Data Processing Bounds for Scalar Lossy Source Codes with Side Information at the Decoder *

Avraham Reani and Neri Merhav
Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
Emails: [avire@tx, merhav@ee].technion.ac.il

September 10, 2012

Abstract

In this paper, we introduce new lower bounds on the distortion of scalar fixed-rate codes for lossy compression with side information available at the receiver. These bounds are derived by presenting the relevant random variables as a Markov chain and applying generalized data processing inequalities a la Ziv and Zakai. We show that by replacing the logarithmic function with other functions, in the data processing theorem we formulate, we obtain new lower bounds on the distortion of scalar coding with side information at the decoder. These bounds are shown to be better than one can obtain from the Wyner-Ziv rate-distortion function.

Index Terms: side information, Wyner-Ziv problem, Ziv-Zakai bounds, source coding, on-line schemes, scalar coding, Rényi entropy, Rate-Distortion theory

1 Introduction

The Wyner–Ziv (WZ) problem has received very much attention during the last three decades. There were several attempts to develop practical schemes for lossy coding in the WZ setting, by using codes with certain structures that facilitate the encoding and the decoding. Most notably, these studies include nested structures of linear coset codes (in the role of bins) for discrete sources, and nested lattice structures for continuous valued sources, see e.g., [2], [3]. Other directions of introducing structure into WZ coding are associated with trellis/turbo/LDPC designs ([4] and references therein) and with progressive coding, i.e., successive refinement with layered code design [5], [6]. The case of scalar source codes for the WZ problem was also handled in several papers, e.g. [7] and [8]. Zero-delay coding strategies for the WZ problem, were introduced in [9], where structure theorems for fixed-rate codes, under the assumption of a Markov source, were given. These results were later extended in [10], to include variable-rate coding. In [11] and [12] it was conjectured that under the high-resolution assumption, the optimal quantization level density is periodic. In addition, zero-delay schemes for specific source-side information correlation were presented in [11], [12] and [13]. Zero-delay coding of individual sequences under the conditions of the WZ problem was considered in [14], where existence of universal schemes for fixed-rate and variable-rate coding was established.

In this paper, we develop lower bounds on the distortion in the scalar WZ setting. We generalize the results of [15] and [16], concerning functionals satisfying a data-processing theorem, to this setting. In [15] it was shown that the rate-distortion (RD) bound ($R(D) \leq C$) remains true when the negative logarithm function, in the definition of mutual information, is replaced by an arbitrary convex, non–increasing function satisfying some technical conditions. For certain choices of this convex function, the bounds obtained were better than the classical RD bounds. These results were substantially generalized in [16] to apply to even more general information measures. The methods of [15] were also used in [17], [18] and [19]. In these papers, lower bounds on the distortion of delay-constrained joint source-channel coding were given. These bounds were obtained by combining the Rényi information measure [20] with the generalized data processing theorem of [15], and under high-resolution and high SNR approximations. Another related work is [21], where certain degrees of freedom of the Ziv-Zakai generalized mutual information were further exploited

in order to get better bounds.

We start by presenting the relevant random variables of the WZ problem as a Markov chain. Then, using a data processing theorem, we obtain lower bounds on the distortion. We show that replacing the logarithmic function by other functions, may give better bounds on the distortion of delay–limited coding (in particular, for scalar coding) in the WZ setting. Examples of non-trivial lower bounds for scalar coding, in this setting, are obtained using the convex function $Q(t) = t^{1-\alpha}$, $\alpha > 1$, which is equivalent to using the Rényi information measure. The importance of such bounds stems from the fact that finding the optimal scalar code in the WZ setting is, in general, a hard problem. In fact, it is a problem of finding an optimal partition of the source alphabet and this partition does not necessarily correspond to intervals. A main objective will be to use these bounds for studying the performance of concrete coding schemes.

The remainder of the paper is organized as follows. In Section 2, we present our formulation of the WZ problem and establish a generalized data processing theorem (DPT) for this setting. In Subsection 2.1, we define the fixed-rate scalar coding case. We then give an upper bound on the generalized capacity, which is one component of the above DPT. In Subsection 2.2, we handle the second component of the generalized DPT, i.e., the generalized RD function. We start with a general characterization of this function. Then, we introduce a closed-form expression of the generalized RD function for uniformly distributed sources w.r.t. general symmetric distortion measures. In Section 3, we use the results of Section 2 to obtain non-trivial lower bounds on the distortion of scalar coding in the WZ setting in several cases. Finally, we demonstrate that for large alphabets, non-trivial bounds can be derived for various channels and as a result, the performance range for scalar coding can be given.

2 Problem Formulation and Results

In this section, we present the relevant random variables of the WZ problem as a Markov chain and establish a generalized data processing theorem (DPT) for this setting, using the method of [15].

We begin with notation conventions. Capital letters represent scalar random variables, specific realizations of them are denoted by the corresponding lower case letters and their alphabets - by calligraphic letters. The inner product of the two vectors $\vec{a}$ and $\vec{b}$ will be

denoted by $\vec{a} \cdot \vec{b}$. Logarithms are defined to the base 2.

We consider a memoryless source producing a random sequence $X_1, X_2, \ldots X_i \in \mathcal{X}$, $i = 1, 2, \ldots$, where $\mathcal{X}$ is a finite alphabet with cardinality K. Without loss of generality, we define this alphabet to be the set $\{1, 2, \ldots, K\}$. The probability mass function of X, $p(x)$, is known. A fixed-rate scalar source code with rate $R = \log M$,[1] partitions $\mathcal{X}$ into M disjoint subsets $(A_1, A_2, \ldots, A_M)$, $M \leq K$. The encoder maps X_i into a channel symbol Z_i, using a function $f : \mathcal{X} \to \{1, 2, \ldots, M\}$, that is, $Z_i = f(X_i)$. The decoder, in addition to Z_i, has access to a random variable Y_i, which is dependent on X_i via a known discrete memoryless channel (DMC), defined by the single-letter transition probability matrix $\{p(y|x)\}$, whose entries are the conditional probabilities of the different channel output symbols given the channel input symbols. Based on Z_i and Y_i, the decoder produces the reconstruction $\hat{X}_i$, using a decoding function $g : \{1, 2, \ldots, M\} \times \mathcal{X} \to \mathcal{X}$, i.e., $\hat{X}_i = g(Z_i, Y_i)$. This setting is depicted in Fig. 1. For simplicity, we assume that X_i, Y_i and $\hat{X}_i$, all take on values in the same finite alphabet $\mathcal{X}$. The distortion in this setting is defined to be:

$$D = \mathbb{E}\rho(X_i, \hat{X}_i) = \sum_{x,y} p(x,y)\rho(x,\hat{x}) \tag{1}$$

where $p(x, y)$ is the joint distribution of x and y and $\rho(x, \hat{x})$ is a distortion measure.

Let $Q(t)$, $0 \leq t < \infty$, be a real-valued convex function, where $\lim_{t \to 0} t \cdot Q(1/t) = 0$. This requirement implies that $Q(t)$ is non-increasing, as was shown in [15]. We define $0 \cdot Q(r/0) = 0$, for all $0 \leq r < \infty$. The generalized mutual information relative to the function Q is defined as

$$I^Q(X;Y) = \sum_{x,y} p(x,y)Q\left(\frac{p(y)}{p(y|x)}\right). \tag{2}$$

We apply the generalized DPT [15, Theorem 3] in the following way:

$$I^Q(X;\hat{X}) \leq I^Q(X;Y,Z) \tag{3}$$

where we have used the fact that $X \leftrightarrow (Y, Z) \leftrightarrow \hat{X}$ is a Markov chain. Since $Z \leftrightarrow X \leftrightarrow Y$ is also a Markov chain, we have:

$$p(x,y,z) = p(x)p(y|x)p(z|x) \tag{4}$$

---

[1] Through this paper, the symbols of Z are not necessarily transformed into bits. Therefore, $\log M$ need not necessarily be an integer.
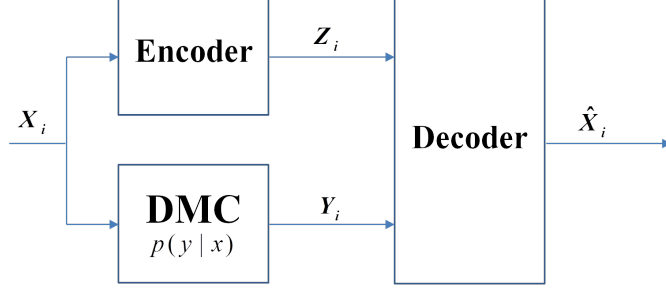
Figure 1: The WZ setting

and $I^Q(X;Y,Z)$ is given by:

$$I^Q(X;Y,Z) = \sum_{x,y,z} p(x)p(y|x)p(z|x)Q\left(\frac{p(y)p(z|y)}{p(y,z|x)}\right)$$

$$= \sum_{x,y,z} p(x)p(y|x)p(z|x)Q\left(\frac{p(y)p(z|y)}{p(y|x)p(z|x)}\right)$$

$$= \sum_{x,y,z} p(x)p(y|x)p(z|x)Q\left(\frac{\displaystyle\sum_{\tilde{x}} p(\tilde{x})p(y|\tilde{x})p(z|\tilde{x})}{p(y|x)p(z|x)}\right)$$

$$= \sum_{y}\sum_{z}\sum_{x} p(x)p(y|x)p(z|x)Q\left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x)p(z|x)}\right) \tag{5}$$

where we have defined the following K-dimensional vectors $\{\vec{p}_z\}_{z=1}^{M}$:

$$\vec{p}_z = [p(z|x), x \in \mathcal{X}] \tag{6}$$

and the following K-dimensional vectors $\{\vec{p}_y\}$, $y \in \mathcal{X}$:

$$\vec{p}_y = [p(x,y), x \in \mathcal{X}]. \tag{7}$$

By definition of $\{\vec{p}_z\}_{z=1}^{M}$, we have the following property:

$$\sum_{z=1}^{M} \vec{p}_z = [1,1,\ldots,1]. \tag{8}$$

We now define the following functions $\{G_y(\vec{p}_z)\}$, $y \in \mathcal{X}$:

$$G_y(\vec{p}_z) = \sum_{x} p(x)p(y|x)p(z|x)Q\left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x)p(z|x)}\right). \tag{9}$$

