Information Rates of Pre/Post Filtered Dithered Quantizers *

Ram Zamir† Meir Feder‡

December 18, 1997

Abstract

We consider encoding of a source with a pre-specified second order statistics, but otherwise arbitrary, by Entropy Coded Dithered (lattice) Quantization (ECDQ) incorporating linear pre- and post-filters. In the design and analysis of this scheme we utilize the equivalent additive noise channel model of the ECDQ. For Gaussian sources and square error distortion measure, the coding performance of the pre/post filtered ECDQ approaches the rate-distortion function, as the dimension of the (optimal) lattice quantizer becomes large; actually, in this case the proposed coding scheme simulates the optimal *forward channel realization* of the rate-distortion function. For non-Gaussian sources and finite dimensional lattice quantizers, the coding rate exceeds the rate-distortion function by at most the sum of two terms: the "information divergence of the source from Gaussianity" and the "information divergence of the quantization noise from Gaussianity". Additional bounds on the excess rate of the scheme from the rate distortion function are also provided.

Key words: entropy coded dithered quantization, pre/post filtering, forward channel realization, divergence from Gaussianity.

*This research was supported in part by the Wolfson Research Awards administered by the Israel Academy of Science and Humanities. This work was partially presented at the 1993 IEEE Int. Symp. on Information Theory, San Antonio, TX, January 1993

†Dept. of Elect. & Comp. Eng., University of California at Santa Barbara, CA 93106 USA. This research was performed while this author was at Tel Aviv University, Israel.

‡Department of Electrical Engineering - Systems, Tel-Aviv University, Tel-Aviv, 69978, ISRAEL

I Introduction

Many coding techniques for continuous sources incorporate linear operations such as sampling, filtering, prediction, error-feedback or spectral transformations (see, e.g. [11] for a good survey). In this context it is usually assumed that some of the spectral properties of the source (e.g., its bandwidth) are known, and a mean squared error (MSE) criterion or, more generally, a frequency weighted squared error distortion measure is used. A common justification for choosing the various operations that are incorporated in the coding process comes from a heuristic analysis, in which the quantization effect is assumed to be equivalent to adding an independent (or uncorrelated) white noise to the signal. This approximated model, while not accurate [10], is very useful in the design of coding schemes and the analysis of their rate-distortion performance; see e.g. [5], [19], [4] and [12].

In this paper the problem of combining optimally filtering and quantization is revisited and further analyzed in a different context, in which Entropy Coded Dithered (lattice) Quantization (ECDQ) is used as the basic coding component. An illustration of such a scheme, incorporating pre- and post-filters with the ECDQ is depicted in Figure 1(a), where Q_K is a lattice quantizer, E and D are a lossless encoder-decoder pair, and Z is a subtractive dither. The ECDQ has been introduced originally as a tool for universal quantization in [26], and its properties were developed further in [24] and [23]. One interesting property is that for the ECDQ the additive noise model is accurate. In fact, the rate-distortion performance of the entire coding scheme of Figure 1(a) may be expressed in terms of information theoretic quantities associated with the channel illustrated qualitatively in Figure 1(b). This channel is intuitively appealing, since with the appropriate filters and when the additive noise is Gaussian, it is the channel that attains the rate-distortion function of a Gaussian source [2].

Motivated by that, we provide in this paper a technique for encoding analog sources whose power spectrum is given, by a variable rate code. Our principal result is that the resulting coding scheme exceeds the rate-distortion function by at most the sum of two terms: the divergence of the source from Gaussianity and the divergence of the quantization noise from Gaussianity. Our work also suggests a simple analytic approach for the design of source coding schemes which combine filtering, vector quantization and lossless coding.

In Section II below we give a simple scalar example which illustrates the basic idea of the paper. Then, we show in Section III how this scheme is generalized to other cases, and specifically discuss its vector version. Section IV presents and investigates a pre/post-filtered ECDQ scheme for encoding time processes. Other possible implementations are considered in Section V. Throughout the paper we use the terminology and utilize results associated with ECDQ, lattice quantization and quantization noise spectral shaping that have been obtained in our other work on the subject. The reader is referred to [24]-[25] for further background.

II A Simple Scalar Example

In order to motivate and demonstrate the basic idea presented in this paper, we begin with the following simple scalar case. Let X be a memoryless source with mean μ_x and variance σ_x^2, to be encoded with some target distortion level $0 < D \leq \sigma_x^2$ by a scalar ECDQ. Throughout the paper we use the mean squared error (MSE) distortion measure. Consider the following two schemes:

Scheme I: A direct scalar ECDQ of the source, i.e., the source is quantized by a dithered uniform (unbounded) scalar quantizer $Q_1(\cdot)$ with step size $\Delta = \sqrt{12D}$, where D is the target distortion level, followed by lossless ("entropy") coding. More specifically, let $Z \sim \mathcal{U}(-\Delta/2, \Delta/2)$, the "dither", be a (pseudo) random variable uniformly distributed over the basic cell of Q_1 and known to both transmitter and receiver. A dither sample is added to the source before the quantization, and a (uniquely decodable) binary code is assigned to the quantizer output. At reconstruction, the dither is subtracted from the decoded codeword, so that the reproduction value is $\widehat{X} = Q_1(X + Z) - Z$.

An efficient lossless code, conditioned on the dither Z, can be designed so that its average code length approaches the minimal possible value

$$H(Q_1|Z) = H(Q_1(X+Z)|Z) = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} dz \sum_{i=-\infty}^{\infty} p_i(z)\log\left(\frac{1}{p_i(z)}\right) , \tag{1}$$

where $H(\cdot)$ denotes the entropy in bits, and $p_i(z) = \Pr(Q_1(X+z) = i\Delta)$. As discussed in [23], to achieve the conditional entropy in practice we must discretize the dither, and use a different lossless codebook for each value of the dither. Note also that this coding rate may be achieved asymptotically (for a large block of source samples), even if the source statistics are not known, using a universal coding algorithm, provided that $\sigma_x^2 < \infty$ [22].

Following the results in [24] and [23], the above simple coding scheme is equivalent to the additive noise channel $X \to \widehat{X} = X + N$, where $N \sim \mathcal{U}(-\Delta/2, \Delta/2)$ is independent of X. The equivalence is in the sense that first, $\widehat{X} - X$ is independent of X and is distributed as N, and second,

$$H(Q_1|Z) = I(X; X+N) , \tag{2}$$

where $I(\cdot;\cdot)$ denotes mutual information[1]. The first property asserts that the distortion is $E(\widehat{X} - X)^2 = EN^2 = D$. As shown in [26], [24] and [15], for small D and "smooth" sources the coding rate (2) is about $\frac{1}{2}\log 2\pi e/12 \approx 0.254$ bits higher than the rate-distortion function of the source, defined as [2]

$$R(D) = \inf_{\{U: E(U-X)^2 \leq D\}} I(X;U) . \tag{3}$$

Furthermore, for every D and all sources

$$H(Q_1|Z) - R(D) \leq \frac{1}{2}\log\left(\frac{4\pi e}{12}\right) \approx 0.754 \text{ bits}. \tag{4}$$

[1]A similar expression has been obtained for the coding rate (as an approximation, up to higher order terms) in classical high resolution quantization theory. However, for ECDQ this expression is accurate and applies at any resolution of the quantizer.

A disadvantage of this scheme is its significant redundancy over the rate-distortion function in the low coding rate region. It requires, for example, a positive rate even for $D = \sigma_x^2$.

Scheme II: Consider now the modification of the scheme above shown in Figure 2(a). Set $\Delta = \sqrt{12D}$ as in Scheme I, and apply the ECDQ to the random variable $\alpha(X - \mu_x)$, where

$$\alpha = \sqrt{1 - \frac{D}{\sigma_x^2}} \, , \tag{5}$$

while μ_x and σ_x^2 are the source's mean and variance, respectively. At reconstruction, multiply by $\beta = \alpha$, and add the constant μ_x. This scheme depends on μ_x and σ_x^2 which are assumed to be known, so it is not as universal as Scheme I. However, we next show that it has a smaller rate for the same target distortion level.

Clearly, the subtraction and addition of μ_x before and after the encoding, respectively, is equivalent to encoding the source $X - \mu_x$, which has a zero mean. Thus, for further analysis we may assume that $\mu_x = 0$, without loss of generality. We get $\widehat{X} = \beta\left(Q_1(\alpha X + Z) - Z\right)$, and an average code length of $H\left(Q_1(\alpha X + Z)|Z\right)$ bits. Using the results of [24], we replace the ECDQ block in Figure 2(a) by an additive noise channel, as shown in Figure 2(b). Thus $\widehat{X} - X$ is distributed as $\beta\alpha X + \beta N - X$, implying that the distortion of Scheme II is

$$E(\widehat{X} - X)^2 = (\beta\alpha - 1)^2\sigma_x^2 + \beta^2\frac{\Delta^2}{12} = D \, , \tag{6}$$

as our desired target. Also, from (2), the coding rate of Scheme II is

$$H(Q_1|Z) = I(\alpha X; \alpha X + N) \, . \tag{7}$$

Observe that the mutual information in (7) is associated with the same additive noise channel as the mutual information in (2), but the channel input in (7) is attenuated by $\alpha = \sqrt{1 - \frac{D}{\sigma_x^2}} < 1$. Thus, the coding rate of Scheme II is usually strictly smaller than that of Scheme I[2].

