NTF}(e^{2\pi j f_s / f_c}) \right|^2. \)

\[
\text{SNR}(f) = f_s S_X(f) / \left| \text{STF}(e^{2\pi j f_s / f_c}) \right|^2 \left| \text{NTF}(e^{2\pi j f_s / f_c}) \right|^2.
\]

**Proof:** We only give here a sketch of the proof. The details can be found in [13]. For \( 0 \leq \Delta < 1 \) define

\[
X_\Delta[n] = X((n+\Delta)/f_s), \quad n \in \mathbb{Z}.
\]

It follows that the minimal MSE in (5) can be written as

\[
\text{mmse} = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} E[(X_\Delta[n] - \hat{X}_\Delta[n])^2] \, d\Delta
\]

where \( \hat{X}_\Delta[n] \) is the optimal MSE estimator of \( X_\Delta[n] \) from \( Y[\cdot] \). Since \( X_\Delta[\cdot] \) and \( Y[\cdot] \) are jointly stationary, the result follows by using the error in a non-causal Wiener filter for estimating \( X_\Delta[n] \) from \( Y[\cdot] \) in (8), and integrating over \( \Delta \).

In Theorem 1 we only gave an expression for the optimal MSE. A linear estimator \( \hat{X}(\cdot) \) that achieves the minimal MSE (6) can be obtained in one of the following ways:

(i) Digital low-pass filter with cutoff frequency \( f_B/f_s \) followed by a discrete-to-analog interpolation, followed by an analog Wiener filter to recover \( X(\cdot) \) from the analog process.

(ii) Digital Wiener filter to recover the sampled source \( \hat{X}[\cdot] \) from \( Y[\cdot] \) followed by discrete-to-analog interpolation.

The proof of this assertion can be found in [13].

By using the STF and NTF from (4) in Theorem 1, we obtain the following expression for the minimal MSE in an \( L \)th order \( \Sigma\Delta \)M:

\[
\text{mmse}_{\Sigma\Delta} = \int_{-f_B}^{f_B} \frac{S_X(f)}{1 + \text{SNR}_{\Sigma\Delta}(f)} df,
\]

where

\[
\text{SNR}_{\Sigma\Delta}(f) = f_s S_X(f) / \left| \text{STF}(e^{2\pi j f_s / f_c}) \right|^2 \left| \text{NTF}(e^{2\pi j f_s / f_c}) \right|^2.
\]

Note that \( \text{SNR}_{\Sigma\Delta}(f) \) can be approximated by \( \left( \pi f_s / f_c \right)^2 \) provided \( f_s \) is much bigger than the Nyquist rate \( f_c \). This implies that as long as \( \sigma_\eta^2 \) remains constant, \( \text{SNR}(f) \) increases as \( f_s^{2L+1} \) with increasing \( f_s \). In the following subsection we will study the behavior of (9) when the variance of the quantization noise increases with \( f_s \), so as to keep a prescribed bitrate \( R = q f_s \).

IV. \( \Sigma\Delta \) MODULATION UNDER A FIXED BITRATE

In a sampling and quantization system with sampling rate \( f_s \) and a quantizer resolution of \( q \) bits per sample, the amount of memory per time unit, or the bitrate, at the output of the system is \( R \triangleq q f_s \) bits per time unit. In this section we fix the output bitrate \( R \) and study the dependency of the MSE in the sampling rate \( f_s \). This means that the bit resolution \( q \) of the ADC decreases as the sampling rate increases. Indeed, under the fixed bitrate assumption, the variance of the quantization noise from (2) satisfies

\[
\sigma_\eta^2 = \frac{c_0}{2^{2q}} = \frac{c_0}{2^{2R/f_s}},
\]

which is an increasing function of \( f_s \). Since in this model the ADC must use at least one bit per sample, we limit \( f_s \) to be smaller than the bitrate \( R \). We also assume that \( X(\cdot) \) is band-limited to \( f_B \) and that \( f_s \) is bigger than the Nyquist rate \( 2f_B \). That is, our interval of interest for \( f_s \) is \( f_s \in [2f_B, R] \). From (9) and (10), we obtain the following expression for the minimal MSE under a fixed bitrate \( R \):

