Performance Analysis and Optimal Filter Design for Sigma-Delta Modulation via Duality with DPCM

Or Ordentlich
Tel Aviv University
ordent@eng.tau.ac.il

Uri Erez
Tel Aviv University
uri@eng.tau.ac.il

***Abstract*—Sampling above the Nyquist-rate is at the heart of sigma-delta modulation, where the increase in sampling rate is translated to a reduction in the overall (minimum mean-squared-error) reconstruction distortion. This is attained by using a feedback filter at the encoder, in conjunction with a low-pass filter at the decoder. The goal of this work is to characterize the optimal trade-off between the per-sample quantization rate and the resulting mean-squared-error distortion, under various restrictions on the feedback filter. To this end, we establish a duality relation between the performance of sigma-delta modulation, and that of differential pulse-code modulation when applied to (discrete-time) band-limited inputs. As the optimal trade-off for the latter scheme is fully understood, the full characterization for sigma-delta modulation, as well as the optimal feedback filters, immediately follow.**

I. Introduction

Analog-to-digital (A/D) and digital-to-analog (D/A) convertors are an integral part of almost all electric devices in use. Often, the same A/D (or D/A) component is applied to a variety of signals with distinct characterizations. For this reason, it is desirable to design the data-converter to be robust to the characteristics of its input signal.

One assumption that cannot be avoided is knowledge of the bandwidth of the signal to be converted (or at least an upper bound on the bandwidth), which dictates the minimal sampling rate, according to Nyquist's Theorem. Beyond the bandwidth, however, one would like to assume as little as possible about the input signal. A reasonable model for the input signal is therefore a *stochastic* one, where the input signal is assumed to be a stationary Gaussian process with a given variance and an arbitrary *unknown* power spectral density (PSD) within the assumed bandwidth, and zero otherwise. In this paper, we adopt this *compound* model which is rich enough to include a wide variety of processes. The robustness requirement from the A/D (or D/A) convertor translates to requiring that it induces a small average distortion simultaneously for all processes within our compound model.

Sigma-delta modulation is a widely used technique for A/D as well as D/A conversion. The main advantage offered by this type of modulation is the ability to trade-off the sampling rate and the number of bits per sample required for achieving a target mean-squared error (MSE) distortion. The input to the sigma-delta modulator is a signal sampled at L times the Nyquist rate ($L > 1$). This over-sampled signal is then quantized using an R-bit quantizer, where the task of exploiting the benefits of oversampling is performed by a feedback filter whose role is to "push" the quantization noise into high frequencies which are eventually filtered out at the receiver, see Figure 1.

Another technique for compressing sources with memory is differential pulse-code modulation (DPCM). In DPCM, a prediction filter is applied to the quantized signal. The output of this filter is then subtracted from the source and the result is fed to the quantizer, see Figure 2. At the decoder, the quantized signal is simply passed through the inverse of the prediction filter.

The connection between DPCM and sigma-delta modulation, as two instances of predictive coding, was known from the outset. In fact, both paradigms emerged from two Bell-Labs patents authored by C. C. Cutler in 1952 and 1954. Nevertheless, the techniques used for the performance analysis of DPCM and sigma-delta modulation are quite different. One explanation for the divergence in the analysis methods is that DPCM was developed as a prediction scheme for a stochastic signal, whereas sigma-delta modulation was originally invented as a noise-shaping technique aimed at achieving a more desirable noise spectrum, rather than reducing compression rate. However, through the years the most prominent use of sigma-delta modulation has become reducing compression rate at the expense of increasing sampling rate, as discussed above. Clearly, one can pursue the same goal by applying the DPCM scheme to an over-sampled signal.

The performance of DPCM under the assumption of *high-resolution* quantization is well understood since as early as the mid 60's. Under this assumption, the prediction filter should be chosen as the optimal linear minimum mean-squared-error (MMSE) prediction filter of the source process from its past [1], and the effect of the filtered quantization noise can be neglected in the prediction process. While in most cases where DPCM is traditionally used, the high resolution assumption is well justified, it totally breaks down for the class of band-limited processes, which includes the input signals to sigma-delta modulators.