[2]The mutual information with attenuated input is always smaller if X is Gaussian, since then we can write $I(X; X + N) = h(X + N) - h(N) = h(\alpha X_1 + \sqrt{1 - \alpha^2}X_2 + N) - h(N) \geq h(\alpha X_1 + N) - h(N) = I(\alpha X; \alpha X + N)$,

This gain in coding rate is explained intuitively by the fact that in scheme II we need effectively a smaller number of quantization levels, since due to the attenuation $\alpha < 1$ we quantize the source in a coarser way. In spite of that, we still get the same distortion level as in Scheme I due to the post filter, which effectively reduces the quantization noise power. Also, Scheme II does not suffer from the undesired property of Scheme I of a positive coding rate at $D = \sigma_x^2$.

The mutual information formulas (2) and (7) for the average code length may be calculated for specific sources via the relation

$$I(X; X + N) = h(X + N) - h(N) = h(X + N) - \log(\Delta) \,, \tag{8}$$

where $h(\cdot)$ denotes differential entropy, and since the p.d.f of N is $f_N(n) = 1/\Delta, \; -\Delta/2 \le n \le \Delta/2$, $h(N) = -\int f_N(n) \log f_N(n) dn = \int_{-\Delta/2}^{\Delta/2} \frac{1}{\Delta} \log \Delta = \log \Delta$. Note that $h(X + N)$ exists even if X does not have a density. In Figure 3 we illustrate the rates of Scheme I (Eq. (2)) and of scheme II (Eq. (7)) as a function of the mean-squared distortion D, in coding the Gaussian source $X^* \sim \mathcal{N}(0, \sigma_x^2)$. We have also drawn, for comparison, the rate-distortion function of the source $R^*(D) = \frac{1}{2} \log \sigma_x^2 / D$. As seen from the plots in Figure 3, the improvement in Scheme II is mainly in the high distortion / low rate range, while in the high resolution extreme (small D) both schemes have the same redundancy of ~ 0.254 bits above the rate-distortion function. It is interesting to note that in this Gaussian source example, the redundancy of the modified scheme is always below ~ 0.254 bits.

Scheme II, as specified by the triplet (α, Δ, β), is actually the scalar version of pre/post-filtered ECDQ which is the subject of this paper. Theorem 1 below sheds more light on the general behavior of this scheme, and (as will be seen later on) provides guidelines for the specific selection of (α, Δ, β). To state the theorem we need to introduce some notations. Let U^* denote a Gaussian variable having the same mean and variance as the random variable U. Thus, $X^* \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $N^* \sim \mathcal{N}(0, \Delta^2/12)$ are the Gaussian counterparts of the random variables X and N defined above.

where X_1 and X_2 are independent and identically distributed as X [7, Lemma 16.2.1]. In general, however, this statement requires some regularity conditions on X and/or N.

The "divergence of U from Gaussianity" is defined as $\mathcal{D}(U;U^*) = \int f_U(\alpha)\log(f_U(\alpha)/f_{U^*}(\alpha))d\alpha$, where $f_U(\cdot)$ and $f_{U^*}(\cdot)$ are the density functions of U and U^*, respectively. It is easy to verify [25], [7, p. 234] that

$$\mathcal{D}(U;U^*) = h(U^*) - h(U) = \frac{1}{2}\log 2\pi e\sigma_u^2 - h(U)\ , \tag{9}$$

where σ_u^2 is the variance of U. For example, the divergence from Gaussianity of the "quantization noise" N is

$$\mathcal{D}(N;N^*) = \frac{1}{2}\log 2\pi e\frac{\Delta^2}{12} - \log\Delta = \frac{1}{2}\log\left(\frac{2\pi e}{12}\right)\ . \tag{10}$$

If U does not have a density, then by definition $\mathcal{D}(U;U^*) = \infty$.

Theorem 1 (Divergence Bound - Scalar Case) *Let $R^*(D) = \frac{1}{2}\log\left(\frac{\sigma_x^2}{D}\right)$ be the rate-distortion function of $X^* \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Let $H(Q_1|Z)$ be the average code length (7) of Scheme II. Then,*

$$\begin{aligned} H\left(Q_1|Z\right) &= R^*(D) + \frac{1}{2}\log\left(\frac{2\pi e}{12}\right) - \mathcal{D}(\widehat{X};\widehat{X}^*) && (11)\\ &\le R^*(D) + \frac{1}{2}\log\left(\frac{2\pi e}{12}\right) && (12)\\ &\le R(D) + \mathcal{D}(X;X^*) + \frac{1}{2}\log\left(\frac{2\pi e}{12}\right) && (13)\end{aligned}$$

where $R(D)$ is the rate-distortion function of the source defined in (3), and $\mathcal{D}(X;X^*)$ is the divergence of the source from Gaussianity.

For a Gaussian source, (12) asserts that the coding rate never exceeds the rate-distortion function by more than $\frac{1}{2}\log\left(\frac{2\pi e}{12}\right) \approx 0.254$ bits. It is further shown in Appendix A, that for a Gaussian source $H(Q_1|Z) - R^*(D)$ decreases to zero *monotonically* as D goes from zero to σ_x^2. For a general source, it follows from (10) and (13) that the coding rate redundancy $H(Q_1|Z) - R(D)$ is upper bounded by the sum of the divergences from Gaussianity of the source and the quantization noise. For example, the divergence of a Laplacian source from Gaussianity is ≈ 0.104 bits, so when encoded by Scheme II the rate redundancy does not exceed $0.104 + 0.254 \approx 0.36$ bits for all distortion levels.

Proof: The rate of Scheme II, given by the mutual information expression (7), can be written as

$$H(Q_1|Z) = I(\alpha X; \alpha X + N) = I(\alpha X^*; \alpha X^* + N^*) + \mathcal{D}(N; N^*) - \mathcal{D}(\alpha X + N; \alpha X^* + N^*) \ . \quad (14)$$

This equality follows straightforwardly by combining (8) and (9). Now, since $E(\alpha X^* + N^*)^2 = E(\alpha X + N)^2 = \sigma_x^2$ and $E(N^*)^2 = E(N)^2 = D$, we have $h(\alpha X^* + N^*) = \frac{1}{2}\log(2\pi e\sigma_x^2)$ and $h(N^*) = \frac{1}{2}\log(2\pi eD)$, so by (8),

$$I(\alpha X^*; \alpha X^* + N^*) = \frac{1}{2}\log\left(\frac{\sigma_x^2}{D}\right) \ . \quad (15)$$

Also

$$\mathcal{D}(\alpha X + N; \alpha X^* + N^*) = \mathcal{D}\left(\beta(\alpha X + N); \beta(\alpha X^* + N^*)\right) = \mathcal{D}(\widehat{X}; \widehat{X}^*) \ ,$$

since the divergence does not change by applying the same invertible transformation to its arguments. Substituting this, together with (10) and (15), into (14) completes the proof of (11).

Inequality (12) follows simply from the non-negativity of the information divergence [7]. As for the last inequality in the theorem, we use the Shannon lower bound for the rate-distortion function [2], together with (9), to write

$$R(D) \geq h(X) - \frac{1}{2}\log 2\pi eD = \frac{1}{2}\log\left(\frac{\sigma_x^2}{D}\right) - \mathcal{D}(X; X^*) \ . \quad (16)$$

Substituting (16) in (12) results (13), and the proof is completed. □

Remarks:

1. In the general case of lattice-ECDQ, it is shown in Section III that the term $\frac{1}{2}\log\left(\frac{2\pi e}{12}\right)$ above should be replaced by $\frac{1}{2}\log\left(2\pi eG_k\right)$, where G_k is the *normalized second moment* of the k-dimensional lattice quantizer. Theorem 1 corresponds to the special case $G_1 = 1/12$.

2. Scheme II may also be used to encode a general stationary source with power σ_x^2. It is easy to check that the resulting MSE is still D, as in the memoryless source case. Moreover, efficient

entropy coding in this case will take into account the dependence between successive outputs of the scalar quantizer, and will reduce the coding rate beyond that predicted by Theorem 1. As shown in Section IV, even better results are achieved by replacing the scalar gains α and β in Scheme II by filters that are tuned to the source spectrum.

3. Theorem 1 does not reflect the fact that, similarly to the bound (4) for Scheme I, the redundancy above the rate-distortion function of Scheme II is upper bounded by a universal constant. But, as follows from Theorem 6 in Section IV the redundancy of scheme II also satisfies

$$H(Q_1|Z) - R(D) \leq \frac{1}{2}\log\left(\left(2 - \frac{D}{\sigma_x^2}\right)\frac{2\pi e}{12}\right) \leq \frac{1}{2}\log\left(\frac{4\pi e}{12}\right) \tag{17}$$

for all sources with $EX^2 = \sigma_x^2$.

