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 inter-symbol interference 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. Indeed, for capacity, Cioffi et al [2] showed that \( C \) is equal to the scalar mutual-information over a slicer embedded in a decision-feedback loop. We show that a parallel result holds for the rate-distortion function: \( R(D) \) is equal to the scalar mutual-information over an 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 [5], [4].

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 \( X \to Y \) such that \( \frac{1}{n} \| Y - X \|^2 \leq D \). 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

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

where \( R_k = E\{X_nX_{n+k}\} \) is the autocorrelation function. The water-filling solution, illustrated in Figure 1, gives a parametric formula for the Gaussian RDF in terms of a parameter \( \theta \):

\[
R(D) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{2} \log \left( \frac{S(e^{j2\pi f})}{D(e^{j2\pi f})} \right) df = \int_{f: S(e^{j2\pi f}) > \theta} \frac{1}{2} \log \left( \frac{S(e^{j2\pi f})}{\theta} \right) df
\]

(1)

Fig. 1. The water-filling solution.

where \( D(e^{j2\pi f}) \) is the distortion spectrum

\[
D(e^{j2\pi f}) = \left\{ \begin{array}{ll}
\theta, & \text{if } S(e^{j2\pi f}) > \theta \\
S(e^{j2\pi f}), & \text{otherwise,}
\end{array} \right.
\]

(2)

and where we choose the water level \( \theta \) so that the total distortion is \( D \):

\[
D = \int_{-\frac{1}{2}}^{\frac{1}{2}} D(e^{j2\pi f}) df.
\]

(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

\[
\frac{1}{2} \log \left( \frac{\sigma^2}{D} \right).
\]

(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 [1, Section 4.5]. The most useful one for our purpose replaces \( \alpha \) and \( \beta \) above by linear time-invariant filters, while keeping the noise \( N \) as AWGN [12]:

\[
Y_n = h_{2,n} * (h_{1,n} * X_n + N_n)
\]

(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 whose frequency responses are given in (11) and (16) in the next section.

If we take a discrete approximation of (1),

\[
\sum \frac{1}{2} \log \left( \frac{S(e^{j2\pi f})}{D(e^{j2\pi f})} \right),
\]

then each component has the memoryless form of (4). Hence the frequency domain formula (1) amounts to encoding of parallel (independent) Gaussian sources, where source \( i \) is a memoryless Gaussian source \( X_i \sim N(0, S(e^{j2\pi f})) \) encoded at distortion level \( D(e^{j2\pi f}) \); see [3]. Indeed, practical frequency domain coding schemes such as Transform Coding and Sub-band Coding [7] 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 [3]; for a Gaussian source the entropy-power is given by

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

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

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

Note that the orthogonality principle of MMSE estimation implies that the error process associated with the optimum predictor,

\[
E_n = X_n - \sum_{i=1}^{\infty} a_i X_{n-i},
\]

known also as the innovation process, has zero mean and is white. See, e.g., [5]. From an information theoretic perspective, the entropy-power plays a role in the SLB,

\[
R(D) \geq \log \left( \frac{P_e(X)}{D} \right).
\]

Equality in the SLB holds if the distortion level is smaller than or equal to the lowest value of the power spectrum:

\[
D \leq \min_f S(e^{j2\pi f}) \quad \text{(in which case } D(e^{j2\pi f}) = \theta = D) \quad [1].
\]

Hence, in view of (4), for small distortion levels the RDF of a Gaussian source with memory is given by

\[
\frac{1}{2} \log \left( \frac{\text{var}(E_n)}{D} \right),
\]

the RDF of its memoryless innovation process.

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) [9], [7]. 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 [10]. 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 [9], [6]. 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 our main result in Section II, we provide reflections on the result and its operational implications. Section III discusses the spectral features of the solution. Section IV gives the solution an operational meaning in terms of vector-quantized DPCM of parallel sources. Finally, in Section V we relate prediction in source coding to prediction for channel equalization and to recent observations or Forney [5]. As in [5], the 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

\[
|H_1(e^{j2\pi f})|^2 = 1 - \frac{D(e^{j2\pi f})}{S(e^{j2\pi f})}.
\]

The pre-filter output, denoted \( U_n \), is being fed to the central block which generates a process \( V_n \) according to the following recursion equations:

\[
\hat{U}_n = f(V_{n-1}, V_{n-2}, \ldots, V_{n-L}) \quad (12)
\]

\[
Z_n = U_n - \hat{U}_n \quad (13)
\]

\[
Z q_n = Z_n + N_n \quad (14)
\]

\[
V_n = \hat{U}_n + Z q_n \quad (15)
\]

where \( N_n \sim N(0, \theta) \) is zero-mean additive white Gaussian noise (AWGN), independent of the input process \( \{U_n\} \), whose variance is equal to the water level \( \theta \); 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,

$$H_2(e^{j2\pi f}) = H_1^*(e^{j2\pi f}).$$  \hspace{1cm} (16)

