The Rate Loss in the Wyner-Ziv Problem *

Ram Zamir
Dept. of ECE, University of California at Santa Barbara *zamir@kane.ece.ucsb.edu*

June 21, 1996

Abstract

The rate-distortion function for source coding with side information at the decoder (the "Wyner-Ziv problem") is given in terms of an auxiliary random variable, which forms a Markov chain with the source and the side information. This Markov chain structure, typical to the solution of multiterminal source coding problems, corresponds to a loss in coding rate with respect to the conditional rate-distortion function, i.e., to the case where the encoder is fully informed. We show that for difference (or balanced) distortion measures, this loss is bounded by a universal constant, which is the *minimax capacity* of a suitable additive noise channel. Furthermore, in the worst case this loss is equal to the maximin redundancy over the rate distortion function of the additive noise "test" channel. For example, the loss in the Wyner-Ziv problem is less than 0.5 bit per sample in the squared-error distortion case, and it is less than 0.22 bits for a binary source with Hamming-distance. These results have implications also in universal quantization with side information, and in more general multiterminal source coding problems.

Key words: conditional rate-distortion, side information, Markov chain, additive noise test channel, minimax capacity.

*This research was performed while the author was at Cornell University, and it was supported in part by the Wolfson Research Awards administered by the Israel Academy of Science and Humanities. This work was presented in part in the Information Theory Workshops in Rydzyna, Poland, June 1995, and Haifa, Israel, June 1996

I. Introduction

A sequence $\{(X_i, Y_i)\}_{i=1}^{\infty}$ of independent replica of a pair of dependent random variables $(X, Y) \in (\mathcal{X}, \mathcal{Y})$ is generated. We refer to X as the *source* and to Y as the *side information*. Suppose a source block $X_1 \ldots X_n$ is to be encoded with an average error D under some *difference distortion measure* $d(\cdot)$, i.e., the reconstruction $\widehat{X}_1 \ldots \widehat{X}_n$ needs to satisfy

$$E \frac{1}{n} \sum_{i=1}^{n} d(\widehat{X}_i - X_i) \leq D \ . \tag{1}$$

Both X and $\widehat{X}$ belong to the same alphabet space $\mathcal{X}$, and the distortion measure is a function $d: \ \mathcal{X} \to [0, \infty)$. In the continuous alphabet case $\mathcal{X} = \mathcal{R}^1$, and the subtraction in (1) is in the usual sense, e.g.,

$$d(\hat{x} - x) = |\hat{x} - x|^r \ , \quad r > 0 \tag{2}$$

is the absolute r-th power distortion measure. In the discrete alphabet case $\mathcal{X} = \{0, 1, \ldots, M-1\}$, and the "$-$" sign in (1) denotes modulo-M subtraction, e.g.,

$$d(\hat{x} - x) = \begin{cases} 0, & \hat{x} = x \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

is the Hamming-distance (or error frequency) distortion measure. Note that a *balanced* distortion measure [1] effectively falls in the latter category, since it becomes a difference distortion measure after a proper re-indexing of the source alphabet. Throughout the paper we consider the continuous and the discrete cases interchangeably and provide an indication when necessary.

When the side information is not available, the minimal rate of encoding a block of source samples under the error constraint in (1) is given by the rate-distortion function [1]

$$R_X(D) = \inf_{\{U \in \mathcal{X} \,:\, E\{d(U-X)\} \leq D\}} I(X; U) \ . \tag{4}$$

When the side information $Y_1 \ldots Y_n$ is available to both the encoder and the decoder, the minimal coding rate is reduced to the conditional rate-distortion function [1, Sec. 6.1], [4]

$$R_{X|Y}(D) = \inf_{\{U \in \mathcal{X} \,:\, E\{d(U-X)\} \le D\}} I(X;U|Y) \,. \tag{5}$$

In the Wyner-Ziv problem the side information $Y_1 \ldots Y_n$ is available to the decoder but is not available to the encoder. In this case the minimal coding rate is given by [10], [9]

$$R^{WZ}_{X|Y}(D) = \inf I(X;Z|Y) \,, \tag{6}$$

where the infimum is taken over all random variables Z such that $Y - X - Z$ form a Markov chain (i.e., Y and Z are conditionally independent given X), and there exists a function $\widehat{X} = \widehat{X}(Z,Y)$ such that $E\{d(\widehat{X} - X)\} \le D$. (In the continuous case (6) requires some additional conditions upon the distortion measure $d(\cdot)$, which are satisfied, e.g., by the r-th power distortion measure; see [9].)

From the setting of these three problems it should be clear that

$$R_{X|Y}(D) \le R^{WZ}_{X|Y}(D) \le R_X(D) \,. \tag{7}$$

It is easy to construct cases for which the excess of the the right most term over the left most term in (7) is large; in fact, $R_X(D) - R_{X|Y}(D)$ may be as large as $\log|\mathcal{X}|$, i.e., unbounded for continuous sources[1]. However, the redundancy of the middle term, $R^{WZ}_{X|Y}(D)$, over $R_{X|Y}(D)$ is a more interesting quantity. It is *zero* for the discrete lossless case; Slepian and Wolf [8] showed that even without access to Y the encoder can transmit a message at a rate close to the conditional entropy $H(X|Y)$, from which the decoder, who has access to Y, can decode the source arbitrarily reliably. On the other hand, for $D > 0$ Wyner and Ziv [10] observed that $R^{WZ}_{X|Y}(D)$ may be strictly greater than $R_{X|Y}(D)$. Thus, in lossy coding there might be a cost in rate for not having access to the side information at the encoder.

Mathematically this rate loss follows from the Markov chain constraint on $Y - X - Z$ in the expression for $R^{WZ}_{X|Y}(D)$. Without this constraint, we could substitute the U that minimizes (5)

[1] Set $Y = X$, and $D = 0$.

as "Z" and "$\widehat{X}$" in (6), and get $R^{WZ}_{X|Y}(D) = R_{X|Y}(D)$. So, what is the cost of this Markov chain constraint? Can the rate loss of the uninformed encoder be arbitrarily large?

Our main observation in this paper is that it *cannot*; the rate loss in the Wyner Ziv problem is bounded from above by a universal constant which depends only on the alphabet spaces of the source and the side information, and on the distortion measure. This universal bound is the *minimax capacity* of a suitable additive noise channel. We also *lower* bound the maximum possible rate loss in the Wyner Ziv problem, defined as

$$L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D) = \sup_{(X,Y)\in(\mathcal{X},\mathcal{Y})} \left\{ R^{WZ}_{X|Y}(D) - R_{X|Y}(D) \right\} ,$$

where the supremum is taken over all possible joint distributions on $(\mathcal{X}, \mathcal{Y})$ satisfying $EX^2 < \infty$. Our results show that $L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D)$ is essentially the same as the maximin redundancy over the rate distortion function of the additive noise test channel.

The minimax capacity bound on the rate loss in the Wyner Ziv problem is presented in Theorem 1 in the next section. In Section III we define the *additive (noise) rate distortion function*, a quantity which is useful for bounding $R^{WZ}_{X|Y}(D)$. In particular, for doubly-symmetric source - side information pairs the additive rate distortion function coincides with $R^{WZ}_{X|Y}(D)$, thus providing a simple formula for computing the WZ function; see [7] for a possible application. In Theorem 2 we then establish a relation between $L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D)$ and the redundancy of the additive rate distortion function, a relation which leads to tighter (though less explicit) bounds on $L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D)$. In Section IV we provide examples which illustrate the different mechanisms of rate loss in the Wyner-Ziv problem. In Section V we propose a *universal* scheme for coding with side information at the decoder, whose rate loss is bounded by the minimax capacity, i.e., by the same constant that bounds the loss when the statistics are known. Finally we discuss the extension of this theory to multiterminal source coding.

