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 (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 paper 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 the performance 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 follows.

Index Terms—Quantization, analog-to-digital conversion, sigma-delta modulation, differential pulse-code modulation.

I. INTRODUCTION

ANALOG-TO-DIGITAL (A/D) and digital-to-analog (D/A) converters are essential in modern electronics. 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 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 desirable to design the data converter to be robust to the characteristics of the 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 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) converter 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. Nevertheless, 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, as the rate-distortion theorem is based on coding over large blocklengths [2], guaranteeing a constant product of the number of bits per sample needed to achieve distortion D, and the over-sampling ratio L, requires in principle vector-quantization of a long block of samples. In A/D and D/A conversion, vector-quantization in high dimensions is a prohibitively complex 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 [3], 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
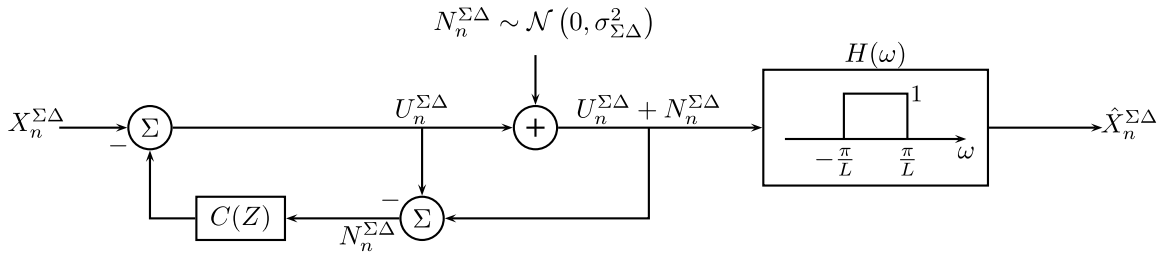
Fig. 1.   The test channel corresponding to the sigma-delta modulation architecture, with the sigma-delta quantizer replaced by an AWGN channel. The input is assumed to be over-sampled at L times the Nyquist rate.
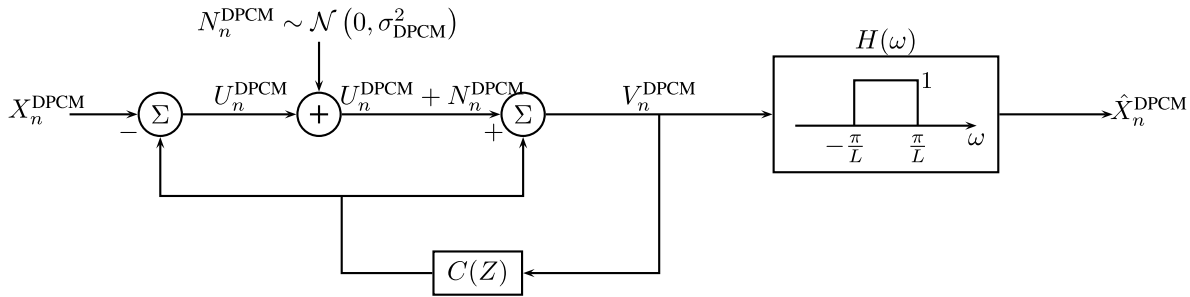

Fig. 2.   The test channel corresponding to the DPCM architecture, with the DPCM quantizer replaced by an AWGN channel. The input is assumed to be over-sampled at L times the Nyquist rate.

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" [4] 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 [4]–[6]. 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 [4], 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. 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.

A. Connection to Previous Work

The connection between DPCM and sigma-delta modulation, as two instances of predictive coding, was known from the outset. Indeed, both paradigms emerged from two Bell-Labs patents authored by CC Cutler [7], [8] in 1952 and 1954.

In fact, by adding appropriate pre- and post-filters to the sigma-delta modulator, as depicted in Figure 3, the input to the quantizer, as well as the final reconstruction of the signal, become identical to those in the DPCM architecture of Figure 2, see [9, Sec. II], [10, Ch. 3.2.4]. Thus, one may implement DPCM via either of the architectures of Figure 2 or Figure 3.

However, an important aspect of our interest in sigma-delta modulators as a means of data-conversion rather than data-compression, is that it dictates that the assumptions one can make on the statistics of the input signal must be minimal. Consequently, we consider a *compound* class of sources that consists of all stationary Gaussian processes with variance σ_X^2 whose PSD is limited to some predefined frequency band. Unfortunately, for this compound class, DPCM is unsatisfactory, as its performance depends not only on the variance and bandwidth, but rather, on the explicit form of the PSD. On the other hand, for any choice of $C(Z)$, sigma-delta modulation, as depicted in Figure 1, attains the same performance for all sources within the class. The duality result we establish here, is that the performance of sigma-delta modulation for *any* source in the compound class is equal to that of DPCM designed for a band-limited stationary process with a flat PSD.
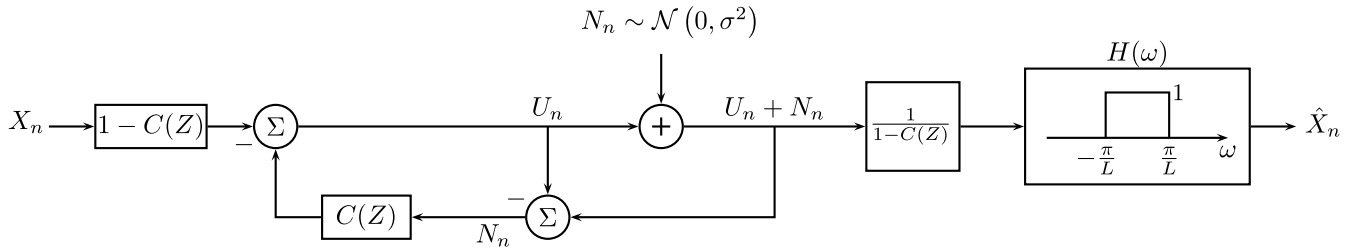
Fig. 3.   A test channel corresponding to the sigma-delta modulation with pre-filter $1 - C(Z)$ and post-filter $\frac{1}{1-C(Z)}$. This test-channel is equivalent to that from Figure 2.

