Achieving the Gaussian Rate-Distortion Function by Prediction

Ram Zamir, Yuval Kochman and Uri Erez
Dept. Electrical Engineering-Systems, Tel Aviv University

***Abstract*— The "water-filling" solution for the quadratic rate-distortion function of a stationary Gaussian source is given in terms of its power spectrum. This formula naturally lends itself to a frequency domain "test-channel" realization. We provide an alternative time-domain realization for the rate-distortion function, based on linear prediction. This solution has some interesting implications, including the optimality at all distortion levels of pre/post filtered vector-quantized differential pulse code modulation (DPCM), and a duality relationship with decision-feedback equalization (DFE) for inter-symbol interference (ISI) channels.**

I. Introduction

The *water-filling* solution for the quadratic rate-distortion function $R(D)$ of a stationary Gaussian source is given in terms of the spectrum of the source. Similarly, the capacity C of a power-constrained ISI channel with Gaussian noise is given by a water-filling solution relative to the effective noise spectrum. Both these formulas amount to mutual-information between vectors in the frequency domain. In contrast, linear prediction along the time domain can translate these vector mutual-information quantities into scalar ones. Indeed, for capacity, Cioffi *et al* [3] showed that C is equal to the *scalar* mutual-information over a slicer embedded in a decision-feedback noise-prediction loop. We show that a parallel result holds for the rate-distortion function: $R(D)$ is equal to the *scalar* mutual-information over an additive white Gaussian noise (AWGN) channel embedded in a source prediction loop. This result implies that $R(D)$ can be realized in a sequential manner, and it joins other observations regarding the role of minimum mean-square error (MMSE) estimation in successive encoding and decoding of Gaussian channels and sources [6], [5], [2].

The Quadratic-Gaussian Rate-Distortion Function

The rate-distortion function (RDF) of a stationary source with memory is given as a limit of normalized mutual information associated with vectors of source samples. For a real valued source $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$, and mean-squared distortion level D, the RDF can be written as, [1],

$$R(D) = \lim_{n\to\infty} \frac{1}{n} \inf I(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$$

where the infimum is over all channels $\mathbf{X} \to \mathbf{Y}$ such that $\frac{1}{n}\|\mathbf{Y} - \mathbf{X}\|^2 \leq D$. A channel which realizes this infimum is called an *optimum test-channel*. When the source is Gaussian, the RDF takes an explicit form in the frequency domain in terms of the power-spectrum

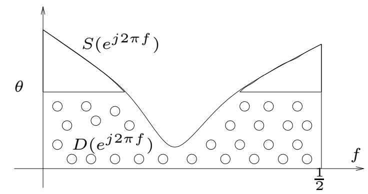

Fig. 1. The water filling solution.

$$S(e^{j2\pi f}) = \sum_k R[k] e^{-jk2\pi f}, \quad -1/2 < f < 1/2,$$

where $R[k] = E\{X_n X_{n+k}\}$ is the auto-correlation function. The water filling solution, illustrated in Figure 1, gives a parametric formula for the Gaussian RDF in terms of a parameter θ:

$$\begin{aligned} R(D) &= \int_{-1/2}^{1/2} \frac{1}{2} \log\left(\frac{S(e^{j2\pi f})}{D(e^{j2\pi f})}\right) df \\ &= \int_{f:S(e^{j2\pi f})>\theta} \frac{1}{2}\log\left(\frac{S(e^{j2\pi f})}{\theta}\right) df \end{aligned} \quad (1)$$

where the *distortion spectrum* is given by

$$D(e^{j2\pi f}) = \begin{cases} \theta, & \text{if } S(e^{j2\pi f}) > \theta \\ S(e^{j2\pi f}), & \text{otherwise,} \end{cases} \quad (2)$$

and where we choose the *water level* θ so that the total distortion is D:

$$D = \int_{-1/2}^{1/2} D(e^{j2\pi f}) df. \quad (3)$$

In the special case of a memoryless (white) Gaussian source $\sim N(0, \sigma^2)$, the power-spectrum is flat $S(e^{j2\pi f}) = \sigma^2$, so $\theta = D$ and the RDF is simplified to

$$\frac{1}{2}\log\left(\frac{\sigma^2}{D}\right). \quad (4)$$

The optimum test-channel can be written in this case in a *backward* additive-noise form $X = Y + N$, with $N \sim N(0, D)$, or in a *forward* linear additive-noise form:

$$Y = \beta(\alpha X + N)$$

with $\alpha = \beta = \sqrt{1 - D/\sigma^2}$ and $N \sim N(0, D)$. In the general source case, the forward channel realization of the RDF has several equivalent forms [7, Sec. 9.7], [1, Sec. 4.5]. The one which is more useful for our purpose replaces α and β above by linear time-invariant filters, while keeping the noise N as AWGN [17]:

$$Y_n = h_{2,n} * (h_{1,n} * X_n + N_n) \tag{5}$$

where $N_n \sim N(0, \theta)$ is AWGN, $*$ denotes convolution, and $h_{1,n}$ and $h_{2,n}$ are the impulse responses of the pre- and post-filters, respectively, whose frequency response are given in (12) and (17) in the next section.

If we take a discrete approximation of (1),

$$\sum_i \frac{1}{2} \log \left(\frac{S(e^{j2\pi f_i})}{D(e^{j2\pi f_i})} \right), \tag{6}$$

then each component has the memoryless form of (4). Hence, we can think of the frequency domain formula (1) as encoding of *parallel* (independent) Gaussian sources, where source i is a memoryless Gaussian source $X_i \sim N(0, S(e^{j2\pi f_i}))$ encoded at distortion level $D(e^{j2\pi f_i})$; see [4]. Indeed, practical frequency domain source coding schemes such as Transform Coding and Sub-band Coding [9] get close to the RDF of a stationary Gaussian source using an "array" of parallel *scalar* quantizers.