II. An Upper Bound on the Rate Loss of the Un-informed Encoder

Before stating our main theorem we define a certain quantity. For a random variable $N \in \mathcal{X}$, the capacity of the additive noise channel $W \to W + N$ under a $d(\cdot)$-constraint is defined as

$$C(D,N) = \sup_{\{W \in \mathcal{X} \,:\, W \perp\!\!\!\perp N,\ E\{d(W)\} \le D\}} I(W; W+N)\,, \tag{8}$$

where $\perp\!\!\!\perp$ denotes statistical independence, and the "+" sign in (8) means real or modulo-M addition as appropriate. We define the minimax (or worst noise) capacity as

$$C_{\mathcal{X}}(D) = \inf_{\{N \in \mathcal{X} \,:\, E\{d(N)\} \le D\}} C(D,N)\,. \tag{9}$$

$C_{\mathcal{X}}(D)$ may be interpreted as the capacity at equilibrium in a "communicator-jammer game", played over an additive noise channel whose noise $d(\cdot)$-moment and whose input $d(\cdot)$-moment are limited to D; see [2, p. 263]. Note that $C_{\mathcal{X}}(D)$ depends on $\mathcal{X}$, $d(\cdot)$ and D, but not on the side information space $\mathcal{Y}$. We further discuss this quantity after proving the following theorem:

Theorem 1 (Capacity Bound) *The rate loss in the Wyner-Ziv problem is bounded by*

$$R^{WZ}_{X|Y}(D) - R_{X|Y}(D) \le C_{\mathcal{X}}(D)\,. \tag{10}$$

Proof: Let U and N be random variables in $\mathcal{X}$ such that N is independent of (U,X,Y). We have

$$\begin{aligned}
I(X;X+N|Y) - I(X;U|Y) &= I(X;X+N|U,Y) - I(X;U|X+N,Y) &(11)\\
&\le I(X;X+N|U,Y) &(12)\\
&= I(X-U;X-U+N|U,Y) &(13)\\
&\le I(X-U,U,Y;X-U+N) &(14)\\
&= I(X-U;X-U+N) + I(U,Y;X-U+N|X-U) &(15)\\
&= I(X-U;X-U+N)\,, &(16)
\end{aligned}$$

where (11) follows by using twice the chain rule for $I(X;U,X+N|Y)$; (12) follows from the non negativity of the mutual information; (13) follows since given U, the variables $X-U$ and X contain the same information; (14) and (15) follow from the chain rule and the non negativity of the mutual information; and finally (16) follows since the independence of N and (U,X,Y) implies that $(U,Y)-(X-U)-(X-U+N)$ form a Markov chain, so $I(U,Y;X-U+N|X-U)=0$.

Suppose now that U achieves the minimum in (5), i.e., $I(X;U|Y) = R_{X|Y}(D)$[2]. We have $Ed(U-X) \leq D$. Thus, combining with (11)-(16) and definition (8) (viewing $X-U$ as "W") we get

$$I(X;X+N|Y) - R_{X|Y}(D) \leq C(D,N)\,. \tag{17}$$

On the other hand, since N is independent of (X,Y), the random variables $Y-X-(X+N)$ form a Markov chain. Thus, if we assume further that $Ed(N) \leq D$, then the random variable $X+N$ satisfies the conditions on Z in the minimization (6) (with $\widehat{X}=Z=X+N$), implying that

$$R^{WZ}_{X|Y}(D) \leq I(X;X+N|Y)\,. \tag{18}$$

The proof now follows by combining (17), (18) and the definition of $C_{\mathcal{X}}(D)$, since (17) and (18) hold for any N which satisfies $Ed(N) \leq D$. $\square$

The main importance of the minimax capacity bound $C_{\mathcal{X}}(D)$ is in showing that the rate loss of the uninformed encoder is bounded. Another nice property of $C_{\mathcal{X}}(D)$ is that it computes easily. Here are a few examples for evaluating $C_{\mathcal{X}}(D)$. In the discrete source case $C_{\mathcal{X}}(D)$ generally depends on the allowed distortion D, e.g., for a binary source and the Hamming distance it is given by

$$C(binary, Hamm, D) = H_B(D*D) - H_B(D) \leq 0.22 \text{ bits}\,, \quad 0 \leq D \leq 1/2, \tag{19}$$

where $H_B(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy, and $*$ denotes the binary-convolution, i.e., $D*D = 2D(1-D)$; see Figure 2.

[2] If the infimum is not achieved, we can always pick a U that approaches $R_{X|Y}(D)$ arbitrarily closely.

For continuous sources and the r-th power distortion measure (2) the bound $C_{\mathcal{X}}(D)$ is a constant that does not depend on D. To see that, note that the r-th power distortion measure satisfies $E\{d(aX)\} = |a|^r E\{d(X)\}$, while $I(aW; aW + aN) = I(W; W + N)$, so $C_{\mathcal{X}}(aD) = C_{\mathcal{X}}(D)$. For example, for $r = 2$, i.e., in the case of mean-squared error (MSE), C is achieved by a Gaussian source and a Gaussian noise, and is given by

$$C(MSE) = 1/2 \text{ bit}. \tag{20}$$

Good estimates for C for general r-th power distortion measures may be obtained by the method of Maximum-Entropy, as done in [11, Appendix C].

Although quite useful, the minimax capacity bound is not tight. We demonstrate that by a few examples. First, in the quadratic Gaussian case Wyner [9] showed that $R_{X|Y}^{WZ}(D) = R_{X|Y}(D)$ for all D, so the loss is actually zero. Second, for any source the point $(R, D) = (0, D_{max})$ is common to both $R_{X|Y}^{WZ}$ and $R_{X|Y}$, since at zero rate it makes no difference whether the encoder has access to the side information or not. Thus, the rate loss in high distortion is zero, while the capacity bound is positive. Finally, for the r-th power distortion measure, the rate loss for smooth continuous sources vanishes as $D \to 0$[3], while the capacity bound is a positive constant (e.g., half a bit for MSE) for all D.

It seems that $C_{\mathcal{X}}(D)$ provides a loose bound that cannot be achieved. In the next two sections we develop alternative estimates for the rate loss in the Wyner-Ziv problem. Although less explicit than $C_{\mathcal{X}}(D)$, these estimates are much tighter.

[3] By (7), (18) and the conditional Shanon lower bound [4], we have for MSE (and similarly for other error criteria)

$$I(X; X + N|Y) \geq R_{X|Y}^{WZ}(D) \geq R_{X|Y}(D) \geq h(X|Y) - \frac{1}{2}\log 2\pi e D ,$$

where $h(\cdot)$ denotes differential entropy. But if $h(X|Y)$ is finite and the noise N is Gaussian $I(X; X + N|Y) = h(X|Y) - \frac{1}{2}\log 2\pi e D + o(1)$, where $o(1)$ goes to zero as D goes to zero [5]. Thus, also $R_{X|Y}^{WZ}(D) - R_{X|Y}(D) \to 0$.

III. The Additive Noise Rate Distortion Function

A. Definition

In this section we introduce the additive (noise) rate distortion function, which provides bounds and explicit expressions for the Wyner-Ziv rate distortion function (6). We show that the redundancy of the additive rate distortion function explains the mechanism of rate loss in the Wyner-Ziv problem. Beyond its role in this paper, this new quantity proves to be useful in other problems; see [7].