Fortunately, the high-resolution assumption in DPCM analysis has been overcome in [2], where it was shown that

The work of O. Ordentlich was supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities, a fellowship from The Yitzhak and Chaya Weinstein Research Institute for Signal Processing at Tel Aviv University and the Feder Family Award. The work of U. Erez was supported by by the ISF under Grant 1557/13.

for any distortion level and any stationary Gaussian source, the DPCM architecture induces a rate-distortion optimal test channel, provided that the prediction filter is chosen as the optimal filter for predicting the source from its *qunatized past*, and in addition water-filling pre- and post-filters are applied. The analysis of [2], which takes into account the effect of the quantization noise, can therefore be used to obtain the optimal feedback filter and its corresponding performance for a DPCM system applied to an over-sampled stationary Gaussian source.

Our main result, derived in Section II, is that for over-sampled band-limited stationary Gaussian processes, the test channel induced by the sigma-delta modulator (Figure 1) achieves precisely the same rate-distortion function as that of the DPCM test channel (Figure 2) with a Gaussian stationary input with the same variance, whose spectrum is flat within the same frequency band. More specifically, for such processes, for any choice of σ^2_{DPCM} and prediction filter $C(Z)$ in the test channel of Figure 2, the same choice of $C(Z)$ together with the choice

$$\sigma^2_{\Sigma\Delta} = \frac{\sigma^2_{\text{DPCM}}}{L \cdot \frac{1}{2\pi}\int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega} \tag{1}$$

in Figure 1, yields the same compression rate and the same distortion.

While this result is simple to derive, it has a very pleasing consequence: the problem of optimizing the filter $C(Z)$ in sigma-delta modulation, under any set of constraints, can be cast as an equivalent problem of optimizing the DPCM prediction filter under the same set of constraints. Using results from linear time-invariant prediction theory, we can then easily find the optimal filter for sigma-delta modulation under constraints for which an explicit solution was lacking in the literature, or was cumbersome to derive.

Finally, in Section III we show that the rate-distortion trade-off derived for sigma-delta modulation in Section II, which is based on analyzing the test-channel from Figure 1, remains valid for a sigma-delta modulator with a scalar uniform quantizer of finite support. Applying such a scalar quantizer incurs a constant additive rate penalty, whose purpose is to ensure that an overload event, which jeopardizes the stability of the system, occurs with low probability. Our treatment tackles the issue of stability, which is treated rather heuristically in much of the sigma-delta literature, in a systematic and rigourous manner, and the trade-off between the rate penalty and the overload probability is analytically determined.

II. Main Result

For a discrete signal $\{c_n\}$, the Z-transform $C(Z)$ and the discrete-time Fourier transform $C(\omega)$ are defined in the usual manner. For a discrete stationary process $\{X_n\}$ with zero-mean and autocorrelation function $R_X[k] \triangleq \mathbb{E}(X_{n+k}X_n)$ we define the power-spectral density (PSD) as the Fourier transform of the autocorrelation function

$$S_X(\omega) \triangleq \sum_{k=-\infty}^{\infty} R_X[k]e^{-j\omega k}.$$

The PSD of a continuous stationary process is defined in an analogous manner.

Assume $X^{\Sigma\Delta}(t)$ is a continuous stationary band-limited Gaussian process with zero mean and variance σ^2_X, whose PSD is zero for all frequencies $|f| > f_{\text{max}}$, but is otherwise unknown. The Nyquist sampling rate for this process is $2f_{\text{max}}$ samples per second. Since our focus here is on quantization of over-sampled signals, we assume that $X^{\Sigma\Delta}(t)$ is sampled uniformly with rate of $2Lf_{\text{max}}$ samples per second for some $L > 1$. The obtained sampled process $\{X^{\Sigma\Delta}_n\}$ is therefore a discrete stationary Gaussian process with zero mean and variance σ^2_X whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$, but is otherwise unknown. Our goal is to characterize the rate-distortion trade-off obtained by a sigma-delta modulator, modeled as the test channel from Figure 1, whose input is $\{X^{\Sigma\Delta}_n\}$. To that end, we establish an equivalence between the performance obtained by this test channel for any stationary band-limited Gaussian process with variance σ^2_X and the performance obtained by the test channel from Figure 2, which models a DPCM compression system, for a stationary *flat* band-limited Gaussian process with variance σ^2_X. The performance of the latter is now well understood [2], and, as we shall show, can be translated to a simple characterization of the sigma-delta modulation performance.