Rate-Distortion and Prediction

Our main result is a predictive channel realization for the quadratic-Gaussian RDF (1), which can be viewed as the time-domain counterpart of the frequency domain formulation above. The notions of *entropy-power* and *Shannon lower bound* (SLB) provide a simple relation between the Gaussian RDF and prediction, and motivate our result. Recall that the entropy-power is the variance of a *white* Gaussian process having the same entropy-rate as the source [4]; for a Gaussian source the entropy-power is given by

$$P_e(X) = \exp \left(\int_{-1/2}^{1/2} \log \left(S(e^{j2\pi f}) \right) df \right). \tag{7}$$

In the context of Wiener's spectral-factorization theory, the entropy-power quantifies the MMSE in one-step linear prediction of a Gaussian source from its infinite past [1]:

$$P_e(X) = \inf_{\{a_i\}} E \left(X_n - \sum_{i=1}^{\infty} a_i X_{n-i} \right)^2. \tag{8}$$

The error process associated with the infinite-order optimum predictor,

$$Z_n = X_n - \sum_{i=1}^{\infty} a_i X_{n-i}, \tag{9}$$

is called the *innovation process*. The *orthogonality principle* of MMSE estimation implies that the innovation process has zero mean and is *white*; in the Gaussian case this implies that

$$Z_n \sim \mathcal{N}(0, P_e(X))$$

is a *memoryless* process. See, e.g., [6]. From information theoretic perspective, the entropy-power plays a role in the SLB:

$$R(D) \geq \frac{1}{2} \log \left(\frac{P_e(X)}{D} \right). \tag{10}$$

Equality in the SLB holds if the distortion level is smaller than or equal to the lowest value of the power spectrum: $D \leq S_{min} \stackrel{\Delta}{=} \min_f S(e^{j2\pi f})$ (in which case $D(e^{j2\pi f}) = \theta = D$) [1]. It follows that for distortion levels below S_{min} the RDF of a Gaussian source with memory is equal to the RDF of its memoryless innovation process:

$$R(D) = R_Z(D) = \frac{1}{2} \log \left(\frac{\sigma_Z^2}{D} \right), \quad D \leq S_{min}. \tag{11}$$

We shall see later in Section II how the observation above translates into a predictive test-channel, which can realize the RDF not only for small but for *all* distortion levels. This test channel is motivated by the sequential structure of Differential Pulse Code Modulation (DPCM) [12], [9]. The goal of DPCM is to translate the encoding of dependent source samples into a series of *independent* encodings. The task of removing the time dependence is achieved by linear prediction.

A negative result along this direction was recently given by Kim and Berger [13]. They showed that the RDF of an auto-regressive (AR) Gaussian process cannot be achieved by directly encoding its innovation process. This can be viewed as *open-loop* prediction, because the innovation process is extracted from the clean source rather than from the quantized source [12], [8]. Here we give a positive result, showing that the RDF can be achieved if we embed the quantizer inside the prediction loop, i.e., by *closed-loop* prediction. Specifically, we construct a system consisting of pre- and post-filters, and an AWGN channel embedded in a source prediction loop, such that the *scalar* (un-conditional) mutual information over this channel is equal to the RDF.

After presenting and proving our main result in Sections II and III, respectively, we discuss its characteristics and operational implications. Section IV discusses the spectral features of the solution. Section V relates the solution to vector-quantized DPCM of parallel sources. Section VI discusses implementation by Entropy Coded Dithered Quantization (ECDQ). Finally, in Section VII we relate prediction in source coding to prediction for channel equalization and to recent observations by Forney [6]. Like in [6], our analysis is based on the properties of information measures; the only necessary result from Wiener estimation theory is the orthogonality principle.

II. Main Result

Consider the system in Figure 2, which consists of three basic blocks: pre-filter $H_1(e^{j2\pi f})$, a noisy channel embedded in a closed loop, and a post-filter $H_2(e^{j2\pi f})$. The system parameters are derived from the water-filling solution (1)-(2). The source samples $\{X_n\}$ are passed through a pre-filter, whose phase is arbitrary and its absolute squared frequency response is given by

$$|H_1(e^{j2\pi f})|^2 = 1 - \frac{D(e^{j2\pi f})}{S(e^{j2\pi f})} \tag{12}$$

where $\frac{0}{0}$ is taken as 1. The pre-filter output, denoted U_n, is being fed to the central block which generates a process V_n

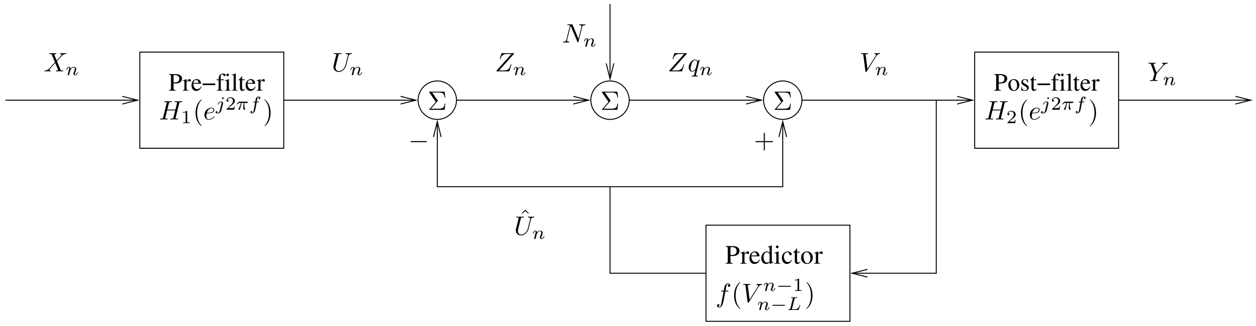

Fig. 2. Predictive Test Channel

according to the following recursion equations:

$$\hat{U}_n = f(V_{n-1}, V_{n-2}, \ldots, V_{n-L}) \quad (13)$$

$$Z_n = U_n - \hat{U}_n \quad (14)$$

$$Zq_n = Z_n + N_n \quad (15)$$

$$V_n = \hat{U}_n + Zq_n \quad (16)$$

where $N_n \sim \mathcal{N}(0, \theta)$ is a zero-mean AWGN, independent of the input process $\{U_n\}$, whose variance is equal to the water level θ; and $f(\cdot)$ is some prediction function for the input U_n given the L past samples of the output process $(V_{n-1}, V_{n-2}, \ldots, V_{n-L})$. Finally, the post-filter frequency response is the complex conjugate of the frequency response of the pre-filter,

$$H_2(e^{j2\pi f}) = H_1^*(e^{j2\pi f}). \quad (17)$$

The block from U_n to V_n is equivalent to the configuration of DPCM, [12], [9], with the DPCM quantizer replaced by the AWGN channel $Zq_n = Z_n + N_n$. In particular, the recursion equations (13)-(16) imply that this block satisfies the well known "DPCM error identity", [12],

$$V_n = U_n + (Zq_n - Z_n) = U_n + N_n. \quad (18)$$

That is, the output V_n is a noisy version of the input U_n via the AWGN channel $V_n = U_n + N_n$. In DPCM the prediction function f is linear:

$$f(V_{n-1}, \ldots, V_{n-L}) = \sum_{i=1}^{L} a_i V_{n-i} \quad (19)$$

where $a_1, \ldots, a_L$ minimize the mean-squared error

$$E\left(U_n - \sum_{i=1}^{L} a_i V_{n-i}\right)^2. \quad (20)$$

Because V_n is the result of passing U_n through an AWGN channel, we call that "noisy prediction".