Let

$$\mathcal{E}_d(X|Y) = \min_g E\left\{d(g(Y) - X)\right\} \tag{21}$$

denote the minimum possible average error in estimating X from Y, under the predetermined distortion measure $d(\cdot)$. For example, for the squared error distortion measure ($r = 2$ in (2)),

$$\mathcal{E}_d(X|Y) = VAR(X|Y) = E\left\{(E(X|Y) - X)^2\right\}$$

is the conditional variance, i.e., the variance of the conditional-mean estimator $\widehat{X} = E(X|Y)$, averaged over Y. For discrete X and the error frequency distortion measure (3),

$$\mathcal{E}_d(X|Y) = \sum_y \Pr(y) \cdot \{1 - \max_x \Pr(x|y)\}$$

is the error probability of the maximum a-posteriori (MAP) estimator. Note that using the notation $\mathcal{E}_d(\cdot|\cdot)$, the Wyner-Ziv rate distortion function (6) can be rewritten as

$$R^{WZ}_{X|Y}(D) = \inf_{\{Z\,:\,Y - X - Z,\ \mathcal{E}_d(X|Y,Z) \le D\}} I(X;Z|Y)\ . \tag{22}$$

When no side information is available, we define

$$f_X(D) = \inf_{\left\{N \in \mathcal{X}\,:\,N \perp\!\!\!\perp X,\ \mathcal{E}_d(X|X+N) \le D\right\}} I(X;X+N) \tag{23}$$

(where, as usual, the $+$ sign denotes real or mod-M addition as appropriate, and $\perp\!\!\!\perp$ denotes statistical independence), and

$$R^{add}_X(D) = \text{LOWER CONVEX ENVELOPE}\left\{f_X(D)\right\}\ . \tag{24}$$

We call the latter function the *additive (noise) rate distortion function.* Let $\mathcal{S} = \{1, 2\}$. Since for each D the function $R_X^{add}(D)$ is a convex combination of (at most) two points on $f_X(D)$, we can write it in terms of conditional mutual information and an auxiliary "time sharing" variable S, as

$$R_X^{add}(D) \quad = \quad \inf I(X; X + N|S) = \inf I(X; X + N, S) \,, \tag{25}$$

where the infima are taken over the set

$$\left\{ N \in \mathcal{X}, S \in \mathcal{S} \; : \; (N, S) \underline{\|} X, \;\; \mathcal{E}_d(X|X + N, S) \leq D \right\} \,. \tag{26}$$

The second equality in (25) follows by the chain rule for mutual information, since $I(X; S) = 0$.

$f_X(D)$ may be viewed as the minimal rate that can be achieved by a "test" channel composed of an additive noise channel followed by an optimal estimator, when the mutual information is measured over the additive noise channel; see Figure 1. Let $\mathcal{C}_f$ denote the set of all such test channels. $R_X^{add}(D)$ is thus the minimal rate achieved by a convex combination of channels in $\mathcal{C}_f$. Since the set $\mathcal{C}_f$ is in general not convex, $R_X^{add}(D)$ is sometimes strictly less than $f_X(D)$.

Figure 3 shows the functions $f(D)$, $R^{add}(D)$ and $R(D)$ for a Bernulli(0.3) source, encoded subject to the Hamming distortion (3). This example is discussed in more detail in Section IV.

Clearly, $f_X(D) \geq R_X^{add}(D) \geq R_X(D)$. Sometimes, however, $R_X^{add}(D) = f_X(D) = R_X(D)$ for all D; one example is the quadratic Gaussian case, where an additive noise channel followed by a filter achieves $R(D)$ [1, pp. 101, 143]; another example is the uniformly distributed discrete source, where (for any balanced distortion measure) $R(D)$ is achieved by a modulo-M additive noise channel [1, p. 62]. Equality holds also at the extreme distortion values. At the maximum distortion $D_{max} = \mathcal{E}_d(X)$ we have $R_X^{add}(D_{max}) = R_X(D_{max}) = 0$. At the other extreme, assuming that $d(x) = 0$ if and only if $x = 0$, we have for a discrete source $R_X^{add}(0) = R_X(0) = H(X)$, while for a smooth continuous source $R_X^{add}(D) - R_X(D) \to 0$, as $D \to 0$ [5].

We now generalize these definitions for the case where side information is available. Following the structure of definitions (23) - (25), we define the *conditional additive rate distortion function* as

$$R^{add}_{X|Y}(D) = \inf_{\{N \in \mathcal{X}, S \in \mathcal{S}\,:\,(N,S) \perp\!\!\!\perp (X,Y),\ \mathcal{E}_d(X|X+N,Y,S) \leq D\}} I(X; X+N, S|Y)\,. \tag{27}$$

Figure 4 shows the functions $R^{add}_{V|W}(D)$ and $R_{V|W}(D)$ for a composite Gaussian source $V \sim \mathcal{N}(0, \sigma^2_W)$, where W is a binary random variable. This example, which is discussed in more detail in Section IV, shows again the redundancy of the additive rate distortion function above the regular rate distortion function.

Observe that if a pair of random variables (N, S) satisfies the conditions of the minimization in (27), then $Y - X - (X + N, S)$ form a Markov chain, and the error of the optimum estimate of X from $(X + N, S, Y)$ is less than or equal to D. Thus, $(X + N, S)$ satisfies the conditions on Z in the minimization (6), so

$$R^{WZ}_{X|Y}(D) \leq R^{add}_{X|Y}(D)\,. \tag{28}$$

Sometimes (28) holds with equality; following the Gaussian example in [9], if X and Y are jointly Gaussian encoded under the squared error distortion measure,

$$R^{add}_{X|Y}(D) = R^{WZ}_{X|Y}(D) = R_{X|Y}(D) \quad \text{for all } D. \tag{29}$$

Other examples for equality in (28) are given in Subsection C below.

B. Equivalence of the Worst Case Rate Losses

In this subsection we show that in the worst case, the rate loss in the Wyner-Ziv problem is essentially the same as the redundancy of the additive rate distortion function above the regular rate distortion function.

Suppose the source and side information spaces $\mathcal{X}$ and $\mathcal{Y}$, and the distortion measure $d(\cdot)$, are fixed. The side information alphabet space $\mathcal{Y}$ may be finite, countably infinite, i.e., $\mathcal{Y} = \mathcal{Z}$, or continuous, i.e. $\mathcal{Y} = \mathcal{R}^k$ for some integer k. The source alphabet may be either finite or continuous, as discussed in the Introduction. Recall the definition given in the Introduction of the maximum loss in the Wyner-Ziv problem

$$L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D) = \sup_{(X,Y)\in(\mathcal{X},\mathcal{Y})} \left\{ R^{WZ}_{X|Y}(D) - R_{X|Y}(D) \right\} \ , \tag{30}$$

where the supremum is taken over all possible joint distributions on $(\mathcal{X},\mathcal{Y})$ satisfying $EX^2 < \infty$. To avoid unnecessary confusion, when $\mathcal{X}$, $\mathcal{Y}$ or both are continuous we assume a (standard) Borel sigma algebra for the probability space of (X,Y). In analogy to (30), we define with respect to the conditional additive rate distortion function

$$L^{add}_{\mathcal{X}|\mathcal{Y}}(D) = \sup_{(X,Y)\in(\mathcal{X},\mathcal{Y})} \left\{ R^{add}_{X|Y}(D) - R_{X|Y}(D) \right\} \ , \tag{31}$$

where the sepremum is taken over the same (X,Y) pairs as in (30). We call $L^{add}_{\mathcal{X}|\mathcal{Y}}(D)$ the *maximin redundancy of the additive noise test channel*, since it is obtained by minimizing $I(X;X+N|Y)$ over N and then maximizing (the redundancy) over (X,Y).

We make the following observations with respect to $L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D)$ and $L^{add}_{\mathcal{X}|\mathcal{Y}}(D)$:

1. Following Theorem 1 we have

$$L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D) \leq C_{\mathcal{X}}(D) \quad \text{for any } \mathcal{Y}. \tag{32}$$

   In fact, it is easy to verify that the proof of Theorem 1 applies also to $R^{add}_{X|Y}(D)$, so also $L^{add}_{\mathcal{X}|\mathcal{Y}}(D) \leq C_{\mathcal{X}}(D)$ for any $\mathcal{Y}$.