5

Using these functions, Eq. (6) becomes:

$$I^Q(X; Y, Z) = \sum_y \sum_z G_y(\vec{p}_z).$$ (10)

The functions $G_y(\vec{p}_z)$ have the following property:

Lemma 1. *For any convex function Q, the functions $\{G_y(\vec{p}_z)\}$, $y \in \mathcal{X}$, are convex.*

The proof is given in Appendix A. This convexity property has important implications in the optimization of $I^Q(X; Y, Z)$, as will be discussed later. Assuming the encoder is given by a deterministic function $f : \mathcal{X} \to \{1, \ldots, M\}$, Eq. (6) becomes the following:

$$
\begin{aligned}
I^Q(X; Y, Z) &= \sum_y \sum_x p(x)p(y|x) \sum_z p(z|x) Q\left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x)p(z|x)} \right) \\
&= \sum_y \sum_x p(x)p(y|x) Q\left(\frac{\vec{p}_{f(x)} \cdot \vec{p}_y}{p(y|x)} \right) \\
&= \sum_{x,y} p(x,y) Q\left(\frac{\sum\limits_{\tilde{x} \in A_z} p(\tilde{x}, y)}{p(y|x)} \right)
\end{aligned}
$$ (11)

where $z = f(x)$ and $A_z \equiv \{\tilde{x} : f(\tilde{x}) = z\}$. Remember that we have defined $0 \cdot Q(r/0) = 0$. Using $Q(t) = -\log t$ in (12), thus turning back to the classical DPT, we next show the following result:

$$R(D) - I(X; Y) \leq \sup H(Z|Y)$$ (12)

where $R(D)$ is the classical RD function and the supremum is taken over all partitions of $\mathcal{X}$ into M disjoint subsets. This inequality stems from the Markov properties of the WZ problem we discussed before. We see that given a rate $R = \log M$, we should find the encoder that maximizes $H(Z|Y)$. This is not surprising as, intuitively, we want the amount of information that Y has on Z to be as little as possible, to decrease the redundancy. Ideally, we want the encoder output and the side information to be independent. This is indeed achieved by the block coding scheme of Wyner and Ziv, in the limit of infinite block length. The term $\sup \{H(Z|Y)\} + I(X; Y)$ will be referred to as the "capacity" of the generalized channel between (X, Z) and Y. This channel is composed of the DMC between X and Y and a noiseless channel with capacity $\log M$ for the encoder's output. Since the source distribution is given, the maximum rate of reliable communication over this channel is indeed $I(X; Y) + \sup H(Z|Y)$.

6

Proof of Eq. (12). Using the function $Q(t) = -\log t$ in (12), we get:

$$
\begin{aligned}
I^Q(X;Y,Z) &= H(Y,Z) - H(Y,Z|X) \\
&= H(Y) + H(Z|Y) - H(Y|X) - H(Z|Y,X) \\
&= I(X;Y) + H(Z|Y) \tag{13}
\end{aligned}
$$

where we have used the fact that $H(Z|Y,X) = H(Z|X) = 0$ since Z is a deterministic function of X. On substituting into (3), we get:

$$
\begin{aligned}
R(D) &\leq I^Q(X;\hat{X}) \\
&\leq I^Q(X;Y,Z) \\
&= I(X;Y) + H(Z|Y) \\
&\leq I(X;Y) + \sup H(Z|Y), \tag{14}
\end{aligned}
$$

which is equivalent to Eq. (12). $\qquad\square$

Notice that if we allow non-deterministic encoders, as in Eq. (6), we get the following:

$$
R(D) \leq I(X;Y) + \sup\left\{H(Z|Y) - H(Z|X)\right\}, \tag{15}
$$

where the supremum is taken over the same set as in (12). Although randomizing the encoder can increase $H(Z|Y)$, it will also increase $H(Z|X)$. Due to the convexity property presented in Lemma 1, the supremum is achieved by a deterministic encoder, as will be discussed in the next section. Therefore, randomizing the encoder cannot improve the bound in this setting.

In Section 3, we show some examples of scalar coding, where this result gives us lower bounds on the distortion, which are better than the bounds obtained from the classical inequality $R_{WZ}(D) \leq \log M$, where $R_{WZ}(D)$ is the WZ RD function.

2.1 Generalized DPT for fixed-rate scalar coding

Assuming a deterministic encoder, the vectors $\{\vec{p}_z\}_{z=1}^{M}$, defined in (6), become:

$$
\vec{p}_z = [1_{1\in A_z}, 1_{2\in A_z}, \ldots, 1_{K\in A_z}], \tag{16}
$$

where 1_B is the indicator function of the event B. The jth coordinate of $\vec{p}_z$ is 1 if $j \in A_z$ and 0 elsewhere. Using these vectors, we can rewrite (12) in the following way:

$$
\begin{aligned}
I^Q(X;Y,Z) &= \sum_{x,y} p(x,y) Q\left(\frac{\vec{p}_{z(x)} \cdot \vec{p}_y}{p(y|x)} \right) \\
&= \sum_y \sum_z \sum_{x \in A_z} p(x,y) Q\left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x)} \right) \\
&= \sum_y \sum_z \vec{p}_z \cdot \vec{q}_{z,y} \\
&= \sum_y \sum_z \Gamma_y(\vec{p}_z),
\end{aligned}
\tag{17}
$$

where we have defined the following K-dimensional vectors:

$$
\vec{q}_{z,y} = \left[p(x_1,y) Q\left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x_1)} \right), p(x_2,y) Q\left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x_2)} \right), \dots \right]
\tag{18}
$$

and the set of functions $\{\Gamma_y\}_{y=1}^K$, $\Gamma_y : \mathbb{R}^K \to \mathbb{R}$:

$$
\Gamma_y(\vec{p}_z) = \vec{p}_z \cdot \vec{q}_{z,y}.
\tag{19}
$$

Notice that the vector $\vec{p}_y$ depends only on y and that the inner product $\vec{p}_z \cdot \vec{p}_y$ is a function of z and y. Applying the RD bound [15, Theorem 4], we get:

$$
R^Q(D) \leq I^Q(X;\hat{X}) \leq I^Q(X;Y,Z) \leq C^Q,
\tag{20}
$$

where

$$
R^Q(D) = \inf I^Q(X;\hat{X})
\tag{21}
$$

and

$$
\begin{aligned}
C^Q &= \sup I^Q(X;Y,Z) \\
&= \sup \sum_y p(y) \sum_z \Gamma_y(\vec{p}_z).
\end{aligned}
\tag{22}
$$

This gives us the following lower bound on the distortion D:

$$
D \geq D^Q(C^Q),
\tag{23}
$$

where $D^Q(R)$ is the inverse function of $R^Q(D)$. The infimum is taken over all conditional distributions $\{p(\hat{x}|x)\}$ that satisfy the distortion constraint $\mathbb{E}\rho(X,\hat{X}) \leq D$. The supremum should be taken over all scalar encoders with a fixed rate $R = \log M$. Alternatively, we can

carry out a continuous optimization by taking the supremum over all sets of positive vectors $\{\vec{p}_z\}_{z=1}^{M}$ that satisfy (8), i.e., all conditional distributions $\{p(z|x)\}$. Whereas the original optimization problem may require exhaustive search over all encoders, and in this case, our mechanism is useless, the continuous problem may have analytic solution. The result of the continuous optimization will, of course, be greater than or equal to C^Q. However, the functions $\{\Gamma_y(\vec{p}_z)\}$ might be neither convex nor concave. In this case, we can carry out the optimization using the general form of $I^Q(X; Y, Z)$ given in (10), which is convex in $\vec{p}_z$.

Until now, we only handled fixed-rate codes. However, distortion lower bounds for codes created by time-sharing fixed-rate codes are readily obtained from the above. This can be seen as follows: For a given rate $R \in \mathbf{R} \triangleq \{\log 1, \log 2, \ldots, \log K\}$, let $D(R)$ be the minimum distortion achievable by fixed-rate scalar codes with encoders $f : \mathcal{X} \rightarrow \{1, 2, \ldots, 2^R\}$ and let $\underline{D}(R)$ be a lower bound on this distortion. We construct a variable-rate code whose rate at time t is R_t, $R_t \in \mathbf{R}$, $t = 1, \ldots, n$, by time-sharing scalar fixed-rate codes, under the constraint:

$$\frac{1}{n}\sum_{t=1}^{n} R_t \leq R. \tag{24}$$

The distortion of this time-sharing code is lower bounded by:

$$
\begin{aligned}
D &\geq \frac{1}{n}\sum_{t=1}^{n} D(R_t) \\
&\geq \frac{1}{n}\sum_{t=1}^{n} \underline{D}(R_t) \\
&\geq \frac{1}{n}\sum_{t=1}^{n} \underline{D}^*(R_t) \\
&\geq \underline{D}^* \left(\frac{1}{n}\sum_{t=1}^{n} R_t \right) \\
&\geq \underline{D}^*(R), \tag{25}
\end{aligned}
$$

where $\underline{D}^*(R)$ is the lower convex envelope of the set $\{\underline{D}(R)\}_{R \in \mathbf{R}}$ and is defined by:

$$\underline{D}^*(R) \triangleq \min \sum_{i=1}^{K} \beta_i \underline{D}(\log i) \tag{26}$$

where the minimum is taken over the following set:

$$\{\beta_1, \beta_2, \ldots, \beta_K : \quad \beta_i \geq 0 \quad \forall i, \quad \sum_{i=1}^{K} \beta_i = 1, \quad \sum_{i=1}^{K} \beta_i \log i \leq R\} \tag{27}$$

9

We see that $\underline{D}^*(R)$ lower bounds the distortion of any such time-sharing code with rate no more than R. Concrete examples of this result will be given in the next section.