If $\{U_n\}$ and $\{V_n\}$ are jointly Gaussian, then the best predictor of any order is linear, so the minimum of (20) is also the MMSE in estimating U_n from the vector $(V_{n-1}, \ldots, V_{n-L})$. In the limit as $L \to \infty$, this becomes the infinite order prediction error in U_n, given the infinite past

$$V_n^- \triangleq \{V_{n-1}, V_{n-2}, \ldots\}. \quad (21)$$

We shall further elaborate on the relationship with DPCM later.

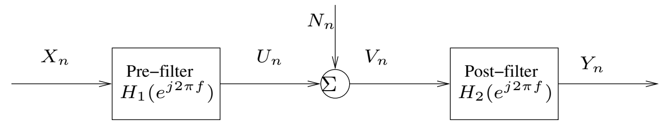

Fig. 3. Equivalent Channel

Note that while the central block is sequential and hence causal, the pre- and post-filters are non-causal and therefore their realization in practice requires large delay. Our main result is the following.

Theorem 1: For any stationary source with power spectrum $S(e^{j2\pi f})$, the system of Figure 2 satisfies

$$E(Y_n - X_n)^2 = D. \quad (22)$$

Furthermore, if the source X_n is Gaussian and $f(V_n^-)$ is the optimum infinite order predictor, then

$$I(Z_n; Z_n + N_n) = R(D), \quad (23)$$

i.e., the *scalar* mutual information over the channel (15) is equal to the RDF.

The proof is given in Section III. The result above is in sharp contrast to the classical realization of the RDF (5), which involves mutual information rate over a test-channel with memory. In a sense, the core of the encoding process in the system of Figure 2 amounts to a *memoryless AWGN test-channel*, although, as we discuss in the sequel, the channel (15) is not quite memoryless nor additive. From a practical perspective, this system provides a bridge between DPCM and rate-distortion theory for a general distortion level $D > 0$.

Another interesting feature of the system is the relationship between the prediction error process Z_n and the original process X_n. If X_n is an auto-regressive (AR) process, then in the limit of small distortion ($D \to 0$), Z_n is roughly its innovation process. Hence, unlike in open-loop prediction [13], encoding the innovations in a closed-loop system is optimal in the limit of high-resolution encoding. We shall return to that, as well as discuss the case of general resolution, in Section IV.

III. Proof of Main Result

By the DPCM error identity (18), the entire system of Figure 2 is equivalent to the system depicted in Figure 3, consisting of a pre-filter (12), an AWGN channel with noise variance θ, and a post-filter (17). This is also the forward channel realization (5) of the RDF [1], [17]. In particular, as simple spectral analysis shows, the power spectrum of the overall error process $Y_n - X_n$ is equal to the water filling distortion spectrum $D(e^{j2\pi f})$ in (2). Hence, by (3) the total distortion is D, and (22) follows.

For the second part, since the system of Figure 3 coincides with the forward channel realization (5), for a Gaussian source we have

$$\overline{I}(\{X_n\};\{Y_n\}) = \overline{I}(\{U_n\};\{V_n\}) = R(D) \tag{24}$$

where $\overline{I}$ denotes mutual information-rate between jointly stationary sources:

$$\overline{I}(\{X_n\};\{Y_n\}) = \lim_{n\to\infty} \frac{1}{n} I(X_1,\ldots,X_n;Y_1,\ldots,Y_n). \tag{25}$$

Note that stationarity implies that $I(\{U_n\};V_i|V_i^-)$ is independent of i, where $V_i^- = \{V_{i-1},V_{i-2},\ldots\}$ is the infinite past of the ith sample. Hence, by the chain rule for mutual information,

$$\overline{I}(\{U_n\};\{V_n\}) = I(\{U_n\};V_i|V_i^-) \tag{26}$$

for any i, so the second part of the theorem is equivalent to

$$I(\{U_n\};V_i|V_i^-) = I(Z_n;Z_n+N_n). \tag{27}$$

To show that, we write

$$\begin{aligned} I(\{U_n\};V_i|V_i^-) &= I(\{U_n\},U_i - f(V_i^-);V_i - f(V_i^-)|V_i^-) \\ &= I(\{U_n\},Z_i;Z_i+N_i|V_i^-) \\ &= I(Z_i;Z_i+N_i|V_i^-) \qquad (28) \\ &= I(Z_i;Z_i+N_i) \qquad (29) \end{aligned}$$

where (28) follows since N_i is independent of $(\{U_n\},V_i^-)$ and therefore $Z_i+N_i \Leftrightarrow (Z_i,V_i^-) \Leftrightarrow \{U_n\}$ form a Markov chain; and (29) follows since by the orthogonality principle, the prediction error Z_i is orthogonal to the measurements V_i^-, so by Gaussianity they are also independent. In view of (24) and (26) the proof is completed.

An alternative proof of Theorem 1, based only on spectral considerations, is given in the end of the next section.

IV. Properties of the Predictive Test-Channel

The following observations shed light on the behavior of the test channel of Figure 2.

Prediction in the high resolution regime. If the power-spectrum $S(e^{j2\pi f})$ is everywhere positive (e.g., if $\{X_n\}$ can be represented as an AR process), then in the limit of small distortion $D \to 0$, the pre- and post-filters (12), (17) converge to all-pass filters, and the power spectrum of U_n becomes the power spectrum of the source X_n. Furthermore, noisy prediction of U_n (from the "noisy past" V_n^-, where $V_n = U_n + N_n$) becomes equivalent to clean prediction of U_n (from its own past U_n^-). Hence, in this limit the prediction error Z_n is equivalent to the innovation process of X_n. In particular, Z_n is an i.i.d. process whose variance is $P_e(X)$ = the entropy-power of the source (7).

Prediction in the general case. Interestingly, for general distortion $D > 0$, the prediction error Z_n is *not white*, as the noisiness of the past does not allow the predictor f to remove all the source memory. Nevertheless, the noisy version of the prediction error $Zq_n = Z_n + N_n$ is white for every $D > 0$, because it amounts to predicting V_n from its *own* infinite past: since N_n has zero-mean and is white (and therefore independent of the past), $\hat{U}_n$ is also the optimal predictor for $V_n = U_n + N_n$.