2. When the side information is degenerate (e.g., when $|\mathcal{Y}| = 0$ or $|\mathcal{Y}| = 1$), the Wyner-Ziv function coincides with the rate distortion function, i.e., $L^{WZ}_{\mathcal{X}}(D) = 0$. However, the additive noise test channel is in general suboptimal even with no side information, i.e., $L^{add}_{\mathcal{X}}(D) > 0$.

3. Like the minimax capacity $C_{\mathcal{X}}$, also the quantities $L^{WZ}_{\mathcal{X}|\mathcal{Y}}$ and $L^{add}_{\mathcal{X}|\mathcal{Y}}$ are *constants* that do not depend on D for continuous sources and the r-th power distortion measure. To see that, note that the r-th power distortion measure (2) satisfies $\mathcal{E}_d(aX|Y) = |a|^r \cdot \mathcal{E}_d(X|Y)$. This implies that the conditional rate distortion function satisfies $R_{aX|Y}(D) = R_{X|Y}(D/|a|^r)$, while

$R^{WZ}_{aX|Y}(D)$ and $R^{add}_{aX|Y}(D)$ satisfy similar relations, so $L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D) = L^{WZ}_{\mathcal{X}|\mathcal{Y}}(aD)$, and $L^{add}_{\mathcal{X}|\mathcal{Y}}(D) = L^{add}_{\mathcal{X}|\mathcal{Y}}(aD)$.

4. The dependence of $L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D)$ and $L^{add}_{\mathcal{X}|\mathcal{Y}}(D)$ on the alphabet space of the side information is only through its *cardinality* $|\mathcal{Y}|$. This fact follows by observing that if there is a one-to-one mapping between the elements of two alphabet spaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$, then every probability distribution on $(\mathcal{X}, \mathcal{Y}_1)$ can be mapped into an identical probability distribution on $(\mathcal{X}, \mathcal{Y}_2)$, and vice versa. We conclude that if $|\mathcal{Y}_1| = |\mathcal{Y}_2|$ (in the sense that a one-to-one mapping exists), then $L^{WZ}_{\mathcal{X}|\mathcal{Y}_1}(D) = L^{WZ}_{\mathcal{X}|\mathcal{Y}_2}(D)$ and $L^{add}_{\mathcal{X}|\mathcal{Y}_1}(D) = L^{add}_{\mathcal{X}|\mathcal{Y}_2}(D)$.

5. $L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D)$ and $L^{add}_{\mathcal{X}|\mathcal{Y}}(D)$ are *non-decreasing functions* of $|\mathcal{Y}|$. This fact follows again by the mapping argument above, since if $|\mathcal{Y}_2| \geq |\mathcal{Y}_1|$ then there exists a mapping of every element in $\mathcal{Y}_1$ to a distinct element in $\mathcal{Y}_2$ (but the opposite may not be true).

We are now in a position to state our main result in this section:

Theorem 2

$$L^{WZ}_{\mathcal{X}|\mathcal{Y}}(D) \leq L^{add}_{\mathcal{X}|\mathcal{Y}}(D) \leq L^{WZ}_{\mathcal{X}|\mathcal{X}\times\mathcal{Y}}(D)\,, \tag{33}$$

where $L^{WZ}_{\mathcal{X}|\mathcal{X}\times\mathcal{Y}}(D)$ is the maximum loss when the side information space is the Cartesian product $\mathcal{X} \times \mathcal{Y}$.

The proof relays on Theorem 3 below, and is given in the forth part of this section. We have a simple corollary from Theorem 2. Let $\mathcal{Z}$ denote a countably infinite set, and let $\mathcal{C}$ denote an uncountable set, e.g., $\mathcal{R}^k$ for some k. Clearly, if $\mathcal{X}$ is finite, then $|\mathcal{X} \times \mathcal{Z}| = |\mathcal{Z}|$, while if $\mathcal{X}$ is either finite or continuous, then by the Borel equivalence $|\mathcal{X} \times \mathcal{C}| = |\mathcal{C}|$. Combining this simple fact with Theorem 2 and Remarks 4 and 5 above, we conclude the following corollary:

Corollary 1 *If $\mathcal{X}$ is finite, then*

$$L^{WZ}_{\mathcal{X}|\mathcal{Z}}(D) = L^{add}_{\mathcal{X}|\mathcal{Z}}(D)\,, \tag{34}$$

while if $\mathcal{X}$ is finite or $\mathcal{X} = \mathcal{R}^1$, then

$$L^{WZ}_{\mathcal{X}|\mathcal{C}}(D) = L^{add}_{\mathcal{X}|\mathcal{C}}(D)\,. \tag{35}$$

Thus, if we do not restrict the cardinality of the side information alphabet space, the maximum loss in the Wyner-Ziv problem is *equal* to the maximin redundancy of the additive noise test channel.

By constructing an example (see the second example in the next section), we could show that for the MSE criterion and $|\mathcal{Y}| = 2$, $L^{add}(MSE) \geq 0.108$ bits. By Remark 5 above, this is also a lower bound for $L^{add}(MSE)$ if the cardinality of $\mathcal{Y}$ is *unrestricted*. Combining with (20) and (35), we thus conclude that for unrestricted side information space

$$0.108 \leq L^{WZ}(MSE) \leq 0.5 \text{ bits.}$$

We were not able to get better estimates of $L^{WZ}(MSE)$. We tend to believe, however, that the lower bound above, achieved by the composite Gaussian source example in Section IV below, is actually tight.

C. Relations for symmetric side information

We now develop some relations among the conditional-, the Wyner-Ziv-, and the additive rate distortion functions for the special case where the side information is obtained by passing the source through an additive noise channel. Theorem 3 below summarizes the main result of this subsection, whose two corollaries serve as tools for proving Theorem 2.

Assume that $Y = X - V$, where V is independent of X. The minus sign of the additive noise $(-V)$ eases the notation in the sequel, but is in fact not necessary if either the random variable V or the distortion measure function $d(\cdot)$ are symmetric. Since subtracting the condition does not affect the conditional rate distortion function relative to a difference distortion measure, we have $R_{X|Y}(D) = R_{(X-Y)|Y}(D) = R_{V|Y}(D)$. Applying well known bounds on the conditional rate distortion function [4, Thm. 3.1], we obtain

$$R_V(D) \geq R_{X|Y}(D) \geq R_V(D) - I(V;Y)\,, \tag{36}$$

where $R_V(D)$ is the rate distortion function subject to the distortion measure d of the noise V in the side information channel.

As shown in the following theorem, the Wyner-Ziv rate distortion function satisfies similar inequalities, but with respect to the *additive* rate distortion function of V. The theorem actually considers a slightly more general situation, in which the decoder has access also to another random variable W, which is correlated with the noise in the side information channel but not with the source:

Theorem 3 *Suppose $Y = (X - V, W)$, where the pair (V, W) is independent of X. Then*

$$R^{add}_{V|W}(D) \geq R^{WZ}_{X|Y}(D) \geq R^{add}_{V|W}(D) - I(V; X - V|W) \,, \tag{37}$$

where $R^{add}_{V|W}(D)$ is the conditional additive rate distortion function (27) (subject to the distortion measure d) of the noise in the side information channel. In particular, Ineq. (37) holds for the case $Y = X - V$, i.e., without conditioning on W.

We give the proof in Appendix II. In the rest of this subsection we comment on and present a few important corollaries of Theorem 3.