We end this subsection with an upper bound on C^Q for the specific convex function $Q(t) = t^{1-\alpha}$, $\alpha > 1$. Using this choice of Q is equivalent to using the Rényi mutual information of order α, which is defined as [20]:

$$
\begin{aligned}
I_\alpha^r(X;Y) &\equiv \frac{1}{\alpha - 1} \log \sum_{x,y} p(x,y) \left[\frac{p(x)}{p(x|y)} \right]^{1-\alpha} \\
&= \frac{1}{\alpha - 1} \log I^Q(X;Y).
\end{aligned}
\tag{28}
$$

Thus, Eq. (20) can be written in the following equivalent form:

$$
R_\alpha^r(d) \leq I_\alpha^r(X;\hat{X}) \leq I_\alpha^r(X;Y,Z) \leq C_\alpha^r,
\tag{29}
$$

where

$$
R_\alpha^r(d) = \frac{1}{\alpha - 1} \log R^Q(D)
\tag{30}
$$

and

$$
C_\alpha^r = \frac{1}{\alpha - 1} \log C^Q.
\tag{31}
$$

The logarithmic measure is a special case, obtained for $\alpha \to 1$. Thus, optimizing over α can only improve the classical bounds. In addition, the function $Q(t) = t^{1-\alpha}$ is relatively convenient to work with.

Lemma 2. *For the convex function $Q(t) = t^{1-\alpha}$, $1 < \alpha < 2$, we have the following upper bound:*

$$
C^Q \leq M^{\alpha-1} \sum_y \left(\sum_x p(x) \cdot p(y|x)^{\frac{1}{2-\alpha}} \right)^{2-\alpha}.
\tag{32}
$$

Proof. Using the function $Q(t) = t^{1-\alpha}$ in (17), we get:

$$
\begin{aligned}
I^Q(X;Y,Z) &= \sum_y \sum_z \sum_{x \in A_z} p(x,y) \left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x)} \right)^{1-\alpha} \\
&= \sum_y \sum_z \sum_{x \in A_z} p(x)p(y|x)^\alpha \cdot [\vec{p}_z \cdot \vec{p}_y]^{1-\alpha} \\
&= \sum_y \sum_z [\vec{p}_z \cdot \vec{p}_y]^{1-\alpha} \sum_{x \in A_z} p(x)p(y|x)^\alpha \\
&= \sum_y \sum_z \left[\sum_x p(x,y)1_{x \in A_z} \right]^{1-\alpha} \left[\sum_x p(x)p(y|x)^\alpha 1_{x \in A_z} \right].
\end{aligned}
\tag{33}
$$

10

In order to get an upper bound on $I^Q(X; Y, Z)$, we define:

$$\begin{aligned} q &= 1/(\alpha - 1) \\ r &= 1/(2 - \alpha) \end{aligned} \tag{34}$$

and the following K-dimensional vectors:

$$\begin{aligned} \vec{a}_y &= \left[p(x_1, y)^{\alpha-1} \cdot 1_{1 \in A_z}, p(x_2, y)^{\alpha-1} \cdot 1_{2 \in A_z}, \ldots \right], \\ \vec{b}_y &= \left[p(x_1)^{2-\alpha} \cdot p(y|x_1) \cdot 1_{1 \in A_z}, p(x_2)^{2-\alpha} \cdot p(y|x_2) \cdot 1_{2 \in A_z}, \ldots \right]. \end{aligned} \tag{35}$$

Applying these definitions to (33), we have:

$$I^Q(X; Y, Z) = \sum_y \sum_z \left(\sum_{k=1}^K a_{y,k}^q \right)^{-1/q} \cdot \left(\vec{a}_y \cdot \vec{b}_y \right). \tag{36}$$

Assuming α is in the range $1 < \alpha < 2$, we have $1 \le q, r < \infty$ and $1/q + 1/r = 1$. Thus we can apply the Hölder inequality to each term in the sum, in the following way:

$$\left(\vec{a}_y \cdot \vec{b}_y \right) \cdot \left(\sum_{k=1}^K a_{y,k}^q \right)^{-1/q} \le \left(\sum_{k=1}^K b_{y,k}^r \right)^{1/r} \tag{37}$$

We then have:

$$\begin{aligned} I^Q(X; Y, Z) &\le \sum_y \sum_z \left(\sum_{k=1}^K b_{y,k}^r \right)^{1/r} \\ &= \sum_y \sum_z \left(\vec{p}_z \cdot [p(x_1) \cdot p(y|x_1)^{\frac{1}{2-\alpha}}, p(x_2) \cdot p(y|x_2)^{\frac{1}{2-\alpha}}, \ldots] \right)^{2-\alpha} \\ &= \sum_y \sum_z M \cdot \frac{1}{M} \left(\vec{p}_z \cdot [p(x_1) \cdot p(y|x_1)^{\frac{1}{2-\alpha}}, p(x_2) \cdot p(y|x_2)^{\frac{1}{2-\alpha}}, \ldots] \right)^{2-\alpha} \\ &\le \sum_y M \cdot \left(\frac{1}{M} \sum_z \vec{p}_z \cdot [p(x_1) \cdot p(y|x_1)^{\frac{1}{2-\alpha}}, p(x_2) \cdot p(y|x_2)^{\frac{1}{2-\alpha}}, \ldots] \right)^{2-\alpha} \\ &= M^{\alpha-1} \sum_y \left(\sum_x p(x) \cdot p(y|x)^{\frac{1}{2-\alpha}} \right)^{2-\alpha}, \end{aligned} \tag{38}$$

where the second inequality is due to Jensen, using the fact that the function $q(t) = t^{2-\alpha}$ is concave for $1 < \alpha < 2$. The last equality follows from the constraint (8). $\qquad \square$

The usefulness of this result stems from its generality. It holds for any source distribution and any transition probability matrix $\{p(y|x)\}$. This result is used in Section 3, along with tighter bounds on the capacity that can be achieved in several special cases. It will also be

shown that the application of this result to large alphabets yields non-trivial bounds. For $Q(t) = -\log t$, Eq. (32) is equivalent to the following:

$$
\begin{aligned}
C & \leq \lim_{\alpha \to 1} \frac{1}{\alpha - 1} \log \left(M^{\alpha - 1} \sum_y \left(\sum_x p(x) \cdot p(y|x)^{\frac{1}{2-\alpha}} \right)^{2-\alpha} \right) \\
& = \log M + I(X;Y),
\end{aligned}
\tag{39}
$$

where C is the classical capacity. Therefore, Eq. (32) can be viewed as generalization of (39). Notice that the bound in (39) can be derived easily from (12). This simple bound states that a maximum amount of information is transferred to the decoder when the output of the deterministic encoder is uniformly distributed and independent of the side information. The proof of (39) is given in Appendix B.

2.2 The generalized rate-distortion function for the uniform source distribution

In this subsection, we handle the left-hand side of (20), i.e., the generalized RD function. We start with a general characterization. Then, in Lemma 3, we give a closed-form expression for the generalized RD function of uniformly distributed sources w.r.t. general symmetric distortion measures. Finally, in Lemma 4, we provide an explicit expression of this function for the special case of the Hamming distortion measure. These results will be used in the next section to derive concrete lower bounds on the distortion, from the DPT we formulated in (20).

By definition of the generalized mutual information:

$$
\begin{aligned}
I^Q(X; \hat{X}) & = \sum_{x, \hat{x}} p(x, \hat{x}) Q \left(\frac{p(\hat{x})}{p(\hat{x}|x)} \right) \\
& = \sum_{x, \hat{x}} p(x) p(\hat{x}|x) Q \left(\frac{\sum_{x'} p(x') p(\hat{x}|x')}{p(\hat{x}|x)} \right) \\
& = \sum_{\hat{x}} \sum_x p(x) p(\hat{x}|x) Q \left(\frac{\vec{p} \cdot \vec{p}_{\hat{x}}}{p(\hat{x}|x)} \right) \\
& = \sum_{\hat{x}} \Psi(\vec{p}_{\hat{x}}),
\end{aligned}
\tag{40}
$$

where we have defined the following K-dimensional vectors:

$$
\begin{aligned}
\vec{p} & = [p(x), x \in \mathcal{X}] \\
\vec{p}_{\hat{x}} & = [p(\hat{x}|x), x \in \mathcal{X}]
\end{aligned}
\tag{41}
$$

and the following function:

$$\Psi(\vec{p}_{\hat{x}}) = \sum_x p(x)p(\hat{x}|x)Q\left(\frac{\vec{p}\cdot\vec{p}_{\hat{x}}}{p(\hat{x}|x)}\right).$$ (42)

For any convex function Q, $\Psi(\vec{p}_{\hat{x}})$ is a convex function. This can be shown easily by the same method we used to prove Lemma 1. The generalized RD function is given by:

$$R^Q(D) = \inf\left\{\sum_{\hat{x}} \Psi(\vec{p}_{\hat{x}})\right\},$$ (43)

where the infimum is taken over all conditional distributions $\{p(\hat{x}|x)\}$ under the constraint:

$$\sum_{\hat{x}}\sum_x p(x)p(\hat{x}|x)\rho(x,\hat{x}) \le D.$$ (44)

$I^Q(X;\hat{X})$ is, of course, convex in the set $\{\vec{p}_{\hat{x}}\}_{\hat{x}=1}^K$. Thus, this is a standard problem of minimizing a convex function over a convex set under linear constraints. Generally, this optimization problem can be solved numerically by various algorithms (see, e.g., [23, Chap. 3]).

In the next steps we will give analytic expressions to the generalized RD function under certain conditions. We refer to a distortion measure $\rho(x,\hat{x})$ as symmetric if the rows of the distortion matrix, $\{\rho(x,\hat{x})\}$, are permutations of each other and the columns are permutations of each other. A uniformly distributed source is a source for which $p(x) = \dfrac{1}{K}$, $\forall x \in \mathcal{X}$.

Lemma 3. *Consider a discrete source X, uniformly distributed over a finite alphabet $\mathcal{X}$, and let $Q(t)$, $0 \le t < \infty$, be any real-valued differentiable convex function. Then, $R^Q(D)$ w.r.t. any symmetric distortion measure is given by:*

$$R^Q(D) = \sum_{k=1}^K p_k \cdot Q\left(\frac{1}{Kp_k}\right),$$ (45)

where $\{p_k\}_{k=1}^K$ is a probability distribution, which is given by the following equations ($k = 1,\ldots,K$):

$$Q\left(\frac{1}{Kp_k}\right) - \frac{1}{Kp_k}Q'\left(\frac{1}{Kp_k}\right) + \lambda_1 + \lambda_2\rho_k - \mu_k = 0,$$ (46)

where $\{\rho_k\}_{k=1}^K$ are the elements of each row of the matrix $\{\rho(x,\hat{x})\}$ and λ_1, λ_2, $\{\mu_k\}_{k=1}^K$ are constants, chosen such that:

$$\begin{aligned}\sum_{k=1}^K p_k = 1, \quad &\sum_{k=1}^K p_k\rho_k = D\\ \mu_k \ge 0, \quad &\mu_k\cdot p_k = 0, \qquad k = 1,\ldots,K\end{aligned}$$ (47)

13

Notice that the equations (46) are decoupled, thus each p_k can be calculated separately. The proof of Lemma 3 is given in Appendix C.