III Simulating the Forward Channel Realization of $R^*(D)$ Using the ECDQ

We next consider in brief the more general case of encoding vector sources, and later on in section IV, we consider in detail encoding time processes, using ECDQ. In both more general cases we actually suggest to use generalized forms of Scheme II above. The main idea behind the proposed schemes is to simulate, by the pre/post-filtered ECDQ, the forward channel realization of the rate-distortion function of Gaussian sources.

This idea has already been utilized above, in the simple scalar case, as follows. For a Gaussian source X^* and when the quantization noise is Gaussian (which is the case for a lattice quantizer having a large dimension [25]), the coding rate of Scheme II is given by the mutual information (15) associated with its equivalent channel. However, this mutual information is equal to the rate-distortion function of X^*. Thus, the equivalent channel shown in Figure 2(b) becomes in this case the *forward channel realization* of the rate-distortion function [2, pp. 101, 143], simulated by

Scheme II.

All Gaussian sources have this unique property that their rate-distortion function, for squared error distortion, can be achieved by a simple forward channel, composed of linear transformations (filters) and additive Gaussian noise. The explicit form of the transformations and the spectrum of the additive Gaussian noise for vector sources, and for discrete and continuous time stationary processes, can be found, e.g., in [2]. Thus, similarly to the approach taken in the design of Scheme II, these forward channels can be simulated by an ECDQ block with the appropriate pre- and post-filters.

Specifically, the generalization of Scheme II to the case where a zero mean vector source $\underline{X} \in \mathcal{R}^n$ is encoded by an (unbounded) lattice quantizer, consists of replacing the linear gains α and β by pre-matrix A and post-matrix B, so that the channel

$$\underline{X}^* \longrightarrow \underline{\widehat{X}}^* = B(A\underline{X}^* + \underline{N}^*) \tag{18}$$

is a forward channel realization of the rate-distortion function of the zero mean Gaussian source $\underline{X}^*$ which has the same covariance as the source $\underline{X}$. The matrices A and B, and the covariance matrix R_N of the Gaussian vector $\underline{N}^*$, depend on the covariance matrix of the source $\underline{X}^*$ (i.e., of $\underline{X}$) and on the target distortion level D, but they are not unique. Yet, all triplets (A, R_N, B) which realize the rate distortion function of $\underline{X}^*$ lead to the same output $\underline{\widehat{X}}^*$. Let k, where $0 \leq k \leq n$, be the rank of the covariance matrix of $\underline{\widehat{X}}^*$ (k is a non-increasing function of D). We can always pick a triplet (A, R_N, B) such that A and B are $k \times n$ and $n \times k$ matrices, respectively, and R_N is a $k \times k$ non-singular matrix[3].

The proposed vector scheme (A, Q_k, B) simulates a forward channel realization with the intermediate dimension k defined above. This scheme uses a k-dimensional lattice quantizer Q_k, such

[3]By the "water pouring" law [2, pp. 108-123], the rate distortion function of $\underline{X}^*$ depends only on the k eigenvalues of the covariance matrix of the source which are "above the water level". Thus, only the projection of $\underline{N}^*$ on the corresponding k eignvectors is effective.

that

$$E\{\underline{Z} \cdot \underline{Z}^t\} = R_N \ , \tag{19}$$

where $\underline{Z}$ is the dither vector which is uniformly distributed over the basic cell of Q_k. Note that since R_N is not singular, it is possible to satisfy (19) by *shaping* an arbitrary k-dimensional lattice quantizer [25]. The reconstruction of the vector $\underline{X}$ when coded by (A, Q_k, B) is thus $\widehat{\underline{X}} = B(Q_k(A\underline{X} + \underline{Z}) - \underline{Z})$, while the coding rate is the conditional entropy of the quantizer $\frac{1}{n}H(Q_k|\underline{Z}) = \frac{1}{n}H(Q_k(A\underline{X} + \underline{Z})|\underline{Z})$ bits per sample.

We now state the extension of Theorem 1 to the vector case, for the scheme described above. But before this, we recall some notations. Let G_k be the *normalized second moment* of Q_k [6],

$$G_k = \frac{1}{k}\frac{E\left\{\|\underline{Z}\|^2\right\}}{V^{2/k}} \ , \tag{20}$$

where V is the volume of the basic cell of Q_k. Note that $G_1 = 1/12$ corresponds to the uniform scalar quantizer considered in the previous section. Let [2]

$$R_n(D) = \inf_{\{\underline{U}:\ \frac{1}{n}E\|\underline{U}-\underline{X}\|^2 \leq D\}} \frac{1}{n} I(\underline{X}; \underline{U}) \tag{21}$$

be the rate-distortion function in bits per sample of $\underline{X}$, and let $R_n^*(D)$ denote the rate-distortion function of $\underline{X}^*$. Finally, let [17], [9]

$$\mathcal{D}(\underline{X}; \underline{X}^*) = E_{\underline{X}}\left\{\log\frac{dF(\underline{X})}{dF(\underline{X}^*)}\right\} = \int d\underline{x} f_{\underline{X}}(\underline{x}) \log\frac{f_{\underline{X}}(\underline{x})}{f_{\underline{X}^*}(\underline{x})} = h(\underline{X}^*) - h(\underline{X}) \tag{22}$$

denote the divergence from Gaussianity of the vector source $\underline{X}$, where $\frac{dF(\underline{X})}{dF(\underline{X}^*)}$ is the Radon-Nikodym derivative between the distribution functions of $\underline{X}$ and $\underline{X}^*$, and the two last equalities hold (as in (9) for the scalar case) if $\underline{X}$ has a density function $f_{\underline{X}}(\underline{x})$. The vector extension of Theorem 1 states:

Theorem 2 (Divergence Bound - Vector Case) *The distortion of the vector coding scheme above is equal to the distortion associated with the forward channel realization (18), and its coding* *rate satisfies*

$$\begin{aligned}\frac{1}{n}H(Q_k|\underline{Z}) &\leq R_n^*(D) + \frac{k}{2n}\log(2\pi e G_k) \\ &\leq R_n(D) + \frac{1}{n}\mathcal{D}(\underline{X};\underline{X}^*) + \frac{k}{2n}\log(2\pi e G_k)\ . \end{aligned} \tag{23}$$

Proof: The distortion part of the theorem follows from the additive nature of subtractive dithered quantization [23], and since the mean squared error of a linear additive noise channel depends only on the second order statistics of the source and the noise. The proof of the rate part (23) follows the steps in the proof of the scalar case (Theorem 1), while noting that the divergence of the dither from Gaussianity satisfies [25]

$$\mathcal{D}(\underline{Z};\underline{Z}^*) \leq \frac{k}{2}\log(2\pi e G_k)\ . \tag{24}$$

We have omitted the details of the proof since it is analogous to the proof of Theorem 1, and to its process version in Theorem 3. □

Equality in (24) holds if and only if the lattice quantizer Q_k is *white* [25], i.e., iff

$$E\{\underline{Z}\cdot\underline{Z}^t\} = \epsilon\cdot I\ , \tag{25}$$

where I is the identity matrix, and ϵ is the *second moment of* Q_k. A white lattice quantizer is, in fact, a favorable choice. Note, first, that if the triplet (A, R_N, B) composes a forward channel realization of $R_n^*(D)$, then so does the triplet

$$(TA, TR_NT^t, BT^{-1}) \tag{26}$$

for any non-singular $k\times k$ matrix T. Thus, we can always pick a triplet with a white noise (by setting $T = \sqrt{\epsilon}R_N^{-1/2}$), and simulate it using a white lattice quantizer. Furthermore, from [25] we know that the *optimal* lattice quantizer at each dimension, i.e., the one that minimizes G_k, is white, and it satisfies

$$\frac{1}{2}\log(2\pi e G_k) = O\left(\frac{\log k}{k}\right) \to 0\ ,\ \ \text{as } k\to\infty\ . \tag{27}$$

Thus, by using white lattice quantizers in the vector scheme above, we can approach $R_n^*(D)$ as closely as desired, with a rate of convergence in the order of $\log n/n$. (Note that by (23) and (27), the redundancy term is $\frac{k}{n} \cdot O\left(\frac{\log k}{k}\right) \leq O\left(\frac{\log n}{n}\right)$.)

It is interesting to compare the behavior of the ECDQ-based scheme as $n \to \infty$ with that of fixed rate lossy source coding schemes. In [20] an $O(\sqrt{\log n/n})$ rate of convergence to $D^*(R)$, the inverse function of $R^*(D) \triangleq \lim_{n\to\infty} R_n^*(D)$, as a function of the block size n, has been obtained for fixed rate encoding of a correlated stationary Gaussian source. A similar result follows from [13] for memoryless (not necessarily Gaussian) sources. Since under some mild conditions on $S_x(f)$, $R_n^*(D) - R^*(D) = O(1/n)$ as $n \to \infty$ (see [21, Eq. (38)-(40)]), it follows from (23) and (27) that a better rate of convergence of $O(\log n/n)$ may be achieved, at least for Gaussian sources, using fixed to variable coding.

