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

Or Ordentlich and Uri Erez, Member, IEEE

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) converters are an integral part of almost all electric devices in use. In many cases, it is the quality of these converters that constitutes the main bottleneck in the system, and consequently, dictates its entire performance. On the other hand, as digital circuits are now considered relatively cheap to implement, the interface between the analog and digital domains is often one of the most expensive components in the system. Developing A/D and D/A components that are on the one hand relatively simple, and on the other hand introduce little distortion, is therefore of great interest.

Often, the same A/D (or D/A) component is applied to a variety of signals with distinct characterizations. For this reason, it is desired to design the data-converter to be robust to the characteristics of its input signal. One assumption that cannot be avoided is the bandwidth of the signal to be converted, 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. In much of the literature about sigma-delta modulation, no stochastic model is assumed for the input signal. However, when such a model is assumed, the benefit of over-sampling can be easily understood from basic rate-distortion theoretic principles: the (per-sample) rate required to achieve distortion $D$ for the over-sampled signal is $L$ times smaller than the rate required to achieve the same distortion for the signal obtained by sampling at the Nyquist rate. Thus, in principle, increasing the sampling rate should allow one to use quantizers with lower resolution, which is desirable in many applications.

However, the rate-distortion theoretical property that guarantees a constant product of the number of bits per sample needed to achieve distortion $D$, and the over-sampling ratio $L$, is only valid when a very long block of samples is vector-quantized. In A/D and D/A conversion, vector-quantization in high dimensions is a prohibitively complicated operation, and quantization is invariably done via scalar uniform quantizers. Scalar quantizers alone cannot translate the increase of sampling rate to a significant reduction in the necessary resolution, but fortunately this problem can be circumvented with the aid of appropriate signal processing.

In sigma-delta based converters, the quantization noise is shaped using a causal shaping filter embedded within a feedback loop, see Figure 1. The filter coefficients are chosen in a manner that ensures that most of the energy of the shaped quantization noise lies outside the frequency band occupied by the over-sampled signal. At the decoder, the quantized signal is low-pass filtered, cancelling out the high-frequencies of the quantization noise process without effecting the signal, such that the decoder’s output is composed of the original signal corrupted by a low-pass noise process.

Another technique for compressing sources with memory, which explicitly models the source as a stochastic process, 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 well-known “DPCM error identity” \[1\] states that the output of the decoder is equal to the source plus the quantization error, just like in simple non-predictive quantization. The benefit of using DPCM, however, is that the signal fed to the quantizer is the error in predicting the source from its quantized past, rather than the source itself. If the coefficients of the prediction filter are chosen appropriately, the variance of this error should be smaller than the variance of the original source, which translates to a reduction in the number of bits required from the quantizer for achieving a certain distortion.

The performance of DPCM under the assumption of high-resolution quantization is well understood since as early as the mid 60’s \[1\]–\[3\]. 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 the sigma-delta modulators. Indeed, the prediction error of such a process from its infinite past has zero-variance, rendering the DPCM high-resolution rate-distortion formulas completely useless.

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 CC Cutler \[4\], \[5\] 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. However, as discussed above, the high-resolution assumption collapses for over-sampled signals, making the standard DPCM analysis inadequate for such signals.

Fortunately, the high-resolution assumption in DPCM analysis has been overcome in \[6\], where it was shown that 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 quantized past, and in addition water-filling pre- and post-filters are applied. The analysis of \[6\], 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 exactly 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 \(\sigma_{DPCM}^2\) and prediction filter \(C(Z)\) in the test channel of Figure \[2\] the same choice of \(C(Z)\) together with the choice

\[
\sigma_{DPCM}^2 = \frac{\sigma_{DPCM}^2}{L \cdot \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega}
\]

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.

One may question the relevance of the test channel from Figure \[1\] and its information-theoretic analysis to the practical, resource limited, problem of A/D and D/A conversion via sigma-delta modulators. To that end, in Section \[III\] we replace the AWGN channel from Figure \[1\] with a simple scalar uniform (dithered) quantizer of finite support, which is suitable for implementation within A/D and D/A converters. As long as overload does not occur, the effect of applying the scalar quantizer is equivalent to that of an additive noise channel. We show that the rate-distortion trade-off derived for sigma-delta modulation in Section \[II\] remains valid with high probability, with a constant additive rate penalty for using scalar quantization. The purpose of this rate penalty is to ensure that an overload event, which jeopardizes the stability of the system, occurs with low probability. The stochastic model we assume for the input process allows us to tackle the issue of stability in a systematic and rigorous manner, and the trade-off between the rate penalty and the overload probability is analytically determined. In Section \[IV\] we demonstrate the introduced ideas via several examples.

**II. MAIN RESULT**

We begin by introducing some basic notation that will be used in the sequel. For a discrete signal \(\{c_n\}\), the \(Z\)-transform is defined as

\[
C(Z) \triangleq \sum_{n=-\infty}^{\infty} c_n Z^{-n},
\]

and the Fourier transform as

\[
C(\omega) \triangleq C(Z)|_{Z=e^{j\omega}} = \sum_{n=-\infty}^{\infty} c_n e^{-j\omega n}.
\]
For a discrete stationary process \( \{X_n\} \) with zero-mean and autocorrelation function \( R_X[k] \equiv E(X_{n+k}X_n) \) we define the power-spectral density (PSD) as the Fourier transform of the autocorrelation function

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

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 \( \sigma_X^2 \), whose PSD is zero for all frequencies \( |f| > f_{\text{max}} \), but 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 \( \{X_n^{\Sigma\Delta}\} \) is therefore a discrete stationary Gaussian process with zero mean and variance \( \sigma_X^2 \), whose PSD is zero for all \( \omega \notin [-\pi/L, \pi/L] \), but 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 \( \{X_n^{\Sigma\Delta}\} \). To do that, we establish an equivalence between the performance obtained by this test channel for any stationary band-limited Gaussian process with variance \( \sigma_X^2 \), 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 \( \sigma_X^2 \). The performance of the latter is now well understood [6]–[9], and, as we shall show, can be translated to a simple characterization of the sigma-delta modulation performance.