Data converters often operate at very high rates, and it therefore makes sense to impose various constraints on the sigma-delta feedback filter $C(Z)$, such as confining it to be a finite impulse response (FIR) filter with a limited number of taps. For a given desired MSE distortion level, our goal is to find the constrained sigma-delta feedback filter $C(Z)$ that minimizes the quantization rate w.r.t. all sources in the compound class, and to characterize the attained rate. We note that this goal is different than the one pursued in [11], where the optimal *unconstrained* filters w.r.t. a known PSD were found.

The problem of finding the optimal N-tap FIR sigma-delta feedback filter $C(Z)$ for a compound family of sources similar to ours, was considered in [9]. The optimal filter was claimed in [9] to be the Nth order MMSE prediction filter $C(Z) = (1 - Z^{-1})^N$ of a bandpass stationary process from its past, and for a fixed target MSE distortion the required quantization rate was found to decrease linearly with N. Such as statement is obviously inaccurate, as it violates Shannon's rate-distortion theorem. The major shortcoming of the analysis in [9] is that it (implicitly) makes the high-resolution assumption that the variance of the quantizer's input is solely dictated by the target signal $\{X_n\}$, whereas the contribution of the quantization noise to this variance can be neglected. As discussed above, for over-sampled processes this assumption may not be valid even when the quantizer's resolution is very high. In particular, using the filter $C(Z) = (1 - Z^{-1})^N$ from [9], the energy of the quantization noise within the frequency band occupied by the signal indeed decreases exponentially with N. However, the noise's energy outside this band increases rapidly with N, and for any quantization resolution it will become much greater than σ_X^2 for N large enough, making the high-resolution assumption inapplicable. In this case, the dynamic range of the quantizer will be exceeded and overload errors would frequently occur.

It therefore follows that in the analysis of sigma-delta modulators one should not make high-resolution assumptions, but rather must take into account the effect of the filtered quantization noise on the variance of the quantizer's input. Fortunately, in the analysis of DPCM modulators the high-resolution assumption has been overcome in [12]. 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 [12], 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. In this paper, we leverage the results from [12] to the analysis of sigma-delta modulators, by establishing an appropriate duality between the two architectures.

B. Contributions

Let $\mathcal{S}$ be the compound class of all discrete-time stationary Gaussian sources with variance σ_X^2 and PSD that is zero for all $\omega \notin [-\pi/L, \pi/L]$, $L \geq 1$. Note that this class corresponds to uniformly sampling a compound class of continuous-time stationary Gaussian processes with variance σ_X^2 and PSD that is zero for all $|f| > f_{\max}$, at a sampling rate of $2Lf_{\max}$ samples/per second. Let $\{X_n^{\text{DPCM}}\}$ be a discrete-time stationary Gaussian process with PSD

$$S_X^{\text{DPCM}}(\omega) = \begin{cases} L\sigma_X^2 & \text{for } |\omega| \leq \pi/L \\ 0 & \text{for } \pi/L < |\omega| < \pi, \end{cases} \quad (1)$$

and note that $\{X_n^{\text{DPCM}}\} \in \mathcal{S}$.

Our main result, derived in Section II, is that for any process $\{X_n^{\Sigma\Delta}\}$ from the compound class $\mathcal{S}$, 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 input $\{X_n^{\text{DPCM}}\}$. More specifically, for such processes, for any choice of σ_{DPCM}^2 and prediction filter $C(Z)$ in the test channel of Figure 2, the same choice of $C(Z)$ together with the choice

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

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 w.r.t. any signal in $\mathcal{S}$, under any set of constraints, can be cast as an equivalent problem of optimizing the DPCM prediction filter w.r.t. input $\{X_n^{\text{DPCM}}\}$ under the same set of constraints. Furthermore, in Section II-A, we formalize a similar duality between DPCM and sigma-delta modulation for a frequency-weighted-mean-squared-error distortion measure. In this case $S_X^{\text{DPCM}}(\omega)$ is replaced with a PSD that depends on the distortion's weight function.

In principle, recasting the sigma-delta optimization problem as an MMSE prediction problem may be derived directly

from the formulas characterizing its performance, as given in Proposition 1. Nevertheless, establishing the equivalence between sigma-delta modulation and DPCM, in the specific form described above, is insightful as it allows to borrow known results from the literature about the latter.

Having recast the filter optimization problem for sigma-delta as that of optimal linear prediction, we can readily obtain the solution 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 of 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 excess-rate penalty for using scalar quantization. The purpose of this excess-rate 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 excess-rate and the overload probability is analytically determined.

Clearly, a sigma-delta modulator can only perform well if overload errors are rather rare. The stability analysis in Section III is based on avoiding overload events w.h.p., and does not aim to consider the effect of such events on the distortion once they occur. In general, the overload probability of the scheme described in Section III decreases double exponentially with the excess-rate of the quantizer w.r.t. the mutual information. Thus, taking an excess rate of $1 - 2$ bits will usually yield a sufficiently low overload probability. However, sigma-delta quantizers are often employed with a one-bit quantizer. In this case, the overload error probability cannot be very low. Consequently, the designer would need to guarantee that the effect of overload errors is local in time, and does not drive the system out of stability. There are various restrictions one can place on $C(Z)$ in pursuit of the latter goal. The issue of maintaining stability when overload errors are unavoidable is outside the scope of this paper. Nevertheless, we stress that our main result is of great relevance to this setting, as it shows that the filter $C(Z)$ should be chosen as the optimal MMSE prediction filter of $\{X_n^{\text{DPCM}}\}$ from its noisy past under the stability ensuring restrictions.