The whiteness of Zq_n might seem at first a contradiction, because Zq_n is the sum of a non-white process Z_n and a white process N_n; nevertheless, $\{Z_n\}$ and $\{N_n\}$ are *not* independent, because Z_n depends on past values of N_n through the past of V_n. Thus, the channel $Zq_n = Z_n + N_n$ is not quite additive, but "sequentially additive": each new noise sample is independent of the past but not necessarily of the future. In particular, this channel satisfies:

$$I(Z_n;Z_n{+}N_n|Z_1{+}N_1,\ldots,Z_{n-1}{+}N_{n-1}) = I(Z_n;Z_n{+}N_n)\,, \tag{30}$$

so by the chain rule for mutual information

$$\bar{I}(\{Z_n\};\{Z_n+N_n\}) > I(Z_n;Z_n+N_n)\ .$$

Later in Section VI we rewrite (30) in terms of directed mutual information.

The channel when the SLB is tight. As long as D is smaller than the lowest point of the source power spectrum, we have $D(e^{j2\pi f}) = \theta = D$ in (1), and the quadratic-Gaussian RDF coincides with the SLB (10). In this case, the following properties hold for the predictive test channel:

- The power spectra of U_n and Y_n are the same and are equal to $S(e^{j2\pi f}) - D$.
- The power spectrum of V_n is equal to the power spectrum of the source $S(e^{j2\pi f})$.
- Since as discussed above the (white) process $Zq_n = Z_n + N_n$ is the optimal prediction error of the process V_n from its *own* infinite past, the variance of Zq_n is equal to the entropy-power (7) of V_n, which is equal to $P_e(X)$.
- As a consequence we have

  $$\begin{aligned} I(Z_n;Z_n+N_n) &= h\Big(\mathcal{N}(0,P_e(X))\Big) - h\Big(\mathcal{N}(0,D)\Big) \\ &= \frac{1}{2}\log\Big(\frac{P_e(X)}{D}\Big) \end{aligned}$$

  which is indeed the SLB (10).

As discussed in the Introduction, the SLB is also the RDF of the innovation process (11), i.e., the conditional RDF of the source X_n given its infinite *clean* past X_n^-.

An alternative derivation of Theorem 1 in the spectral domain. For a general D, we can use the observation above and the equivalent channel of Figure 3 to re-derive the scalar mutual information - RDF idnetity (23). Note that

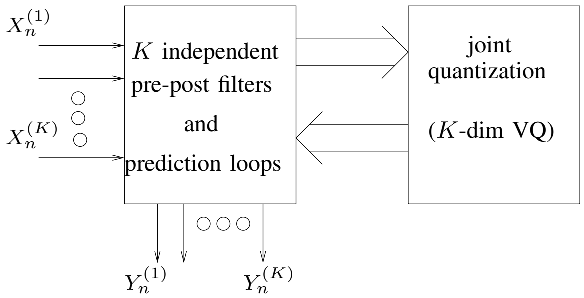

Fig. 4. DPCM of parallel sources.

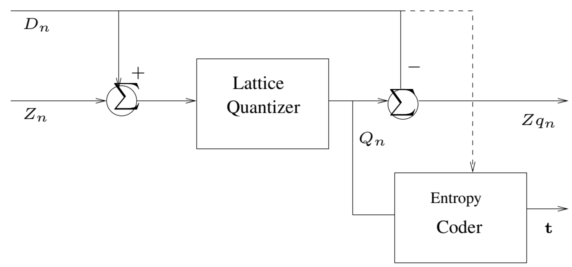

Fig. 5. ECDQ Structure

for any D the power spectrum of U_n and Y_n is equal to $\max\{0, S(e^{j2\pi f}) - \theta\}$. Thus the power spectrum of $V_n = U_n + N_n$ is given by $\max\{\theta, S(e^{j2\pi f})\}$. Since as discussed above the variance of $Zq_n = Z_n + N_n$ is given by the entropy power of the process V_n, we have

$$\begin{aligned} I(Z_n; Z_n + N_n) &= \frac{1}{2}\log\Big(\frac{P_e(\max\{\theta, S(e^{j2\pi f})\})}{\theta}\Big) \\ &= R(D) \end{aligned}$$

where $P_e(\cdot)$ as a function of the spectrum is given in (7), and the second equality follows from (1).

V. Vector-Quantized DPCM and D*PCM

As mentioned earlier, the structure of the central block of the channel of Figure 2 is of a DPCM quantizer[1], with the scalar quantizer replaced by the AWGN channel $Zq_n = Z_n + N_n$. However, if we wish to implement the additive noise by a quantizer whose rate is the mutual information $I(Z_n; Z_n + N_n)$, we must use *vector* quantization (VQ). For example, good high dimensional lattices generate near Gaussian quantization noise [18]. Yet how can we combine VQ and DPCM without violating the sequential nature of the system? In particular, the quantized sample Zq_n must be available before the system can predict U_{n+1} and generate Z_{n+1}.

One way we can achieve the VQ gain is by adding a "spatial" dimension, i.e., by jointly encoding a large number of *parallel* sources, as happens, e.g., in video coding. Figure 4 shows DPCM encoding of K parallel sources. The spectral shaping and prediction are done in the time domain for each source separately. Then, the resulting vector of K prediction errors is quantized jointly at each time instant by a vector quantizer. The desired properties of additive quantization error, and rate which is equal to K times the mutual information $I(Z_n; Z_n + N_n)$, can be approached in the limit of large K by a suitable choice of the quantizer. In the next section we discuss one way to do that using lattice ECDQ.

If we have only one source with decaying memory, we can still approximate the parallel source coding approach above, at the cost of large delay, by using interleaving. The source is divided into K long blocks which are jointly encoded, as if they were parallel sources as above. This is analogous to the method used in [11] for combining coding-decoding and decision-feedback equalization (DFE).

If we do not use any of the above, but restrict ourselves to scalar quantization ($K = 1$), then we have a pre/post filtered DPCM scheme. By known bounds on the performance of entropy-constrained scalar quantizers (e.g., [17]), we have

$$H(Q^{opt}(Z_n)) \le R(D) + \frac{1}{2}\log\Big(\frac{2\pi e}{12}\Big)$$

where $1/2\log(2\pi e/12) \approx 0.254$ bit. Hence, Theorem 1 implies that in principle, a pre/post filtered DPCM scheme is optimal, up to the loss of the VQ gain, at all distortion levels and not only at the high resolution regime.

It is interesting to mention that "open loop" prediction for quantization, which we mentioned earlier regarding the model of [13], is known in the quantization literature as D*PCM [12]. The best pre-filter for D*PCM under a high resolution assumption turns out to be the "half-whitening filter": $|H_1(e^{j2\pi f})|^2 = 1/\sqrt{S(e^{j2\pi f})}$. But even with this filter, D*PCM is still inferior to DPCM.