The test channels in Figure 1 and Figure 2 model a sigma-delta modulator and a DPCM system, respectively, where in both systems the filter $C(Z)$ is assumed strictly causal and the quantizer was replaced by an AWGN channel. We analyze the distortions attained by the test channels and the scalar mutual information $I(U_n; U_n + N_n)$ between the input and output of the additive white Gaussian noise (AWGN) channels embedded within the two test channels. The test channels in Figure 1 and Figure 2 do not immediately induce an output distribution from which a random quantization codebook with rate $I(U_n; U_n + N_n)$ and MSE distortion D can be drawn. The reason for this is the sequential nature of the compression, which seems to conflict with the need of using high-dimensional quantizers, as required for attaining a quantization error distributed as N_n with compression rate $I(U_n; U_n+N_n)$. Fortunately, this difficulty, which is also present in decision–feedback equalization for intersymbol interference channels, can be overcome with the help of an interleaver [2]–[4] (see discussion in [5, Section II.B]). Thus, the scalar mutual information $I(U_n; U_n + N_n)$ can indeed be interpreted as the compression rate needed to achieve the distortion attained by the test channels in Figure 1 and Figure 2. Moreover, in Section III we show that $I(U_n; U_n + N_n)$ is closely related to the required quantization rate in a sigma-delta modulator that applies a *uniform scalar quantizer* of finite support.

The proofs of the following two propositions are straightforward and can be found in [5].

Proposition 1: For a Gaussian stationary process $\{X^{\Sigma\Delta}_n\}$ with variance σ^2_X whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$,

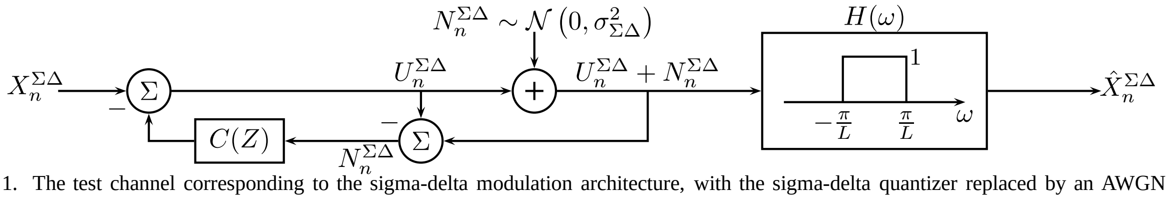
$N_n^{\Sigma\Delta} \sim \mathcal{N}\left(0, \sigma^2_{\Sigma\Delta}\right)$
$H(\omega)$
$X_n^{\Sigma\Delta}$
Σ
$-$
$U_n^{\Sigma\Delta}$
$+$
$U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}$
1
$-\frac{\pi}{L}$
$\frac{\pi}{L}$
ω
$\hat{X}_n^{\Sigma\Delta}$
$C(Z)$
$-$
Σ
$N_n^{\Sigma\Delta}$

Fig. 1. The test channel corresponding to the sigma-delta modulation architecture, with the sigma-delta quantizer replaced by an AWGN channel.

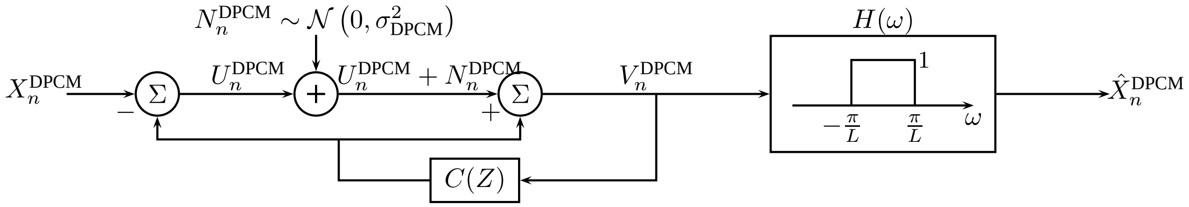

Fig. 2. The test channel corresponding to the DPCM architecture, with the DPCM quantizer replaced by an AWGN channel.

the test channel from Figure 1 achieves MSE distortion

$$D = \sigma^2_{\Sigma\Delta} \cdot \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega,$$

and its scalar mutual information satisfies[1]

$$\begin{aligned} I(&U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \\ &= \frac{1}{2} \log \left(1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega + \frac{\sigma_X^2}{\sigma^2_{\Sigma\Delta}} \right). \end{aligned}$$