II. Main Results

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 (real) 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_X^2, whose PSD is zero for all frequencies $|f| > f_{\max}$, but otherwise unknown. The Nyquist sampling rate for this process is $2 f_{\max}$ 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 $2Lf_{\max}$ 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 σ_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 that end, we establish an equivalence between the performance obtained by this test channel for any stationary band-limited Gaussian process with variance σ_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 σ_X^2. The performance of the latter is now well understood [12], and, as we shall show, can be translated to a simple characterization of the performance of sigma-delta modulation.

First, we recall the derivation of the distortions attained by the test channels from Figure 1 and Figure 2, and the scalar mutual information $I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta})$ and $I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}})$ 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 [12]–[14]. 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 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.

We begin by considering 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 to be strictly causal. The derivations in the proof of Proposition 1 below, as well as in that of Proposition 2, are straightforward, and are based on, see [11], [12], [15]. We nevertheless present the proofs in order to set forward some basic ideas to be used later.

Proposition 1: For any Gaussian stationary process $\{X_n^{\Sigma\Delta}\}$ with variance σ_X^2 whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$, the test channel from Figure 1, corresponding to sigma-delta modulation, achieves MSE distortion

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

and its scalar mutual information satisfies[1]

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

Proof: From Figure 1, we have that

$$U_n^{\Sigma\Delta} = X_n^{\Sigma\Delta} - c_n * N_n^{\Sigma\Delta}, \tag{3}$$

and therefore

$$U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta} = X_n^{\Sigma\Delta} + (\delta_n - c_n) * N_n^{\Sigma\Delta},$$

where δ_n is the discrete identity filter. Using the fact that $\{X_n^{\Sigma\Delta}\}$ is a low-pass process, passing it through the filter $H(\omega)$ has no effect, and hence

$$\hat{X}_n^{\Sigma\Delta} = h_n * (U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta})$$
$$= X_n^{\Sigma\Delta} + h_n * (\delta_n - c_n) * N_n^{\Sigma\Delta}.$$

The MSE distortion attained by the test channel from Figure 1 is therefore

$$D = \mathbb{E}(X_n^{\Sigma\Delta} - \hat{X}_n^{\Sigma\Delta})^2 = \sigma_{\Sigma\Delta}^2 \cdot \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega.$$

The scalar mutual information between the "quantizer's" input $U_n^{\Sigma\Delta}$ and output $U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}$ is given by

$$I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) = h(U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) - h(N_n^{\Sigma\Delta}) \tag{4}$$

$$= \frac{1}{2} \log\left(1 + \frac{\mathbb{E}(U_n^{\Sigma\Delta})^2}{\sigma_{\Sigma\Delta}^2}\right), \tag{5}$$

where (4), as well as (5), follow from the statistical independence of $N_n^{\Sigma\Delta}$ and $U_n^{\Sigma\Delta}$. Using (3), the variance of $U_n^{\Sigma\Delta}$ is

$$\mathbb{E}(U_n^{\Sigma\Delta})^2 = \sigma_X^2 + \sigma_{\Sigma\Delta}^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega. \tag{6}$$

Substituting (6) into (5) 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

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

replaced by an AWGN channel with zero mean and variance σ_{DPCM}^2. 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 $I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}})$, were already found in [12, Th. 1] for the special case where $C(Z)$ is the optimal MMSE infinite length prediction filter of X_n^{DPCM} from all past samples of the process $\{X_n^{\text{DPCM}} + N_n^{\text{DPCM}}\}$. The following straightforward proposition characterizes the rate and distortion for any choice of the causal filter $C(Z)$ and any value of σ_{DPCM}^2.

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

$$S_X^{\text{DPCM}}(\omega) = \begin{cases} L\sigma_X^2 & \text{for } |\omega| \leq \pi/L \\ 0 & \text{for } \pi/L < |\omega| < \pi, \end{cases} \tag{7}$$

the test channel from Figure 2, corresponding to DPCM, achieves MSE distortion

$$D = \frac{\sigma_{\text{DPCM}}^2}{L},$$

and its scalar mutual information satisfies

$$I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}})$$
$$= \frac{1}{2} \log\left(1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega\right.$$
$$\left. + \frac{L\sigma_X^2}{\sigma_{\text{DPCM}}^2} \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega\right).$$

Proof: From Figure 2, we have that

$$U_n^{\text{DPCM}} = X_n^{\text{DPCM}} - c_n * V_n^{\text{DPCM}} \tag{8}$$
$$V_n^{\text{DPCM}} = U_n^{\text{DPCM}} + N_n^{\text{DPCM}} + c_n * V_n^{\text{DPCM}} \tag{9}$$

Substituting (8) in (9) yields

$$V_n^{\text{DPCM}} = X_n^{\text{DPCM}} + N_n^{\text{DPCM}}. \tag{10}$$

Using the fact that $\{X_n^{\text{DPCM}}\}$ is a low-pass process, as before, we obtain

$$\hat{X}_n^{\text{DPCM}} = h_n * (X_n^{\text{DPCM}} + N_n^{\text{DPCM}})$$
$$= X_n^{\text{DPCM}} + h_n * N_n^{\text{DPCM}}. \tag{11}$$