\[
\text{mmse}(f_s, L, R) = \int_{-f_B}^{f_B} S_X(f) \left( f_s / 1 + \text{SNR}(f) \right) \, df
\]
where

\[ SNR(f) = SNR_{L,f_s,R}(f) = \frac{S_X(f)}{c_0} \frac{f_s 2^{R/f_s}}{2^L \sin^2 \left( \frac{\pi f}{f_s} \right)} \]  

(13)

It follows that the contribution of each frequency \( f \in (-f_B, f_B) \) to the MSE is inversely proportional to the SNR at that frequency. This SNR is the result of the interplay between the overall in-band noise which is shaped by a factor of \( \left( 2 \sin \left( \frac{\pi f}{f_s} \right) \right)^{2L} / f_s \), and the magnitude of that noise which is attenuated by \( 2^{-2R/f_s} \).

We can bound \( \text{mmse}(f_s, L, R) \) from above by replacing \( 1 + \text{SNR}(f) \) with \( \text{SNR}(f) \) in the denominator of (12). This leads to

\[ \text{mmse}(f_s, L, R) \leq \frac{c_0 2^{2L}}{f_s 2^{2R/f_s}} \int_{-f_B}^{f_B} \sin^2 \left( \frac{\pi f}{f_s} \right) df, \]  

(14)

which is independent of the particular PSD \( S_X(f) \). The MMSE as a function of \( f_s \) for a given \( R \) as well as the bound (14) are depicted in Fig. 2.

**Optimal Sampling Rate**

For a given \( R, L \) and \( S_X(f) \), we are interested in the sampling rate \( f_s^* \) that minimizes \( \text{mmse}(f_s, L, R) \) over \( f_s \in [2f_B, R] \). This requires an optimization over (12) and (13), and in general cannot be obtained in a closed form. By considering the ratio between \( \text{SNR}(f) \) at these two sampling rates, we obtain the following two statements:

**Proposition 1.** Fix \( S_X(f) \).

(i) For any \( L \in \mathbb{N} \), there exists \( R \) large enough such that \( f_s^* = 2f_B \).

(ii) For any \( R > 0 \) there exists \( L \in \mathbb{N} \) large enough such that \( f_s^* > 2f_B \).

Proposition 1 shows that we can divide the \( R-L \) plane into two regions: one that contains high values of \( R \) in which Nyquist rate sampling is optimal, and the other which contains high values of \( L \) in which oversampling achieves lower MMSE than Nyquist rate sampling. We will see below that for PSDs of the form (15), the optimal sampling rate is either the Nyquist rate or the maximal sampling rate \( f_s = R \).

**Flat Power Spectral Density**

In this subsection we consider the case where the PSD of the source is of the form

\[ S_X(f) = \sigma^2 \begin{cases} 1 & |f| \leq f_B, \\ 0 & |f| > f_B. \end{cases} \]  

(15)

where \( f_B > 0 \). Under the PSD (15), the SNR (13) can be written as

\[ \text{SNR}(f) = \frac{\sigma^2}{c_0} \frac{f_s 2^{R/f_s}}{2 \sin^2 \left( \frac{\pi f}{f_s} \right)} \]  

(16)

The second derivative of (16) with respect to \( f_s \) can be obtained in a closed form, and is found to be strictly positive for all \( L \in \mathbb{N}, R > 0, f \in (-f_s/2, f_s/2) \) and \( f_s \in [2f_B, R] \). That is, (16) is concave with respect to \( f_s \) in this domain and the maximal value of (16) is obtained at one of the endpoints of the interval \([2f_B, R]\) (see Fig. 2). We conclude that \( f_s^* \) is either the Nyquist rate or the maximal sampling rate \( f_s = R \). Since \( \text{SNR}(f) \) is maximal at these two values of \( f_s \), (14) provides a relatively good approximation for these two possible values of \( f_s^* \). At \( f_s = 2f_B \), (14) leads to

\[ \text{mmse}(2f_B, L, R) \leq \frac{c_0 2^{2L}}{4R} \int_{-f_B}^{f_B} \sin^2 \left( \frac{\pi f}{f_s} \right) df \leq \frac{c_0 2^{-R/2f_B} (2L)!}{(2^L)!}. \]  