The block from $U_n$ to $V_n$ is equivalent to the configuration of DPCM, [9], [7], with the DPCM quantizer replaced by the AWGN channel $Z_q = Z_n + N_n$. In particular, the recursion equations (12)-(15) imply that this block satisfies the well known “DPCM error identity”, [9],

$$V_n = U_n + (Z_q - Z_n) = U_n + N_n.$$  \hspace{1cm} (17)

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:

$$f(V_{n-1}, \ldots, V_{n-L}) = \sum_{i=1}^{L} a_i V_{n-i}.$$  \hspace{1cm} (18)

where $a_1, \ldots, a_L$ may be chosen to minimize the mean-squared error

$$E \left( U_n - \sum_{i=1}^{L} a_i V_{n-i} \right)^2.$$  \hspace{1cm} (19)

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 (19) is also the the MMSE in estimating $U_n$ from the vector $(V_{n-1}, \ldots, V_{n-L})$. We shall further elaborate on the relationship with DPCM later.

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

$$E(Y_n - X_n)^2 = D.$$  \hspace{1cm} (20)

Furthermore, if the source $X_n$ is Gaussian, and if the function $f(\cdot)$ is the optimum linear predictor of $U_n$ given the infinite past $(V_{n-1}, V_{n-2}, \ldots)$, i.e., $L \to \infty$ in (12), then for all $n$

$$I(Z_n; Z_n + N_n) = R(D).$$  \hspace{1cm} (21)

The proof will be given in the full paper. The main feature of Theorem 1 is the fact that the left-hand side of (21) is a *single letter* mutual information. This is in sharp contrast to the classical realization of the RDF (5), where both the input and the test-channel have memory. In a sense, the core of the encoding process in the system of Figure 2 amounts to a *memoryless input AWGN test-channel*. From a practical perspective, this system provides a bridge between DPCM and rate-distortion theory for a general distortion level $D > 0$.

### III. 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 $D \to 0$, noisy prediction of $U_n$ (from the “noisy past” $V_{n-}^-$) is equivalent to clean prediction of $U_n$ (from its own past $U_{n-}$). Hence, in this limit the prediction error $Z_n$ is equal to the innovation process associated with $U_n$. Since for small distortion the pre- and post-filters (11), (16) are roughly all-pass filters, $U_n$ has roughly the same power spectrum as $X_n$, hence $Z_n$ is in fact equivalent to the innovation process of $X_n$. In particular, $Z_n$ is an i.i.d. process whose variance is $P_e(X) = \text{entropy-power of the source}$.

**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 $Z_q = Z_n + N_n$ is white for every $D > 0$, because it amounts to predicting $V_n$ from its own infinite past: since $N_n$ is white and has zero-mean, the optimal predictor for $U_n$ is also the optimal predictor for $V_n = U_n + N_n$. (The whiteness of $Z_q$ might seem at first a contradiction, because $Z_q$ 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 $V_n$.) These observations imply that the channel $Z_q = Z_n + N_n$ satisfies

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

while in general

$$I(\{Z_n\}; \{Z_n + N_n\}) > I(Z_n; Z_n + N_n).$$
The channel when the Shannon lower bound holds. As long as $D$ is smaller than the lowest point of the source spectrum (i.e., $D(e^{j2\pi f}) = \theta = D$ in (1)), the quadratic-Gaussian RDF coincides with the Shannon Lower Bound (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 the (white) process $Z_{q,n} = Z_n + N_n$ is the optimal prediction error of the process $V_n$ from its own infinite past, the variance of $Z_{q,n}$ is equal to the entropy-power (7) of $V$, which is equal to $P_e(X)$.
- As a consequence we have

$$I(Z_n; Z_n + N_n) = h\left(N(0, P_e(X))\right) - h\left(N(0, D)\right) = 1/2 \log\left(\frac{P_e(X)}{D}\right)$$

which is indeed the Shannon lower bound (10).

Note that as long as the SLB is tight, the RDF of the source is equal to the conditional RDF of the sample $X_n$ given the infinite clean past of the source $X_n -$

$$R(D) = R_{X_n|X_n^\infty}(D).$$

Interestingly, the predictive test-channel achieves that while conditioning on the noisy past $V_n$.