Since $\{N_n^{\text{DPCM}}\}$ is AWGN with variance σ_{DPCM}^2, the variance of the filtered process $h_n * N_n^{\text{DPCM}}$ is σ_{DPCM}^2/L. Thus,

$$D = \mathbb{E}(X_n^{\text{DPCM}} - \hat{X}_n^{\text{DPCM}})^2 = \frac{\sigma_{\text{DPCM}}^2}{L}.$$

As in the analysis of the test channel from Figure 1, the scalar mutual information between U_n^{DPCM} and $U_n^{\text{DPCM}} + N_n^{\text{DPCM}}$ is given by

$$I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}}) = \frac{1}{2} \log\left(1 + \frac{\mathbb{E}(U_n^{\text{DPCM}})^2}{\sigma_{\text{DPCM}}^2}\right). \tag{12}$$

Now, substituting (10) in (8) gives

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

and the variance of U_n is therefore

$$
\mathbb{E}(U_n^{\mathrm{DPCM}})^2
$$
$$
= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_X^{\mathrm{DPCM}}(\omega)|1 - C(\omega)|^2 d\omega
$$
$$
+ \frac{1}{2\pi} \int_{-\pi}^{\pi} S_N^{\mathrm{DPCM}}(\omega)|C(\omega)|^2 d\omega
$$
$$
= \frac{L\sigma_X^2}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega + \frac{\sigma_{\mathrm{DPCM}}^2}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega. \tag{13}
$$

Substituting (13) into (12) 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 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 [12] and [14], 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 any Gaussian stationary process with variance σ_X^2 whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$, let $\{X_n^{\mathrm{DPCM}}\}$ be a flat low-pass Gaussian stationary process with PSD as in (7), and let $C(Z)$ be a strictly causal filter. The test channel from Figure 1 with

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

and the test channel from Figure 2 with

$$
\sigma_{\mathrm{DPCM}}^2 = L \cdot D,
$$

both achieve MSE distortion D. Furthermore, in both cases, the scalar mutual information satisfies

$$
I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) = I(U_n^{\mathrm{DPCM}}; U_n^{\mathrm{DPCM}} + N_n^{\mathrm{DPCM}})
$$
$$
= \frac{1}{2} \log \left(1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} |C(\omega)|^2 d\omega \right.
$$
$$
\left. + \frac{\sigma_X^2}{D} \frac{1}{2\pi} \int_{-\pi/L}^{\pi/L} |1 - C(\omega)|^2 d\omega \right).
$$

This theorem indicates that for any stationary band-limited Gaussian process with variance σ_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 (2). Thus, Theorem 1 provides a unified framework for analyzing the performance of sigma-delta modulation and DPCM. The benefit 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 Corollary 1 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 (7), 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

$$
\sigma_D^{*2} = \min_{C(Z) \in \mathcal{C}} \mathbb{E}(S_n - c_n * (S_n + W_n))^2
$$
$$
C_D^*(Z) = \underset{C(Z) \in \mathcal{C}}{\mathrm{argmin}}\, \mathbb{E}(S_n - c_n * (S_n + W_n))^2.
$$

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

$$
I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \geq \frac{1}{2} \log \left(1 + \frac{\sigma_D^{*2}}{L \cdot D} \right), \tag{14}
$$

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 [4], [9], [11], [15]–[18]. 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 [11], [15], [18]. Spang, III, and Schultheiss [18] 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 [11] and [15] the design of an optimal *unconstrained* sigma-delta filter was studied, under the assumption of a fixed scalar quantizer that 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 [16]. 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

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

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 $X_n^{\text{DPCM}} = S_n$, $N_n^{\text{DPCM}} = W_n$ and $\sigma_{\text{DPCM}}^2 = L \cdot D$. Thus,

$$\begin{aligned}
I(U_n^{\Sigma\Delta}; & U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \\
&= I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}}) \\
&= \frac{1}{2} \log\left(1 + \frac{\mathbb{E}(S_n - c_n * (S_n + W_n))^2}{L \cdot D}\right),
\end{aligned} \quad (15)$$

where we have used (8), (10), and (12), to arrive at (15). It follows that among all filters in $\mathcal{C}$, the filter that minimizes (15) is $C_D^*(Z)$, and that it attains (14) with equality. ∎

It is interesting to note [12] that since $\{W_n\}$ is an i.i.d. process with variance $L \cdot D$ and $C(Z)$ is strictly causal, the mutual information (15) can also be written as

$$\begin{aligned}
I(U_n^{\Sigma\Delta}; & U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \\
&= \frac{1}{2} \log\left(\frac{\mathbb{E}(S_n + W_n - c_n * (S_n + W_n))^2}{L \cdot D}\right).
\end{aligned} \quad (16)$$

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 $\{S_n + W_n\}$ [19], which equals

$$2^{\frac{1}{2\pi} \int_{-\pi}^{\pi} \log(S_S(\omega) + L \cdot D) d\omega} = (L \cdot D)\left(1 + \frac{\sigma_X^2}{D}\right)^{1/L}. \quad (17)$$

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 (17) implies that for the optimal unconstrained sigma-delta filter $C(Z)$ we must have

$$\begin{aligned}
S_{E^{\text{pred}}}(\omega) &\triangleq |1 - C(\omega)|^2 (L \cdot D + S_S(\omega)) \\
&= (L \cdot D)\left(1 + \frac{\sigma_X^2}{D}\right)^{1/L}, \quad \forall \omega \in [-\pi, \pi) \quad (18)
\end{aligned}$$

Combining (16), (17), and (18) yields the following corollary.