Proposition 2: For a Gaussian stationary process $\{X_n^{\text{DPCM}}\}$ with variance σ_X^2 and PSD

$$S_X^{\text{DPCM}}(\omega) = \begin{cases} L\sigma_X^2 & \text{for } |\omega| \leq \pi/L \\ 0 & \text{for } \pi/L < |\omega| < \pi \end{cases}, \tag{2}$$

the test channel from Figure 2 achieves MSE distortion $D = \frac{\sigma^2_{\text{DPCM}}}{L}$ and its scalar mutual information satisfies

$$\begin{aligned} I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}}) &= \frac{1}{2} \log \Bigg(1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega \\ &+ \frac{L\sigma_X^2}{\sigma^2_{\text{DPCM}}} \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega \Bigg). \end{aligned}$$

Remark 1: In Propositions 1 and 2 we derived the *scalar* mutual information between the input and output of the AWGN test channels embedded in Figures 1 and 2, respectively. As will become clear in Section III, the scalar mutual information is closely related to the required quantization rate when a scalar memoryless quantizer is used within the sigma-delta or DPCM modulator. In [2], [4], the directed information was shown to be related to the required quantization rate when the quantizer is followed by an entropy coder. Here, we do not consider applying entropy coding to the quantizer's output as we require that the designed modulator be robust to the statistics of the input process, whereas entropy coding heavily relies on the statistics of the process. Furthermore, entropy coding is undesirable in A/D converters.

Our main result now follows immediately from Propositions 1 and 2.

Theorem 1: Let $\{X_n^{\Sigma\Delta}\}$ be a Gaussian stationary process with variance σ_X^2 whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$, let $\{X_n^{\text{DPCM}}\}$ be a Gaussian stationary process with PSD as in (2), and let $C(Z)$ be a strictly causal filter. The test channel from Figure 1 with

$$\sigma^2_{\Sigma\Delta} = \frac{D}{\frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega},$$

and the test channel from Figure 2 with $\sigma^2_{\text{DPCM}} = L \cdot D$ both achieve MSE distortion D and their scalar mutual information satisfy

$$\begin{aligned} I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) &= I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}}) \\ &= \frac{1}{2} \log \Bigg(1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega \\ &\qquad + \frac{\sigma_X^2}{D} \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega \Bigg). \end{aligned}$$

This theorem indicates that for any stationary band-limited Gaussian process with variance σ_X^2, the sigma-delta test channel from Figure 1 achieves precisely the same rate-distortion trade-off as that of the DPCM test channel from Figure 2 with a stationary flat band-limited Gaussian input with the same variance, provided that the AWGN variances are scaled according to (1). Thus, Theorem 1 provides a unified framework for analyzing the performance of sigma-delta modulation and DPCM. A great advantage afforded by such a unified framework, is that any result known for DPCM can be translated to a corresponding result for sigma-delta modulation, and vice versa. Theorems 2 and 3 below constitute two important examples of such results.

Theorem 2: Let $\{X_n^{\Sigma\Delta}\}$ be a Gaussian stationary process with variance σ_X^2 whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$ and let $\mathcal{C}$ be a family of strictly causal filters. Define the "virtual" process $\{S_n\}$ as a Gaussian stationary process with PSD as in (2), and the "virtual" process $\{W_n\}$ as a Gaussian i.i.d. random process statistically independent of $\{S_n\}$ with

[1] All logarithms in this paper are taken with base 2.

variance $L \cdot D$, $D > 0$. Let

$$\sigma_D^{*2} = \min_{C(Z)\in\mathcal{C}} \mathbb{E}\left(S_n - c_n * (S_n + W_n)\right)^2$$

$$C_D^*(Z) = \operatorname*{argmin}_{C(Z)\in\mathcal{C}} \mathbb{E}\left(S_n - c_n * (S_n + W_n)\right)^2.$$

If the filter $C(Z)$ in the sigma-delta test channel from Figure 1 belongs to $\mathcal{C}$ and the MSE distortion attained by this test channel is D, then

$$I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \geq \frac{1}{2}\log\left(1 + \frac{\sigma_D^{*2}}{L\cdot D}\right), \tag{3}$$

with equality if $C(Z) = C_D^*(Z)$.