IV. 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, with the scalar quantizer replaced by AWGN channel. 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). It is not possible to do that along the time domain, due to the sequential nature of the system above. Nevertheless, we can achieve the VQ gain by adding a “spatial” dimension, i.e., by jointly encoding a large number of parallel sources, as happens, e.g., in video coding.

Figure 3 shows DPCM encoding of $K$ parallel sources. The spectral shaping and prediction are done in the time domain for each of the $K$ sources 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 such as lattice entropy-coded dithered quantizer ECDFQ.

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. This is analogous to the method used in [8] 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. It follows from Theorem 1 and from known bounds on the performance of scalar quantizers (e.g., [12]), 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 high resolution. For example, in the context of entropy-coded quantization,

$$H\left(Q(Z_n)\right) \leq R(D) + \frac{1}{2} \log\left(\frac{2\pi e}{12}\right).$$

It is interesting to mention that in the quantization literature, the “open loop” prediction approach investigated in [10] is referred to as $D^1$-PCM [9].

V. A DUAL RELATIONSHIP WITH DECISION-FEEDBACK EQUALIZATION

Consider the (real-valued) discrete-time time-invariant linear Gaussian channel arising at the output of a sampled matched followed by a zero-forcing (ZF) linear equalizer,

$$Y_n = X_n + N_n,$$

where the noise $N_n$ has power spectral density $S_{NN}(e^{j2\pi f})$ and $X_n$ represents the data stream, which we here assume is an i.i.d. Gaussian random process with power $\sigma^2$. This channel model arises when $X_n$ is passed through a spectral shaping filter, the output of which is the input of a linear channel that introduces ISI and possibly colored Gaussian noise (and satisfies a prescribed power constraint); the channel output passes through a sampled matched filter then a ZF linear equalizer is applied.

The mutual information (normalized per symbol) between the input and output of the channel (22) is

$$I(X_n; Y_n) = \int_{-1/2}^{1/2} \frac{1}{2} \log \left(1 + \frac{\sigma^2}{S_{NN}(e^{j2\pi f})}\right) df.$$  (23)

We note that if the spectral shaping filter satisfies the water-filling condition, then (23) will equal the channel capacity.

Similarly to the observations made in Section I with respect to the RDF, we note (as reflected in (23)) 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-DFE equalization structure used in single-carrier transmission. It is well known that this scheme,
The channel (25), which describes the input/output relation of ISI and Gaussian noise, by orthogonality principle [5].

It follows that

\[
\int_{-1/2}^{1/2} \frac{1}{2} \log \left( 1 + \frac{\sigma^2}{\sigma^2_x + S_{NN}(\epsilon^{2\pi f})} \right) df = \frac{1}{2} \log \left( \frac{\sigma^2}{P_e(D)} \right). \tag{28}
\]

As a corollary from (23), Theorem 2 and (28), we obtain the following well known result from Wiener theory,

\[
P_e(D) = \exp \left( \int_{-1/2}^{1/2} \frac{1}{2} \log \left( \frac{\sigma^2}{\sigma^2_x + S_{NN}(\epsilon^{2\pi f})} \right) df \right). 
\]

REFERENCES

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

[2] 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.

[3] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, New York, 1991.

[4] U. Erez, S. Shamai (Shitz), and R. Zamir. Capacity and lattice strategies for cancelling known interference. IEEE Trans. Information Theory, submitted, June 2003.

[5] G. D. Forney, Jr. meets Wiener II: On MMSE estimation in successive decoding schemes. In 42nd Annual Allerton Conference on Communication, Control, and Computing, Monticello, Illinois, Oct. 2004.

[6] A. Gersho and R. M. Gray Vector Quantization and Signal Compression'. Kluwer Academic Pub., Boston, 1992.

[7] 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.

[8] 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.

[9] N. S. Jayant and P. Noll. Digital Coding of Waveforms. Prentice-Hall, Englewood Cliffs, NJ, 1984.

[10] 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.

[11] E. A. Lee and D. G. Messerschmitt. Digital Communication. Kluwer, 2nd ed. Boston, MA, 1994.

[12] R. Zamir and M. Feder. Information rates of pre/post filtered dithered quantizers. IEEE Trans. Info. Theory, pages 1340–1353, Sept. 1996.

Theorem 2: The mutual information of the channel (22) is equal to the scalar mutual information

\[
I(\hat{X}_n; \tilde{X}_n + E_n)
\]

of the memoryless channel (25).

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

\[
\mathbb{I}(X_n, Y_n) = \mathbb{I}(X_n) - \mathbb{I}(Y_n | X_n)
\]

where (27) follows from successive application of the orthogonality principle [5].