On the Role of Encoder Side-Information in Source Coding for Multiple Decoders

EPFL Technical Report – 2006 [LICOS-REPORT-2006-003]

Etienne Perron[*], Suhas N. Diggavi, İ. Emre Telatar

August 30, 2006

Abstract

We consider a lossy source coding problem where the description of a source is going to be used by two decoders, each having access to information correlated with the source. This side-information is also present at the encoder. We give inner and outer bounds to the set of achievable rate and distortion triples. For the special case of Gaussian sources with degraded side-information and squared error distortions, the two bounds coincide and we obtain the true rate-distortion region. As a further specialization, we obtain the rate-distortion region of the Gaussian version of a problem previously solved by Kaspi for discrete memoryless sources. Using this result, we quantify how much revealing the side-information to the encoder helps in such a Gaussian setup.

1 Introduction

Consider a scenario depicted in Figure 1 where a (discrete memoryless) source X^n needs to be conveyed to two decoders which have access to sequences Y^n and Z^n correlated to the source. This scenario can for instance occur in a

[*]E. Perron was supported in part by SNSF Grant # 200021-105640/1.

file-distribution system, where two clients have different versions of the same file and a server tries to communicate a third version of the file to both clients by sending only one message. If the reconstruction is to be lossless, then by a binning strategy as done in [6] (which does not use encoder side-information) it is enough to send at a rate of $\max\{H(X|Y), H(X|Z)\}$ [5]. Note that encoder side-information cannot improve this rate since we need at least the conditional entropy of the source. In fact for the case of multiple decoders, it is unclear how to use the knowledge of the side-information at the encoder. When we need only approximate reconstructions, the situation might be different in that encoder side-information could make a difference in compression rate. This question is the focus of the investigation in this paper.

In [3], Kaspi studied a related question where one of the side-information sources was absent. His question was motivated by coding for a case where the encoder does not know whether the decoder has access to the side-information or not. For this case, a complete characterization of the rate-distortion function was found in [3] for discrete memoryless sources. In related work, Heegard and Berger studied the same problem as Kaspi when the encoder was uninformed and established the rate-distortion function [2]. This work built upon the seminal work of Wyner and Ziv [8] who had established the rate-distortion function when there is an uninformed encoder and a single decoder with side-information.

In investigating the question illustrated by Figure 1, we pose two questions: (i) how can we use the knowledge of side-information at the encoder? (ii) how much does it help in terms of compression rate? Our first contribution is to provide an achievable rate and therefore an upper bound to the rate-distortion function, as well as a trivial lower bound. The upper bound reduces to the rate-distortion function of the known result by Kaspi [3]. Our achievability result partially answers the questions posed above, by giving a coding scheme that utilizes encoder side-information which improves the compression rate over the uninformed case. Further, we explicitly find the optimal coding strategy for the Kaspi setup when there is a Gaussian source and a jointly Gaussian side-information with the squared-error distortion measure. This explicit characterization allows us to quantify the benefit of encoder side-information for this case. This is in contrast to the Gaussian Wyner-Ziv setup where encoder side-information does not help in terms of compression rate [8]. In addition, this Gaussian Kaspi result is used to show that our achievable rate-distortion region is optimal to the case of jointly

Gaussian sources with squared-error distortion measure and physically degraded side-information.

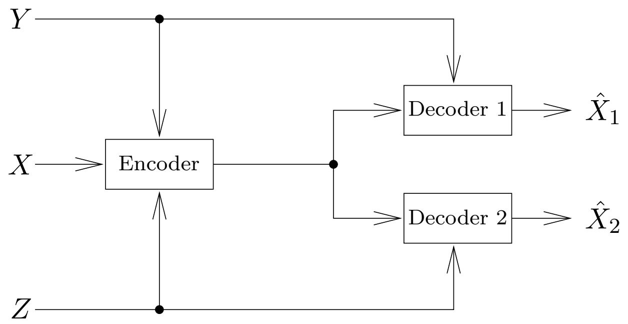

Figure 1: Problem setup studied in the paper.