Theorem 2 states that for a target distortion D, the sigma-delta filter which minimizes the required compression rate is the optimal linear time-invariant MMSE estimator, within the class of constraints $\mathcal{C}$, for S_n from the past of the noisy process $\{S_n + W_n\}$. For example, if $\mathcal{C}$ consists of all strictly causal finite-impulse response (FIR) filters of length p, the optimal filter $C(Z)$ is the optimal predictor of S_n from the samples $\{S_{n-1} + W_{n-1}, \ldots, S_{n-p} + W_{n-p}\}$, which can be easily calculated in closed-form.

The optimal sigma-delta filter design problem was studied by several authors, under various assumptions [6]–[9]. However, to the best of our knowledge, the simple expression from Theorem 2 for the optimal filter as the optimal predictor of S_n from the past of $\{S_n + W_n\}$ is novel.

Proof of Theorem 2. By Proposition 1, if the test channel from Figure 1 achieves MSE distortion D, we must have

$$\sigma_{\Sigma\Delta}^2 = \frac{D}{\frac{1}{2\pi}\int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega}.$$

By Theorem 1, the corresponding mutual information $I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta})$ is equal to the mutual information $I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}})$ in the DPCM test channel from Figure 2 with $X_n^{\text{DPCM}} = S_n$ and $N_n^{\text{DPCM}} = W_n$. It is shown in [2], [5] that

$$I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}}) = \frac{1}{2}\log\left(1 + \frac{\mathbb{E}(U_n^{\text{DPCM}})^2}{\sigma_{\text{DPCM}}^2}\right)$$

and that

$$U_n^{\text{DPCM}} = X_n^{\text{DPCM}} - c_n * (X_n^{\text{DPCM}} + N_n^{\text{DPCM}}).$$

Therefore, we have

$$\begin{aligned} &I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \\ &= \frac{1}{2}\log\left(1 + \frac{\mathbb{E}\left(S_n - c_n * (S_n + W_n)\right)^2}{L\cdot D}\right). \end{aligned} \tag{4}$$

It follows that among all filters in $\mathcal{C}$, the filter that minimizes (4) is $C_D^*(Z)$, and that it attains (3) with equality. ■

It is interesting to note [2] that since $\{W_n\}$ is an i.i.d. process with variance $L \cdot D$ and $C(Z)$ is strictly causal, the mutual information (4) can also be written as

$$\begin{aligned} &I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \\ &= \frac{1}{2}\log\left(\frac{\mathbb{E}\left(S_n + W_n - c_n * (S_n + W_n)\right)^2}{L\cdot D}\right). \end{aligned} \tag{5}$$

Thus, the optimal predictor of S_n from the past of $\{S_n + W_n\}$ is identical to the optimal predictor of $S_n + W_n$ from its past samples. When $C(Z)$ is taken as the (unique) infinite order optimal one-step prediction filter of $S_n + W_n$ from its past samples, the prediction error variance is the entropy power of the process $\{S_n + W_n\}$ [10], which equals

$$2^{\frac{1}{2\pi}\int_{-\pi}^{\pi}\log(L\cdot D + S_S(\omega))d\omega} = (L\cdot D)\left(1 + \frac{\sigma_X^2}{D}\right)^{1/L}. \tag{6}$$

Moreover, the infinite order prediction error

$$E_n^{\text{pred}} \triangleq S_n + W_n - c_n * (S_n + W_n)$$

is in this case a white process. This, together with (6) implies that for the optimal unconstrained sigma-delta filter $C(Z)$ we must have

$$\begin{aligned} S_{E^{\text{pred}}}(\omega) &\triangleq |1 - C(\omega)|^2 \left(L\cdot D + S_S(\omega)\right) \\ &= (L\cdot D)\left(1 + \frac{\sigma_X^2}{D}\right)^{1/L}, \ \forall\omega \in [-\pi, \pi) \end{aligned} \tag{7}$$

Combining (5), (6), and (7) yields the following theorem.