Corollary 1: Let $\{X_n^{\Sigma\Delta}\}$ be a Gaussian stationary process with variance σ_X^2 whose PSD is zero for all $\omega \notin [-\pi/L, \pi/L]$.

If the test channel from Figure 1 attains MSE distortion D, then

$$I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) \geq \frac{1}{2L} \log\left(1 + \frac{\sigma_X^2}{D}\right). \quad (19)$$

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

$$|1 - C(\omega)|^2 = \begin{cases} \left(1 + \frac{\sigma_X^2}{D}\right)^{-(L-1)/L} & \omega \in [-\frac{\pi}{L}, \frac{\pi}{L}] \\ \left(1 + \frac{\sigma_X^2}{D}\right)^{1/L} & \omega \notin [-\frac{\pi}{L}, \frac{\pi}{L}], \end{cases} \quad (20)$$

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: Note that the existence of a strictly causal filter $C(Z)$ which satisfies (20) is guaranteed by Wiener's spectral-factorization theory [19] due to the readily verified fact that

$$2^{\frac{1}{2\pi} \int_{-\pi}^{\pi} \log|1 - C(\omega)|^2 d\omega} = 1.$$

The optimal filter induces a two-level frequency response for $|1 - C(\omega)|^2$. Østergaard and Zamir [14] 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 $|1 - C(\omega)|^2$. We also note that the optimality of the unconstrained filter specified by (20) can be deduced as a special case of [11, Sec. IV].

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 $\tilde{\hat{X}}_n^{\Sigma\Delta} = \alpha \hat{X}_n^{\Sigma\Delta}$, where

$$\alpha = \frac{\sigma_X^2}{\sigma_X^2 + D}.$$

Consequently, the obtained MSE distortion is reduced to

$$\tilde{D} = \mathbb{E}(X_n^{\Sigma\Delta} - \alpha \hat{X}_n^{\Sigma\Delta})^2 = \frac{\sigma_X^2 \cdot D}{\sigma_X^2 + D}.$$

It is straightforward to verify [20] that with this improvement, the sigma-delta test channel from Figure 1 with $C(Z)$ and $\sigma_{\Sigma\Delta}^2$ as specified in Corollary 1 attains

$$I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + V_n^{\Sigma\Delta}) = \frac{1}{2L} \log\left(\frac{\sigma_X^2}{\tilde{D}}\right),$$

which is the optimal rate-distortion function for a stationary Gaussian source $\{X_n^{\Sigma\Delta}\}$ with PSD as in (7). It follows that the sigma-delta test channel from Figure 1 with $C(Z)$ and $\sigma_{\Sigma\Delta}^2$ as specified in Corollary 1 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 $\tilde{D}$ 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

$$D_{\text{FWMSE}} \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) S_E(\omega) d\omega, \tag{21}$$

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

$$S_X^{\text{FWMSE}}(\omega) = \begin{cases} L\sigma_X^2 P(\omega) & \text{for } |\omega| \leq \pi/L \\ 0 & \text{for } \pi/L < |\omega| < \pi, \end{cases} \tag{22}$$

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

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

$$C_{D_{\text{FWMSE}}}^*(Z) = \operatorname*{argmin}_{C(Z) \in \mathcal{C}} \mathbb{E}(S_n - c_n * (S_n + W_n))^2.$$

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

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

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 (22), rather than (7). Note that this changes $I(U_n^{\text{DPCM}}; U_n^{\text{DPCM}} + N_n^{\text{DPCM}})$, 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_{FWMSE} (under the appropriate distortion measure), then

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

∎

B. Sigma-Delta Modulation With an Interleaved Vector 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 [12]–[14], 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

$$\mathbf{x}^{\Sigma\Delta} = [X_1^{\Sigma\Delta}, \dots, X_N^{\Sigma\Delta}],$$

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

$$\mathbf{x}_k^{\Sigma\Delta} = [X_{(k-1)M+1}^{\Sigma\Delta}, \dots, X_{kM}^{\Sigma\Delta}], \quad k = 1, \dots, K,$$

where $M \triangleq N/K$. Now, we can apply K parallel sigma-delta modulators, 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 4. By our assumption that $\{X_n^{\Sigma\Delta}\}$ has decaying memory, if M is large enough the K inputs that enter the quantizer $\underline{Q}(\cdot) = [Q_1(\cdot), \dots, Q_K(\cdot)]$ 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 arbitrarily closely, in KL divergence, to the AWGN $N_n^{\Sigma\Delta}$ [20].

III. SIGMA-DELTA MODULATION WITH A SCALAR UNIFORM 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 distribution is close to that of $N_n^{\Sigma\Delta}$. 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
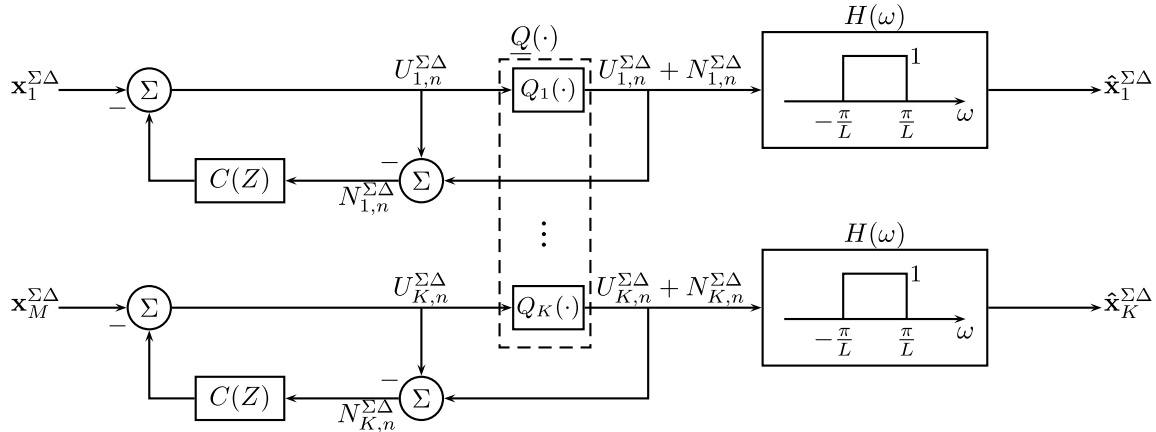
Fig. 4. K parallel sigma-delta modulators coupled by an K-dimensional quantizer $\underline{Q}(\cdot)$.