Example. Taking $Q(t) = t^{-s}$, we get:

$$(s+1)p_k^s + \lambda_1 + \lambda_2 \rho_k - \mu_k = 0, \tag{48}$$

which is equivalent to the following:

$$p_k = c(\mu_k - \lambda - \rho_k)^{\frac{1}{s}}, \tag{49}$$

where specific value of λ matches to a point on the generalized RD curve and c is a normalization factor.

For the Hamming distortion measure, defined by:

$$\rho(x, \hat{x}) = \begin{cases} 0 & x = \hat{x} \\ 1 & x \neq \hat{x} \end{cases} \tag{50}$$

we have the following closed-form expression:

Lemma 4. *Consider a discrete source X, uniformly distributed over a finite alphabet $\mathcal{X}$, and let $Q(t)$, $0 \leq t < \infty$ be any real-valued convex function. Then, $R^Q(D)$ w.r.t. the Hamming distortion measure is given by:*

$$R^Q(D) = (1-D) \cdot Q\left[\frac{1}{K(1-D)}\right] + D \cdot Q\left[\frac{K-1}{KD}\right]. \tag{51}$$

Notice that Lemma 4 does not require the differentiability of the convex function Q. The proof is given in Appendix D.

The general form of $R^Q(D)$ enables the use of any convex function Q. These results make the Ziv-Zakai mechanism much more tractable, at least for the case of uniform sources. In addition, they provide direct solutions for a broad class of classical RD functions. We use these results in the next section to derive non-trivial bounds on the distortion of scalar coding in several cases.

3 Applications

In this section, we use the results of the Section 2 to derive lower bounds on the distortion in several cases. Non-trivial bounds are obtained using the convex function $Q(t) = t^{1-\alpha}$,

$\alpha > 1$, which was mentioned above. We assume that the source is uniformly distributed. Under these conditions, Eq. (19) becomes:

$$\Gamma_y(\vec{p}_z) = K^{\alpha-1} \frac{\vec{p}_z \cdot \vec{p}_y^{\alpha}}{(\vec{p}_z \cdot \vec{p}_y)^{\alpha-1}}, \tag{52}$$

where we have defined the following K-dimensional vectors:

$$
\begin{aligned}
\vec{p}_y &= [p(y|x), x \in \mathcal{X}], \\
\vec{p}_y^{\alpha} &= [p(y|x)^{\alpha}, x \in \mathcal{X}].
\end{aligned}
\tag{53}
$$

Applying (51) and (52) to (20), we get:

$$
\begin{aligned}
R^Q(D) &= K^{\alpha-1}\left[(1-d)^{\alpha} + \frac{d^{\alpha}}{(K-1)^{\alpha-1}}\right] \\
&\leq K^{\alpha-2}\sup\left\{\sum_{y,z} \frac{\vec{p}_z \cdot \vec{p}_y^{\alpha}}{(\vec{p}_z \cdot \vec{p}_y)^{\alpha-1}}\right\} \\
&= C^Q,
\end{aligned}
\tag{54}
$$

where the supremum is taken over all sets of positive vectors $\{\vec{p}_z\}_{z=1}^M$ that satisfy (8), in order to carry out continuous optimization. It is easy to see that the optimization is done over a convex set. The functions $\Gamma_y(\vec{p}_z)$ are neither convex nor concave in this case, but we can use the general form of $I^Q(X;Y,Z)$ given in (10). By simple substitution under the above conditions, $G_y(\vec{p}_z)$ has the following form:

$$G_y(\vec{p}_z) = K^{\alpha-1}\frac{\vec{p}_z^{\alpha} \cdot \vec{p}_y^{\alpha}}{(\vec{p}_z \cdot \vec{p}_y)^{\alpha-1}}, \tag{55}$$

where $\vec{p}_z^{\alpha}$ is the vector obtained from $\vec{p}_z$ by raising each element to the power of α. Clearly, $G_y(\vec{p}_z) = \Gamma_y(\vec{p}_z)$ for any deterministic encoder. The functions $\{G_y(\vec{p}_z)\}$ are convex as shown in Lemma 1. Therefore, the supremum of $I^Q(X;Y,Z)$ is attained on the boundary of the convex set. Finding the supremum on the boundary requires searching over all vertices of the set, i.e, over all sets of binary vectors $\{\vec{p}_z\}_{z=1}^M$ that satisfy (8). The meaning is, of course, returning to discrete optimization and performing it over all deterministic encoders. Seemingly, this makes the mechanism above useless. However, at least for some cases, C^Q can be calculated directly, as shown in the following examples. In addition, we can upper bound C^Q by using (32). This upper bound may give us non-trivial bounds, as shown in Example 2. It is also shown to be very useful when handling large alphabets, as demonstrated in Section 3. Finally, notice that optimizing over α, separately for each rate, will produce the best lower bound on the achievable distortion at this rate.

15

Example 1

The symmetric DMC is defined by:

$$p(y|x) = \begin{cases} \mu & y = x \\ \epsilon & else \end{cases} \tag{56}$$

where $\mu, \epsilon \in [0,1]$, $\mu > \epsilon$, and $\mu + (K-1)\epsilon = 1$. The distortion measure we use is the Hamming distortion, defined in (50). In these conditions, the minimal achievable distortion of a scalar source code with a fixed-rate $R = \log M$, is:

$$D(M) = \epsilon(K - M). \tag{57}$$

The proof is given in Appendix E. Knowing the best achievable distortion in this case, we can compare it to the bounds we get from (54) to examine their quality. The generalized capacity (22) for this channel is given by:

$$\begin{aligned}
C^Q &= K^{\alpha-2} \cdot \sup\left\{ \sum_{y,z} \frac{\vec{p}_z \cdot \vec{p}_y^{\alpha}}{(\vec{p}_z \cdot \vec{p}_y)^{\alpha-1}} \right\} \\
&= K^{\alpha-2} \epsilon \cdot \sup\left\{ \sum_z M_z \cdot \frac{(M_z + \mu^{\alpha}/\epsilon^{\alpha} - 1)}{(M_z + \mu/\epsilon - 1)^{\alpha-1}} \right. \\
&\quad \left. + (K - M_z)M_z^{2-\alpha} \right\} \\
&= K^{\alpha-2} \epsilon \cdot \sup\left\{ \sum_z q_\alpha(M_z) \right\}, \tag{58}
\end{aligned}$$

where M_z, $M_z \in \{1, \ldots, K - M + 1\}$, is the cardinality of A_z, i.e., the number of source symbols that are encoded to z. Obviously, $\sum_z M_z = K$. Notice that the supremum is taken over all deterministic encoders, where each encoder is represented by a specific set $\{\vec{p}_z\}_{z=1}^M$ as defined in (16). The second equality is proved in Appendix F. The function $q_\alpha(M_z)$ is concave for $1 < \alpha \leq 2$, and may be concave also out of this range, with dependence on the channel parameters, as shown in Appendix F. When $q_\alpha(M_z)$ is concave, we can bound the supremum by taking equal M_z's, i.e., $M_z = K/M$, $\forall z$, and we get:

$$\begin{aligned}
C^Q \leq\ & K^{\alpha-1}\epsilon \left[\frac{(K/M + \mu^{\alpha}/\epsilon^{\alpha} - 1)}{(K/M + \mu/\epsilon - 1)^{\alpha-1}} \right. \\
&\left. + (M - 1)\left(\frac{K}{M}\right)^{2-\alpha} \right]. \tag{59}
\end{aligned}$$

If M divides K, this bound is achieved by any feasible encoder that partitions the source alphabet into equally sized subsets, thus the optimization is exact. An example for specific values of μ and ϵ is presented in Fig. 2. The bound is compared with the bound obtained from the classical DPT (12), the bound obtained from the classical inequality $R_{WZ}(D) \leq \log M$, the bound obtained by using (32) and the exact solution of Eq. 57. The WZ RD function was calculated using the Blahut-Arimoto-type algorithm presented in [22]. Eq. (54) was optimized over α, for each $M \leq K$, so as to get the best lower bound on the distortion. We see that even the classical DPT gives us non-trivial lower bounds and that
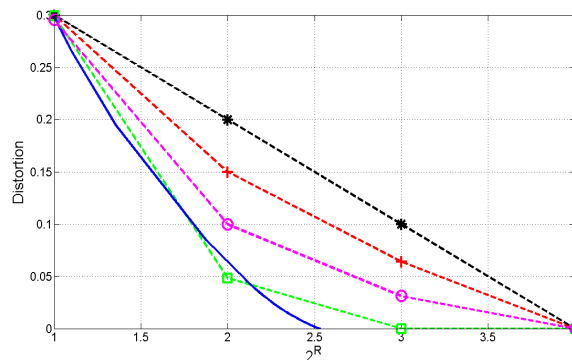

Figure 2: $K = 4$, $\mu = 0.7$. Plus - the lower bound obtained from (59). Circle - the lower bound obtained from the classical DPT (12). Star - the exact solution. Solid line - the lower bound obtained from $R_{WZ}(D)$. Square - the lower bound obtained from (32).

the lower bound obtained from (59) is much better than the trivial bound obtained from $R_{WZ}(D)$. The lower bound obtained from (32) is not useful in this case. There is a gap between the exact solution and the best bound, even for $M = 2$, where the optimization (22) is exact.