First, we analyze the distortions attained by the test channels from Figure 1 and Figure 2 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 [6]–[9]. 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. We elaborate further about this in subsection II-B. Moreover, in Sections III and IV 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.

We begin with the test channel in Figure 1 that corresponds to sigma-delta modulation with the sigma-delta quantizer replaced by an AWGN channel with zero mean and variance \( \sigma_{\Sigma\Delta}^2 \). The filter \( C(Z) \) is assumed strictly causal.

**Proposition 1:** For a Gaussian stationary process \( \{X_n^{\Sigma\Delta}\} \) with variance \( \sigma_X^2 \), whose PSD is zero for all \( \omega \notin [-\pi/L, \pi/L] \),
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\(^1\)
\[ I(U^\Sigma_n; U^\Sigma_n + N^\Sigma_n) = \frac{1}{2} \log \left( 1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega + \frac{\sigma^2_X}{\sigma^2_{\Sigma \Delta}} \right). \]

**Proof:** From Figure 1 we have that
\[ U^\Sigma_n = X^\Sigma_n - c_n \ast N^\Sigma_n, \]
and therefore
\[ U^\Sigma_n + N^\Sigma_n = X^\Sigma_n - c_n \ast N^\Sigma_n, \]
where \( c_n \) is the discrete identity filter. Using the fact that \( \{X^\Sigma_n\} \) is a low-pass process, passing it through the filter \( H(\omega) \) has no effect, and hence
\[ \hat{X}^\Sigma_n = h_n \ast (U^\Sigma_n + N^\Sigma_n) = X^\Sigma_n + h_n \ast (c_n \ast N^\Sigma_n). \]
The MSE distortion attained by the test channel from Figure 1 is therefore
\[ D = E(\hat{X}^\Sigma_n - X^\Sigma_n)^2 = \sigma^2_{\Sigma \Delta} \cdot \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega. \]
The scalar mutual information between the “quantizer’s” input \( U^\Sigma_n \) and output \( U^\Sigma_n + N^\Sigma_n \) is given by
\[ I(U^\Sigma_n; U^\Sigma_n + N^\Sigma_n) = h(U^\Sigma_n) + N^\Sigma_n) - h(N^\Sigma_n) \]
\[ = \frac{1}{2} \log \left( 1 + \frac{\mathbb{E}(U^\Sigma_n)^2}{\sigma^2_{\Sigma \Delta}} \right), \]
where (3), as well as (4), follow from the statistical independence of \( N^\Sigma_n \) and \( U^\Sigma_n \). Using (2), the variance of \( U^\Sigma_n \) is
\[ \mathbb{E}(U^\Sigma_n)^2 = \sigma^2_X + \frac{\sigma^2_{\Sigma \Delta} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega. }{2\pi} \]
Substituting (5) into (4) establishes the second part of the proposition.

Next, we analyze the test channel in Figure 2 that corresponds to DPCM compression with the DPCM quantizer replaced by an AWGN channel with zero mean and variance \( \sigma^2_{\text{DPCM}} \). As in the test channel of Figure 1 the filter \( C(Z) \) is strictly causal. The distortion corresponding to this test channel, as well as \( I(U^\text{DPCM}; U^\text{DPCM} + N^\text{DPCM}) \), were already found in [6] Theorem 1 for the special case where \( C(Z) \) is the optimal MMSE infinite length prediction filter of \( X^\text{DPCM} \) from all past samples of the process \( \{X^\text{DPCM} + N^\text{DPCM}\} \). The following straightforward proposition characterizes the rate and distortion for any choice of the causal filter \( C(Z) \) and any value of \( \sigma^2_{\text{DPCM}} \).

**Proposition 2:** For a Gaussian stationary process \( \{X^\text{DPCM}\} \) with variance \( \sigma^2_X \) and PSD
\[ S_X^{\text{DPCM}}(\omega) = \begin{cases} \sigma^2_X & \text{for } |\omega| \leq \pi/L \\ 0 & \text{for } \pi/L < |\omega| < \pi \end{cases} \]
the test channel from Figure 2 achieves MSE distortion
\[ D = \frac{\sigma^2_{\text{DPCM}}}{L}, \]
and its scalar mutual information satisfies
\[ I(U^\text{DPCM}; U^\text{DPCM} + N^\text{DPCM}) = \frac{1}{2} \log \left( 1 + \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega + \frac{\sigma^2_{\text{DPCM}}}{\sigma^2_{\text{DPCM}}} \right) \]
\[ + \frac{\log 2}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega. \]

**Proof:** From Figure 2 we have that
\[ U^\text{DPCM} = X^\text{DPCM} - c_n \ast V^\text{DPCM} \]
\[ V^\text{DPCM} = U^\text{DPCM} + N^\text{DPCM} + c_n \ast V^\text{DPCM} \]
Substituting (7) in (8) yields
\[ V^\text{DPCM} = X^\text{DPCM} + N^\text{DPCM}. \]
Using the fact that \( \{X^\text{DPCM}\} \) is a low-pass process, as before, we obtain
\[ \hat{X}^\text{DPCM} = h_n \ast (X^\text{DPCM} + N^\text{DPCM}) = X^\text{DPCM} + h_n \ast N^\text{DPCM}. \]
Since \( \{N^\text{DPCM}\} \) is AWGN with variance \( \sigma^2_{\text{DPCM}} \), the variance of the filtered process \( h_n \ast N^\text{DPCM} \) is \( \sigma^2_{\text{DPCM}} \). Thus,
\[ D = \mathbb{E}(\hat{X}^\text{DPCM} - X^\text{DPCM})^2 = \frac{\sigma^2_{\text{DPCM}}}{L}. \]
As in the analysis of the test channel from Figure 1 the scalar mutual information between \( U^\text{DPCM} \) and \( U^\text{DPCM} + N^\text{DPCM} \) is given by
\[ I(U^\text{DPCM}, U^\text{DPCM} + N^\text{DPCM}) = \frac{1}{2} \log \left( 1 + \frac{\mathbb{E}(U^\text{DPCM})^2}{\sigma^2_{\text{DPCM}}} \right). \]
Now, substituting (9) in (7) gives
\[ U^\text{DPCM} = (\delta_n - c_n) \ast X^\text{DPCM} - c_n \ast N^\text{DPCM}, \]
and the variance of \( U_n \) is therefore
\[ \mathbb{E}(U^\text{DPCM})^2 = \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} S_X^{\text{DPCM}}(\omega) |1 - C(\omega)|^2 d\omega \]
\[ + \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} S_N^{\text{DPCM}}(\omega) C(\omega)^2 d\omega \]
\[ = \frac{L \sigma^2_X}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega + \frac{\sigma^2_{\text{DPCM}}}{2\pi} \int_{-\pi/L}^{\pi/L} |C(\omega)|^2 d\omega. \]
Substituting (12) into (11) establishes the second part of the proposition.