Equality in (37) occurs at the extreme distortion levels. The upper bound in (37) is tight at the maximum distortion $D_{max} = \mathcal{E}_d(V|W)$, where $R^{add}_{V|W}(D_{max}) = 0$. The lower bound in (37) is tight in the discrete case at $D = 0$, where $R^{WZ}_{X|Y}(0) = H(X|Y) = H(V|Y)$ and $R^{add}_{V|W}(0) = H(V|W)$. In the continuous case the lower bound is asymptotically tight as $D \to 0$, due to the asymptotic tightness of the Shannon lower bound [5].

We obtain a slight variation on the bound (37): Verify from (27), that like the regular rate distortion function [4], the additive rate distortion function also decreases by further conditioning, and (for difference distortion measures) is not affected by adding or subtracting the condition. So $R^{add}_{V|W}(D) \geq R^{add}_{V|(W,X-V)}(D) = R^{add}_{X|Y}(D)$. Thus, from (28) and (37)

$$R^{add}_{X|Y}(D) \geq R^{WZ}_{X|Y}(D) \geq R^{add}_{X|Y}(D) - I(V; X - V|W) \,. \tag{38}$$

Two simple corollaries follow from Theorem 3. The first is a generalization of Wyner-Ziv's doubly symmetric binary source example (see [10, Section II]) to any finite alphabet, while the second is its continuous analog:

Corollary 2 (Doubly Symmetric Discrete Source) *For the same Y as in Theorem 3, if X is uniform over $\{0,\ldots,M-1\}$, then*

$$R^{WZ}_{X|Y}(D) = R^{add}_{X|Y}(D) = R^{add}_{V|W}(D) \,. \tag{39}$$

Corollary 3 (Asymptotically Doubly Symmetric Continuous Source) *Suppose $Y = (X - V, W)$ where the pair (V, W) is independent of X, as in Theorem 3. Assume further that V is real valued with $EV^2 < \infty$, and that $X = \alpha Z$, where α is a parameter and Z is a continuous random variable with $EZ^2 < \infty$ and $h(Z) > -\infty$. Then, as $\alpha \to \infty$*

$$R^{WZ}_{X|Y}(D)\,,\ R^{add}_{X|Y}(D) \to R^{add}_{V|W}(D)\,, \tag{40}$$

i.e., both the WZ- and the additive rate distortion functions of X given Y converge to $R^{add}_{V|W}(D)$.

The first corollary follows by noting that for a uniform discrete source the output of the side information channel (given by the modulo-M sum of the source and the additive noise) is conditionally independent of the noise V given W, so $I(V; X - V|W) = 0$. As shown in Appendix I, the same thing happens for the continuous source case asymptotically, as the signal to noise ratio at the side information channel goes to infinity, i.e., $I(V; \alpha Z - V|W) \to 0$ as $\alpha \to \infty$.

In parallel with these two corollaries, we have for the (regular) conditional rate distortion function

$$R_{X|Y}(D) = R_{V|W}(D)\,, \quad \text{and} \quad R_{X|Y}(D) \to R_{V|W}(D) \ \text{ as } \alpha \to \infty, \tag{41}$$

for a doubly symmetric discrete source and for an asymptotically doubly symmetric continuous source, respectively. Eq. (41) follows from the inequality $R_{V|W}(D) \geq R_{X|Y}(D) \geq R_{V|W}(D) - I(V; X - V|W)$, which is a conditional version of (36). Note that in the continuous case, the *un*conditional rate distortion function of $X = \alpha Z$ diverges to infinity as $\alpha \to \infty$; yet, as (40) and (41) show, the conditional-, the conditional-additive- and the WZ-rate distortion functions of X given Y converge to finite quantities.

Mackenthun [6] studied the discrete case $Y = X + V$, where X is uniform over $\mathcal{X}$, and

$$\Pr(V = v) = \begin{cases} 1 - P_e, & v = 0 \\ P_e/(M-1), & v \neq 0. \end{cases}$$

For this special case, Mackenthun derived lower and upper bounds on $R^{WZ}_{X|Y}(D)$ relative to the Hamming distortion measure for $0 \leq D \leq P_e$. Corollary 2, when specialized to this case, yields $R^{WZ}_{X|Y}(D) = R^{add}_{V}(D) = \text{L.C.E.}\{f_V(D)\}$, where $f_V(D)$ (see (23)) is achieved by an additive noise N which is distributed like V above but with a parameter D instead of P_e. This latter expression seems to confirm Mackenthun's conjecture that the upper bound in [6] is tight.

Finally we note that a similar technique was used in [7] to show that $R^{WZ}_{X|Y}(D) = R^{add}_{X|Y}(D)$ in a mixed discrete-continuous case, where the source is binary symmetric, while the side information is given by passing the source through a Gaussian channel.

D. Proof of Theorem 2

The left inequality in (33) follows immediately from (28). The right inequality is shown by construction. Specifically, we show that for *any* pair of random variables $(V, W) \in (\mathcal{X}, \mathcal{Y})$ with $EV^2 < \infty$, we can find a pair of random variables $(X, Y) \in (\mathcal{X}, \mathcal{X} \times \mathcal{Y})$ with $EX^2 < \infty$, such that $R^{WZ}_{X|Y}(D) - R_{X|Y}(D)$ is equal (in the discrete source case) or close (in the continuous source case) to $R^{add}_{V|W}(D) - R_{V|W}(D)$. This will imply that $L^{add}_{\mathcal{X}|\mathcal{Y}}(D) \leq L^{WZ}_{\mathcal{X}|\mathcal{X}\times\mathcal{Y}}(D)$, as desired.

The construction of (X, Y) is done as follows. In the discrete case we let X be independent of (V, W) and uniform over $\mathcal{X}$, and we let $Y = (X - V, W)$. From Corollary 2 and (41) we then have $R_{X|Y}(D) = R_{V|W}(D)$ and $R^{WZ}_{X|Y}(D) = R^{add}_{V|W}(D)$, which implies the desired equation

$$R^{WZ}_{X|Y}(D) - R_{X|Y}(D) = R^{add}_{V|W}(D) - R_{V|W}(D)\,. \tag{42}$$

In the continuous case we let X be Gaussian with variance σ_x^2 and independent of (V, W), and we let $Y = (X - V, W)$. From Corollary 3 and (41) we then know that for any $\epsilon > 0$, we can pick σ_x^2

large enough but finite, such that $0 \leq R_{V|W}(D) - R_{X|Y}(D) \leq \epsilon$ and $0 \leq R^{add}_{V|W}(D) - R^{WZ}_{X|Y}(D) \leq \epsilon$. Thus, (42) can be approached as closely as desired, and the theorem is proved. □

Remark: By symmetry, we could have used a dual construction, in which the source is obtained by passing the side information through an additive noise channel, i.e., $Y = (U, W)$, where U is independent of (V, W) and is uniformly distributed over $\mathcal{X}$ (or is a large Gaussian in the continuous case), and $X = U + V$. This formulation explains better the idea behind the construction in the proof of Theorem 2, which we call "spreading"; we obtain the source by randomizing the variable V using a "highly random" variable U, while providing U as side information to the decoder.

IV. Two Examples

Theorem 2 and Corollary 1 above indicate that in the worst case, the optimal strategy of the uninformed encoder is to adopt the structure of an additive noise test channel (or convex combination of such channels). In light of this observation, we find two principal mechanisms of rate losses in coding with an uninformed encoder. The first is associated with the sub optimality of the additive noise test channel for any value of the side information. We can study the effect of this mechanism by considering the redundancy of $R^{add}_X(D)$ above $R_X(D)$.

The second is associated with the Markov chain constraint in (6), which means that the *same channel* is applied to different modes of a composite source (i.e., for different values of the side information). This is in contrast with the structure of the optimal solution, i.e., the channel that realizes the conditional rate-distortion function, which combines the optimal solutions of the different modes at *points of equal slope* of their R-D curves [1, Sec. 6.1]. The loss due to this mechanism may be studied by considering the redundancy of $R^{add}_{X|Y}(D)$ above $R_{X|Y}(D)$ for a composite source, where Y indicates the source mode. We provide below two examples which illustrate the two rate loss mechanisms discussed above.