VI. ECDQ in a Closed Loop System

An ECDQ (entropy-coded dithered quantizer) operating on the source Z_n is depicted in Figure 5. A dither sequence D_n, independent of the input sequence Z_n, is added before the quantization and subtracted after. If the quantizer has a lattice structure of dimension $K \ge 1$, then we assume that the sequence length is

$$L = MK$$

for some integer M, so the quantizer is activated M times. At each quantizer operation instant m, a dither vector $\mathbf{D}_m$ is independently and uniformly distributed over the basic lattice cell. The lattice points at the quantizer output $\mathbf{Q}_m,\ m = 1, \ldots, M$ are fed into an entropy coder which is allowed to do joint coding of the sequence, and has knowledge of the dither as well, thus for an input sequence of length L it achieves a rate of:

$$R_{ECDQ} \triangleq \frac{1}{L} H(\mathbf{Q}_1^M | \mathbf{D}_1^M) \tag{31}$$

bit per source sample. The entropy coder produces a sequence $\mathbf{s}$ of LR_{ECDQ} bits, from which the decoder can recover $\mathbf{Q}_1, \ldots \mathbf{Q}_M$, and then subtract the dither to obtain the reconstruction sequence $Zq_n = Q_n - D_n,\ n = 1, \ldots L$. The reconstruction error sequence

$$N_n = Zq_n - Z_n$$

[1] According to R.M. Gray, [10], DPCM was first introduced in a U.S. patent by C.C. Cutler in 1952.

has K-blocks which are uniformly distributed over the mirror image of the basic lattice cell and mutually i.i.d. [16]. It is further stated in [16, Thm.1] that the sequences $\mathbf{Z}$ and $\mathbf{N}$ are statistically independent, and that the ECDQ rate is equal to the mutual information over an additive noise channel with the input $\mathbf{Z}$ and the noise $\mathbf{N}$:

$$\begin{aligned} R_{ECDQ} &= \frac{1}{L} I(Z_1^L; Zq_1^L) \\ &= \frac{1}{L} I(Z_1^L; Z_1^L + N_1^L) \quad . \end{aligned} \tag{32}$$

However, the derivation of [16, Thm. 1] depends on the implicit assumption that the quantizer is *used without feedback*, that is, the current input is conditionally independent of past outputs given the past inputs. When there is feedback, this condition does not hold, which implies that even with dither the sequences $\mathbf{Z}$ and $\mathbf{N}$ become dependent. Specifically, for the realistic setting of causal feedback, the input can depend on past values of the ECDQ noise, therefore the joint distribution of the ECDQ input and noise may take any joint distribution of the form:

$$f_{\mathbf{Z},\mathbf{N}}(z_1^L, n_1^L) = \prod_{m=0}^{M-1} f_{\mathbf{N}}(n_{mK+1}^{(m+1)K}) f_{\mathbf{Z}|\mathbf{N}}(z_{mK}^{(m+1)K-1} | n_1^{mK-1}) \quad . \tag{33}$$

In this case, the mutual information rate of (32) over-estimates the true rate of the ECDQ. Massey shows in [15] that for DMCs with feedback, traditional mutual information is not a suitable measure, and should be replaced by *directed information*. The directed information between the sequences $\mathbf{Z}$ and $\mathbf{Zq}$ is defined as

$$I(\mathbf{Z} \to \mathbf{Zq}) \triangleq \sum_{n=1}^{L} I(Z_n; Zq_n | Zq_1^{n-1}) \quad . \tag{34}$$

For our purposes, we will define the K-block directed information:

$$I_K(\mathbf{Z} \to \mathbf{Zq}) \triangleq \sum_{m=1}^{M} I(\mathbf{Z}_m; \mathbf{Zq}_m | \mathbf{Z}q_1^{m-1}) \tag{35}$$

where $\mathbf{Z}_m = Z_{(m-1)K+1}^{mK}$ denotes the mth K-block, and similarly for $\mathbf{Zq}_m$. The following result, proven in Appendix A, extends Massey's observation to ECDQ with feedback, and generalizes the result of [16, Thm. 1]:

Theorem 2: The ECDQ system with causal feedback defined by (33) satisfies:

$$R_{ECDQ} = \frac{1}{L} I_K(\mathbf{Z} \to \mathbf{Zq}) \quad . \tag{36}$$

Remarks:

1. When there is no feedback, the past and future input blocks $(\mathbf{Z}_1^{m-1}, \mathbf{Z}_{m+1}^{M})$ are conditionally independent of the current output block $\mathbf{Z}q_m$ given the current input block $\mathbf{Z}_m$, implying by the chain rule that

$$I_K(\mathbf{Z} \to \mathbf{Zq}) = I(\mathbf{Z}_1^M; \mathbf{Z}q_1^M)$$

and Theorem 2 reduces to [16, Thm. 1].

2. Even for scalar quantization ($K = 1$), the ECDQ rate (31) refers to joint entropy coding of the whole input vector. This does not contradict the sequential nature of the system, since the entropy coder is not a part of the feedback. Furthermore, it follows from the chain rule for entropy that it is enough to encode the instantaneous quantizer output $\mathbf{Q}_m$ conditioned on past quantizer outputs $\mathbf{Q}_1^{m-1}$ and on past and present dither samples $\mathbf{D}_1^m$.

3. We can embed a K-dimensional lattice ECDQ for $K > 1$ in the predictive test channel of Figure 2, instead of the additive noise channel, using the Vector-DPCM configuration discussed in the previous section. For good lattices, when the quantizer dimension $K \to \infty$, the noise process $\mathbf{N}$ in fact becomes white Gaussian and the scheme achieves the rate-distortion function. Combining Theorems 1 and 2, we see that:

$$R(D) = R_{ECDQ} = I(Z_n; Z_n + N_n) \quad ,$$

thus the entropy coder does not need to be conditioned on the past at all, as the predictor handles all the memory. However, when the quantization noise is not Gaussian, or the predictor is not optimal, the entropy coder uses the residual time-dependence after prediction. The resulting rate of the ECDQ would be the average directed information between the source and its reconstruction as stated in the Theorem.

VII. A Dual Relationship with Decision-Feedback Equalization

We consider the (real-valued) discrete-time time-invariant linear Gaussian channel,

$$R_n = c_n * S_n + Z_n, \tag{37}$$

where the transmitted signal S_n is subject to a power constraint $E[S_n^2] \leq P$, and where Z_n is (possibly colored) Gaussian noise.

Let X_n represent the data stream which we model as an i.i.d. zero-mean Gaussian random process with variance σ_x^2. Further, let h_n be a spectral shaping filter, satisfying

$$\sigma_x^2 \int_{-1/2}^{1/2} |H(e^{j2\pi f})|^2 df \leq P$$

so the channel input $S_n = h_n * X_n$ indeed satisfies the power constraint. For the moment, we make no further assumption on h_n.