**Remark 1:** In propositions 1 and 2 we derived the scalar mutual information between the input and output of the

---

\(^1\) All logarithms in this paper are taken with base 2.
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 (6), 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
is very sensitive to the process statistics. Moreover, if the design of an A/D (or D/A) is considered, the appropriate merit
for the modulator’s complexity is the number of quantization levels within the scalar quantizer, which are not reduced by
incorporating an entropy coder.

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 \( \sigma_X^2 \) whose PSD is zero for all \( \omega \in \{-\pi/L, \pi/L\} \), let \( \{X_n^{DPCM}\} \) be a Gaussian stationary process with PSD as in (6), and let \( C(Z) \) be a strictly causal filter. The test channel from Figure 1 with

\[
\sigma_{\Sigma \Delta}^2 = \frac{D}{4} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega,
\]

and the test channel from Figure 2 with

\[
\sigma_{DPCM}^2 = L \cdot D,
\]

both achieve MSE distortion \( D \) and their scalar mutual information satisfy

\[
I(U_n^{\Sigma \Delta}; U_n^{\Sigma \Delta} + N_n^{\Sigma \Delta}) = I(U_n^{DPCM}; U_n^{DPCM} + N_n^{DPCM})
\]

\[
= \frac{1}{2} \log \left( 1 + \frac{1}{\pi \sigma_X^2} \int_{-\pi/L}^{\pi/L} |C(\omega)|^2 d\omega \right.
\]

\[
+ \frac{\sigma_X^2}{D} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega \right).
\]

This theorem indicates that for any stationary band-limited Gaussian process with variance \( \sigma_X^2 \), the sigma-delta test channel from Figure 1 achieves exactly 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 offered 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 \( \sigma_X^2 \) whose PSD is zero for all \( \omega \in \{-\pi/L, \pi/L\} \) and let \( C \) be a family of strictly causal filters. Define the “virtual” process \( \{S_n\} \) as a Gaussian stationary process with PSD as in (6), and the “virtual” process \( \{W_n\} \) as a Gaussian i.i.d. random process statistically independent of \( \{S_n\} \) with variance \( L \cdot D, D > 0 \). Let

\[
\sigma_B^2 = \min_{C(Z) \in \mathcal{C}} \mathbb{E}(S_n - c_n * (S_n + W_n))^2
\]

\[
C_Z^*(Z) = \arg \min_{C(Z) \in \mathcal{C}} \mathbb{E}(S_n - c_n * (S_n + W_n))^2.
\]

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

\[
I(U_n^{\Sigma \Delta}; U_n^{\Sigma \Delta} + N_n^{\Sigma \Delta}) \geq \frac{1}{2} \log \left( 1 + \frac{\sigma_B^2}{L \cdot D} \right),
\]

with equality if \( C(Z) = C_Z^*(Z) \).

**Theorem 3:** states that for a target distortion \( D \), the sigma-delta filter which minimizes the required compression rate is the optimal linear time-invariant MSE estimator, within the class of constraints \( C \), for \( S_n \) from the past of the noisy process \( \{S_n + W_n\} \). For example, if \( C \) 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 (1, 9–14). 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. The references most relevant to Theorem 2 are perhaps (12) and (13, 14).

In (12), Spang and Schultheiss formulated an optimization problem for finding the best FIR filter with \( p \) coefficients in a sigma-delta modulator with a scalar quantizer, under a fixed overload probability. Their optimization problem can be solved numerically, but no closed form solution was given. In (13) and (14) the design of an optimal unconstrained sigma-delta filter was studied, under the assumption of a fixed scalar quantizer which can only be scaled in order to control the overload probability. Equations that characterize the optimal filter were derived. However, the obtained expressions usually yield filters with an infinite number of taps, and do not provide the solution to the constrained problem. It is also worth mentioning that for the case of a stationary Gaussian process \( \{X_n\} \) with \( L = 1 \) (sampling at the Nyquist rate) and known PSD, the optimal infinite length filter under the assumption of high-resolution quantization is known to equal the optimal prediction filter of \( X_n \) from its (clean) past (10). As already mentioned in the introduction, the high-resolution assumption never holds when \( L > 1 \) and therefore this result is inapplicable for over-sampled signals.

**Proof of Theorem 2:** By Proposition 1 if the test channel from Figure 1 achieves MSE distortion \( D \), we must have

\[
\sigma_{\Sigma \Delta}^2 = \frac{D}{4} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega.
\]

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.

...
The process $\mathbf{S}$, the prediction error variance is the entropy power of $C$ satisfying

Thus, the optimal predictor of $C$ is in this case a white process. This, together with (16) implies

Moreover, the infinite order prediction error 

Thus, 

It is interesting to note that since $\{W_n\}$ is an i.i.d. process with variance $L \cdot D$ and $C(Z)$ is strictly causal, the mutual information (14) can also be written as 

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 $\{S_n + W_n\}$ [15], which equals

Moreover, the infinite order prediction error 

is in this case a white process. This, together with (16) implies that for the optimal unconstrained sigma-delta filter $C(Z)$ we must have

Combining (15), (16), and (17) yields the following theorem.

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

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

and

Remark 2: Note that the existence of a strictly causal filter $C(Z)$ which satisfies (19) is guaranteed by Wiener’s spectral-factorization theory [15] due to the readily verified fact that

The optimal filter induces a two-level frequency response for $|1 - C(\omega)|^2$. In [8] Østergaard and Zamir used sigma-delta modulation to attain the optimal multiple-description rate-distortion region. Interestingly, the optimal filter $C(Z)$ in their scheme also induced a two-level response for $|1 - C(\omega)|^2$.