Example 2

The symmetric DMC is defined by:

$$p(y|x) = \begin{cases} 1/l & y \in \{x \bmod K, \ldots, (x + l - 1) \bmod K\} \\ 0 & else \end{cases} \tag{60}$$

where l is an integer, $0 < l < K$, and $K \bmod K$ is defined to be K. Given an input x, the channel produces one of l values with equal probability. The generalized capacity (22) for

this channel is given by:

$$
\begin{aligned}
C^Q &= K^{\alpha-2} \cdot \sup \left\{ \sum_{y,z} \frac{\vec{p}_z \cdot \vec{p}_y^{\alpha}}{(\vec{p}_z \cdot \vec{p}_y)^{\alpha-1}} \right\} \\
&= K^{\alpha-2} \cdot \sup \left\{ \sum_{y,z} \frac{M_{y,z}(1/l)^{\alpha}}{(M_{y,z}(1/l))^{\alpha-1}} \right\} \\
&= K^{\alpha-2} \cdot l^{-1} \cdot \sup \left\{ \sum_{y,z} M_{y,z}^{2-\alpha} \right\},
\end{aligned}
\tag{61}
$$

where $M_{y,z} = l \cdot [\vec{p}_z \cdot \vec{p}_y]$. It is easy to see that $\sum_z M_{y,z} = l$. For $1 < \alpha < 2$, the function $M_{y,z}^{2-\alpha}$ is concave in M_z. Thus, the supremum is achieved by setting $M_{y,z} = l/M$, $\forall \{y, z\}$:

$$
C^Q = K^{\alpha-2} \cdot l^{-1} \cdot K \cdot M \cdot (l/M)^{2-\alpha} = K^{\alpha-1} \cdot (M/l)^{\alpha-1}.
\tag{62}
$$

If M divides l, equal $M_{y,z}$'s can be obtained by the following feasible encoder:

$$
z = f(x) = 1 + x \bmod M.
\tag{63}
$$

Therefore, in this case, the optimization is exact. For $\alpha > 2$, C^Q is infinite, because we can always set some $M_{y,z}$ to 0 by an appropriate choice of the encoder. Thus, this range of α does not lead to a useful bound. An example for specific values of K and l is presented in Fig. 3. The lower bound on the distortion, which coincides with the bound obtained from (32) (the upper bound on C^Q is tight for this channel), is compared with the bound obtained from the classical DPT (12) and the bound obtained by the classical inequality $R_{WZ}(D) \leq \log M$. Eq. (54) was optimized over α, for each $M \leq K$, so as to get the best lower bound on the distortion. We see that in this case, the generalized DPT leads to bounds that are better than the trivial bound, whereas the classical DPT does not lead to a useful bound. We also present the exact distortion of the encoder defined in (63), which is, of course, an upper bound on the distortion. Thus, the distortion of the optimal encoder must be in the range between this upper bound and our highest lower bound. For $M = l$, zero distortion can indeed be achieved using the encoder defined in (63), thus our lower bound at this point is tight.

Large alphabets

In this part we show that as the alphabet size increases, we obtain interesting bounds on the performance of scalar coding. These useful bounds can be obtained for a large
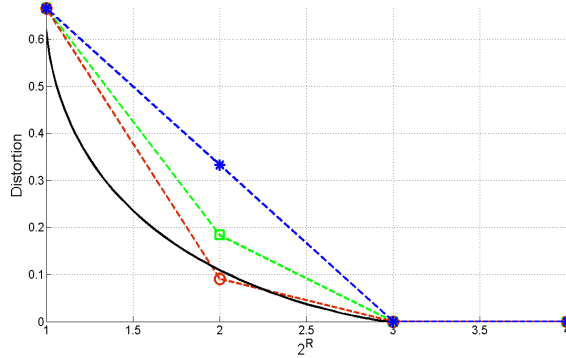
18

Figure 3: $K = 4$, $l = 3$. Square - the lower bound obtained from (59) and (32). Circle - the lower bound obtained from the classical DPT (12). Solid line - the lower bound obtained from $R_{WZ}(D)$. Star - the exact distortion of the encoder (63).

variety of channels, without any symmetry requirements. The results are obtained using the upper bound on the "capacity", presented in Lemma 2. This bound becomes tighter as the alphabet size increases, for various channels. For these channels, we can get close to the last upper bound in Eq. (38), i.e., to achieve $\{\vec{p}_z \cdot [p(x_1) \cdot p(y|x_1)^{\frac{1}{2-\alpha}}, p(x_2) \cdot p(y|x_2)^{\frac{1}{2-\alpha}}, \ldots]\}_{z=1}^{M}$ which are almost equal to each other, by a suitable choice of encoder. This is because $\{p(y|x)\}$ is composed of large number of probabilities with small values. Using the bound of Lemma 2, we bypass the problem of optimizing the capacity for general channels. As was mentioned earlier in Subsection 2.3, this optimization is in general a convex maximization problem which requires searching over all possible encoders. Using our lower bounds along with simple upper bounds, we give the performance range for scalar coding. These results are, of course, interesting from the practical point of view.

In the following examples we assume that the sources are uniform. This is because analytic expression for the generalized RD function of general sources is not available. We use the Hamming distortion measure for convenience. Bounds for more general distortion measures can be calculated using the result of Lemma 3. In the two former examples, we compare our results to the WZ RD function. This function was calculated using the algorithm presented in [22]. However, the computational complexity of this algorithm is of order K^K. Therefore, this algorithm is not practical for large alphabets. Since no other efficient algorithms are known, the computation of the WZ RD function for large alphabets is problematic. As a result, even the trivial bound obtained from $R_{WZ}(D) \leq \log M$ does not lead to a closed-form expression. This makes our results even more interesting.

Instead of comparing our distortion bounds to the bounds obtained from $R_{WZ}(D) \leq \log M$, we compare them to the following linear function:

$$\overline{D}(R) = D_{max}\left(1 - \frac{R}{H(X|Y)}\right), \tag{64}$$

where $d_{max} = R_{WZ}^{-1}(0)$, i.e., the lowest achievable distortion for rate $R = 0$. The function $\overline{D}(R)$ is simply the straight line obtained by time-sharing the two known endpoints of the $R_{WZ}^{-1}(R)$ curve, $(0, D_{max})$ and $(H(X|Y), 0)$. This line is, of course, a trivial upper bound on $R_{WZ}^{-1}(R)$. As an upper bound on the best achievable distortion of a fixed-rate code, we use the performance of the code composed of the encoder defined in (63), along with the corresponding optimal decoder, given by:

$$\hat{x} = g(z, y) = \operatorname*{argmax}_{x \in A_z}\{p(y|x)\}. \tag{65}$$

Remember that the optimal decoding strategy is maximum likelihood because we use the Hamming distortion measure. The choice of the encoder (63) seems natural when handling channels with transition probabilities that decrease with the distance. In this case, we want adjacent symbols to be in different subsets of the encoder. Obviously, any code that performs better will improve the performance range.

In the first example, the DMC is defined by:

$$p(y|x) = c_x \exp\left\{-\frac{(x-y)^2}{2\sigma^2}\right\}, \tag{66}$$

where c_x is a normalization factor such that $\sum_y p(y|x) = 1$. This is a 'Gaussian'-like channel. Notice that this channel is not symmetric. Performance ranges for different values of K are presented in Fig. 4. We see that we get bounds that are higher than $\overline{D}(R)$ and therefore, higher than $R_{WZ}^{-1}(R)$. These bounds also show that the time-sharing of $\overline{D}(R)$ performs better than any fixed-rate code for considerable range of rates. Similar results can be presented for various channels, not necessarily additive.

In the second example, the DMC is the same as in Example 2 and is defined by (60). In this case, $H(X|Y) = \log l$. We now present the performance range for large alphabets where in all cases we take $l = K/4$. The results are shown in Fig. 5. Again, we see that our bounds are higher than $\overline{D}(R)$. As the alphabet size increases, the gap between our bound and $\overline{D}(R)$ also increases.
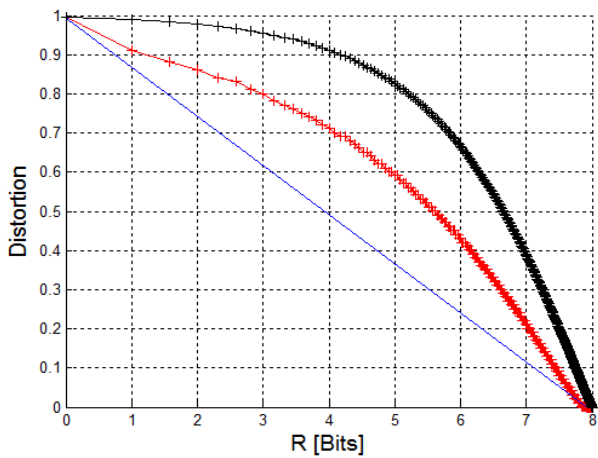
In all examples, one can easily notice that the lower convex envelope of our lower bounds, is very close to the straight line $\overline{D}(R)$. As was shown in Section 2, this convex envelope is a lower bound on the distortion of time-sharing fixed-rate codes.
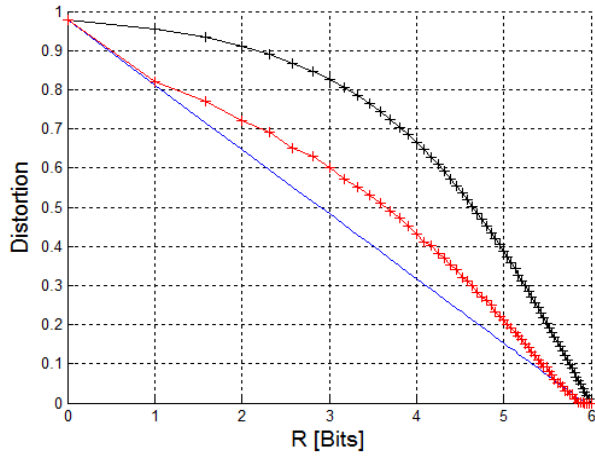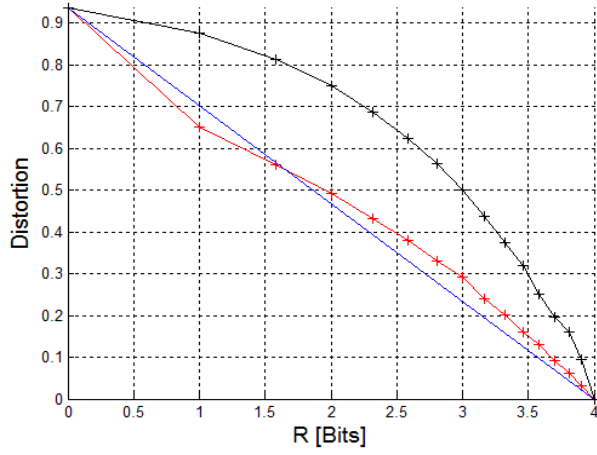
(a) $K = 64$, $\sigma = 0.5$
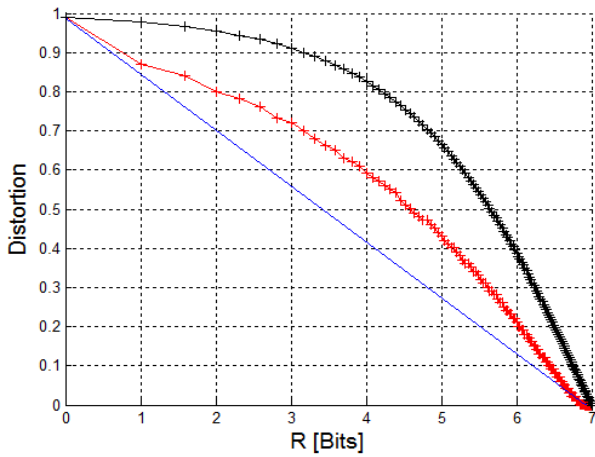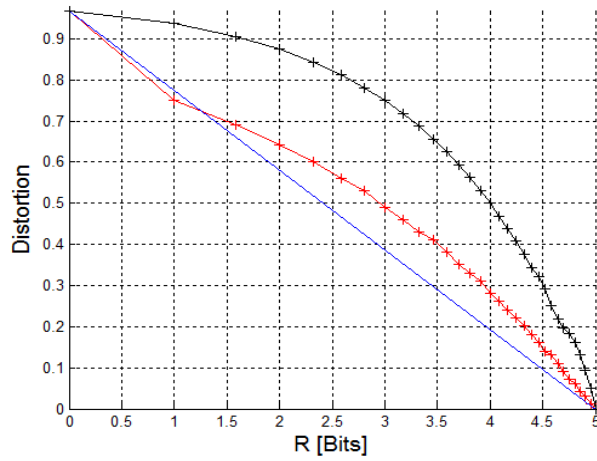