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 uniform scalar quantizers of finite support are used. For such quantizers, the quantization error is composed of two main factors [4]: *granular errors* that correspond 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 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.[2]

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 exponentially decreasing function of $R - I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta})$, 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 $I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) + \delta(P_{ol})$, where $\delta(P_{ol}) = \mathcal{O}\left(\log\log\left(\frac{1}{P_{ol}}\right)\right)$.

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 5. Our goal is to analyze the distortion and overload probability
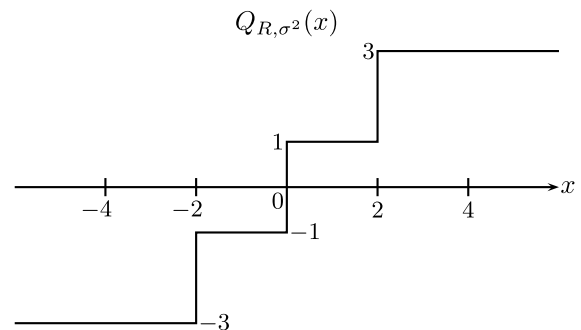
Fig. 5. An illustration of $Q_{R,\sigma^2}(\cdot)$ for $R = 2$ and $\sigma^2 = 1/3$.

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 ensures that the effect of an overload error in the original system is restricted to the block where it occurs.

Remark 4: In [11], the disastrous effect of overload error amplification by the feedback loop was elegantly mitigated by clipping the input to the quantizer, and feeding the quantization errors of the clipped process to the noise-shaping filter $C(Z)$. This technique ensures that the system remains stable, but requires to add an analog clipping component to the sigma-delta modulator. Our treatment here follows the "classic" sigma-delta modulation paradigm, which does not include such a clipper. Therefore, we take the conservative approach that assumes that once an overload error occurred we can guarantee nothing about the fidelity of the system's output, and limit the effect of overload errors to at most N samples, by operating in independent blocks.

The analysis is made much simpler by introducing a subtractive *dither* [20]. Namely, let $\{Z_n\}$ be a sequence of
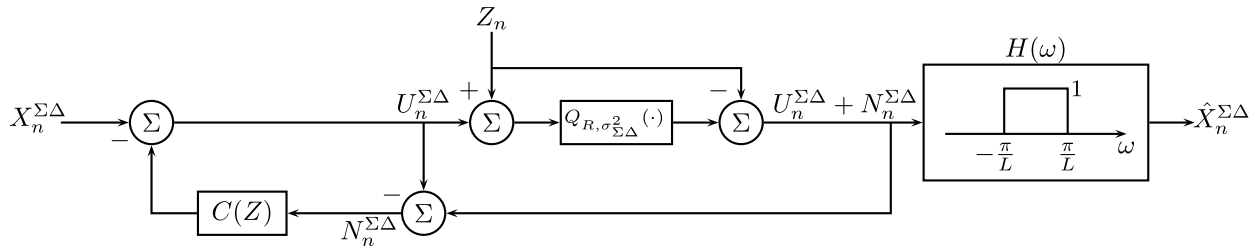
Fig. 6. A sigma-delta modulator with a dithered scalar uniform quantizer. The input is assumed to be over-sampled at L times the Nyquist rate, and the dither sequence $\{Z_n\}$ is assumed to be an i.i.d. sequence of random variables uniformly distributed over the interval $\left[-\sqrt{12\sigma_{\Sigma\Delta}^2}/2, \sqrt{12\sigma_{\Sigma\Delta}^2}/2\right)$ and statistically independent $\{X_n^{\Sigma\Delta}\}$.

i.i.d. random variables uniformly distributed over the interval $[-\sqrt{12\sigma_{\Sigma\Delta}^2}/2, \sqrt{12\sigma_{\Sigma\Delta}^2}/2)$. In order to quantize $U_n^{\Sigma\Delta}$, we add Z_n to it before applying the quantizer, and subtract Z_n afterwards, such that the obtained result is $Q_{R,\sigma_{\Sigma\Delta}^2}(U_n^{\Sigma\Delta} + Z_n) - Z_n$. Adding and subtracting $U_n^{\Sigma\Delta}$, we get $U_n^{\Sigma\Delta} + \left(Q_{R,\sigma_{\Sigma\Delta}^2}(U_n^{\Sigma\Delta} + Z_n) - (U_n^{\Sigma\Delta} + Z_n)\right)$, and the quantization error is therefore

$$N_n \triangleq Q_{R,\sigma_{\Sigma\Delta}^2}(U_n^{\Sigma\Delta} + Z_n) - (U_n^{\Sigma\Delta} + Z_n) \qquad (23)$$

The main result in this section is the following.