Remark 3: 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 $\hat{X}_n^{\Sigma \Delta}$. This boils down to producing the estimate $\tilde{X}_n^{\Sigma \Delta} = \alpha \hat{X}_n^{\Sigma \Delta}$, where

Consequently, the obtained MSE distortion is reduced to

It is straightforward to verify (18) that with this improvement, the sigma-delta test channel from Figure 1 with $C(Z)$ and $\sigma_{\Sigma \Delta}^2$ as specified in Theorem 3 attains

which is the optimal rate-distortion function for a stationary Gaussian source $\{X_n^{\Sigma \Delta}\}$ with PSD as in (18). 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 $\sigma_X^2$ 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.

A. Extension to Frequency-Weighted Mean Squared Error Distortion

In many applications, higher values of distortion are acceptable in certain frequency bands while smaller distortion is permitted in other bands. The MSE distortion measure is inadequate for such scenarios, and a commonly used distortion measure, that (partially) captures such perceptual effects, is the frequency-weighted mean squared error (FWMSE) criterion. Under this criterion, the distortion is measured as

where

and


where $P(\omega)$ is a non-negative weight function, and $S_E(\omega)$ is the PSD of the error process $E_n \triangleq X_n^{\Sigma_\Delta} - \hat{X}_n^{\Sigma_\Delta}$. Note that for $P(\omega) = 1, \forall \omega \in [-\pi, \pi]$ the FWMSE criterion reduces to the MSE one. The next theorem shows that the constrained optimal sigma-delta filter under the FWMSE criterion is the optimal constrained prediction filter of a noisy process defined according to the weight function $P(\omega)$.

**Theorem 4:** Let $\{X_n^{\Sigma_\Delta}\}$ be a Gaussian stationary process with variance $\sigma_X^{2}$ whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$, $P(\omega)$ a weighting function which forms a valid PSD, and $C$ a family of strictly causal filters. Define the “virtual” process $\{S_n\}$ as a Gaussian stationary process with PSD

$$S_X^{\text{FWMSE}}(\omega) = \begin{cases} L \sigma_X^2 P(\omega) & \text{for } |\omega| \leq \pi/L \\ 0 & \text{for } \pi/L < |\omega| < \pi \end{cases}, \quad (21)$$

and the “virtual” process $\{W_n\}$ as a Gaussian i.i.d. random process statistically independent of $\{S_n\}$ with variance $L \cdot D_{\text{FWMSE}}^{\Sigma_\Delta} > 0$. Let

$$\sigma_{\text{FWMSE}}^2 = \min_{C(\cdot) \in C} \mathbb{E} (S_n - c_n*(S_n + W_n))^2$$

$$C^*_d^{\text{FWMSE}}(Z) = \arg \min_{C(\cdot) \in C} \mathbb{E} (S_n - c_n*(S_n + W_n))^2.$$ 

If the filter $C(Z)$ in the sigma-delta test channel from Figure 1 belongs to $C$ and the FWMSE distortion w.r.t. $P(\omega)$ attained by this test channel is $D_{\text{FWMSE}}^{\Sigma_\Delta}$, then

$$I(U_n^{\Sigma_\Delta}; U_n^{\Sigma_\Delta} + N_n^{\Sigma_\Delta}) \geq \frac{1}{2} \log \left( 1 + \frac{\sigma_{\text{FWMSE}}^2}{L \cdot D_{\text{FWMSE}}^{\Sigma_\Delta}} \right),$$

with equality if $C(Z) = C^*_d^{\text{FWMSE}}(Z)$.

**Sketch of proof:** The proof is fairly similar to that of Theorem 2. Thus, for brevity, we omit the full proof and only highlight its main steps:

- Repeat the derivation of Proposition 1 where now the MSE distortion is replaced by FWMSE distortion. Note that this has no effect on $I(U_n^{\Sigma_\Delta}; U_n^{\Sigma_\Delta} + N_n^{\Sigma_\Delta})$.

- Repeat the derivation of Proposition 2 where the PSD of the input process is $\{\Sigma_\Delta\}$, rather than $\{\Sigma\}$. Note that this changes $I(U_n^{\Sigma_\Delta}; U_n^{\Sigma_\Delta} + N_n^{\Sigma_\Delta})$, but has no effect on the attained distortion.

- It follows that the DPCM test channel for the process $\{S_n\}$ under MSE distortion is equivalent to the sigma-delta test channel with input $\{X_n^{\Sigma_\Delta}\}$ under FWMSE distortion, in the sense that in both channels if the attained distortion is $D_{\text{FWMSE}}^{\Sigma_\Delta}$ (under the appropriate distortion measure), then

$$I(U_n^{\Sigma_\Delta}; U_n^{\Sigma_\Delta} + N_n^{\Sigma_\Delta}) = I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}}) \quad \frac{1}{2} \log \left( 1 + \frac{\mathbb{E} (S_n - c_n*(S_n + W_n))^2}{L \cdot D_{\text{FWMSE}}^{\Sigma_\Delta}} \right).$$

### III. Sigma-Delta Modulation with a Scalar Uniform Quantizer

The goal of this short subsection is to give the test channel from Figure 1 an operational meaning, i.e., to show how the AWGN from the figure can be replaced with a lossy source code of rate $R = I(U_n^{\Sigma_\Delta}; U_n^{\Sigma_\Delta} + N_n^{\Sigma_\Delta})$ whose incurred quantization noise is distributed as $N_n^{\Sigma_\Delta}$. As already mentioned, the key idea is to use an interleaver [6]–[8], as we now recall.