Theorem 3: Let $\{X_n^{\Sigma\Delta}\}$ be a Gaussian stationary process with variance σ_X^2 whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$. If the test channel from Figure 1 attains MSE distortion D, then

$$I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \geq \frac{1}{2L}\log\left(1 + \frac{\sigma_X^2}{D}\right). \tag{8}$$

with equality if and only if $C(Z)$ is a strictly causal filter satisfying[2]

$$|1 - C(\omega)|^2 = \begin{cases} \left(1 + \frac{\sigma_X^2}{D}\right)^{-(L-1)/L} & \omega \in [-\frac{\pi}{L}, \frac{\pi}{L}] \\ \left(1 + \frac{\sigma_X^2}{D}\right)^{1/L} & \omega \notin [-\frac{\pi}{L}, \frac{\pi}{L}], \end{cases} \tag{9}$$

and

$$\sigma_{\Sigma\Delta}^2 = \frac{D}{\frac{1}{2\pi}\int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega} = \frac{L\cdot D}{\left(1 + \frac{\sigma_X^2}{D}\right)^{-(L-1)/L}}.$$

Remark 2: The output of the test channel from Figure 1 (as well as that from Figure 2) is of the form $\hat{X}_n^{\Sigma\Delta} = X_n^{\Sigma\Delta} + E_n^{\Sigma\Delta}$, where $E_n^{\Sigma\Delta}$ has zero mean and variance D, and is statistically independent of $X_n^{\Sigma\Delta}$. This estimate can be further improved by applying scalar MMSE estimation for $X_n^{\Sigma\Delta}$ from $\hat{X}_n^{\Sigma\Delta}$. In this case the mutual information from (8) is further reduced to $\frac{1}{2L}\log\left(\frac{\sigma_X^2}{D}\right)$ which is the optimal rate-distortion function for a stationary Gaussian source $\{X_n^{\Sigma\Delta}\}$ with PSD as in (2). It follows that the sigma-delta test channel from Figure 1 with $C(Z)$ and $\sigma_{\Sigma\Delta}^2$ as specified in Theorem 3 is minimax optimal for the class of all stationary Gaussian sources with variance σ_X^2 and PSD that equals zero for all $\omega \notin [-\pi/L, \pi/L]$, i.e., no other system can achieve MSE distortion D with a smaller compression rate, universally for all sources in this class.

[2] The existence of a strictly causal filter $C(Z)$ which satisfies (9) is guaranteed by Wiener's spectral-factorization theorem.

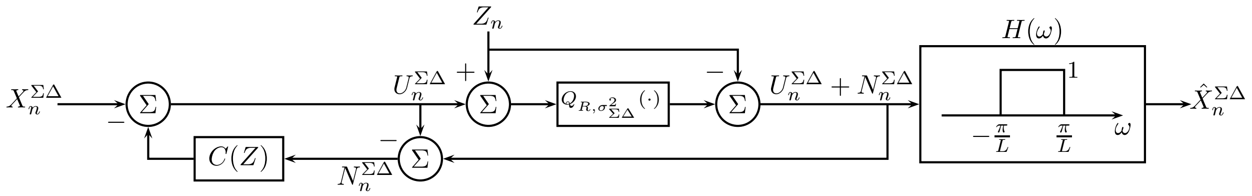

Fig. 3. A sigma-delta modulator with a dithered scalar uniform quantizer.

III. Sigma-Delta Modulation with a Scalar Uniform Quantizer

The sigma-delta modulation architecture is mainly used for A/D and D/A conversion. In such applications, vector quantization is typically out of the question and simple scalar quantizers of finite support are used instead. For such quantizers, the quantization error is composed of two main factors [1]: *granular errors* that corresponds to the quantization error in the case where the input signal falls within the quantizer's support, and *overload errors* that correspond to the case where the input signal falls outside the quantizer's support. Due to the feedback loop, inherent to the sigma-delta modulator, errors of the latter kind, whose magnitude is not bounded, may have a disastrous effect as they jeopardize the system's stability. In order to avoid such errors, the support of the quantizer has to be large enough, which translates to a constraint on the quantizer rate.

We shall show that, given that overload errors did not occur, the quantization noise can be modeled as an additive noise. Thus, the test channel from Figure 1 accurately predicts the total distortion incurred by a sigma-delta A/D (or D/A) in this case. Further, the overload probability can be controlled by taking the quantization rate greater than $I(U_n^{\Sigma\Delta};U_n^{\Sigma\Delta}+N_n^{\Sigma\Delta})$.