(b) $K = 128$, $\sigma = 0.5$
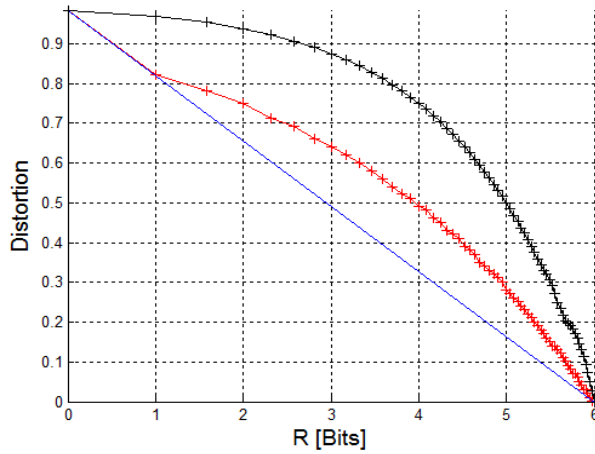

(c) $K = 256$, $\sigma = 0.5$

Figure 4: The performance range for the channel (66) and different sizes of alphabets. Straight line - $\overline{D}(R)$. Lower curve - our lower bound. Higher curve - the distortion of the code (63).

(a) $K = 64$, $l = 16$

(b) $K = 128$, $l = 32$

(c) $K = 256$, $l = 64$

Figure 5: The performance range for the channel (60) and different sizes of alphabets and l's. Straight line - $\overline{D}(R)$. Lower curve - our lower bound. Higher curve - the distortion of the code (63).

4 Summary, Conclusions and Future Directions

In this paper, we presented the relevant random variables of the WZ problem as a Markov chain. As far as we know, this is a new formulation in the context of the WZ setting, which enabled us to use the DPT and its generalizations. We then focused on the generalized DPT presented in [15]. We found the generalized RD function of uniform sources, for any convex function Q and a broad class of distortion measures. We also found a useful upper bound for the generalized capacity in the setting above and calculated this capacity exactly for several interesting channels. We then showed that replacing the logarithmic function with other functions, in the DPT we formulated, yield better bounds on the distortion of scalar coding in the WZ setting. Examples of non-trivial lower bounds for scalar coding in this setting were given. For large alphabets, we demonstrated that non-trivial bounds can be achieved for various channels and as a result, the performance range for scalar coding can be given. We also saw that simple time-sharing between the two endpoints of the WZ RD curve, may perform better than any fixed-rate code in various cases. As far as we know, these bounds are the only existing non-trivial bounds for this situation. Clearly, these results are relevant from the practical point of view.

Analytic expressions for the generalized RD function of general sources are not apparent to be available. Improving the bounds above by using, for example, the techniques of [21], is yet to be explored. In the next step, a possible direction will be to extend our setting to more general scenarios. The first interesting scenario is the variable-rate coding case, where Z is encoded by a variable-length code. In this setting, the generalized capacity (22) will be maximized under the constraint $\mathbb{E}\{L(Z)\} \leq R$, where $L(Z)$ is the length of the codeword Z. The best variable-rate code that can be used is the Huffman code for $p(z)$, provided that $p(x, y) > 0$ for all $x, y \in \mathcal{X}$. The challenge is to perform this optimization. Another interesting scenario is when the output of the encoder is transmitted over a noisy channel (instead of the noiseless one in the WZ setting). In this case, the generalized capacity will be of the form $I^Q(X, Z; Z', Y)$, where Z' is the output of the encoder noisy channel. Again, the challenge will be to optimize this capacity. Another direction is to extend the mechanism above to settings of coding with memory. In these scenarios, the states of the encoder and the decoder will be allowed to depend on past inputs. Using generalized DPTs, we hope to get useful bounds for this case.

Appendix A - Proof of Lemma 1

Proof. We defined:

$$G_y(\vec{p}_z) = \sum_x p(x)p(y|x)p(z|x)Q\left(\frac{\vec{p}_z \cdot \vec{p}_y}{p(y|x)p(z|x)}\right). \tag{A.1}$$

In order to prove the convexity of $G_y(\vec{p}_z)$, we use the following known result (cf., e.g., [23]): $G_y(\vec{p}_z)$ is convex if and only if $\mathrm{dom}(G_y)$ is convex, and the function $g : \mathbb{R} \to \mathbb{R}$, defined by:

$$g(t) = G_y(\vec{r} + \vec{v}t) \quad , \quad \mathrm{dom}(g) = \{t : \vec{r} + \vec{v}t \in \mathrm{dom}(G_y)\}, \tag{A.2}$$

is convex in t, for any $\vec{r} \in \mathrm{dom}(G_y)$, $\vec{v} \in \mathbb{R}^K$. Substituting $\vec{p}_z = \vec{r} + \vec{v}t$, we get:

$$
\begin{aligned}
g(t) &= \sum_{x=1}^K p(x)p(y|x)(r_x + v_x t)Q\left(\frac{[\vec{r} \cdot \vec{p}_y] + [\vec{v} \cdot \vec{p}_y]t}{p(y|x)(r_x + v_x t)}\right) \\
&= \sum_{x=1}^K p(x)p(y|x)(r_x + v_x t)Q\left(\frac{a_y + b_y t}{p(y|x)(r_x + v_x t)}\right),
\end{aligned} \tag{A.3}
$$

where $\{r_x\}_{x=1}^K$ and $\{v_x\}_{x=1}^K$ are the elements of $\vec{r}$ and $\vec{v}$, respectively, and $a_y \equiv \vec{r} \cdot \vec{p}_y$, $b_y \equiv \vec{v} \cdot \vec{p}_y$ are constants. Linear combination of convex functions is convex. Thus, it is enough to show the convexity of the following functions, $x = \{1, \dots, K\}$:

$$
\begin{aligned}
f_x(t) &= p(y|x)(r_x + v_x t)Q\left(\frac{a_y + b_y t}{p(y|x)(r_x + v_x t)}\right) \\
&= h(a_y + b_y t, p(y|x)(r_x + v_x t)),
\end{aligned} \tag{A.4}
$$

where $h(u, s) = sQ(u/s)$ is the perspective (cf., e.g., [23]) of the convex function $Q(u)$, and thus convex. The function $f_x(t)$ is the restriction of the convex function $h(u, s)$ to the straight line

$\{u = p(y|x)(r_x + v_x t), s = a_y + b_y t\}$. Therefore, $f_x(t)$ is convex. $\quad\square$

Appendix B - Proof of Eq. (39)

Proof. It is enough to calculate the limit:

$$
\begin{aligned}
&\lim_{\alpha \to 1} \frac{1}{\alpha - 1} \log\left(\sum_y \left(\sum_x p(x) \cdot p(y|x)^{\frac{1}{2-\alpha}}\right)^{2-\alpha}\right) \\
&= \lim_{\alpha \to 1} \frac{d}{d\alpha}\left[\log\left(\sum_y \left(\sum_x p(x) \cdot p(y|x)^{\frac{1}{2-\alpha}}\right)^{2-\alpha}\right)\right],
\end{aligned} \tag{B.1}
$$

where the equality is due to L'Hôpital's rule, using the fact that both the numerator and denominator tend to 0 as $\alpha \to 1$. Noticing that the expression in the brackets has the form of the Gallager function (see, e.g., [24, Chap. 5]) for the given DMC, defined as

$$E_0(\rho, \mathbf{p}) = -\log \left(\sum_y \left(\sum_x p(x) \cdot p(y|x)^{\frac{1}{1+\rho}} \right)^{1+\rho} \right), \tag{B.2}$$

we can use the following property of $E_0(\rho, \mathbf{p})$:

$$\left. \frac{\partial E_0(\rho, \mathbf{p})}{\partial \rho} \right|_{\rho=0} = I(X;Y) \tag{B.3}$$

to get:

$$\left. \frac{d}{d\alpha} \left[\log \left(\sum_y \left(\sum_x p(x) \cdot p(y|x)^{\frac{1}{2-\alpha}} \right)^{2-\alpha} \right) \right] \right|_{\alpha=1} = I(X;Y), \tag{B.4}$$

which completes the proof. $\qquad\square$

Appendix C - Proof of Lemma 3

The idea of the proof is to exploit the symmetry of the distortion matrix. From symmetry, we expect that each input symbol will have the same set of transition probabilities.