(17)

For \( f_s = R \) we have

\[ \text{mmse}(R, L, R) \leq \frac{c_0 2^{2L}}{4R} \int_{-f_B}^{f_B} \sin^2 \left( \frac{\pi f}{f_s} \right) df \leq \frac{c_0 2^{2L}}{4R} \int_{-f_B}^{f_B} \left( \frac{\pi f}{R} \right)^{2L} df = \frac{c_0 \pi 2^{2L}}{4(1+2L)} \left( \frac{R}{2f_B} \right)^{-2L-1}, \]  

(18)

where this bound becomes tight as \( R \gg f_B \). From (17) and (18) we conclude that the value of \( \text{mmse}(f_s^*, L, R) \) is approximately a function of the maximal oversampling ratio

\[ \bar{R} \triangleq \frac{R}{2f_B}, \]

which can also be interpreted as the number of bits per unit bandwidth. Fig. 3 shows the \( \bar{R}-L \) plane that represents the optimal sampling rate for the PSD (15), where the dashed line approximates the border between the two regions by points at which (17) is equal to (18). The minimal MSE obtained under the optimal sampling rate is plotted in Fig. 4.
Discussion

By increasing the bitrate $R$ and sampling at the Nyquist rate, we increase the accuracy of the quantizer which leads to an exponential decrease in the MMSE as a function of $R$. On the other hand, the $\Sigma\Delta M$ with a 1-bit quantizer leads to a polynomial MMSE reduction in $R$. The optimal sampling rate $f_s^*$ and, as a result, the optimal bit allocation $q = R/f_s^*$ is determined by the interplay between these two behaviors, which must be traded off for a fixed bitrate.

The first part of Proposition 1 essentially says that if $R$ is large enough then the exponential MMSE decrease obtained by sampling at the Nyquist rate eventually leads to lower MMSE. This implies that for high bitrates $\Sigma\Delta M$ with a 1-bit quantizer does not lead to an efficient bit allocation, which is explained by the high correlation of the samples at high sampling rates. The second part of Proposition 1 implies that if $R$ is low compared to $L$, then using a 1-bit quantizer and oversampling at the maximal rate, and thus exploiting the noise shaping mechanism of the $\Sigma\Delta M$, is the preferred strategy. If $S_x(f)$ is of the form (15), the optimal sampling rate is always one of these two cases, i.e. either sampling at the Nyquist rate or oversampling at the maximal possible rate $R$. We note that an intermediate case in which $2f_B < f_s^* < R$ is possible depending on the particular form of the PSD, $R$ and $L$. In this case, sacrificing quantization resolution in order to sample at rates slightly higher than the Nyquist rate and thus exploit the feedback system of the modulator reduces the MSE.

V. Conclusions

We have considered A/D conversion using $\Sigma\Delta M$ with a fixed bitrate at the output, using a statistical model of a wide-sense stationary input and mean squared error under optimal linear estimation as the error metric. We showed that the optimal sampling rate $f_s^*$ and the optimal bit-allocation strategy is a function of the spectrum of the input signal, the modulator order and the bitrate $R$. Our analysis shows that the optimal bit allocation corresponds to oversampling when the bitrate is low and to Nyquist rate sampling when the bitrate is high.

ACKNOWLEDGMENT

This work was supported in part by the NSF Center for Science of Information (CSoI) under grant CCF-0939370, the BSF Transformative Science Grant 2010505 and the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

REFERENCES