Example 1: A binary source V with probabilities $(p, 1-p)$, where $p < 0.5$, encoded with probability D of coding error, where $0 \leq D \leq p$. The rate-distortion function of this source is

$$R_V(D) = H_B(p) - H_B(D) \ ,$$

while the function $f_V(D)$ defined in (23) is given by

$$f_V(D) = \begin{cases} H_B(p * D) - H_B(D) \ , & D < p \\ 0 \ , & D = p \ . \end{cases} \tag{43}$$

Note that the point $(R, D) = (0, p)$ is achieved by a binary symmetric noise, i.e., $\Pr(N = 1) = 0.5$, but in order to achieve any distortion which is (even slightly) smaller than p the noise must have $\Pr(N = 1) = D < p < 0.5$, so $f_V(D)$ is discontinuous at $D = p$. Finally, by definition (24), $R_V^{add}(D)$ is the lower convex envelope of (43), given by

$$R_V^{add}(D) = \min_{\{0 \leq \theta \leq 1, \, 0 \leq \beta \leq p \, : \, \theta\beta + (1-\theta)p = D\}} \theta \cdot f_V(\beta) \ .$$

Figure 3 illustrates these three curves, where the lower curve is $R_V(D)$, and the upper convex curve is $f_V(D)$. The upper portion of $R_V^{add}(D)$ coincides with $f_V(D)$, while its lower portion is the straight line segment which starts at $(R, D) = (0, p)$ and is tangent to $f_V(D)$.

As shown in Section III.C, if X is a binary symmetric source independent of V, and $Y = X + V$, then

$$R_{X|Y}(D) = R_V(D) \ , \quad \text{and} \quad R_{X|Y}^{WZ}(D) = R_V^{add}(D) \ .$$

This (X, Y) - source / side information pair is in fact the "doubly symmetric binary source" example given in [10, Sec. II][4].

We believe that the concavity $\cap$ of $f(D)$ near D_{max} as shown in the example above can be observed also in other cases, e.g., when encoding under MSE a doubly-modal Gaussian source, $X \sim (1 - p) \cdot \mathcal{N}(0, \sigma^2) + p \cdot \mathcal{N}(\mu, \sigma^2)$, where $\mu \gg \sigma$.

[4] To the best of the author's knowledge, the doubly symmetric binary source example was established during private communication between Wyner, Ziv and Berger.

Example 2: A composite Gaussian source $V \sim \mathcal{N}(0, \sigma_W^2)$, where $W \in \{1, 2\}$ is a binary random variable with $Pr(W = 1) = p$, encoded with mean squared error D. The variance of V is $\sigma^2 = p\sigma_1^2 + (1-p)\sigma_2^2$, and its conditional entropy power is

$$P(V|W) = \frac{2^{2h(V|W)}}{2\pi e} = \sigma_1^{2p} \cdot \sigma_2^{2(1-p)} .$$

Assume $\sigma_1^2 \leq \sigma_2^2$. The conditional rate-distortion function of V given W is [1, Sec. 6.1]

$$R_{V|W}(D) = \begin{cases} \frac{1}{2}\log\left(\frac{P(V|W)}{D}\right) , & 0 < D \leq \sigma_1^2 \\ \frac{1-p}{2}\log\left(\frac{\sigma_2^2(1-p)}{D - p\sigma_1^2}\right) , & \sigma_1^2 \leq D \leq \sigma^2 . \end{cases}$$

It is easy to verify that $R^{add}_{V|W}(D)$ defined in (27) is achieved by Gaussian noise, and is given parametrically by

$$R^{add}_{V|W} = \frac{p}{2} \cdot \log\left(1 + \frac{\sigma_1^2}{\sigma_n^2}\right) + \frac{1-p}{2} \cdot \log\left(1 + \frac{\sigma_2^2}{\sigma_n^2}\right)$$

and

$$D = p \cdot \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_n^2}\right)^{-1} + (1-p) \cdot \left(\frac{1}{\sigma_2^2} + \frac{1}{\sigma_n^2}\right)^{-1} ,$$

where the entire R-D curve, $0 < D \leq \sigma^2$, is obtained as the parameter σ_n^2 (the variance of the additive Gaussian noise) varies from zero to infinity.

Figure 4 illustrates the above two R-D curves (on a logarithmic scale for the distortion) for $\sigma_1^2 = 1$, $\sigma_2^2 = 1000$, and $p = 0.82$. It can be seen from the figure that the maximum gap between the curves is around $D \approx 1.1\sigma_1^2 = 1.1$. Further analysis shows that the maximal rate loss of the additive noise channel realization occurs for $\sigma_2/\sigma_1^2 \to \infty$ and $p \approx 0.82$, and it is about 0.108 bits.

Following Corollary 3, we let $X \sim \mathcal{N}(0, \sigma_x^2)$ and $Y = (X + V, W)$. The source / side information pair (X, Y) satisfies, as $\sigma_x^2 \to \infty$,

$$R_{X|Y}(D) \to R_{W|V}(D) , \quad \text{and} \quad R^{WZ}_{X|Y}(D) \to R^{add}_{W|V}(D) .$$

Figure 4, thus, reflects also a rate loss curve for an uninformed encoder in coding a composite Gaussian source[5].

[5]Note that the loss of $R^{WZ}_{V|W}(D)$ with respect to $R_{V|W}(D)$ is much smaller than $R^{add}_{V|W}(D) - R_{V|W}(D)$. Without

Returning to the second rate loss mechanism discussed in the beginning of this section, we show in Figure 5 the additive noise channel realizations of $R_{V|W}(D)$ and $R^{add}_{V|W}(D)$. As seen in the figure, the channel associated with the informed encoder uses two different noise powers, σ^2_{n1} and σ^2_{n2}, for each one of the two modes of the source, while due to the lack of side information, the channel associated with the uninformed encoder must use the same noise power, σ^2_n, for both modes. For example, if $p = 0.85$, $\sigma^2_2/\sigma^2_1 \to \infty$, and $D = \sigma^2_1$, we have $\sigma^2_{n1} = \infty$ and $\sigma^2_{n2} = D$, while $\sigma^2_n = D/\sqrt{1-p} \approx 2.6D$.

V. On Universal Quantization with Side Information

There is a striking similarity between the minimax capacity bound C in (10) on the rate loss of the uninformed encoder, and the bound due to Ziv [14] on the redundancy of *entropy coded dithered quantization* ("universal quantization"); see also [11]. This resemblance is actually not so surprising, since both problems are associated with the rate redundancy of the additive noise test channel over the optimal test channel.

In this section we further develop the lines of similarity between universal quantization and the Wyner-Ziv problem. We consider coding with side information in the case where the joint statistics of the source and the side information are unknown a priori. We limit the discussion to continuous sources with a (known) bounded support, finite side information space, and the MSE criterion.

Let $N^* \sim \mathcal{N}(0, D)$ be a Gaussian variable independent of the source and the side information, and let

$$R^{add*}_{X|Y}(D) = I(X; X + N^*|Y) \tag{44}$$

be a possible (suboptimal) solution of the minimization in (27). We show below that there exists

the "spreading" (i.e., the transformation of the pair (V, W) to the pair (X, Y)), the uninformed encoder of V can simply encode in a coarse way the region near the origin where the distribution of V given $W = 1$ (the narrow Gaussian mode) is concentrated, and so approach the performance as if it had access to W.

a fixed (universal) coding scheme, operating at a fixed rate R_U, whose decoding distortion is close to D for any source / side information pair (X, Y) for which

$$R^{add*}_{X|Y}(D) < R_U \, . \tag{45}$$

Since N^* is the additive noise achieving $C(MSE)$, the minimax capacity in the MSE case, it follows from the proof of Theorem 1 that the redundancy of $R^{add*}_{X|Y}$ above the conditional rate-distortion function is upperbounded by (20)

$$R^{add*}_{X|Y}(D) - R_{X|Y}(D) \leq C(MSE) = 0.5 \ \text{bit}. \tag{46}$$

Thus, the redundancy of this universal scheme is upperbounded by the same constant which upperbounds the loss of $R^{WZ}_{X|Y}(D)$, i.e., the loss in the non-universal case.