We note, however, that dithered quantization corresponds to a random code, whose structure is randomized by the dither, and whose ensemble average performance (rate and distortion) is given in Theorem 2. In order to draw a conclusion on the rate of convergence for *deterministic* codes, it may be useful to consider the weighted sum

$$\frac{1}{n} H(Q_k|\underline{Z}) + sD \ , \quad \text{where } s = \frac{dR^*(D)}{dD} \ ,$$

which by (23) and (27) converges (vertically) to $R^*(D) + sD$ as $O(\log n/n)$. Thus, for at least one dither realization, i.e., for one code in the ensemble, the weighted sum of the coding-rate and the distortion converges to its minimal possible value at the above rate.

A few more remarks are yet in order regarding the vector pre/post filtered ECDQ scheme:

1. The reason we chose the minimal possible dimension k for the ECDQ in the vector scheme above is to avoid the extra rate that was observed in [23], and affected the rate in ECDQ encoding of an oversampled bandlimited process. In the next section, where we deal with the case of encoding a time process, this requirement is taken care of by sampling the pre-filtered

source at exactly the Nyquist rate.

2. The vector coding scheme above, which simulates the structure of the forward channel realization, is generally not the optimal linear solution, i.e., the optimal combination of linear pre- and post-filters for a finite dimensional ECDQ and a non Gaussian source[4]. For example, it may be seen from Figure 3 that at high distortion, in the range $\sigma_x^2/D \approx 0 \div 1$ dB, the rate-distortion curve of the scheme is concave $(\cap)$[5], so it may be "straightened up" by time-sharing. Nevertheless, in Section V we give a *min-max* argument to justify this choice of structure for the coding scheme.

3. As discussed in Section V, the vector form of Theorem 1 may also be extended to include a frequency weighted squared error distortion measure.

IV Information Rates in Coding Stationary Time Processes

In this section we present and analyze a specific pre/post-filtered ECDQ scheme for encoding a stationary bandlimited time process, which provides the most physically motivated example of using pre/post filters. Let $X = \{X(t), -\infty < t < \infty\}$ be a zero mean stationary bandlimited source, whose (one-sided) power spectrum is $S_x(f)$ for $f \geq 0$, and where $S_x(f) = 0$ for $f > B$. X is to be encoded with an average mean-squared error D. It is assumed that

$$X(t) = \sum_n X_n \frac{\sin \pi(2Bt - n)}{\pi(2Bt - n)} \quad , -\infty < t < \infty \ , \tag{28}$$

where $X_n = X(n/2B)$, $n = 0, \pm 1, \pm 2, \ldots$. Assumption (28) means that *every* sample function of X is bandlimited, and may be interpolated from its samples taken at the Nyquist sampling rate $2B$

[4]For a given pre-filter, our proposed post-filter is optimal, since the filter obtaining the minimum MSE depends only on the second order statistics of the source and the quantization noise. However, the optimal pre-filter may be different from our pre-filter.

[5]The x axis in this figure is logarithmic. However, the curve remains concave even in linear scaling.

(this may be the case when X is the output of an ideal low pass filter). We further assume that $\int_0^B S_x(f)df = \sigma_x^2 < \infty$.

Let the rate-distortion function of this bandlimited process, normalized to bits per second, be [2]

$$R(D) = 2B \cdot \lim_{n\to\infty} R_n(D) \ , \tag{29}$$

where $R_n(D)$ is the rate-distortion function per sample (21), of the vector of Nyquist samples $\underline{X} = (X_1, \ldots, X_n)$. In accordance with the above notation, let $R^*(D)$ denote the rate-distortion function of the Gaussian process $X^*(t)$, having the same power spectrum as $X(t)$. By the "water pouring" law [2, pp. 108-123]

$$R^*(D) = \int_{\mathcal{B}_D} \log\left(\frac{S_x(f)}{\theta}\right) df \ , \tag{30}$$

where the "in-band" $\mathcal{B}_D$ is defined by $\mathcal{B}_D = \{f : S_x(f) > \theta\}$, the "water level" θ is chosen to satisfy

$$D = \int_{f\notin\mathcal{B}_D} S_x(f)df + \theta\cdot B_D \stackrel{\Delta}{=} D_s + D_q \ , \tag{31}$$

and where $B_D = \mu(\mathcal{B}_D)$ is the Lebesgue measure of the set $\mathcal{B}_D$ (see Figure 4). The mean-squared errors D_s and D_q are sometimes called the "sampling error" and the "quantization error", respectively [2, p. 143].

From the discussion in Section III it follows that for each n, we can use the correlation function of the source to design a vector scheme (which combines pre- and post-matrices and an n-dimensional ECDQ), for coding the source block $X_1 \ldots X_n$. Let us denote this scheme by S_n^*. The proposed coding scheme introduced in this section, denoted S^*, may be thought of as the limit as $n\to\infty$ of S_n^*.

Before we define S^* explicitly, we note that as discussed in Section III, the optimal forward channel which realizes $R^*(D)$ does not have a unique structure, and can be implemented by various choices of filters and noise spectra [2, p. 101]. The scheme S^* we chose simulates a forward channel having an additive *white* noise. In this implementation, the source is first pre-filtered using a filter whose frequency response $H_1(f)$ satisfies

$$|H_1(f)|^2 = \begin{cases} 1 - \frac{\theta}{S_x(f)} & , \ f \in \mathcal{B}_D \\ 0 & , \ \text{otherwise} \end{cases} \tag{32}$$

where θ and $\mathcal{B}_D$ are defined in (30) and (31). Then, all separate pieces of $\mathcal{B}_D$ (if any) are coupled by appropriate down-conversions to a single base-band process whose bandwidth is B_D. This process is sampled at exactly the Nyquist rate $F_s = 2B_D$, to attain the discrete-time process $X_q = \{X_{q,n}, n = 0, \pm 1, \pm 2, \ldots\}$. The sampled process is then ECDQ-encoded using a K-dimensional lattice quantizer Q_K, where the dimension K is a free parameter in our analysis (unlike the dimension k in the vector scheme of Section III). The lattice quantizer Q_K is *white* (see (25)), with a second moment

$$\epsilon = \theta \cdot B_D = D_q \ . \tag{33}$$

The K-vectors of the dither associated with the ECDQ are drawn independently at each quantization stage, while the lossless coding is applied jointly to successive outputs of the quantizer. At reconstruction, the ECDQ is decoded, i.e., the lossless code is decoded and the dither is subtracted, to yield the process $\widehat{X}_q = \{\widehat{X}_{q,n}, n = 0, \pm 1, \pm 2, \ldots\}$. Then, $\widehat{X}_q$ is interpolated by an ideal discrete-to-continuous converter, the "pieces" of $\mathcal{B}_D$ are placed back at their original position via appropriate up-conversions, and a post-filter

$$H_2(f) = H_1^*(f) \tag{34}$$

where * denotes complex conjugation, is applied to obtain $\widehat{X} = \{\widehat{X}(t), -\infty < t < \infty\}$. This coding procedure is illustrated in Figure 5. The entire linear transformations $X \to X_q$ and $\widehat{X}_q \to \widehat{X}$ are denoted L_1^* and L_2^*, respectively, so S^* is fully specified by the triplet (L_1^*, Q_K, L_2^*).

Our first obligation is to show that the mean squared error of S^*, defined as $E\{E(t)^2\} = E(\widehat{X}(t) - X(t))^2$, is time invariant and equals to the desired distortion D. It is shown in [23] that, similarly to the scalar and the vector cases, ECDQ-encoding of a discrete time process is equivalent to passing the source samples through an additive noise channel, where the additive noise process

$N_q = \{N_{q,n}, n = 0, \pm 1, \pm 2, \dots\}$ is composed of i.i.d. K-blocks, each one is distributed as $-\underline{Z}$, and where $\underline{Z}$ is the dither vector associated with Q_K. Thus, the entire scheme S^* is equivalent to a system composed of the pre-processor L_1^*, this additive noise channel and the post-processor L_2^* in cascade. Let $N(t)$ denote the continuous time output of L_2^* when it is fed by the discrete time noise $N_{q,n}$ defined above. From the equivalent system description of S^*, it is easy to see that

$$E\{E(t)^2\} = \int_{-\infty}^{\infty} \left(|H_2(f)H_1(f) - 1|^2 S_x(f) + |H_2(f)|^2 S_N(f) \right) df \ , \tag{35}$$

where $S_x(f)$ and $S_N(f)$ are the power spectra of the source and the continuous time quantization noise $N(t)$, respectively. Since the lattice quantizer is white, $N(t)$ is a wide sense stationary process, whose power spectrum is

$$S_N(f) = \begin{cases} \epsilon / B_D = \theta & , \text{for } f \in \mathcal{B}_D \\ 0 & , \text{elsewhere} \ . \end{cases} \tag{36}$$

Substituting (36) in (35) and recalling the filters definition (32), (34), we get $E\{E(t)^2\} = D_s + D_q = D$ as desired.