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 6 applied on a sequence of N consecutive source samples with quantization rate $R = I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) + \delta(P_{ol})$, where

$$\delta(P_{ol}) \triangleq \frac{1}{2}\log\left(-\frac{2}{3}\ln\frac{P_{ol}}{2N}\right). \qquad (24)$$

attains MSE distortion smaller than

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

given that overload did not occur, where $o_N(1) \to 0$ as N increases, and the overload probability is smaller than P_{ol}.

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

$$Q_{R,\sigma^2}(x) = \tilde{Q}_{\sqrt{12\sigma^2}\mathbb{Z}}\left(x + \frac{\sqrt{12\sigma^2}}{2}\right) - \frac{\sqrt{12\sigma^2}}{2}. \qquad (25)$$

Applying (23) therefore yields that if overload did not occur in the nth sample, i.e., if $|U_n^{\Sigma\Delta} + Z_n| \le \Gamma/2$, we have

$$N_n = \tilde{Q}_{\sqrt{12\sigma_{\Sigma\Delta}^2}\mathbb{Z}}\left(U_n^{\Sigma\Delta} + Z_n + \sqrt{3\sigma_{\Sigma\Delta}^2}\right)$$
$$- \left(U_n^{\Sigma\Delta} + Z_n + \sqrt{3\sigma_{\Sigma\Delta}^2}\right). \qquad (26)$$

Dealing with the overload event of the quantizer directly is rather involved. Instead, as done in [21], we first consider a *reference system* with an infinite-support quantizer ($R = \infty$)

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 (26) regardless of whether or not $|U_n^{\Sigma\Delta} + Z_n| \le \Gamma/2$, and account for the overload probability later.

Assuming that the dither sequence $\{Z_n\}$ is drawn statistically independent of the process $\{X_n^{\Sigma\Delta}\}$, the Crypto Lemma, see, e.g. [20, 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 $\{X_n^{\Sigma\Delta}\}$. 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 6 (with an infinite-support quantizer) is equivalent to the test channel from Figure 1 with $N_n^{\Sigma\Delta} \sim \text{Uniform}\left([-\sqrt{12\sigma_{\Sigma\Delta}^2}/2, \sqrt{12\sigma_{\Sigma\Delta}^2}/2)\right)$ instead of $N_n^{\Sigma\Delta} \sim \mathcal{N}(0, \sigma_{\Sigma\Delta}^2)$. Thus, the average MSE distortion attained by the reference scalar sigma-delta modulator from Figure 6 is as given in Proposition 1 up to a multiplicative factor of $1 + o_N(1)$ 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 $\{U_n^{\Sigma\Delta}\}$ 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 error occurs within a block of length N, as a function of R and $I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta})$. Since this event is equivalent to the event that at the reference system some input to the quantizer exceeds $\Gamma/2$ in magnitude within the block, it suffices to upper bound the probability of the latter event.

Assume the reference scalar sigma-delta modulator from Figure 6 is applied to a vector $\mathbf{x}^{\Sigma\Delta} = [X_1^{\Sigma\Delta}, \ldots, X_N^{\Sigma\Delta}]$ 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 $\text{OL}_k \triangleq \{|U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta}| > \Gamma/2\}$ and the event $\text{OL} \triangleq \cup_k^N \text{OL}_k$.

By the union bound, we have

$$P_{ol} \triangleq \Pr(\text{OL}) \leq \sum_{k=1}^{N} \Pr(\text{OL}_k). \tag{27}$$

The random variable $U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta} = X_k^{\Sigma\Delta} + (\delta_k - c_k) * N_k^{\Sigma\Delta}$ is a linear combination of a Gaussian random variable $X_k^{\Sigma\Delta}$ and statistically independent uniform random variables $\{N_k^{\Sigma\Delta}\}$. In [22, 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

$$\Pr\left(|U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta}| > \Gamma/2\right)$$
$$\leq 2\exp\left\{-\frac{\Gamma^2}{8\mathbb{E}(U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta})^2}\right\}$$
$$= 2\exp\left\{-\frac{12\sigma_{\Sigma\Delta}^2 2^{2R}}{8\left(\mathbb{E}(U_k^{\Sigma\Delta})^2 + \mathbb{E}(N_k^{\Sigma\Delta})^2\right)}\right\},$$

where in the last equality we have used the definition of Γ and the fact that $U_k^{\Sigma\Delta}$ and $N_k^{\Sigma\Delta}$ are statistically independent. Equivalently, we may write

$$\Pr\left(\text{OL}_k\right) \leq 2\exp\left\{-\frac{12\sigma_{\Sigma\Delta}^2 2^{2R}}{8\sigma_{\Sigma\Delta}^2\left(1 + \frac{\mathbb{E}(U_k^{\Sigma\Delta})^2}{\sigma_{\Sigma\Delta}^2}\right)}\right\}$$
$$= 2\exp\left\{-\frac{3}{2}2^{2\left(R - \frac{1}{2}\log\left(1 + \frac{\mathbb{E}(U_k^{\Sigma\Delta})^2}{\sigma_{\Sigma\Delta}^2}\right)\right)}\right\}$$
$$= 2\exp\left\{-\frac{3}{2}2^{2(R - I(U_k^{\Sigma\Delta}; U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta}))}\right\}, \tag{28}$$

where we have used (5) in the last equality. Substituting (28) into (27) gives

$$P_{ol} \leq 2\sum_{k=1}^{N}\exp\left\{-\frac{3}{2}2^{2(R - I(U_k^{\Sigma\Delta}; U_k^{\Sigma\Delta} + N_k^{\Sigma\Delta}))}\right\}. \tag{29}$$