The proposed universal scheme combines randomized ("dithered") quantization [14], [11], and universal Slepian-Wolf lossless coding [8], [3]. Specifically, let $Q_n(\cdot) : \ \mathcal{R}^n \rightarrow \{1, 2, \ldots, M\}$ be a bounded n-dimensional lattice quantizer covering the source support, and let $\underline{Z}$, the dither, be a random vector known to both encoder and decoder and uniformly distributed over the basic cell of the quantizer. Q_n is scaled so that $E\frac{1}{n}\|\underline{Z}\|^2 = D$. The encoder adds the dither to the source block, and quantize the result using Q_n. The quantizer output is then encoded at rate R_U using the (universal) binning technique described in [3, pp. 264, 265]. This method assures that if

$$R_U > \frac{1}{n} H(Q_n | \underline{Z}, \underline{Y}) \, ,$$

then the decoder (that has access to $\underline{Y}$ and $\underline{Z}$) will be able to decode Q_n with an arbitrarily small error. Finally, the decoder subtracts the dither from Q_n. We therefore have with high probability

$$\underline{\widehat{X}} = Q_n(\underline{X} + \underline{Z}) - \underline{Z} \, . \tag{47}$$

Consider first the resulting distortion. From [11] we know that, independently of the statistics of $\underline{X}$,

$$E\frac{1}{n}\|Q_n(\underline{X} + \underline{Z}) - \underline{Z} - \underline{X}\|^2 = E\frac{1}{n}\|\underline{Z}\|^2 = D \, ,$$

where the expectation is taken with respect to $\underline{X}$ and $\underline{Z}$. Thus $E\frac{1}{n}\|\underline{\widehat{X}} - \underline{X}\|^2 \approx D$. As for the rate condition (45), a slight modification of Theorem 1 in [11] yields

$$H\left(Q_n(\underline{X}+\underline{Z})|\underline{Z},\underline{Y}\right) = I(\underline{X};\underline{X}+\underline{N}|\underline{Y})\,, \tag{48}$$

where $\underline{N}$ is independent of $\underline{X}$ and $\underline{Y}$, and is distributed as $-\underline{Z}$. Theorem 3 in [12], which asserts the asymptotic Gaussianity of lattice quantization noise, implies that by taking the optimal lattice quantizer at each dimension and increasing n, the right member of (48), normalized by $1/n$, approaches the right member of (44). This proves our claim.

One may argue that the decoder, who has access to the side information, may do better than the reconstruction (47). Indeed, by observing both Q_n, $\underline{Z}$ and $\underline{Y}$, i.e., virtually observing independent samples of $X + N^*$ and Y, the decoder may construct an estimator $\underline{\widehat{X}}(Q_n, \underline{Z}, \underline{Y})$ whose distortion approaches $\mathcal{E}_d(X|X+N^*, Y)$. Assuming N^* is close to the optimal noise in (27), this distortion is actually the solution for D of the equation $R^{add}_{X|Y}(D) = \frac{1}{n}H(Q_n|\underline{Z},\underline{Y})$. Thus, if the full rate R_U is exploited, we may even achieve the solution for D of the equation $R^{add}_{X|Y}(D) = R_U$. Unfortunately, the encoder, who does not have access to Y, cannot learn the joint statistics of X and Y and thus cannot assume the enhancement described above when it chooses Q_n.

Yet, we can achieve part of this enhanced performance if we assume some a priori knowledge about the joint statistics of the source and the side information. In [13], Zamir and Feder showed how to incorporate pre/post-filters to reduce the rate of the universal quantizer if the second order statistics of the source are known in advance. Here we use the same idea, assuming the covariance matrix of (X, Y), i.e., σ_x^2, σ_y^2 and the correlation coefficient ρ, are known a priori. We improve the reconstruction (47) by applying at the decoder the linear transformation

$$\underline{\widehat{X}} = a\left(Q_n(\underline{X}+\underline{Z}) - \underline{Z}\right) + b\underline{Y}\,, \tag{49}$$

where $a = 1 - \frac{D}{\sigma_x^2(1-\rho^2)}$, $b = (1-a)\rho\frac{\sigma_x}{\sigma_y}$, and the lattice quantizer is now scaled so that $E\frac{1}{n}\|\underline{Z}\|^2 = D/a$; see Figure 6. According to [11] and [12], in the limit as $n \to \infty$ this scheme becomes equivalent to the channel $\widehat{X} = a(X + N^*) + bY$, which is actually the optimal test channel that realizes both $R_{X|Y}(D)$ and $R^{WZ}_{X|Y}(D)$ in the joint Gaussian case [9, Sec. 3]. Thus, the rate loss of this scheme in the joint Gaussian case is zero. Note, however, that this "ideal" result can only be approached, since according to our assumptions X has a finite support and Y is discrete.

When this scheme is used for a non-Gaussian (X, Y) pair, the distortion is still D, since the mean squared error in a linear additive noise channel depends only on the second order statistics. Following [13, Th. 5], the capacity bound of 0.5 bit on the excess rate over the conditional rate-distortion function still holds. Furthermore, following [13, Ths. 1, 2], we have for large n and an optimal lattice quantizer

$$R^{WZ}_{X|Y}(D) \leq \frac{1}{n} H(Q_n|Z, Y) \leq R_{X|Y}(D) + \mathcal{D}(f_{X|Y} \| f^*_{X|Y}) + \delta(Q_n) \ , \tag{50}$$

where $\delta(Q_n) \rightarrow 0$ as $n \rightarrow \infty$, and $\mathcal{D}(f_{X|Y} \| f^*_{X|Y})$ is the conditional divergence [2, Eq. 2.65] between the density function of X given Y and the density function of X^* given Y^*, where (X^*, Y^*) are jointly Gaussian with the same covariance matrix as X and Y. As a by product of (50), we got another simple (non-universal) bound, $\mathcal{D}(f_{X|Y} \| f^*_{X|Y})$, on the rate loss in the quadratic Wyner-Ziv problem.

VI. Summary and Possible Extensions

We considered the rate loss, with respect to the conditional rate distortion function, in coding with side information at the decoder under difference distortion measures. We first bounded the rate loss by the minimax capacity $C_{\mathcal{X}}(D)$, which is easy to compute (as shown in several examples) but is not tight. We then showed that if the side information space is not restricted, the maximum possible rate loss in the Wyner-Ziv problem L^{WZ} is equal to the maximin redundancy of additive noise test channel. The latter quantity is unfortunately hard to compute, yet it allows to construct simple lower bounds on L^{WZ}, e.g., for the MSE case $0.108 \leq L^{WZ} \leq 0.5$ bits.