The channel (38) has inter-symbol interference (ISI) due to the channel filter c_n, as well as colored Gaussian noise. Let us assume that the channel frequency response is non-zero everywhere, and pass the received signal R_n through a zero-forcing (ZF) linear equalizer $\frac{1}{C(z)H(z)}$, resulting in Y_n. We thus arrive at an equivalent ISI-free channel,

$$Y_n = X_n + N_n, \tag{38}$$

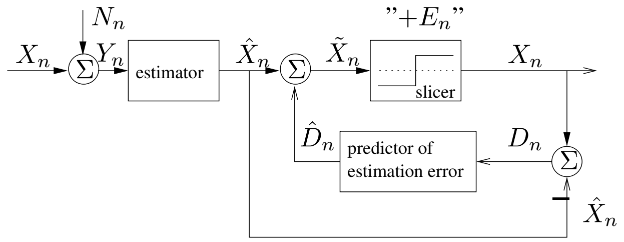

Fig. 6. MMSE-DFE in predictive form.

where the power spectrum of N_n is

$$S_{NN}(e^{j2\pi f}) = \frac{S_{ZZ}(e^{j2\pi f})}{|C(e^{j2\pi f})H(e^{j2\pi f})|^2}.$$

The mutual information (normalized per symbol) between the input and output of the channel (38) is

$$\overline{I}(\{X_n\},\{Y_n\}) = \int_{-1/2}^{1/2} \frac{1}{2}\log\left(1+\frac{\sigma_x^2}{S_{NN}(e^{j2\pi f})}\right) df. \quad (39)$$

We note that if the spectral shaping filter h_n satisfies the water-filling condition, then (39) will equal the channel capacity [4].

Similarly to the observations made in Section I with respect to the RDF, we note (as reflected in (39)) that capacity may be achieved by parallel AWGN coding over narrow frequency bands (as done in practice in Discrete Multitone (DMT)/Orthogonal Frequency-Division Multiplexing (OFDM) systems). An alternative approach, based on time-domain prediction rather than the Fourier transform, is offered by the canonical MMSE - feed forward equalizer - decision feedback equalizer (FFE-DFE) structure used in single-carrier transmission. It is well known that this scheme, coupled with AWGN coding, can achieve the capacity of linear Gaussian channels. This has been shown using different approaches by numerous authors; see [11], [3], [14], [6] and references therein. Our exposition closely follows that of Forney [6]. We now recount this result, based on linear prediction of the error sequence; see Figure 6. In the communication literature, this structure is referred to as "noise prediction". It can be recast into the more familiar FFE-DFE form by absorbing a part of the predictor into the estimator filter, forming the usual FFE.

As a first step, let $\hat{X}_n$ be the optimal MMSE estimator of X_n from the equivalent channel output sequence $\{Y_n\}$ of (38). Since $\{X_n\}$ and $\{Y_n\}$ are jointly Gaussian and stationary this estimator is linear and time invariant. (Note that the combination of the ZF equalizer $\frac{1}{C(z)H(z)}$ at the receiver front-end and the estimator above is equivalent to direct MMSE estimation of X_n from the original channel output R_n.)

Denote the estimation error, which is composed in general of ISI and Gaussian noise, by D_n. Then

$$X_n = \hat{X}_n + D_n \quad (40)$$

where $\{D_n\}$ is independent of $\{\hat{X}_n\}$ due to the orthogonality principle and Gaussianity.

Assuming correct decoding of past symbols[2], the decoder knows the past samples $D_{n-1}, D_{n-2}, \ldots$ and may form an optimal predictor, $\hat{D}_n$, of the current estimation error D_n, which may then be added to $\hat{X}_n$ to form $\tilde{X}_n$. The prediction error $E_n = D_n - \hat{D}_n$ has variance $P_e(D)$, the entropy power of D_n. It follows that

$$\begin{aligned} X_n &= \hat{X}_n + D_n \\ &= \tilde{X}_n - \hat{D}_n + D_n \\ &= \tilde{X}_n + E_n, \end{aligned} \quad (41)$$

where $\{\tilde{X}_n\}$ and $\{E_n\}$ are statistically independent. Hence the residual estimation error satisfies

$$E\{X_n - \tilde{X}_n\}^2 = E\{D_n - \hat{D}_n\}^2 = P_e(D). \quad (42)$$

The channel (41), which describes the input/output relation of the slicer in Figure 6, is often referred to as the *backward channel*. Furthermore, since X_n and E_n are i.i.d Gaussian and since E_n is independent of present and past values of $\tilde{X}_n$ (but dependent of future values), it is a "sequentially additive" AWGN channel. Notice the strong resemblance with the channel (15), $Zq_n = Z_n + N_n$, in the predictive test-channel of the RDF: in both channels the output and the noise are i.i.d. and Gaussian, but the input has memory due to the feedback loop.

We have therefore derived the following.

Theorem 3: The mutual information of the channel (38) is equal to the scalar mutual information

$$I(\tilde{X}_n; \tilde{X}_n + E_n)$$

of the channel (41).

Proof: Let $X_n^- = \{X_{n-1}, X_{n-2}, \ldots\}$ and $D_n^- = \{D_{n-1}, D_{n-2}, \ldots\}$. Using the chain rule of mutual information we have

$$\begin{aligned} \overline{I}(\{X_n\},\{Y_n\}) &= \overline{h}(\{X_n\}) - \overline{h}(\{X_n\}|\{Y_n\}) \\ &= \overline{h}(\{X_n\}) - h(X_n|\{Y_n\}, X_n^-) \\ &= \overline{h}(\{X_n\}) - h(X_n - \hat{X}_n|\{Y_n\}, X_n^-) \\ &= \overline{h}(\{X_n\}) - h(D_n|\{Y_n\}, X_n^-) \\ &= \overline{h}(\{X_n\}) - h(D_n|\{Y_n\}, D_n^-) \\ &= \overline{h}(\{X_n\}) - h(D_n - \hat{D}_n|\{Y_n\}, D_n^-) \\ &= \overline{h}(\{X_n\}) - h(E_n|\{Y_n\}, D_n^-) \\ &= \overline{h}(\{X_n\}) - h(E_n) \quad (43) \\ &= I(\tilde{X}_n; \tilde{X}_n + E_n), \end{aligned}$$

where (43) follows from successive application of the orthogonality principle [6]. ■

[2]Here we must actually break with assumption that X_n is a Gaussian process. We implicitly assume that X_n are symbols of a capacity-achieving AWGN code. The slicer should be viewed as a mnemonic aid where in practice an optimal decoder should be used. Furthermore, the use of an interleaver and long delay is necessary. See [6] and [11].