We now turn to the main observation of this paper, asserting that the rate redundancy of S^* is upper bounded by the sum of two terms: the divergence from Gaussianity of the source and the divergence from Gaussianity of the quantization noise. Let

$$\overline{H}(Q_K|Z) = \lim_{m \to \infty} \frac{1}{mK} H\left(Q_K(\underline{X}_{q,1} + \underline{Z}_1), \dots, Q_K(\underline{X}_{q,m} + \underline{Z}_m) \mid \underline{Z}_1 \dots \underline{Z}_m \right)$$

denote the conditional entropy rate (per sample) of the lattice quantizer, where $\underline{X}_{q,1}, \underline{X}_{q,2}, \dots$ are successive K-blocks of X_q (the ECDQ input), and $\underline{Z}_1, \underline{Z}_2, \dots$ are the corresponding dither K-blocks. The existence of the limit above has been shown in [23]. Let $R_Q^{S^*} = F_s \cdot \overline{H}(Q_K|Z)$ denote the coding rate of S^* in bits per second. Finally, let

$$\overline{\mathcal{D}}(X; X^*) = 2B \cdot \lim_{n \to \infty} \frac{1}{n} \mathcal{D}(\underline{X}; \underline{X}^*) \ ,$$

where $\underline{X} = (X_1 \dots X_n)$ and $\mathcal{D}(\underline{X}; \underline{X}^*)$ is defined in (22), denote the divergence from Gaussianity of X in bits per second[6]. Similarly, $2B_D \cdot \overline{\mathcal{D}}(N_q; N_q^*)$ denotes the divergence from Gaussianity per second of the quantization noise process, which is given by $B_D \cdot \log(2\pi e G_K)$ bits per second [23].

Theorem 3 (Divergence Bound - Process Case) *At all distortion levels*

$$\begin{aligned} R_Q^{S^*}(D) &\leq R^*(D) + B_D \cdot \log(2\pi e G_K) && (37) \\ &\leq R(D) + \overline{\mathcal{D}}(X; X^*) + B_D \cdot \log(2\pi e G_K)\ , && (38) \end{aligned}$$

where the rate-distortion functions $R^*(D)$ and $R(D)$ are defined in (30) and (29), respectively.

Proof: Since the ECDQ is equivalent to an additive noise channel,

$$\overline{H}(Q_K|Z) = \overline{I}(X_q; X_q + N_q) = \lim_{n\to\infty} \frac{1}{n} I(X_{q,1}, \ldots, X_{q,n}\ ;\ X_{q,1} + N_{q,1}, \ldots, X_{q,n} + N_{q,n})\ , \qquad (39)$$

where $\overline{I}(\cdot;\cdot)$ denotes mutual information rate. Thus the coding rate is

$$R_Q^{S^*} = 2B_D \cdot \overline{I}(X_q; X_q + N_q) \text{ bits per second.} \qquad (40)$$

Next we show that if the source and the quantization noise are Gaussian, the rate in (40) is equal to the rate-distortion function of the source. For that, we express the mutual-information rate $\overline{I}(X_q^*; X_q^* + N_q^*)$ by means of power spectral densities of X_q^* and N_q^* (see, e.g., [2] and [17]), where as above we denote by X_q^* and N_q^* Gaussian processes with the same power spectrum as X_q and N_q, respectively. Utilizing the linear structure of the pre-filter, we get

$$2B_D \cdot \overline{I}(X_q^*; X_q^* + N_q^*) = \int_{\mathcal{B}_D} \log\left(1 + \frac{|H_1(f)|^2 \cdot S_x(f)}{S_N(f)}\right) df = \int_{\mathcal{B}_D} \log\left(\frac{S_x(f)}{\theta}\right) df = R^*(D) \quad (41)$$

as desired. Note that (41) asserts that S^* is indeed a "forward channel realization" of $R^*(D)$.

We now turn to consider general sources and finite dimensional quantizers (non-Gaussian quantization noise). By expressing the mutual information rate in (39) as a difference between differential entropy rates (similarly to (8) in the scalar case), we can write

$$R_Q^{S^*} = 2B_D \cdot \{\overline{h}(X_q + N_q) - \overline{h}(N_q)\}\ , \qquad (42)$$

[6] This divergence rate may also be expressed directly in terms of the continuous time process $X(t)$ using the "Pinsker definition" of the divergence; see [17] pp. 76-110 and [9] Chapter 7.

where $\overline{h}(\cdot)$ denotes differential entropy rate, e.g., $\overline{h}(N_q) = \lim_{n\to\infty} h(N_{q,1}, \ldots, N_{q,n})/n$. Note that the decomposition in (42) always exists, since

$$\infty > \frac{1}{2}\log(2\pi e(\sigma_x^2 + D_q)) \geq \overline{h}(X_q + N_q) \geq \overline{h}(N_q) = \frac{1}{2}\log(\epsilon/G_K) > -\infty$$

(see [25]). We may therefore follow the steps in the proof of Theorem 1, and use (41), (42) and the property (22), to get

$$R_Q^{S^*}(D) = R^*(D) + B_D \cdot \log(2\pi e G_K) - 2B_D \cdot \overline{\mathcal{D}}(\widehat{X}_q; \widehat{X}_q^*) \;, \tag{43}$$

where $2B_D \cdot \overline{\mathcal{D}}(\widehat{X}_q; \widehat{X}_q^*)$ is the divergence from Gaussianity per second of the ECDQ output process.

The first bound (37) in Theorem 3 follows from (43) using the non-negativity of the divergence. As for the second bound, we substitute in (37) the lower bound [3, Thm. 3]

$$R(D) \geq R^*(D) - \overline{\mathcal{D}}(X; X^*) \tag{44}$$

on the rate-distortion function (under the mean-squared error criterion), and that completes the derivation of the "divergence bound". $\square$

The following two theorems characterize the asymptotic behavior of S^* at the high and low distortion limits, i.e., in the two extremes $D \to \sigma_x^2$ and $D \to 0$, where tighter estimates for the scheme's performance can be attained. In both cases we scale the quantizer in order to get the desired second moment ϵ, determined by D, while keeping its structure, e.g., its normalized second moment G_K, fixed.

Theorem 4 (Low Resolution) *For any source*

$$\lim_{D\to\sigma_x^2} R_Q^{S^*}(D) = 0 = R(\sigma_x^2) \;. \tag{45}$$

The proof is given in Appendix B. Note that $R_Q^{S^*}(D)$ may be a discontinuous function of D for some (non-Gaussian) processes at $D < \sigma_x^2$.

Theorem 5 (High Resolution) *Assume that the source satisfies* $\overline{h}(X) > -\infty$ *(i.e., the entropy rate of its Nyquist samples process exists and is finite), and* $S_x(f) \geq s_m > 0$ *for* $0 \leq f \leq B$*. Then,*

$$R_Q^{S^*}(D) - R(D) \to B \cdot \log(2\pi e G_K) \qquad as\ D \to 0. \tag{46}$$

The proof of Theorem 4 is given in Appendix C. We believe, actually, that the condition in Theorem 4 that $S_x(f)$ is bounded away from zero is not necessary, since the finite entropy rate condition implies $\int \log S_x(f) > -\infty$.

The "divergence bound" of Theorem 2 above, whose value is at least the rate distortion function of a Gaussian source, may not be tight, especially for sources with high divergence from Gaussianity. As shown in Theorems 4 and 5 it can be improved for specific distortion levels. We provide next an alternative universal bound on the redundancy of S^*, which holds for any source, and can be very useful for sources that are far away from Gaussianity. Let

$$C_D = 2B_D \cdot \sup_{\{X: E\{X(t)^2\} \leq D\}} \overline{I}(X_q; X_q + N_q) \tag{47}$$

be the power constraint capacity of the equivalent additive noise channel. C_D may be also interpreted as the maximal rate of S^* over all inputs whose power is less than or equal to D. Note that C_D is a function of D, $S_x(f)$ and the quantizer Q_K.

Theorem 6 (Capacity Bound) *For any source*

$$R_Q^{S^*}(D) - R(D) \leq C_D\ . \tag{48}$$

The proof is given in Appendix D. The capacity bound reflects the maximal redundancy for non-Gaussian processes at medium distortion. As shown in the next section, this bound is actually attained for some sources with multi-modal probability distribution at $D << \sigma_x^2$.

It is interesting to compare the results presented in Theorems 4-6, to the performance of an ECDQ scheme which does not utilize the knowledge of the power spectrum of the source, and does not use pre- and post-filters. Such a scheme, which may be thought of as the process version of

Scheme I of Section II, was considered in [23]. Theorem 4 implies that at the high distortion region, S^* approaches the rate-distortion function for any source and any quantizer Q_K. This is in contrast with the high (usually the highest) redundancy at the point $D = \sigma_x^2$ of the scheme considered in [23], which does not use filters. In the high rate extreme, however, both schemes coincide, and Theorem 5 indeed show that S^* approaches the performance given by Theorem 4 of [23].