Note that $\mathbb{E}(U_k^{\Sigma\Delta})^2 = \sigma_X^2 + \sigma_{\Sigma\Delta}^2 \sum_{m=1}^{k} c_k^2$ is monotonically nondecreasing in k and is given by (6) for values of k that are greater than the length of the filter c_k. We can therefore further bound (29) as

$$P_{ol} \leq 2N\exp\left\{-\frac{3}{2}2^{2(R - I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}))}\right\}, \tag{30}$$

where $I\left(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}\right)$ 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 $1 + o_N(1)$ multiplicative term, and that the probability that one of the quantizer input samples exceeds $\Gamma/2$ in magnitude within a block of length N, is bounded by (30). 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 right hand side of (30). 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 $D_{\overline{\text{OL}}}$ and the expected distortion conditioned on the event that OL did occur by D_{OL}. For the reference system, we have

$$D(1 + o_N(1)) = \Pr(\overline{\text{OL}})D_{\overline{\text{OL}}} + \Pr(\text{OL})D_{\text{OL}} \geq \Pr(\overline{\text{OL}})D_{\overline{\text{OL}}},$$

and therefore

$$D_{\overline{\text{OL}}} \leq \frac{D(1 + o_N(1))}{1 - P_{ol}}.$$

This shows that the scalar sigma-delta system from Figure 6, whose quantizer has limited support $[-\Gamma/2, \Gamma/2]$, with $R = I(U_n^{\Sigma\Delta}; U_n^{\Sigma\Delta} + N_n^{\Sigma\Delta}) + \delta$ achieves the same average MSE distortion as the test channel from Figure 1 up to a multiplicative factor of $(1 + o_N(1))/(1 - P_{ol})$, with block error probability smaller than $2N\exp\left\{-\frac{3}{2}2^{2\delta}\right\}$. 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 δ, that depends on the target overload error probability. More precisely, for any $0 < P_{ol} < 1$, taking the rate penalty as in (24) guarantees that the overload error probability is smaller than P_{ol}. ∎

REFERENCES

[1] O. Ordentlich and U. Erez, "Performance analysis and optimal filter design for sigma-delta modulation via duality with DPCM," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 321–325.
[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[3] R. Zamir and M. Feder, "Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 141–154, Jan. 1995.
[4] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1984.
[5] R. A. McDonald, "Signal-to-noise and idle channel performance of differential pulse code modulation systems—Particular applications to voice signals," *Bell Labs Tech. J.*, vol. 45, no. 7, pp. 1123–1151, 1966.
[6] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.
[7] C. C. Cutler, "Differential quantization of communication signals," U.S. Patent 2 605 361, Jul. 29, 1952,
[8] C. C Cutler, "Transmission systems employing quantization," U.S. Patent 2 927 962, May 8, 1960.
[9] S. Tewksbury and R. Hallock, "Oversampled, linear predictive and noise-shaping coders of order N & gt; 1," *IEEE Trans. Circuits Syst.*, vol. CS-25, no. 7, pp. 436–447, Jul. 1978.
[10] M. S. Derpich, "Optimal source coding with signal transfer function constraints," Ph.D. dissertation, School Elect. Eng. Comput. Sci., Univ. Newcastle, Callaghan, NSW, Australia, 2009.
[11] M. S. Derpich, E. I. Silva, D. E. Quevedo, and G. C. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3871–3890, Aug. 2008.
[12] 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, Jul. 2008.
[13] T. Guess and M. K. Varanasi, "An information-theoretic framework for deriving canonical decision-feedback receivers in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 173–187, Jan. 2005.

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

[15] M. S. Derpich and J. Ostergaard, "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.

[16] P. Noll, "On predictive quantizing schemes," *Bell Labs Tech. J.*, vol. 57, no. 5, pp. 1499–1532, May 1978.

[17] M. A. Gerzon and P. G. Craven, "Optimal noise shaping and dither of digital signals," in *Proc. 87th Audio Eng. Soc. Conv.*, 1989.

[18] H. Spang, III, and P. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun. Syst.*, vol. 10, no. 4, pp. 373–380, Dec. 1962.

[19] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[20] R. Zamir, *Lattice Coding for Signals Network*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[21] A. Ben-Yishai and O. Shayevitz, "Interactive schemes for the AWGN channel with noisy feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2409–2427, Apr. 2017.

[22] O. Ordentlich and U. Erez, "Precoded integer-forcing universally achieves the MIMO capacity to within a constant gap," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 323–340, Jan. 2015.

Or Ordentlich received the B.Sc. degree (cum laude) in 2010, M.Sc. degree (summa cum laude) in 2011, and completed his PhD studies in 2015, all in electrical engineering in Tel Aviv University, Israel. During the years 2015-2017 he is was a postdoctoral fellow in the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology (MIT), and in the Department of Electrical and Computer Engineering at Boston University. He is currently a senior lecturer (assistant professor) in the School of Computer Science and Engineering at the Hebrew University of Jerusalem.

Uri Erez (M'09) was born in Tel-Aviv, Israel, on October 27, 1971. He eceived the B.Sc. degree in mathematics and physics and the M.Sc. and Ph.D. degrees in electrical engineering from Tel-Aviv University in 1996, 1999, and 2003, respectively During 2003-2004, he was a Postdoctoral Associate at the Signals, Information and Algorithms Laboratory at the Massachusetts Institute of Technology (MIT), Cambridge. Since 2005, he has been with the Department of Electrical Engineering-Systems at Tel-Aviv University. His research interests are in the general areas of information theory and digital communication. He served in the years 2009-2011 as Associate Editor for Coding Techniques for the IEEE TRANSACTIONS ON INFORMATION THEORY.