These notions apply also to more general multiuser lossy source coding problems. Consider, for example, the case where both X and Y must be encoded with distortions D_x and D_y, under difference distortion measures $d_x(\cdot)$ and $d_y(\cdot)$. It is interesting to asses the total rate loss due to separating the encoders with respect to jointly encoding X and Y. The problem of determining the region $\mathcal{R} = \{(R_1, R_2)\}$ of admissible rate pairs for separate encoding has not been solved yet. Nevertheless, we can apply the same technique used in proving the capacity bound in (10), and show that

$$\min_{(R_1,R_2)\in\mathcal{R}}\{R_1 + R_2\} - R_{X,Y}(D_x, D_y) \leq C_{\mathcal{X}}(D_x) + C_{\mathcal{Y}}(D_y) \ , \tag{51}$$

where $R_{X,Y}(D_x, D_y)$ is the (joint) rate distortion function of (X, Y), and $C(\cdot)$ is the capacity defined in (9), relative to the corresponding distortion measure (d_x or d_y). Thus, if both sources are continuous and encoded under the MSE criterion, the rate loss due to separating the encoders is upper bounded by 1 bit.

To summarize the main theme of the paper, coding with an uniformed encoder is usually much more complex than regular conditional coding; yet, we showed that at least the cost in coding rate is not so severe.

Acknowledgments

I thank Toby Berger, Srikant Jayaraman, and Harish Viswanathan from Cornell University for contributing to the basic idea behind this work. I also acknowledge Shlomo Shamai and Tamas Linder for important comments regarding the material presented in Section III.

Appendix I: Proof of Corollary 3

We show that under the conditions of corollary 3, $I(V;\alpha Z - V|W) \to 0$ as $\alpha \to \infty$. This will prove that (40) follows from (37) and (38).

Since $h(Z) > -\infty$, we can expand the mutual information as

$$\begin{aligned} 0 &\leq I(V;\alpha Z - V|W) = I(V; Z - V/\alpha|W) \stackrel{a}{=} h(Z - V/\alpha|W) - h(Z) \\ &\stackrel{b}{\leq} h(Z - V/\alpha) - h(Z)\ , \end{aligned} \tag{52}$$

where a follows since Z is independent of (V, W), and b follows since conditioning can only reduce the entropy. We show below that

$$\limsup_{\alpha\to\infty} h(Z - V/\alpha) \leq h(Z)\ , \tag{53}$$

which together with (52) implies that $I(V;\alpha Z - V|W) \to 0$ as $\alpha \to \infty$, as desired.

To prove (53), define the Gaussian variables $V^* \sim \mathcal{N}(0, \sigma_v^2)$ and $Z^* \sim \mathcal{N}(0, \sigma_z^2)$, where σ_v^2 and σ_z^2 are the variances of V and Z. We have [2, pp. 234]

$$\mathcal{D}(Z - V/\alpha; Z^* - V^*/\alpha) = \frac{1}{2} \log 2\pi e(\sigma_z^2 + \sigma_v^2/\alpha^2) - h(Z - V/\alpha) \tag{54}$$

where $\mathcal{D}$ denotes divergence. Since $Z - V/\alpha$ and $Z^* - V^*/\alpha$ converge in distribution to Z and Z^*, respectively, we have by the semi-continuity of the divergence (see [5])

$$\liminf_{\alpha\to\infty} \mathcal{D}(Z - V/\alpha; Z^* - V^*/\alpha) \geq \mathcal{D}(Z; Z^*) = \frac{1}{2} \log 2\pi e\sigma_z^2 - h(Z)\ . \tag{55}$$

Combining (54) and (55) proves (53).

Appendix II: Proof of Theorem 3

We first state and prove the following simple lemma:

Lemma 1 *Let $Y = (X + V, W)$, where the pair (V, W) is statistically independent of X. Then $Y - X - Z$ form a Markov chain if and only if the pair (X, Z) is statistically independent of the pair (V, W).*

Proof: Consider the following chain of identities:

$$I(Y; Z|X) \stackrel{a}{=} I(V, W; Z|X) \stackrel{b}{=} I(V, W; Z, X) - I(V, W; X) \stackrel{c}{=} I(V, W; Z, X)\,, \tag{56}$$

where (a) follows by subtracting the condition (operation which does not affect the conditional mutual information), (b) follows from the chain rule, and (c) follows since $I(V, W; X) = 0$. By definition, the term in the left hand side is zero if and only if $Y - X - Z$ form a Markov chain, while the term in the right hand side is zero if and only if the pair (V, W) is independent of the pair (X, Z). □

We are now ready to prove Theorem 3. To get the upper bound in (37), we use (28) and the properties of the additive rate distortion function (i.e., not affected by subtracting the condition, and decreases by further conditioning), to obtain

$$R_{X|Y}^{WZ}(D) \leq R_{X|Y}^{add}(D) = R_{V|(X-V,W)}^{add}(D) \leq R_{V|W}^{add}(D)\,. \tag{57}$$

We turn to prove the lower bound. Since $Y = (X - V, W)$, we can write for any random variable Z

$$I(X; Z|Y) = I(V; Z|X - V, W) = I(V; X - V, Z|W) - I(V; X - V|W)\,, \tag{58}$$

where the first equality follows by subtracting the condition from X (operation which does not change the mutual information), and the second one follows by the chain rule. In view of (58) and Lemma 1, we can write the Wyner Ziv rate distortion function of X given Y (22) as

$$R_{X|Y}^{WZ}(D) = -I(V; X - V|W) + \inf_{\{Z\,:\,(Z,X) \underline{||} (V,W),\, \mathcal{E}_d(X|X-V,Z,W) \leq D\}} I(V; X - V, Z|W)\,. \tag{59}$$

View $(X - V, Z, W)$ as noisy measurements for estimating X and V, and observe that every estimator $\widehat{X}$ with an error $\widehat{X} - X$ induces an estimator $\widehat{V} = \widehat{X} - (X - V)$ having an error $\widehat{V} - V = \widehat{X} - X$, and vice versa. Thus

$$\mathcal{E}_d(V|X - V, Z, W) = \mathcal{E}_d(X|X - V, Z, W)\,. \tag{60}$$

Returning to the minimization in the right hand side of (59), we thus write

$$
\begin{aligned}
\inf_{\{Z\,:\,(Z,X)\,\underline{||}\,(V,W),\ \mathcal{E}_d(X|X-V,Z,W)\le D\}} I(V;X-V,Z|W) &\ge \\
\inf_{\{X,Z\,:\,(X,Z)\,\underline{||}\,(V,W),\ \mathcal{E}_d(V|V-X,Z,W)\le D\}} I(V;V-X,Z|W) &= R^{add}_{V|W}(D)\,,
\end{aligned}
\tag{61}
$$

where the inequality follows since the second minimization is taken over a larger set, and the equality follows by interpreting V as the "source", $-X$ as the "noise", and Z as the "time sharing variable" in the definition (27) of the conditional additive rate distortion function of V given W. The lower bound in Theorem 3 now follows by combining (59) and (61).

Note that as discussed before definition (25), the second minimization in (61) is unchanged if the cardinality of the alphabet of Z is restricted to two; see, e.g., [2, Sec. 14.3].

References

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

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

[3] I. Csiszar and J. Korner. *Information Theory - Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[4] R. M. Gray. A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions. *IEEE Trans. Information Theory*, IT-19:480–489, July 1973.

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

[6] K. Mackenthun. An evaluation of the Wyner-Ziv rate region. *IEEE Trans. Information Theory*, IT-26, No. 3:347–350, 1980.

[7] S. Shamai, S. Verdú, and R. Zamir. Systematic lossy source/channel coding. *in preparation.*

[8] D. Slepian and J.K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Information Theory*, IT-19:471–480, July 1973.

[9] A.D. Wyner. The rate-distortion function for source coding with side information at the decoder - II: General sources. *Information and Control*, 38:60–80, 1978.

[10] A.D. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Information Theory*, IT-22:1–10, Jan., 76.

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

[12] R. Zamir and M. Feder. On lattice quantization noise. *IEEE Trans. Information Theory*, to appear July 1996.

[13] R. Zamir and M. Feder. Information rates of pre/post filtered dithered quantizers. *IEEE Trans. Information Theory*, to appear September 1996.

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