It follows that

$$\int_{-1/2}^{1/2} \frac{1}{2}\log\left(1+\frac{\sigma_x^2}{S_{NN}(e^{j2\pi f})}\right)df = \frac{1}{2}\log\left(\frac{\sigma_x^2}{P_e(D)}\right). \quad (44)$$

As a corollary from (39), Theorem 3 and (44), we obtain the following well known result from Wiener theory,

$$P_e(D) = \exp\left(\int_{-1/2}^{1/2}\log\left(\frac{\sigma_x^2 S_{NN}(e^{j2\pi f})}{\sigma_x^2+S_{NN}(e^{j2\pi f})}\right)df\right).$$

Remarks:

Capacity achieving shaping filter. Let us define the equivalent noise $Z_{eq,n}$ as the noise Z_n passed through the filter $\frac{1}{C(Z)}$, so that

$$S_{Z_{eq}Z_{eq}}(e^{j2\pi f}) = \frac{S_{ZZ}(e^{j2\pi f})}{|C(e^{j2\pi f})|^2}.$$

Thus for a given spectral shaping filter h_n, the corresponding mutual information is,

$$\overline{I}(\{X_n\},\{Y_n\}) = \int_{-1/2}^{1/2}\frac{1}{2}\log\left(1+\frac{\sigma_x^2}{S_{NN}(e^{j2\pi f})}\right)df$$

$$= \int_{-1/2}^{1/2}\frac{1}{2}\log\left(\frac{\sigma_x^2|H(e^{j2\pi f})|^2+S_{Z_{eq}Z_{eq}}(e^{j2\pi f})}{S_{Z_{eq}Z_{eq}}(e^{j2\pi f})}\right)df \quad (45)$$

The shaping filter h_n which maximizes the mutual information (and yields capacity) is given by the parametric water-filling formula:

$$\sigma_x^2 H(e^{j2\pi f}) = \lceil \theta - S_{Z_{eq}Z_{eq}}(e^{j2\pi f})\rceil^+, \quad (46)$$

where the "water level" θ is chosen so that the power constraint is met with equality,

$$\sigma_s^2 = \int_{-1/2}^{1/2}\sigma_x^2 H(e^{j2\pi f})df \quad (47)$$

$$= \int_{-1/2}^{1/2}\lceil\theta - S_{Z_{eq}Z_{eq}}(e^{j2\pi f})\rceil^+ df \quad (48)$$

$$= P. \quad (49)$$

The linear estimator is unnecessary at high SNR. At high signal-to-noise ratio(SNR), the FFE estimation filter becomes all-pass and (up to phase variation) we have $\hat{X}_n \approx Y_n$. Specifically, the optimal estimation (Wiener) filter of X_n from $\{Y_n\}$ is

$$\frac{\sigma_x^2}{\sigma_x^2+S_{NN}(e^{j2\pi f})}. \quad (50)$$

Thus, if $\sigma_x^2 \gg S_{NN}(e^{j2\pi f})$ for all f (high SNR), then the magnitude of the estimation filter becomes identity.

The prediction process when the Shannon upper bound is tight. The Shannon upper bound (SUB) on capacity states that

$$C \le \frac{1}{2}\log(2\pi e\sigma_y^2) - \overline{h}(Z_{eq}) \quad (51)$$

$$\le \frac{1}{2}\log\left(\frac{P+\sigma_{Z_{eq}}^2}{P_e(Z_{eq})}\right) \quad (52)$$

$$\triangleq C_{\mathrm{SUB}}, \quad (53)$$

where

$$\sigma_{Z_{eq}}^2 = \int_{-1/2}^{1/2} S_{Z_{eq}Z_{eq}}(e^{j2\pi f})df$$

is the variance of the equivalent noise, and where equality holds if and only if the output Y_n is white. This in turn is satisfied if and only if

$$\theta \ge \max_f S_{Z_{eq}Z_{eq}}(e^{j2\pi f}).$$

The estimation error D_n corresponding to the optimal Wiener estimation filter satisfies,

$$S_{DD}(e^{j2\pi f}) = \frac{\sigma_x^2 S_{NN}(e^{j2\pi f})}{\sigma_x^2+S_{NN}(e^{j2\pi f})} \quad (54)$$

$$= \frac{\sigma_x^2 S_{Z_{eq}Z_{eq}}(e^{j2\pi f})}{\sigma_s^2+S_{Z_{eq}Z_{eq}}(e^{j2\pi f})}.$$

When the SUB is tight $\sigma_s^2 + S_{Z_{eq}Z_{eq}}(e^{j2\pi f})$ is equal to θ for all frequencies f. Hence, the power spectrum of the estimation error D_n is proportional to the equivalent noise spectrum $S_{Z_{eq}Z_{eq}}(e^{j2\pi f})$. We further have that $\sigma_s^2 = P$, and

$$S_{DD}(e^{j2\pi f}) = \frac{\sigma_x^2}{P+\sigma_{Z_{eq}}^2}\cdot S_{Z_{eq}Z_{eq}}(e^{j2\pi f}), \quad (55)$$

and therefore

$$P_e(D) = \frac{\sigma_x^2}{P+\sigma_{Z_{eq}^2}}\cdot P_e(Z_{eq}). \quad (56)$$

It follows that the mutual information of the channel (41), i.e., the mutual information at the slicer, satisfies

$$I(\tilde{X}_n;\tilde{X}_n+E_n) = \frac{1}{2}\log\frac{\sigma_x^2}{\mathrm{Var}(E_n)} \quad (57)$$

$$= \frac{1}{2}\log\left(\frac{\sigma_x^2}{P_e(D)}\right) \quad (58)$$

$$= \frac{1}{2}\log\left(\frac{P+\sigma_{Z_{eq}^2}}{P_e(Z_{eq})}\right) \quad (59)$$

$$= C_{\mathrm{SUB}}. \quad (60)$$

VIII. Summary

We demonstrated the dual role of prediction in rate-distortion theory of Gaussian sources and capacity of ISI channels. These observations shed light on the configurations of DPCM (for source compression) and FFE-DFE (for channel demodulation), and show that in principle they are "information lossless" for any distortion / SNR level. The theoretic bounds, RDF and capacity, can be approached in practice by appropriate use of feedback and linear estimation in the time domain combined with coding across the "spatial" domain.

It is tempting to ask whether the predictive form of the RDF can be extended to more general sources and distortion measures (and similarly for capacity of more general ISI channels). Yet, examination of the arguments in our derivation reveals that it is strongly tied to the quadratic-Gaussian case:

- The orthogonality principle, implied by the MMSE criterion, guarantees whiteness of the noisy prediction error Zq_n and un-correlation with the past.

- Gaussianity implies that un-correlation is equivalent to statistical independence.

For other error criteria and/or non-Gaussian sources, prediction (either linear or non-linear) is in general unable to remove the dependence on the past. Hence the scalar mutual information over the prediction error channel would in general be greater than the mutual information rate of the source before prediction.