Assume that $\{X_n^{\Sigma_\Delta}\}$, the input process to the sigma-delta modulator, has a decaying memory, such that $X_n^{\Sigma_\Delta}$ is essentially independent of all samples of sufficiently distant sampling times. In order to compress an $N$-dimensional vector

$$x^{\Sigma_\Delta} = [X_1^{\Sigma_\Delta}, \ldots, X_N^{\Sigma_\Delta}],$$

containing $N$ consecutive samples of the process $\{X_n^{\Sigma_\Delta}\}$, we first split it into $K$ vectors

$$x_k^{\Sigma_\Delta} = [X_{(k-1)M+1}^{\Sigma_\Delta}, \ldots, X_{kM}^{\Sigma_\Delta}], \quad k = 1, \ldots, K,$$

where $M \triangleq N/K$. Now, we can apply $K$ parallel sigma-delta quantizers, one for each such vector, where the only coupling between the $K$ parallel systems is through the quantization step, which is applied jointly on all of them, as depicted in Figure 3. By our assumption that $\{X_n^{\Sigma_\Delta}\}$ has decaying memory, if $M$ is large enough the $K$ inputs that enter the quantizer $Q(\cdot) = [Q_1(\cdot), \ldots, Q_K(\cdot)]$ are i.i.d. random variables distributed as $U_n^{\Sigma_\Delta}$ from Figure 1. For large enough $K$, standard rate-distortion arguments imply that there exists a vector quantizer with rate $I(U_n^{\Sigma_\Delta}; U_n^{\Sigma_\Delta} + N_n^{\Sigma_\Delta})$ that induces quantization noise distributed as $N_n^{\Sigma_\Delta}$.

### B. Sigma-Delta Modulation with an Interleaved Vector Quantizer

The previous subsection showed how to replace the AWGN channel in Figure 1 with a vector quantizer whose rate is arbitrarily close to $R = I(U_n^{\Sigma_\Delta}; U_n^{\Sigma_\Delta} + N_n^{\Sigma_\Delta})$ and whose induced quantization noise is distributed as $N_n^{\Sigma_\Delta}$. The inputs to the vector quantizer are vectors of i.i.d. Gaussian components. Thus, any “off-the-shelf” rate–distortion optimal vector quantizer for an i.i.d. Gaussian source can be used. The total sigma-delta compression system that is obtained is therefore simple in the sense that it only requires the vector quantizer to be good for quantizing an i.i.d. Gaussian source, which is a standard task, rather than requiring it to be a good quantizer for a band-limited Gaussian source.

However, 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. 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 from 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 chosen appropriately. As the support of the quantizer determines its rate for a given quantization resolution, the overload probability can be controlled by increasing the quantization 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. Moreover, the overload probability is a doubly exponential decreasing function of the difference \( R - I(U^\Sigma_\Delta; U^\Sigma_\Delta + N^\Sigma_\Delta) \), where \( 2^R \) are the number of levels in the scalar quantizer. Thus, fixing the desired overload error probability as \( P_{ol} \), we may achieve the MSE distortion predicted by the test channel from Figure 1 (characterized in Proposition 1) with a scalar quantizer whose rate is \( I(U^\Sigma_\Delta; U^\Sigma_\Delta + N^\Sigma_\Delta) + \delta(P_{ol}) \), where \( \delta(P_{ol}) = \frac{1}{2} \log \left( -\frac{4}{3} \ln P_{ol} \right) \).

Let \( Q_{R,\sigma^2}(\cdot) \) be a uniform quantizer with quantization step \( \sqrt{12\sigma^2} \) and \( 2^R \) quantization levels, such that the quantizer support is \([-\Gamma/2, \Gamma/2]\), where \( \Gamma \triangleq 2^R \sqrt{12\sigma^2} \), see Figure 2.

Our goal is to analyze the distortion and overload probability attained by a sigma-delta modulator that uses a \( Q_{R,\sigma^2}(\cdot) \) quantizer, as a function of \( R \) and \( \sigma^2 \).

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 \( 16 \). 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 \( Q_{R,\sigma^2}(U_n + Z_n) - Z_n \). Adding and subtracting \( U_n \), we get

\[
U_n + \left( Q_{R,\sigma^2}(U_n + Z_n) - (U_n + Z_n) \right),
\]

and the quantization error is therefore

\[
N_n \triangleq Q_{R,\sigma^2}(U_n + Z_n) - (U_n + Z_n)
\]

(22)

The main result in this section is the following.

**Theorem 5:** Let \( D \) be the MSE distortion attained by the test channel in Figure 1 with a filter \( C(Z) \) of finite length, and \( I(U^\Sigma_\Delta; U^\Sigma_\Delta + N^\Sigma_\Delta) \) 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 5 applied on a sequence of \( N \) consecutive source samples with quantization rate \( R = I(U^\Sigma_\Delta; U^\Sigma_\Delta + N^\Sigma_\Delta) \), where \( \sigma^2 \) is the variance of the input, and \( P_{ol} \) the overload probability, we have

\[
D(1 + o_1(N))
\]

with overload probability smaller than \( P_{ol} \), where \( o_1(N) \to 0 \) as \( N \) increases, and

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

(23)

**Proof:** Let \( \tilde{Q}_{\sqrt{12\sigma^2}\Sigma_\Delta}(x) \) be the operation of rounding \( x \) to the nearest point in the (infinite) lattice \( \sqrt{12\sigma^2}\Sigma_\Delta \). It is easy to verify that for any \( x \in [-\Gamma/2, \Gamma/2] \) we have

\[
Q_{R,\sigma^2}(x) = \tilde{Q}_{\sqrt{12\sigma^2}\Sigma_\Delta} \left( x + \frac{\sqrt{12\sigma^2}\Sigma_\Delta}{2} \right) - \frac{\sqrt{12\sigma^2}\Sigma_\Delta}{2}.
\]

(24)
Applying (22) therefore yields that if overload did not occur in the \( n \)th sample, i.e., if \(|U_n + Z_n| \leq \Gamma/2\), we have
\[
N_n = \sqrt[2]{\sigma_{\Sigma\Delta}^2} n \left( U_n + Z_n + \sqrt{3\sigma_{\Sigma\Delta}^2} \right) - \left( U_n + Z_n + \sqrt{3\sigma_{\Sigma\Delta}^2} \right) .
\] (25)

Dealing with the overload event of the quantizer directly is rather involved. Instead, as done in [17], we first consider a reference system with an infinite-support quantizer and analyze its performance. If the magnitude of the input to the infinite-support quantizer never exceeds \( \Gamma/2 \) within the processed block, then clearly the reference system is completely equivalent to the original system within this block. Thus, it suffices to find the average distortion of the reference system and the probability that the input to its quantizer exceeds \( \Gamma/2 \) within a block. In what follows we will therefore assume that the quantization noise is given by (25) regardless of whether or not \(|U_n + Z_n| \leq \Gamma/2\), and account for the overload probability later.