Proof. We define $p(i|j) \equiv P_{\hat{X}|X}(i|j)$. By definition:

$$
\begin{aligned}
I^Q(X;\hat{X}) &= \sum_{x,\hat{x}} p(x,\hat{x}) Q\left(\frac{p(\hat{x})}{p(\hat{x}|x)} \right) \\
&= \frac{1}{K} \sum_{i=1}^{K} p(i|i) Q\left(\frac{p(\hat{x}=i)}{p(i|i)} \right) + \frac{1}{K} \sum_{i=1}^{K} \sum_{j \neq i}^{K} p(j|i) Q\left(\frac{p(\hat{x}=j)}{p(j|i)} \right). \tag{C.1}
\end{aligned}
$$

Each row of the distortion matrix contains the same K values. We enumerate these values as $\{d_1, d_2, \ldots, d_K\}$ where $d_1 \equiv d(x,x) = 0$. Without loss of generality and only for convenience of the proof, we assume that we have K different values. We define:

$$\tilde{x}_k(x) = \{\tilde{x} \in \mathcal{X} : \rho(x,\tilde{x}) = \rho_k\}. \tag{C.2}$$

In words, $\tilde{x}_k(x)$ is the unique alphabet symbol with distortion ρ_k relative to x. Using this definition, we can write $(C.1)$ in the following way:

$$I^Q(X;\hat{X}) = \frac{1}{K} \sum_{i=1}^{K} p(i|i) Q\left(\frac{p(\hat{x}=i)}{p(i|i)} \right) + \frac{1}{K} \sum_{i=1}^{K} \sum_{k=2}^{K} p(\tilde{x}_k(i)|i) Q\left(\frac{p(\hat{x}=\tilde{x}_k(i))}{p(\tilde{x}_k(i)|i)} \right)$$

$$\tag{C.3}$$

We define:

$$p_k = \frac{1}{K}\sum_{i=1}^{K} p(\tilde{x}_k(i)|i), \qquad k \in \{2,\ldots,K\}$$

$$p_1 = \frac{1}{K}\sum_{i=1}^{K} p(i|i) = 1 - \sum_{k=2}^{K} p_k \qquad\qquad (C.4)$$

Using these definitions, the distortion is given by:

$$D = \frac{1}{K}\sum_{k=2}^{K}\sum_{i=1}^{K} p(\tilde{x}_k(i)|i)\rho_k = \sum_{k=2}^{K} p_k \rho_k \qquad\qquad (C.5)$$

Applying the Jensen inequality, we get:

$$
\begin{aligned}
I^Q(X;\hat{X}) &= \frac{1}{K}\sum_{i=1}^{K} p(i|i)Q\left(\frac{p(\hat{x}=i)}{p(i|i)}\right) + \frac{1}{K}\sum_{k=2}^{K}\sum_{i=1}^{K} p(\tilde{x}_k(i)|i)Q\left(\frac{p(\hat{x}=\tilde{x}_k(i))}{p(\tilde{x}_k(i)|i)}\right) \\
&= p_1\sum_{i=1}^{K}\frac{p(i|i)}{Kp_1}Q\left(\frac{p(\hat{x}=i)}{p(i|i)}\right) + \sum_{k=2}^{K} p_k\sum_{i=1}^{K}\frac{p(\tilde{x}_k(i)|i)}{Kp_k}Q\left(\frac{p(\hat{x}=\tilde{x}_k(i))}{p(\tilde{x}_k(i)|i)}\right) \\
&\geq p_1 Q\left(\sum_{i=1}^{K}\frac{p(i|i)p(\hat{x}=i)}{Kp_1 p(i|i)}\right) + \sum_{k=2}^{K} p_k Q\left(\sum_{i=1}^{K}\frac{p(\tilde{x}_k(i)|i)p(\hat{x}=\tilde{x}_k(i))}{Kp_k p(\tilde{x}_k(i)|i)}\right) \\
&= p_1 Q\left(\frac{1}{Kp_1}\right) + \sum_{k=2}^{K} p_k Q\left(\frac{1}{Kp_k}\right) \\
&= \sum_{k=1}^{K} p_k Q\left(\frac{1}{Kp_k}\right). \qquad\qquad (C.6)
\end{aligned}
$$

Notice that the sum $\sum_{i=1}^{K} p(\hat{x}=\tilde{x}_k(i))$ is running over all values of $\hat{x}$ due to the symmetry of the distortion matrix (ρ_k appears in each column) and thus equal to 1. The lower bound in (C.6) is achieved by a channel of the form:

$$p(\tilde{x}_k(i)|i) = p_k, \quad k \in \{2,\ldots,K\}, \quad i \in \{1,2,\ldots,K\}. \qquad\qquad (C.7)$$

This channel achieves, of course, the same distortion D, which depends only on $\{p_k\}$. We also have for this channel:

$$p(\hat{x}=i) = \frac{1}{K}p(i|i) + \frac{1}{K}\sum_{k=2}^{K} p_k = \frac{1}{K} \qquad\qquad (C.8)$$

Substituting (C.7) and (C.8) in (C.3), we get:

$$
\begin{aligned}
I^Q(X;\hat{X}) &= \frac{1}{K}\sum_{i=1}^{K}p(i|i)Q\left(\frac{p(\hat{x}=i)}{p(i|i)}\right) + \frac{1}{K}\sum_{i=1}^{K}\sum_{k=2}^{K}p(\tilde{x}_k(i)|i)Q\left(\frac{p(\hat{x}=\tilde{x}_k(i))}{p(\tilde{x}_k(i)|i)}\right) \\
&= \frac{1}{K}\sum_{i=1}^{K}p_1 Q\left(\frac{1}{Kp_1}\right) + \frac{1}{K}\sum_{i=1}^{K}\sum_{k=2}^{K}p_k Q\left(\frac{1}{Kp_k}\right) \\
&= \sum_{k=1}^{K}p_k Q\left(\frac{1}{Kp_k}\right),
\end{aligned}
\tag{C.9}
$$

which is exactly the lower bound. In summary, we showed that the channel that minimizes $I^Q(X;\hat{X})$, among all channels with the same $\{p_k\}_{k=1}^{K}$, is of the form (C.7). Therefore, to get the RD function, it is enough to optimize (C.9) over all probability measures $\{p_k\}_{k=1}^{K}$, subject to the constraint (C.5). The generalized mutual information $I^Q(X;\hat{X})$ is, of course, convex in $\{p_k\}_{k=1}^{K}$. Thus, this is a standard convex minimization under linear constraints problem and the solution is given by the Karush-Kuhn-Tucker conditions, which are exactly (46) and (47). □

Appendix D - Proof of Lemma 4

Proof. Again, the idea of the proof is to exploit the symmetry of the source and the Hamming distortion. We assume that the source X is uniformly distributed over $\mathcal{X}$, and use the Hamming distortion measure. Under these conditions, the distortion D is given by:

$$
\begin{aligned}
D &= \sum_{i,j=1}^{K}p(i)p(j|i)\rho(i,j) \\
&= \frac{1}{K}\sum_{j}\sum_{i\neq j}p(j|i) \\
&= \frac{1}{K}\sum_{j}D_j,
\end{aligned}
\tag{D.1}
$$

where we have defined $p(i) = P_X(i)$, $p(j|i) = P_{\hat{X}|X}(j|i)$, and:

$$
D_j = \sum_{i\neq j}p(j|i).
\tag{D.2}
$$

Thus we have the following:

$$
p(\hat{x}=j) = \frac{1}{K}\sum_{i}p(j|i) = \frac{1}{K}\sum_{i\neq j}p(j|i) + \frac{1}{K}p(j|j) = \frac{1}{K}\left[D_j + p(j|j)\right]
\tag{D.3}
$$

and

$$1 - D = \frac{1}{K} \sum_i p(i|i)$$

$$D = \frac{1}{K} \sum_j \sum_{i \neq j} p(j|i) = \frac{1}{K} \sum_i \sum_{j \neq i} p(j|i). \tag{D.4}$$

Calculating :

$$
\begin{aligned}
I^Q(X; \hat{X}) &= \sum_{x,\hat{x}} p(x, \hat{x}) Q\left(\frac{p(\hat{x})}{p(\hat{x}|x)} \right) \\
&= \frac{1}{K} \sum_i p(i|i) Q\left[\frac{1}{K} \left(\frac{D_i}{p(i|i)} + 1 \right) \right] + \frac{1}{K} \sum_i \sum_{j \neq i} p(j|i) Q\left[\frac{D_j + p(j|j)}{K p(j|i)} \right] \\
&\geq (1 - D) \cdot Q\left[\frac{1}{(1-D)K^2} \sum_i (D_i + p(i|i)) \right] + \\
&\quad D \cdot Q\left[\frac{1}{DK^2} \sum_i \sum_{j \neq i} p(j|i) \left(\frac{D_j + p(j|j)}{p(j|i)} \right) \right] \\
&= (1 - D) \cdot Q\left[\frac{D + 1 - D}{K(1 - D)} \right] + D \cdot Q\left[\frac{1}{DK} \sum_i \left(D - \frac{D_i}{K} + 1 - D - \frac{p(i|i)}{K} \right) \right] \\
&= (1 - D) \cdot Q\left[\frac{1}{K(1 - D)} \right] + D \cdot Q\left[\frac{K - 1}{KD} \right]. \tag{D.5}
\end{aligned}
$$

The second equality is due to (D.3). The inequality is obtained by applying the Jensen inequality to each one of the two weighted sums, after normalizing the weights using (D.4). The next equalities are obtained by calculating the sums that appear as arguments of the function Q using (D.4) and by simple algebraic manipulations. It is easy to show, by simple substitution in $I(X; \hat{X})$, that this lower bound is achieved , for any convex function $Q(t)$, by the following symmetric channel:

$$
p(\hat{x}|x) = \begin{cases} 1 - D & \hat{x} = x \\ \dfrac{D}{K - 1} & \hat{x} \neq x \end{cases} \tag{D.6}
$$

$\square$

Appendix E - Proof of Eq. (57)

Proof. We assume that the source is uniformly distributed and that the DMC is given by (56). The Hamming distortion is equal to the average probability of error. Therefore, given a scalar encoder of fixed rate $R = \log M$, the optimal decoding strategy is, of course, maximum likelihood:

$$\hat{x} = g(z, y) = \operatorname*{argmax}_{x \in A_z}\{p(y|x)\} \tag{E.1}$$

29

For the channel (56), the decoder gets the form:

$$\hat{x} = \begin{cases} y & y \in A_z \\ choose\ x' \in A_z\ uniformly\ at\ random & y \notin A_z \end{cases} \tag{E.2}$$