Theorem 6 here parallels Theorem 5 of [23], which states that for any source and all distortion levels the redundancy of the scheme without filters is upperbounded by $B \cdot \log(4\pi e G_K)$ bits per second, i.e., B bits per second more than the redundancy at the high rate extreme. (This may be thought of as the process version of (4).) To compare this bound with our bound in Theorem 6 above, we write the following chain of inequalities:

$$
\begin{aligned}
C_D \le C_D^* + B_D \cdot \log 2\pi e G_K \quad &\le \quad B_D \cdot \log\left(\left(1 + \frac{D}{\epsilon}\right) 2\pi e G_K\right) && (49)\\
&\le \quad B_D \cdot \log\left(\left(1 + \frac{B}{B_D}\right) 2\pi e G_K\right) && (50)\\
&\le \quad B \cdot \log(4\pi e G_K)\ , && (51)
\end{aligned}
$$

where C_D^* is the capacity when the additive noise is Gaussian, i.e., when $N_q = N_q^*$. (For example, when $S_x(f)$ is flat over the source band, this capacity is given by $C_D^* = B \cdot \log(2 - D/\sigma_x^2)$, i.e., it monotonically decreases from B to zero as D goes from zero to σ_x^2.) The second upper bound in (49) follows since $|H_1(f)| \le 1$, and thus the capacity with $H_1(f) = 1$ for $f \in \mathcal{B}_D$, which is $B_D \cdot \log(1 + \frac{D}{\epsilon})$, upper bounds C_D^*. Inequality (50) follows since $S_x(f) \le \theta$, $\forall f \notin \mathcal{B}_D$, implying $\frac{D}{\epsilon} \le \frac{B}{B_D}$. Finally, Inequality (51) follows since $B_D \le B$, and it becomes equality if and only if $B_D = B$. Thus, as expected, pre/post filtered quantization is universally superior to a scheme incorporating sampling and quantization but no filters. This is true, of course, provided that the source spectral characteristics are available, so that the filters may be designed appropriately.

To summarize the bounds derived in this section, we combine (38), (45), (46) and (51) and conclude that for non-Gaussian sources the redundancy in bits per sample typically attains its maximal value in the *medium* distortion region. Note that in any case the redundancy is upper bounded by $\frac{1}{2}\log 4\pi e G_K$, i.e., 0.754 bit per sample for $K=1$ and 0.5 bit per sample for $K\to\infty$. A typical R-D curve of S^* for a non-Gaussian "smooth" source, as compared to $R(D)$ and $R^*(D)$, is depicted in Figure 6.

V Modified Schemes

In this section we consider modifications of S^*, the scheme considered in Section IV, which fit better the following specific situations:

1. Frequency weighted squared-error distortion measures: Suppose the source $X(t)$ is to be encoded under a general fidelity criterion of the form

$$\int L(f) S_E(f) df \le D \;, \tag{52}$$

where $S_E(f)$ is the power spectrum of the error signal $E(t) = \widehat{X}(t) - X(t)$, and where we assume that the weighting function $L(f)$ is invertible in the source band. Note that the regular MSE criterion is the special case $L(f)=1$. Let WX denote a source obtained by passing X through a filter W satisfying $|W(f)|^2 = L(f)$, and let S^* be a scheme designed according to the previous section, to encode WX with a mean-squared distortion D. It is easy to verify that (52) is satisfied if we apply S^* to WX, and pass the output process through the inverse filter W^{-1}. Similarly, the rate-distortion function of X under the criterion (52) is given by the rate-distortion function of WX under mean-squared error D; see e.g. [16], [2, Sec. 4.5.4], and [18]. Thus, Theorems 3, 4, 5 and 6 apply to the modified scheme under the modified distortion measure by substituting the power spectrum $L(f)S_x(f)$ of WX, instead of the power spectrum of X. Note that, since $L(f)$ is invertible, the divergence from Gaussianity of WX is the same as the divergence from Gaussianity of X.

2. Partial knowledge of the spectrum: In many cases the exact spectral density of the source might not be known (e.g., when the source is quasi-stationary). Nevertheless, as in [18], suppose we know the power of the source in a set of sub-bands $\mathcal{B}_1 \dots \mathcal{B}_N$, where $\bigcup_{i=1}^{N} \mathcal{B}_i = (0, B)$. A simple modification of S^* may be used to encode the source X in this case. Let W^* be a Gaussian process that has the same power as X in each one of the sub-bands, but whose spectral density is flat over the sub-bands. Let S^*_W denote the pre/post filtered ECDQ which encodes W^* with distortion D. According to the definition in Section IV, the pre- and post- filters of S^*_W have a constant gain in each of the sub-bands, thus S^*_W yields the same distortion, D, in coding the source X. Furthermore, a slight modification of Theorem 3 shows that the rate of S^*_W in encoding X satisfies

$$R_Q^{S^*_W}(D) \leq R^*_W(D) + B_D \cdot \log(2\pi e G_K) \ , \tag{53}$$

where $R^*_W(D)$ is the rate-distortion function of W^* at distortion D, and B_D here is the bandwidth associated with $R^*_W(D)$. The other results of Section IV can be generalized to this case in a similar manner.

3. Sub-band coding and lattice shaping: The scheme S^* corresponds to a direct encoding of the source in the time domain, using a quantizer with many (actually infinitely many) levels. The rate reduction relies heavily on the fact that the quantizer is followed by a lossless encoder. In practice, however, linear techniques such as prediction or transformation are often used for data rate reduction instead of (or in addition to) entropy coding. We show below that our scheme may be easily modified to fit these practical methods. The modification is based on the invariance property of the forward channel realization (26). Specifically, the scheme $\widetilde{S} = (\mathcal{T}L_1, \mathcal{T}Q_k, L_2\mathcal{T}^{-1})$ is equivalent to $S = (L_1, Q_k, L_2)$, for any linear transformation $\mathcal{T}$ which is invertible with respect to the process L_1X. The notation $\mathcal{T}Q_K$ stands for a quantizer obtained from Q_K by *shaping* according to the procedure described in [25] using the transformation $\mathcal{T}$. The resulting modified scheme and its equivalent channel are depicted in Figures 7(a) and 7(b), respectively. Clearly, from the equivalent channel, the reconstructed source is unaffected by this transformation. Furthermore, since $\mathcal{T}$ is invertible, it is shown in Appendix E that

$$\widetilde{R}_Q = \overline{I}(\mathcal{T}X_q; \mathcal{T}X_q + \mathcal{T}N_q) = \overline{I}(X_q; X_q + N_q) \ , \tag{54}$$

i.e., the coding rate is preserved as well. Note that there is an interesting similarity between the notion of scheme transformation, and the general companding model suggested in [8] for (sub-optimal) vector quantization by a lattice quantizer.

The transformation $\mathcal{T}$ can be the Fourier transform, and in this case we get an incorporation of the ECDQ in sub-band coding. In the vector case in general, the modified scheme corresponds to the incorporation of ECDQ in transform coding. If we choose $\mathcal{T}$ to be the appropriate whitening transformation, the modified scheme will correspond to predictive coding. It may be shown that in both cases mentioned above, the uncoded bit rate, i.e., the bit rate in the input of the lossless encoder, is reduced to a value close to $R^*(D)$ for either a Gaussian or a non-Gaussian source. See [22] for the details.

4. The min-max property of S^*: As mentioned in Section II, the forward channel configuration is not necessarily optimal among all possible ECDQ schemes incorporating linear filters. Nevertheless, by using the results of Section IV we can justify the choice of $S^* = \{L_1^*, Q_\infty, L_2^*\}$, associated with large lattice dimension, which has an equivalent additive white Gaussian noise channel. For that, we suggest a min-max argument, associated with the notion of *robust quantization* in the class of sources

$$\mathcal{X}_S = \{X \ : \ X \text{ has a power spectrum } S_x(f)\} \ . \tag{55}$$

More specifically, let $\mathcal{S}_D$ denote the set of all linear pre/post-filtered lattice-ECDQ schemes, i.e., the set of all triplets (L_1, Q_K, L_2), having a mean-squared error D for $X \in \mathcal{X}_S$ (note that a linear scheme has the same error for all $X \in \mathcal{X}_S$). Let R_Q^S denote the coding rate of some $S \in \mathcal{S}_D$. We are looking for that scheme in $\mathcal{S}_D$ that minimizes the coding rate for the "worst source" in $\mathcal{X}_S$. A similar approach was taken in [18].

Since the rate of the ECDQ is equal to the mutual information in an additive noise channel, this problem is equivalent to minimizing the capacity of an additive noise channel under spectral constraints. Now, for a given set of spectral constraints, i.e., source spectrum, pre/post-filters and noise spectrum, it is well known that the channel capacity is minimized by a Gaussian noise, and this minimal capacity is achieved by a Gaussian source; see, e.g., [7], p. 263. This implies that for any fixed filters and noise spectrum (associated with some lattice quantizer), the desirable scheme corresponds to $K \to \infty$, and the "worst source" is $X^* \in \mathcal{X}_S$. Furthermore, given that the noise and the source are Gaussian, the minimal rate is of course the rate-distortion function $R^*(D)$, which is actually achieved by the forward channel configuration. We have thus proved:

Theorem 7

$$\min_{S \in \mathcal{S}_D} \max_{X \in \mathcal{X}_S} R_Q^S(D) = R^*(D) \ , \tag{56}$$

and the minimum, over all pre/post-filtered ECDQ schemes, is attained by $S^* = (L_1^*, Q_\infty, L_2^*)$.