Assuming that the dither sequence \( \{Z_n\} \) is drawn statistically independent of the process \( \{U_n\} \), the Crypto Lemma, see, e.g. [16] Lemma 4.1.1, implies that \( \{N_n\} \) is an i.i.d. sequence of random variables uniformly distributed over the interval \([-\sqrt{12\sigma_{\Sigma\Delta}^2}/2, \sqrt{12\sigma_{\Sigma\Delta}^2}/2]\), statistically independent of \( \{U_n\} \). Note that \( N_n \) has zero mean and variance \( \sigma_{\Sigma\Delta}^2 \). Following this reasoning, the reference sigma-delta data converter depicted in Figure [5] (with an infinite-support quantizer) is equivalent to the test channel from Figure [11] with \( N_{\Sigma\Delta} \sim \text{Uniform} \left( -\sqrt{12\sigma_{\Sigma\Delta}^2}/2, \sqrt{12\sigma_{\Sigma\Delta}^2}/2 \right) \) instead of \( N_{\Sigma\Delta} \sim N(0, \sigma_{\Sigma\Delta}^2) \). Thus, the average MSE distortion attained by the reference scalar sigma-delta modulator from Figure [5] is as given in Proposition [11] up to a multiplicative factor of \( 1 + o_1(N) \) that accounts for edge effects. These effects are the by-product of the operation of nulling the filter memory at the beginning of each new block, which incurs temporal non-stationarities. In particular, if the filter \( C(Z) \) has \( L \) taps, then only after \( L \) samples within the block the statistics of the process \( \{U_n\} \) will converge to its stationary distribution. However, if the block length is sufficiently large w.r.t. the filter length and the inverse of the MSE distortion, the influence of these effects vanishes.

Next, we turn to analyze the probability that an overload event occurs within a block of length \( N \), as a function of \( R \) and \( 1(U_n^\Sigma\Delta; N_n^\Sigma\Delta) \). Since this event is equivalent to the event that at the reference system the input to the quantizer exceeds \( \Gamma/2 \) in magnitude within the block, it suffices to lower bound the probability of the latter event.

Assume the reference scalar sigma-delta modulator from Figure [5] is applied on a vector \( x^{\Sigma\Delta} = [X_1^{\Sigma\Delta}, \ldots, X_N^{\Sigma\Delta}] \) of \( N \) consecutive samples of the process \( \{X_n^{\Sigma\Delta}\} \), where the memory of the filter \( C(Z) \) is initialized with zeros. Define the event \( \text{OL}_k = \{ |U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta}| > \Gamma/2 \} \) and the event \( \text{OL} = \cup_k \text{OL}_k \). By the union bound, we have
\[
Pr(\text{OL}_k) \leq \sum_{k=1}^N Pr(\text{OL}_k) .
\] (26)

The random variable \( U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta} = X_k^{\Sigma\Delta} + (\delta_k - c_k) \) is a linear combination of a Gaussian random variable \( X_k^{\Sigma\Delta} \) and statistically independent uniform random variables \( \{N_k^{\Sigma\Delta}\} \). In [18] Lemma 4] the probability that a random variable of this type exceeds a certain threshold was bounded in terms of its variance. Applying this bound to \( U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta} \) yields
\[
Pr \left( |U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta}| > \Gamma/2 \right) \leq 2 \exp\left\{ -\frac{\Gamma^2}{8 \left( \langle E|U_k^{\Sigma\Delta}|^2 \rangle + \langle E|N_k^{\Sigma\Delta}|^2 \rangle \right)} \right\} .
\]

where in the last equality we have used the definition of \( \Gamma \) and the fact that \( U_k^{\Sigma\Delta} \) and \( N_k^{\Sigma\Delta} \) are statistically independent. Equivalently, we may write
\[
Pr(\text{OL}_k) \leq \exp\left\{ -\frac{12 \sigma_{\Sigma\Delta}^2 2^{2R}}{8 \sigma_{\Sigma\Delta}^2 \left( 1 + \langle E|U_k^{\Sigma\Delta}|^2 \rangle / \sigma_{\Sigma\Delta}^2 \right)} \right\} ,
\]
\[
= 2 \exp\left\{ -\frac{3}{2} \left( R - I(U_k^{\Sigma\Delta}; x_k^{\Sigma\Delta}) \right) \right\} ,
\] (27)
where we have used (1) in the last equality. Substituting (27) into (26) gives

\[
  P_{ol} \leq 2 \sum_{k=1}^{N} \exp \left\{ -\frac{3}{2} \sigma^2 \left( R - 1 - (U_n^\Sigma + N_n^\Sigma) \right) \right\}.
\]

(28)

Note that \( E(U_n^\Sigma)^2 = \sigma_X^2 + \sigma_{\Sigma \Delta}^2 \sum_{k=1}^{k} c_k^2 \) is monotonically nondecreasing in \( k \) and is given by (3) for values of \( k \) that are greater than the length of the filter \( c_k \). We can therefore further bound (28) as

\[
  P_{ol} \leq 2N \exp \left\{ -\frac{3}{2} \sigma^2 \left( R - 1 - (U_n^\Sigma + N_n^\Sigma) \right) \right\}.
\]

(29)

where \( I(U_n^\Sigma; U_n^\Sigma + N_n^\Sigma) \) is as given in Proposition 1. To summarize, we have shown that the reference system achieves the same MSE distortion \( D \) as characterized by Proposition 1 up to a \( 1 + o_1(N) \) multiplicative term, and the probability that one of the inputs to its quantizer exceeds \( \Gamma / 2 \) in magnitude within a block of length \( N \), is bounded by (29).