APPENDIX

A. PROOF OF THEOREM 2

It will be convenient to look at K-blocks, which we denote by bold letters as in Section VI. Substituting the ECDQ rate definition (31), the required result (36) becomes:

$$H(\mathbf{Q}_0^{M-1}|\mathbf{D}_0^{M-1}) = \sum_{m=0}^{M-1} I(\mathbf{Z}_m;\mathbf{Z}q_m|\mathbf{Z}q_0^{m-1}) \quad .$$

Using the chain rule for entropies, it is enough to show that:

$$H(\mathbf{Q}_m|\mathbf{Q}_0^{m-1},\mathbf{D}_0^{M-1}) = I(\mathbf{Z}_m;\mathbf{Z}q_m|\mathbf{Z}q_0^{m-1}) \quad . \tag{61}$$

To that end, we have the following sequence of equalities:

$$\begin{aligned}
& H(\mathbf{Q}_m|\mathbf{Q}_0^{m-1},\mathbf{D}_0^{M-1}) \\
\overset{(a)}{=}\; & H(\mathbf{Q}_m|\mathbf{Q}_0^{m-1},\mathbf{D}_0^{m}) \\
\overset{(b)}{=}\; & H(\mathbf{Q}_m|\mathbf{Q}_0^{m-1},\mathbf{D}_0^{m}) - H(\mathbf{Q}_m|\mathbf{Q}_0^{m-1},\mathbf{Z}_0^{m},\mathbf{D}_0^{m}) \\
=\; & I(\mathbf{Q}_m;\mathbf{Z}_0^{m}|\mathbf{Q}_0^{m-1},\mathbf{D}_0^{m}) \\
\overset{(c)}{=}\; & I(\mathbf{Q}_m - D_m;\mathbf{Z}_0^{m}|\mathbf{Q}_0^{m-1},\mathbf{D}_0^{m}) \\
=\; & I(\mathbf{Z}q_m;\mathbf{Z}_0^{m}|\mathbf{Q}_0^{m-1},\mathbf{D}_0^{m}) \\
=\; & I(\mathbf{Z}q_m;\mathbf{Z}_0^{m}|\mathbf{Q}_0^{m-1}-\mathbf{D}_0^{m-1},\mathbf{D}_0^{m}) \\
=\; & I(\mathbf{Z}q_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m-1},\mathbf{D}_0^{m}) \\
\overset{(d)}{=}\; & I(\mathbf{Z}q_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m-1},\mathbf{D}_m) \\
\overset{(e)}{=}\; & I(\mathbf{Z}q_m, D_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m-1},\mathbf{D}_m) \\
& -I(D_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m-1},\mathbf{D}_m) \\
\overset{(f)}{=}\; & I(\mathbf{Z}q_m, D_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m-1},\mathbf{D}_m) \\
\overset{(g)}{=}\; & I(\mathbf{Z}q_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m-1},\mathbf{D}_m) \\
& +I(D_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m},\mathbf{D}_m) \\
\overset{(h)}{=}\; & I(\mathbf{Z}q_m;\mathbf{Z}_0^{m}|\mathbf{Z}q_0^{m-1}) \quad .
\end{aligned}$$

In this sequence, equality (a) comes from the independent dither generation and causality of feedback. (b) is justified because $\mathbf{Q}_m$ is a deterministic function of the elements on which the subtracted entropy is conditioned, thus entropy is 0. In (c) we subtract from the left hand side argument of the mutual information one of the variables upon which mutual information is conditioned. (d) holds because the independent generation of $\mathbf{D}$ and the feedback structure dictate a Markov relation $\mathbf{D}_0^{m-1} \leftrightarrow \mathbf{Z}q_0^{m-1} \leftrightarrow \mathbf{Z}_0^{m-1}$. For (e) and (g) we applied the chain rule for mutual information. (f) and (h) follow because dither values are independent of any past behavior of the system, and of present input and reconstruction.

REFERENCES

[1] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs, NJ, 1971.
[2] J. Chen, C. Tian, T. Berger, and S. S. Hemami. Multiple Description Quantization via Gram-Schmidt Orthogonalization. *Submitted to IEEE Transactions on Information Theory*.
[3] J.M. Cioffi, G.P. Dudevoir, M.V. Eyuboglu, and G.D. J. Forney. MMSE Decision-Feedback Equalizers and Coding - Part I: Equalization Results. *IEEE Trans. Communications*, COM-43:2582–2594, Oct. 1995.
[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
[5] U. Erez and R. Zamir. Achieving $\frac{1}{2}\log(1+SNR)$ on the AWGN channel with lattice encoding and decoding. *IEEE Trans. Information Theory*, IT-50:2293–2314, Oct. 2004.
[6] G. D. Forney, Jr. Shannon meets Wiener II: On MMSE estimation in successive decoding schemes. In *42st Annual Allerton Conference on Communication, Control, and Computing, Allerton House, Monticello, Illinois*, Oct. 2004.
[7] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, N.Y., 1968.
[8] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression* Kluwer Academic Pub., Boston, 1992.
[9] G.D.Gibson, T.Berger, T.Lookabaugh, D.Lindbergh, and R.L.Baker. *Digital Compression for Multimedia: Principles and Standards*‘. Morgan Kaufmann Pub., San Fansisco, 1998.
[10] R. Gray. Private communication.
[11] T. Guess and M. K. Varanasi. An information-theoretic framework for deriving canonical decision-feedback receivers in gaussian channels. *IEEE Trans. Information Theory*, IT-51:173–187, Jan. 2005.
[12] N. S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, Englewood Cliffs, NJ, 1984.
[13] K. T. Kim and T. Berger. Sending a Lossy Version of the Innovations Process is Suboptimal in QG Rate-Distortion. In *Proceedings of ISIT-2005, Adelaide, Australia*, pages 209–213, 2005.
[14] J. Barry, E. A. Lee and D. G. Messerschmitt. *Digital Communication*. Kluwer Academic Press, 2004 (third edition).
[15] J. Massey. Causality, Feedback and Directed Information. In *Proc. IEEE Int. Symp. on Information Theory*, pages 303–305, 1990.
[16] R. Zamir and M. Feder. On universal quantization by randomized uniform / lattice quantizer. *IEEE Trans. Information Theory*, pages 428–436, March 1992.
[17] R. Zamir and M. Feder. Information rates of pre/post filtered dithered quantizers. *IEEE Trans. Information Theory*, pages 1340–1353, September 1996.
[18] R. Zamir and M. Feder. On lattice quantization noise. *IEEE Trans. Information Theory*, pages 1152–1159, July 1996.