Let $Q_{R,\sigma^2}(\cdot)$ be a uniform mid-riser quantizer [1] with quantization step $\sqrt{12\sigma^2}$ and 2^R quantization levels, such that the quantizer support is $[-\Gamma/2,\Gamma/2)$, where $\Gamma\triangleq 2^R\sqrt{12\sigma^2}$. Our goal is to analyze the distortion and overload probability P_{ol} attained by a sigma-delta modulator that uses a $Q_{R,\sigma^2_{\Sigma\Delta}}(\cdot)$ quantizer, as a function of R and $\sigma^2_{\Sigma\Delta}$.

Clearly, if we employ the scalar sigma-delta modulator on a long enough input sequence, an overload event will eventually occur. As discussed above, the effects of overload errors can be amplified due to the feedback loop, and in this case the average MSE may significantly grow. We therefore split the input sequence into finite blocks of length N, and initialize the memory of the filter $C(Z)$ with zeros before the beginning of each new block. This makes sure that the effect of an overload error in the original system is restricted to the block where it occurs.

The analysis is made much simpler by introducing a subtractive *dither* [11]. Namely, let $\{Z_n\}$ be a sequence of i.i.d. random variables uniformly distributed over the interval $[-\sqrt{12\sigma^2_{\Sigma\Delta}}/2,\sqrt{12\sigma^2_{\Sigma\Delta}}/2)$. In order to quantize a real number U_n, we add Z_n to it before applying the quantizer, and subtract Z_n afterwards, such that the obtained result is $Q_{R,\sigma^2_{\Sigma\Delta}}(U_n+Z_n)-Z_n$.

The following theorem, whose proof can be found in [5], characterizes the trade-off between the distortion, quantization rate and overload probability achieved by the scalar sigma-delta modulator depicted in Figure 3 in terms of the scalar mutual information between the input and output of the AWGN channel from Figure 1.

Theorem 4: Let D be the MSE distortion attained by the test channel in Figure 1 with a filter $C(Z)$ of finite length, and $I(U_n^{\Sigma\Delta};U_n^{\Sigma\Delta}+N_n^{\Sigma\Delta})$ the scalar mutual information between the input and output of the AWGN channel in the same figure. For any $0<P_{ol}<1$ the scalar sigma-delta modulator from Figure 3 applied on a sequence of N consecutive source samples with quantization rate $R=I(U_n^{\Sigma\Delta};U_n^{\Sigma\Delta}+N_n^{\Sigma\Delta})+\delta(P_{ol})$ attains MSE distortion smaller than

$$\frac{D(1+o_1(N))}{1-P_{ol}},$$

with probability greater than $1-P_{ol}$, where $o_1(N)\to 0$ as N increases, and

$$\delta(P_{ol})\triangleq\frac{1}{2}\log\left(-\frac{2}{3}\ln\frac{P_{ol}}{2N}\right).$$

References

[1] N. S. Jayant and P. Noll, *Digital coding of waveforms: principles and applications to speech and video*. Prentice-Hall, 1984.

[2] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate-distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, July 2008.

[3] T. Guess and M. K. Varanasi, "An information-theoretic framework for deriving canonical decision-feedback receivers in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. IT-51, pp. 173–187, Jan 2005.

[4] J. Østergaard and R. Zamir, "Multiple-description coding by dithered delta-sigma quantization," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4661–4675, Oct 2009.

[5] O. Ordentlich and U. Erez, "Performance analysis and optimal filter design for sigma-delta modulation via duality with DPCM," 2015. [Online]. Available: http://arxiv.org/abs/1501.01829

[6] H. Spang III and P. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Tran. Comm. Systems*, vol. 10, no. 4, pp. 373–380, Dec 1962.

[7] P. Noll, "On predictive quantizing schemes," *The Bell System Technical Journal*, vol. 57, no. 5, pp. 1499–1532, May 1978.

[8] M. Derpich, E. Silva, D. Quevedo, and G. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Trans. Signal Processing*, vol. 56, no. 8, pp. 3871–3890, Aug 2008.

[9] M. Derpich and J. Østergaard, "Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3131–3152, May 2012.

[10] T. Berger, *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall, 1971.

[11] R. Zamir, *Lattice Coding for Signals and Networks*. Cambridge University Press, 2014.