For our original system whose quantizer has finite support of \( -\Gamma / 2, \Gamma / 2 \), this means that the overload probability is also upper bounded by the RHS of (29). Moreover, the average distortion it achieves if overload did not occur is the same as that of the reference system conditioned on the event that OL did not occur. Denote this conditioned expected distortion by \( D_{ol} \) and the expected distortion conditioned on the event that OL did occur by \( D_{ol} \). For the reference system, we have

\[
  D(1 + o_1(N)) = \Pr(OL)D_{ol} + \Pr(OL)D_{ol} \geq \Pr(OL)D_{ol},
\]

and therefore

\[
  D_{ol} \leq \frac{D(1 + o_1(N))}{1 - P_{ol}}.
\]

This shows that the scalar sigma-delta system from Figure 5 whose quantizer has limited support \( -\Gamma / 2, \Gamma / 2 \), with \( R = I(U_n^\Sigma; U_n^\Sigma + N_n^\Sigma) + \delta \) achieves the same average MSE distortion as the test channel from Figure 1 up to a multiplicative factor of \( 1 + o_1(N) / (1 - P_{ol}) \), with block error probability smaller than \( 2N \exp \left\{ -\frac{3}{2} \sigma^2 \left( R - 1 - P_{ol} \right) \right\} \). Thus, Proposition 1 characterizes the rate-distortion tradeoff achieved by the scalar sigma-delta system up to the aforementioned factor and a constant rate penalty \( \delta(P_{ol}) \), that depends on the target overload error probability. To be more precise, for any \( 0 < P_{ol} < 1 \), taking the rate penalty as in (23) guarantees that the overload error probability is smaller than \( P_{ol} \).

IV. NUMERICAL EXAMPLES

In this section we demonstrate the performance obtained by the scalar sigma-delta modulator introduced in the previous subsection. More specifically, we consider the scalar sigma-delta modulator from Figure 5 with an FIR feedback filter of length \( p = 8 \) taps. The input to the modulator is a stationary bandlimited Gaussian process with variance 1 that was sampled at twice the Nyquist rate \( (L = 2) \). We consider two different models for the input process:

1) When sampled at the Nyquist rate, the input \( \{X_n^{(1)}\} \) is a discrete white Gaussian process with variance \( \sigma^2_X = 1 \).

2) When sampled at the Nyquist rate, the input \( \{X_n^{(2)}\} \) is a first order autoregressive process with parameter \( \alpha = 0.9 \) and variance \( \sigma^2_X = 1 \), i.e., \( X_n^{(2)} = \alpha X_{n-1}^{(2)} + Z_n \), where \( Z_n \sim N(0,1 - \alpha^2) \).

As evident from our analysis in the preceding sections, the performance guarantees for the A/D from Figure 5 do not vary between the two different models, as in both models the input process has the same variance and bandwidth.

The input to the A/D is processed in blocks of length \( N = 216 = 65,536 \) consecutive samples. We consider different quantization rates for \( Q_{RD_\Sigma \Delta}(\cdot) \), varying from \( R = 3 \) bits per sample to \( R = 7 \) bit per sample, whereas the corresponding value of \( \sigma^2_{\Sigma \Delta} \) and the optimal feedback filter \( C(Z) \) are chosen according to the target overload probability, as described below.

For a target overload probability \( P_{ol} \), set \( \delta(P_{ol}) \) as in (23) and \( R' = R - \delta(P_{ol}) \). By Theorem 3 if \( R' \) is taken to be greater than the mutual information \( I(U_n^\Sigma; U_n^\Sigma + N_n^\Sigma) \) between the input and output of the AWGN channel in Figure 1 the overload probability should not exceed \( P_{ol} \). It therefore remains to choose the values of \( \sigma^2_{\Sigma \Delta} \) and \( C(Z) \) that minimize the MSE distortion under the constraint that the induced \( I(U_n^\Sigma; U_n^\Sigma + N_n^\Sigma) \) in Figure 1 is smaller than \( R' \). To this end, we invoke Theorem 2.

Our class of filters \( C \) consists of all strictly causal FIR filters with \( p = 8 \) taps. Let \( \{S_n\} \) and \( \{W_n\} \) be the virtual processes defined in Theorem 2 with \( \sigma^2_X = 1 \) and some \( D > 0 \), and let \( Y_n = [S_{n-1} + W_{n-1}, \ldots, S_{n-p} + W_{n-p}] \). For any choice of \( D > 0 \) in Theorem 2 the optimal filter \( C_D^*(Z) \in C \) can be expressed in closed form as

\[
  c_D^* = R_{SY} R_{YY}^{-1}
\]

where the entries of \( c_D^* = [c_D^*(1), \ldots, c_D^*(p)] \) are the coefficients of \( C_D^*(Z) \), and

\[
  R_{SY} \triangleq E(S_n Y_n) = \left[ \begin{array}{c} \sin \left( \frac{1}{L} \right) \\ \sin \left( \frac{2}{L} \right) \end{array} \right],
\]

and

\[
  R_{YY} \triangleq E(Y_n^T Y_n) = \left[ \begin{array}{ccc} LD + 1 & \sin \left( \frac{1}{L} \right) & \cdots & \sin \left( \frac{L-1}{L} \right) \\ \sin \left( \frac{1}{L} \right) & LD + 1 & \sin \left( \frac{1}{L} \right) & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \sin \left( \frac{L-1}{L} \right) & \sin \left( \frac{L-2}{L} \right) & \cdots & LD + 1 \end{array} \right].
\]

The corresponding value of \( \sigma_{\Sigma \Delta}^2 \) is given by

\[
  \sigma_{\Sigma \Delta}^2 = 1 - R_{SY} R_{YY}^{-1} R_{SY}^T,
\]

and the induced mutual information is

\[
  I(U_n^\Sigma; U_n^\Sigma + N_n^\Sigma) = \frac{1}{2} \log \left( 1 + \frac{\sigma_{\Sigma \Delta}^2}{L \cdot D} \right).
\]

(30)
becomes very small and the edge effects discussed in Sec-

Table 7a, where the target overload probability we chose was

did not occur. The empirical overload probabilities are given in

the empirical MSE distortion attained by the scalar sigma-delta

modulator in the setting described above, given that overload

MSE distortion is very close to

Fig. 6. Simulation results for the scalar sigma-delta modulator with over-sampling ratio of \( L = 2 \) the Nyquist rate, and an 8-tap FIR feedback filter. The
distortion \( D(R') \) is plotted in solid line, the empirical distortion attained by a system that uses subtractive dithering (given that overload did not occur) in
dashed line, and the empirical distortion attained by a system that does not employ dithering (given that overload did not occur) in dotted line.