Given $x \in A_z$, we have two error events. The first error event is when $y \in A_z$ and $y \neq x$. The probability of such an event is $(M_z - 1)\epsilon$. The other error event is when $y \notin A_z$ and $x' \neq x$. The probability of this event is the product $\Pr\{y \notin A_z|x\} \cdot \Pr\{x' \neq x|x\} = (K - M_z)\epsilon \cdot (M_z - 1)/M_z$. Thus, the distortion is given by:

$$
\begin{aligned}
d &= \frac{1}{K} \sum_x \Pr\{error|x\} \\
&= \frac{1}{K} \sum_z \sum_{x \in A_z} [(M_z - 1)\epsilon + (K - M_z)\epsilon(M_z - 1)/M_z] \\
&= \frac{\epsilon}{K} \sum_z M_z[(M_z - 1) + (K - M_z)(M_z - 1)/M_z] \\
&= \frac{\epsilon}{K} \sum_z [M_z \cdot (M_z - 1) + (K - M_z)(M_z - 1)] \\
&= \frac{\epsilon}{K} \sum_z (M_z - 1)[M_z + K - M_z] \\
&= \epsilon \sum_z (M_z - 1) \\
&= \epsilon(K - M) \\
&= \epsilon(K - 2^R). \tag{E.3}
\end{aligned}
$$

Notice that $d(R)$ is a decreasing concave function. Thus, by time-sharing an encoder with $R = \log K$ and an encoder with $R = 0$, we can achieve the straight line $\tilde{d}(R) = \epsilon(K-1)(1 - R/\log K)$ and outperform any fixed-rate encoder. $\qquad\square$

Appendix F - The concavity of $q_\alpha(x)$ defined in Eq. (58)

We start with explaining the following equality for the channel (56):

$$
\begin{aligned}
\sum_z \sum_y \frac{\vec{p}_z \cdot \vec{p}_y^\alpha}{(\vec{p}_z \cdot \vec{p}_y)^{\alpha-1}} &= \sum_z M_z \cdot \frac{(M_z + \mu^\alpha/\epsilon^\alpha - 1)}{(M_z + \mu/\epsilon - 1)^{\alpha-1}} + (K - M_z)M_z^{2-\alpha} \\
&= \sum_z q_\alpha(M_z). \tag{F.1}
\end{aligned}
$$

The sum over y is calculated by noticing that the product $[\vec{p}_z \cdot \vec{p}_y]$, and, respectively, the product $[\vec{p}_z \cdot \vec{p}_y^\alpha]$, can have one of two results. If the 1's in the binary vector $\vec{p}_z$ overlap only ϵ's in $\vec{p}_y$, we get $[\vec{p}_z \cdot \vec{p}_y] = M_z \cdot \epsilon$ and respectively, $[\vec{p}_z \cdot \vec{p}_y^\alpha] = M_z \cdot \epsilon^\alpha$. Otherwise, we get

$[\vec{p}_z \cdot \vec{p}_y] = \mu + (M_z - 1) \cdot \epsilon$ and respectively, $[\vec{p}_z \cdot \vec{p}_y^\alpha] = \mu^\alpha + (M_z - 1) \cdot \epsilon^\alpha$. It is not hard to see that the second result will occur exactly M_z times in the sum on y, and thus the first will occur exactly $K - M_z$ times. The rest is straightforward. We now prove the concavity of the function $q_\alpha(x)$, $1 \leq x \leq K - M + 1$. The second derivative of $q_\alpha(x)$ is ($c_\alpha \equiv \mu^\alpha/\epsilon^\alpha$):

$$
\begin{aligned}
q_\alpha''(x) \;=\; & 2(2 - \alpha)\,(x + \mu/\epsilon - 1)^{1-\alpha} - x(2 - \alpha)(\alpha - 1)\,(x + \mu/\epsilon - 1)^{-\alpha} \\
& -2c_\alpha(\alpha - 1)\,(x + \mu/\epsilon - 1)^{-\alpha} + x \cdot c_\alpha \cdot \alpha(\alpha - 1)\,(x + \mu/\epsilon - 1)^{-\alpha-1} \\
& -2(2 - \alpha)x^{1-\alpha} - (2 - \alpha)(K - x)(\alpha - 1)x^{-\alpha} \\
\;\leq\; & 2(2 - \alpha)x^{1-\alpha} - x(2 - \alpha)(\alpha - 1)\,(x + \mu/\epsilon - 1)^{-\alpha} \\
& -2c_\alpha(\alpha - 1)\,(x + \mu/\epsilon - 1)^{-\alpha} + x \cdot c_\alpha \cdot \alpha(\alpha - 1)\,(x + \mu/\epsilon - 1)^{-\alpha-1} \\
& -2(2 - \alpha)x^{1-\alpha}.
\end{aligned}
\tag{F.2}
$$

The inequality follows from the assumption $\mu \geq \epsilon$, which leads to $x \leq x + \mu/\epsilon - 1$, and from the fact that the last term in the derivative is negative. Doing some algebraic manipulations, we get:

$$
\begin{aligned}
q_\alpha''(x) \;\leq\; & -(2 - \alpha)(\alpha - 1)x\,(x + \mu/\epsilon - 1)^{-\alpha} \\
& -2c_\alpha(\alpha - 1)\,(x + \mu/\epsilon - 1)^{-\alpha} + \cdot c_\alpha \cdot \alpha(\alpha - 1)x\,(x + \mu/\epsilon - 1)^{-\alpha-1} \\
\;\propto\; & -(2 - \alpha)x\,(x + \mu/\epsilon - 1) - 2c_\alpha\,(x + \mu/\epsilon - 1) + x \cdot c_\alpha \cdot \alpha \\
\;=\; & -[(2 - \alpha)x + 2c_\alpha]\,(x + \mu/\epsilon - 1) + x \cdot c_\alpha \cdot \alpha \\
\;\leq\; & -[(2 - \alpha)x + 2c_\alpha]\,x + x \cdot c_\alpha \cdot \alpha \\
\;\propto\; & -(2 - \alpha)x - 2c_\alpha + c_\alpha \cdot \alpha \\
\;=\; & -(2 - \alpha)x - c_\alpha(2 - \alpha) \\
\;<\; & 0.
\end{aligned}
\tag{F.3}
$$

Notice that the constant of proportionality is positive in all cases. We showed that $q_\alpha(x)$ is concave for $1 < \alpha < 2$. Checking the concavity/convexity out of this range can be done numerically, by simple calculation of $q_\alpha''(x)$.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pt. I, pp. 379–423, 1948; pt. II, pp. 623–656, 1948.

[2] J. Kusuma, "Slepian–Wolf coding and related problems," preprint 2001. Available at: www.mit.edu/∼6.454/www_fall_2001/kusuma/summary.pdf

[3] S. D. Servetto, "Quantization with side information: lattice codes, asymptotics, and applications to sensor networks," vol. 53, no. 2, pp. 714–731, February 2007.

[4] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, pp. 80–94, September 2004.

[5] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner–Ziv problem," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1636–1654, August 2004.

[6] S. Cheng and Z. Xiong, "Successive refinement for the Wyner–Ziv problem and layered code design," in *Proc. DCC 2004*, Snowbird, UT, 2004.

[7] J. Kusuma, L. Doherty and K. Ramchandran, "Distributed compression for sensor networks," in *Proc. ICIP 2001*, vol. 1, pp. 82–85, Thessaloniki, Greece, October 2001.

[8] D. Muresan and M. Effros, "Quantization as histogram segmentation: Optimal scalar quantizer design in network systems," *IEEE Trans. Inform. Theory*, vol. 54, pp. 344–366, January 2008.

[9] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inform. Theory*, vol. 52, pp. 4017–4035, September 2006.

[10] Y. Kaspi and N. Merhav, "Structure theorems for real-time variable rate coding with and without side information," Available at: http://arxiv.org/pdf/1108.2881v1.pdf

[11] J. Nayak, E. Tuncel, "Low-delay quantization for source coding with side information," *Proc. ISIT 2008*, pp. 2732–2736, Toronto, Canada, August 2008.

[12] X. Chen, E. Tuncel, "Low-delay prediction and transform-based Wyner-Ziv Coding," *IEEE Trans. Signal Processing*, vol. 59, no. 2, pp. 653–666 , November 2010.

[13] X. Chen, E. Tuncel, "High-resolution predictive Wyner-Ziv coding of Gaussian sources," *Proc. ISIT 2009*, pp. 1204–1208, Seoul, Korea, August 2009.

[14] A. Reani and N. Merhav, "Efficient on-line schemes for encoding individual sequences with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 57, pp. 6860–6876, October 2011.

[15] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Trans. Inform. Theory*, vol. 19, no. 3, pp. 275–283, May 1973.

[16] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," in: *Information Theory New Trends and Open Problems*, edited by G. Longo, Springer-Verlag, pp. 87-123, 1975.

[17] I. Leibowitz, R. Zamir, "A Ziv-Zakai-Rényi lower bound on distortion at high resolution," *Proc. ITW 2008*, pp. 174–178, Porto, Portugal, May. 2008.

[18] S. Tridenski, R. Zamir, "Bounds for joint source-channel coding at high SNR," *Proc. ISIT 2011*, pp. 771–775, St. Petersburg, Russia, August 2011.

[19] A. Ingber, I. Leibowitz, R. Zamir, M. Feder, "Distortion lower bounds for finite dimensional joint source-channel coding," *Proc. ISIT 2008*, pp. 1183–1187, Toronto, Canada, August 2008.

[20] A. Rényi, "On measures of entropy and information," *Proc. 4th Berk. Symp. Math., Stat. and Prob.*, pp. 547-561, Univ. of Calif. Press, 1961.

[21] N. Merhav, "Data processing theorems and the second law of thermodynamics," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 4926–4939, August 2011.

[22] F. Dupuis, W. Yu, F.M.J. Willems, "Blahut-Arimoto algorithms for computing channel capacity and rate-distortion with side information," *Proc. ISIT 2004*, pp. 179, Chicago, USA, June 2004.

[23] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2009.

[24] R.G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, 1968.