5. Non-linear pre/post processors: If any pre- and post-processors can be incorporated with the ECDQ, then clearly, due to the coding theorem, using complex enough processors the rate-distortion function can be achieved. Interestingly, it turns out that in the case in which the linear scheme S^* suffers from the maximum redundancy, the incorporation of a simple non-linear mechanism can significantly improve the performance, as shown in the following example. Consider an i.i.d. source which is a mixture of two Gaussian sources, i.e., each sample X is distributed as

$$X \sim \frac{1}{2}\left(\phi^*(x-\mu) + \phi^*(x+\mu)\right) \tag{57}$$

where $\phi^*(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal density and $\mu \gg 1$; see Figure 8. The power of X is approximately $\mu^2 + 1$ and suppose the allowed distortion is $D = 1$. By the scalar example of Section II, the pre- and post- filters of S^* are the constant gains $\alpha^2 = \beta^2 \simeq 1 - \frac{1}{\mu^2+1} \simeq 1$, and $\epsilon = D = 1$. Define the binary random variable $V = \mathrm{sign}(X)$. It is easy to see that for Gaussian quantization noise, we get $R_Q^{S^*}(D) \simeq 0.5 + H(V) = 1.5$ bits per sample, while $R(D) \simeq H(V) = 1$, i.e., S^* has the maximal possible redundancy of $\frac{1}{2}\log 4\pi e G_\infty = 0.5$ bit per sample. On the other hand, an almost zero redundancy can be achieved by using the simple scheme $\widehat{X} = \mu V$, whose rate is 1 and its MSE distortion is about 1. More generally, for an allowed $D < 1$, $R(D)$ can be approached by a simple non-linear device in which V selects between two linear sub-systems, each adapted to one of the two "modes" of X. Now, the structure suggested in this simple example can be applied in other "multi-modal" examples, where an enhanced performance can be obtained by utilizing a simple switching device which dynamically selects from a set of linear sub-systems.

Acknowledgments

The authors thank Naftaly Hayat, Tamas Linder and the anonymous reviewers for helpful suggestions.

Appendix

A. Monotonic Behavior of the Redundancy for Gaussian Sources

In this appendix we show that for the Gaussian source $X^* \sim \mathcal{N}(0, \sigma_x^2)$, the coding rate redundancy $H(Q_1|Z) - R^*(D)$ is monotonically decreasing as $D \to \sigma_x^2$. Using (7) and (8), the redundancy may be written as

$$h(\alpha X^* + N) - \log \Delta - \frac{1}{2}\log(\sigma_x^2/D) \tag{A.1}$$

where $N \sim \mathcal{U}(-\Delta/2, \Delta/2)$, $\Delta = \sqrt{12D}$ and $\alpha = \sqrt{1 - D/\sigma_x^2}$. Define $\lambda = D/\sigma_x^2$, and the random variable $M = N/\sqrt{\lambda}$. Note that $E\{M^2\} = \sigma_x^2$, so its Gaussian counterpart M^* is equal to X^* in the distribution sense. Thus, we may rewrite (A.1) as

$$H(Q_1|Z) - R^*(D) = h\left(\sqrt{\lambda}M + \sqrt{1-\lambda}M^*\right) - \frac{1}{2}\log(12\sigma_x^2)\ , \tag{A.2}$$

where as D goes from 0 to σ_x^2, λ goes from 0 to 1. From [1, Lemma 1] it follows that the entropy of the weighted sum in (A.2) is a monotonically decreasing function of λ, where for $\lambda = 0$ it is $h(M^*) = \frac{1}{2}\log 2\pi e\sigma_x^2$, while for $\lambda = 1$ it is $h(M) = \log(\Delta/\sqrt{\lambda}) = \frac{1}{2}\log(12\sigma_x^2)$. This proves our claim.

B. Proof of Theorem 4 (high distortion limit)

We distinguish between two possible cases. First we consider the case where $S_x(f)$ is unbounded, implying $\theta \to \infty$ as $D \to \sigma_x^2$. Since the source power is finite $\sigma_x^2 < \infty$ and since $\sigma_x^2 = \int S_x(f)df = \int B_D d\theta$, it follows that $B_D = o(\frac{1}{\theta}) \to 0$ as $D \to \sigma_x^2$. Inserting $B_D \to 0$ into (37) we get that $\lim_{D\to\sigma_x^2} R_Q^{S^*} \leq \lim_{D\to\sigma_x^2} R^*(D) + \lim_{D\to\sigma_x^2} 2B_D \cdot \frac{1}{2}\log 2\pi e G_k = 0$, since $R^*(D) \to 0$ as $D \to \sigma_x^2$.

The second case is where $S_x(f)$ is bounded by some $S_{max} < \infty$, implying that $\theta \to S_{max}$ as $D \to \sigma_x^2$. In this case we first show that the signal-to-quantization-noise ratio, defined as $SQNR = E\{X_q^2\}/E\{N_q^2\} = \sigma_q^2/\epsilon$, where σ_q^2 is the power of the pre-filtered source at the quantizer input, vanishes at high distortion: We have (32)

$$SQNR = \frac{1}{\epsilon}\int_{\mathcal{B}_D} |H_1(f)|^2 S_x(f) df = \frac{\int_{\mathcal{B}_D}(S_x(f)-\theta)df}{\theta \cdot B_D} \tag{B.1}$$

and, since $S_x \leq S_{max}$ and $\theta \to S_{max}$,

$$SQNR \leq \frac{B_D(S_{max}-\theta)}{\theta B_D} \to 0 \ \text{ as } \ D \to \sigma_x^2 \ . \tag{B.2}$$

We then show that $SQNR \to 0$ implies $\overline{I} \to 0$ in the equivalent channel: Since the mutual-information is invariant to invertible transformation, we may write the coding rate as

$$R_Q^{S^*} = \overline{I}(X_q; X_q + N_q) = \overline{I}(X_q; \frac{1}{\sqrt{\epsilon}}(X_q + N_q)) = F_s \cdot \left\{ \overline{h}\left(\frac{N_q}{\sqrt{\epsilon}} + \frac{X_q}{\sqrt{\epsilon}}\right) - \overline{h}\left(\frac{N_q}{\sqrt{\epsilon}}\right)\right\} \ . \tag{B.3}$$

Now, from the assumption that the lattice quantizer is only scaled when D varies, the process $N_q/\sqrt{\epsilon}$ is independent of ϵ, and thus $\overline{h}(N_q/\sqrt{\epsilon}) = \frac{1}{2}\log(1/G_k)$; see [25]. On the other hand, we have from (B.1) and (B.2) that the power of the process $X_q/\sqrt{\epsilon}$ vanishes as $D \to \sigma_x^2$. Thus, we may apply Corollary 3 in [14] to conclude that, since $\overline{h}(N_q/\sqrt{\epsilon}) > -\infty$,

$$\overline{h}\left(\frac{N_q}{\sqrt{\epsilon}} + \frac{X_q}{\sqrt{\epsilon}}\right) \longrightarrow \overline{h}\left(\frac{N_q}{\sqrt{\epsilon}}\right) \ , \ \text{ as } D \to \sigma_x^2 \ . \tag{B.4}$$

Combining (B.4) with (B.3) and using the fact that $F_s = 2B_D$ is finite for all $D > 0$ completes the proof. $\square$

As we have shown above, either the sampling rate $2B_D$ or the SQNR must vanish at high distortion, but not necessary both. For the flat spectrum source, for instance, the sampling rate does not vanish as $D \to \sigma_x^2$. Also, interestingly, the $SQNR$ may not vanish at high distortion, for a signal with unbounded power spectrum. This implies that the coding rate in *bits per sample* may not vanish in the limit $D \to \sigma_x^2$. For example, consider a bandlimited source with a "$1/f$" power spectral density $S_x(f) \propto \frac{1}{f^a}$, $0 < a < 1$ at some low frequency region, where $\propto$ denotes proportion. Then for small f, $\mathcal{B}_D = (0, B_D)$, $\theta \propto \frac{1}{B_D^a}$, $\epsilon = \theta B_D \propto B_D^{1-a}$, $\sigma_q^2 = \int_{\mathcal{B}_D}(S_x(f)-\theta)df \propto \frac{a}{1-a}B_D^{1-a}$ and thus, for small D, $SQNR = \frac{a}{1-a} > 0$.