(a) Empirical MSE distortion for \( \{X_n^{(1)}\} \)

(b) Empirical MSE distortion for \( \{X_n^{(2)}\} \)

Now, we can find the minimal distortion \( D(R') \) for which the

RHS of (30) is greater than or equal to \( R' \), and set

\[
\sigma_{\Delta}^2 \left( R' \right) \Delta \equiv \frac{D(R')}{\int_{\pi/L}^{\pi/L} [1 - C_D(R')'(\omega)]^2 d\omega}.
\]

Taking \( \sigma_{\Delta}^2 \left( R' \right) \) and \( C_D(R')'(Z) \) as above ensures that the

MSE distortion incurred by the test channel in Figure 1 is

\( D(R') \). Thus, by Theorem 5 the MSE distortion incurred by the

scalar sigma-delta modulator from Figure 5 is upper

bounded by \( D(R')(1 + o_1(N))/(1 - P_{ol}) \).

In Figure 6a we consider the input process \( \{X_n^{(1)}\} \) and show the empirical MSE distortion attained by the scalar sigma-delta

modulator in the setting described above, given that overload
did not occur. The empirical overload probabilities are given in

Table 7a where the target overload probability we chose was

\( P_{ol} = 0.1 \), and the number of blocks processed for each rate

was 1000. It is evident that the empirical overload probability

is indeed smaller than the target \( P_{ol} \) and that the empirical

MSE distortion is very close to \( D(R') \). For high rates, \( D(R') \)

becomes very small and the edge effects discussed in Sec-

tion 11 are more dominant. However, since these edge effects

only increase the distortion in the estimates of samples near the

edges of the block, their influence on the total distortion

can be made negligible if the block length is increased. Note

that as the excess-rate \( \delta(P_{ol}) \) grows with \( N \) at a very slow rate

of \( \frac{1}{2} \log \ln(N) \), there will be a negligible penalty in rate for

increasing the block length. We also considered the effect of

the subtractive dither in the simulations. Specifically, we have

also simulated the same scalar sigma-delta modulator without

incorporating dithers. The input to the quantizer \( Q_{R',\sigma_{\Delta}^2} \) in

this case is simply \( u_n^{(\Delta)} \) and no dither is subtracted from its

output. It is evident from Figures 6b and 7b that the dither has no

significant effect on the performance of the system. Figures 6b

and 7b show the results of the same experiment where now the

input process is \( \{X_n^{(2)}\} \). Indeed, the results for the two

processes are indistinguishable.

Fig. 7. The empirical overload probability attained by a system that uses

subtractive dithering and the empirical overload probability attained by a

system that does not employ dithering. The target overload probability was

\( P_{ol} = 0.1 \).

(b) Empirical overload probability for \( \{X_n^{(2)}\} \)

(a) Empirical overload probability for \( \{X_n^{(1)}\} \)

References

[1] N. S. Jayant and P. Noll, Digital coding of waveforms: principles and

applications to speech and video. Englewood Cliffs, NJ: Prentice-Hall,

1984.

[2] R. A. McDonald, “Signal-to-noise and idle channel performance of
differential pulse code modulation systems particular applications to

voice signals,” Bell System Technical Journal, vol. 45, no. 7, pp. 1125–

1151, 1966.

[3] R. Gray and D. Neuhoff, “Quantization,” IEEE Transactions on Infor-

mation Theory, vol. 44, no. 6, pp. 2325–2383, Oct 1998.

[4] C. Cutler, “Differential quantization of communication signals,” 1952,

US Patent 2,605,361.

[5] ———, “Transmission systems employing quantization,” 1960, US Patent

2,927,962 (filed 1954).

[6] R. Zamir, Y. Kochman, and U. Erez, “Achieving the Gaussian rate-
distortion function by prediction,” IEEE Transactions on Information

Theory, vol. 54, no. 7, pp. 3354–3364, July 2008.

[7] T. Guess and M. K. Varanasi, “An information-theoretic framework for
deriving canonical decision-feedback receivers in Gaussian channels,”

IEEE Transactions on Information Theory, vol. IT-51, pp. 173–187, Jan

2005.

[8] J. Østergaard and R. Zamir, “Multiple-description coding by dithered

delta-sigma quantization,” IEEE Transactions on Information Theory,

vol. 55, no. 10, pp. 4661–4675, Oct 2009.
[9] S. Tewksbury and R. Hallock, “Oversampled, linear predictive and noise-shaping coders of order \( n > 1 \),” IEEE Transactions on Circuits and Systems, vol. 25, no. 7, pp. 436–447, Jul 1978.

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

[11] M. A. Gerzon and P. G. Craven, “Optimal noise shaping and dither of digital signals,” in Audio Engineering Society Convention 87, 1989.

[12] H. Spang III and P. Schultheiss, “Reduction of quantizing noise by use of feedback,” IRE Transactions on Communications Systems, vol. 10, no. 4, pp. 373–380, Dec 1962.

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

[14] M. Derpich and J. Østergaard, “Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources,” IEEE Transactions on Information Theory, vol. 58, no. 5, pp. 3131–3152, May 2012.

[15] T. Berger, Rate distortion theory: A mathematical basis for data compression. Englewood Cliffs, NJ: Prentice-Hall, 1971.

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

[17] A. Ben-Yishai and O. Shayevitz, “The Gaussian channel with noisy feedback: Near-capacity performance via simple interaction,” 2014. [Online]. Available: http://arxiv.org/abs/1407.8022

[18] O. Ordentlich and U. Erez, “Precoded integer-forcing universally achieves the MIMO capacity to within a constant gap,” IEEE Transactions on Information Theory, vol. 61, no. 1, pp. 323–340, Jan 2015.