C. Proof of Theorem 5 (low distortion limit)

For D small enough we get $|H_1(f)| = |H_2(f)| = 1_{\{0,B\}}$, where $1_{\{0,B\}}$ denotes an ideal low pass filter. Now, by (54), the scheme (H_1, Q_K, H_2) has the same rate-distortion performance as the scheme $\left(1_{\{0,B\}},\ H_2 \cdot Q_K,\ 1_{\{0,B\}}\right)$, where $H_2 \cdot Q_K$ denotes a *shaped* lattice quantizer which is white and has the same G_K as Q_K; see [25]. The modified scheme coincides with that of [23], and hence the rest of the proof follows from the proof of Theorem 4 in [23] (the Nyquist sampling rate case). □

D. Proof of Theorem 6 (the Capacity Bound)

The proof follows the technique used in the proof of Theorem 5 in [23]. Let $\{U_n, n = 0, \pm 1, \pm 2, \ldots\}$ be an arbitrary random sequence, jointly stationary with $\{X_n, n = 0, \pm 1, \pm 2\}$, the Nyquist samples of the source X, and statistically independent of the quantization noise N_q. Let

$$U(t) = \sum_n U_n \frac{\sin \pi(2Bt - n)}{\pi(2Bt - n)} \quad , -\infty < t < \infty \ , \tag{D.1}$$

and let $U_q = \{U_{q,n}, n = 0, \pm 1, \pm 2, \ldots\}$ be the output of the pre-processor L_1^* (see Figure 5) when fed by $U = \{U(t), -\infty < t < \infty\}$. It follows from the proof of Theorem 2 of [24] that, for any such U and any block length n, $I(\underline{X}_q; \underline{X}_q + \underline{N}_q) \leq I(\underline{X}_q; \underline{U}_q) + I(\underline{X}_q - \underline{U}_q; \underline{X}_q - \underline{U}_q + \underline{N}_q)$ where $\underline{X}_q$, $\underline{N}_q$ and $\underline{U}_q$ are n-vectors of the corresponding processes. Dividing by n and taking the limit, we get the same inequality for the information rates, i.e.,

$$\overline{I}(X_q; X_q + N_q) - \overline{I}(X_q; U_q) \leq \overline{I}(X_q - U_q; X_q - U_q + N_q) \ , \tag{D.2}$$

provided that $\overline{I}(X_q; U_q)$ exists. Let

$$I_L = \inf_{\{U:\, E(U(t) - X(t))^2 \leq D\}} \overline{I}(X_q; U_q) \ .$$

In Lemma 1 below we prove that $R(D) \geq 2B_D \cdot I_L$. Furthermore, from (40), $R_Q^{S^*}(D) = 2B_D \cdot \overline{I}(X_q; X_q + N_q)$. Thus, by (D.2),

$$R_Q^{S^*}(D) - R(D) \ \leq \ 2B_D \cdot \left(\overline{I}(X_q; X_q + N_q) - I_L\right)$$

$$\leq \; 2B_D \cdot \sup_{\{X,U\colon E(U(t)-X(t))^2\leq D\}} \overline{I}(X_q-U_q; X_q-U_q+N_q)$$

$$= \; C_D \; , \tag{D.3}$$

where the last equality follows from the definition (47) of the capacity C_D. (Note that $X_q - U_q$ is the output of L_1^* when it is fed by $X(t) - U(t)$.) □

Lemma 1

$$R(D) \geq 2B_D \cdot \inf_{\{U\colon E(U(t)-X(t))^2\leq D\}} \overline{I}(X_q; U_q) \; , \tag{D.4}$$

where the infimum is over processes U jointly stationary with X, and where X_q and U_q are defined above.

Proof: Following [9, Thm. 10.6.1], the block definition (29) of $R(D)$ coincides with its process definition, i.e.,

$$R(D) = 2B \cdot \lim_{n\to\infty} \frac{1}{n} \inf_{\{\underline{U}\colon \frac{1}{n}E\|\underline{U}-\underline{X}\|^2\leq D\}} I(\underline{X};\underline{U}) = 2B \cdot \inf_{\{U\colon E(U_n-X_n)^2\leq D\}} \overline{I}^{(g)}(\{X_n\};\{U_n\}) \; , \tag{D.5}$$

where $I^{(g)}$ is the *Pinsker definition of the information rate* (see [9, pp. 135-141]). Note that by the definition of X_n and U_n, $E(U_n - X_n)^2 = E(X(t) - U(t))^2$. Since the processes $X_q \to \{X_n\} \to \{U_n\} \to U_q$ form a Markov chain, and since the Pinsker rate satisfies the data-processing theorem (see [17, p. 95]), it follows that

$$2B \cdot \overline{I}^{(g)}(\{X_n\};\{U_n\}) \geq 2B_D \cdot \overline{I}^{(g)}(X_q;U_q) \; . \tag{D.6}$$

Now, similarly to the proof of [9, Thm. 10.6.1], it may be shown that

$$\inf_{\{U\colon E(U(t)-X(t))^2\leq D\}} \overline{I}^{(g)}(X_q;U_q) = \inf_{\{U\colon E(U(t)-X(t))^2\leq D\}} \overline{I}(X_q;U_q) \; , \tag{D.7}$$

i.e., that the infimum of the Pinsker rate coincides with the infimum of the regular mutual information rate. Combining (D.5), (D.6) and (D.7) yields the proof of the lemma. □

E. Invariance under Transformation

In [23] it is shown that the mutual information rate (39), which is the coding rate of S^*, may be written also as

$$\overline{I}(X_q; X_q + N_q) = \overline{I}^{(g)}(X_q; X_q + N_q) \tag{E.1}$$

where $I^{(g)}$ is the Pinsker rate (see above). The invariance property in (54) follows from the fact that Pinsker rate does not change by applying an invertible transformation to its arguments; see [17, p. 95]. Note that this statement is not true, in general, for the regular definition of the information rate.

References

[1] A.R. Barron. Entropy and the central limit theorem. *The Annals of Probability*, 14, No. 1:336–342, 1986.

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

[3] J. Binia, M. Zakai, and J. Ziv. On the ϵ entropy and the rate-distortion function of certain non-Gaussian processes. *IEEE Trans. Information Theory*, IT-20:517–524, July, 1974.

[4] J. C. Candy and G. C. Temes. *Oversampling Delta-Sigma Data Converters.* IEEE press, New York, 1992.

[5] D. Chan and R. W. Donaldson. Optimum pre- and post-filtering of sampled signals with application to pulse modulation and data compression systems. *IEEE Trans. on Communication Technology*, COM-19:141–156, April, 1971.

[6] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag, New York, N.Y., 1988.

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

[8] A. Gersho. Asymptotically optimal block quantization. *IEEE Trans. Information Theory*, IT-25:373–380, July 1979.

[9] R. M. Gray. *Entropy and Information Theory.* Springer-Verlag, New York, N.Y., 1990.

[10] R. M. Gray. Quantization noise spectra. *IEEE Trans. Information Theory*, IT-36:1220–1244, Nov. 1990.

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

[12] W. C. Kellogg. Information rates in sampling and quantization. *IEEE Trans. Information Theory*, IT-13:506–511, July, 1967.

[13] T. Linder, G. Lugosi, and K. Zeger. Fixed rate universal lossy source coding and rates of convergence for memoryless sources. *IEEE Trans. Information Theory*, IT-41:665–676, May 1995.

[14] T. Linder and R. Zamir. On the asymptotic tightness of the Shannon lower bound. *IEEE Trans. Information Theory*, IT-40:2026–2031, Nov. 1994.

[15] T. Linder and K. Zeger. Asymptotic entropy constrained performance of tesselating and universal randomized lattice quantization. *IEEE Trans. Information Theory*, IT-40:575–579, March 1994.

[16] R.A. McDonald and P.M. Schultheiss. Information rates of Gaussian signals under criteria constraining the error spectrum. *Proc. IEEE*, 52:415–416, 1964.

[17] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes.* Holden Day, San Francis. CA., 1964.

[18] D.J. Sakrison. The rate of a class of random processes. *IEEE Trans. Information Theory*, IT-16:10–16, Jan, 1970.

[19] A. Segall. Bit-allocation and encoding for vector sources. *IEEE Trans. Information Theory*, IT-22:162–169, 1976.

[20] A.D. Wyner. On the transmission of correlated Gaussian data over a noisy channel with finite encoding block length. *Information and Control*, 20:193–215, 1972.

[21] A.D. Wyner and J. Ziv. Bounds on the rate-distortion function for stationary sources with memory. *IEEE Trans. Information Theory*, IT-17:508–513, 1971.

[22] R. Zamir. *Universal Encoding of Signals by Entropy Coded Dithered Quantization*. PhD thesis, Tel-Aviv University, 1994.

[23] R. Zamir and M. Feder. Rate distortion performance in coding band-limited sources by sampling and dithered quantization. *IEEE Trans. Information Theory*, IT-41:141–154, Jan. 1995.

[24] R. Zamir and M. Feder. On universal quantization by randomized uniform / lattice quantizer. *IEEE Trans. Information Theory*, pages 428–436, March 1992.

[25] R. Zamir and M. Feder. On lattice quantization noise. *IEEE Trans. Information Theory*, revised, Jan. 1996.

[26] J. Ziv. On universal quantization. *IEEE Trans. Information Theory*, IT-31:344–347, May 1985.

Figure 1: (a) Pre/post filtered ECDQ, and (b) its equivalent additive noise channel.

Figure 2: Scheme II - (a) Scalar ECDQ with multiplicative factors, and (b) its equivalent channel.

Figure 3: Information rates of a memoryless Gaussian source, encoded by a scalar ECDQ: (A) with filters, (B) without filters, (C) the source's R-D function.

Figure 4: The "water pouring law".

Figure 5: The scheme S^*.

Figure 6: Typical rate-distortion curves for a non-Gaussian source.

Figure 7: (a) Linear transformation of pre/post filtered ECDQ, and (b) its equivalent channel.

Figure 8: A multi-modal Gaussian source which achieves the capacity bound on the redundancy for $K \to \infty$.