This paper is organized as follows. In Section 2 we formally define the problem and establish the notation. In Section 3, we state and prove the achievability result. Section 4 contains a trivial outer bound on the region of achievable rate-distortion triples. As a prelude to the Gaussian case with degraded side-information, given in Section 8, Section 7 develops explicit closed-form expressions for the Gaussian Kaspi scenario and provides an example that illustrates the benefit of encoder side-information. Using the results of Section 7, Section 8 gives inner and outer bounds for the Gaussian case with degraded side-information and shows that these bounds match. Finally we conclude with a brief discussion in Section 9.

2 Notation and Problem Statement

We consider a discrete memoryless setup where the source sequence $X_1, X_2, \ldots$, and the two side-information sources $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$ form an i.i.d. sequence of triples $\{(X_i, Y_i, Z_i) : i \geq 1\}$ taking values in the finite set $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ with distribution p_{XYZ}. We are also given two distortion functions to measure the fidelity of the reconstructions by the two decoders: $d_1 : \mathcal{X} \times \hat{\mathcal{X}}_1 \mapsto [0, \infty)$ and $d_2 : \mathcal{X} \times \hat{\mathcal{X}}_2 \mapsto [0, \infty)$, where $\hat{\mathcal{X}}_j$ is the reconstruction alphabet of decoder j. The distortion between two sequences is computed additively; if $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n) \in \hat{\mathcal{X}}_j^n$, we define $d_j(\mathbf{x}, \hat{\mathbf{x}}) = \sum_i d_j(x_i, \hat{x}_i)$.

Definition 1 *An (n, R, D_1, D_2) rate distortion (RD) code for source X with side-information (Y, Z) consists of an encoding function $f : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \mapsto \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ and two decoding functions $g_1 : \{1, \ldots, \lfloor 2^{nR} \rfloor\} \times \mathcal{Y}^n \mapsto \hat{\mathcal{X}}_1^n$, and $g_2 : \{1, \ldots, \lfloor 2^{nR} \rfloor\} \times \mathcal{Z}^n \mapsto \hat{\mathcal{X}}_2^n$, such that*

$$nD_1 \geq \mathbb{E} d_1\big(X^n, g_1(f(X^n, Y^n, Z^n), Y^n)\big),$$
$$nD_2 \geq \mathbb{E} d_2\big(X^n, g_2(f(X^n, Y^n, Z^n), Z^n)\big).$$

Definition 2 *A triple (R, D_1, D_2) is said to be achievable if for every $\epsilon > 0$ there exists an $(n, R + \epsilon, D_1 + \epsilon, D_2 + \epsilon)$ rate distortion code. We denote the set of all achievable triples by $\mathcal{R}$.*

In [3] it has been shown that if Z^n were absent, then (R, D_1, D_2) is achievable if and only if there exists random variables U and W taking values in $\mathcal{U}$ and $\mathcal{W}$, jointly distributed with X, Y, and two functions $\phi_1' : \mathcal{U} \times \mathcal{W} \times \mathcal{Y} \mapsto \hat{\mathcal{X}}_1$, and $\phi_2' : \mathcal{W} \mapsto \hat{\mathcal{X}}_2$ such that

$$R \geq I(X, Y; W) + I(X; U|W, Y), \tag{1a}$$
$$D_1 \geq \mathbb{E} d_1\big(X, \phi_1'(U, W, Y)\big), \tag{1b}$$
$$D_2 \geq \mathbb{E} d_2\big(X, \phi_2'(W)\big). \tag{1c}$$

3 Achievability

For any random variables U, V, W taking values in $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{W}$ respectively jointly distributed with X, Y, Z and any two functions $\phi_1 : \mathcal{U} \times \mathcal{W} \times \mathcal{Y} \mapsto \hat{\mathcal{X}}_1$ and $\phi_2 : \mathcal{V} \times \mathcal{W} \times \mathcal{Z} \mapsto \hat{\mathcal{X}}_2$ define $\tilde{\mathcal{R}}(U, V, W, \phi_1, \phi_2)$ as the set of triples (R, D_1, D_2) satisfying

$$R \geq I(X, Y, Z; W) - \min\{I(Y; W), I(Z; W)\} + I(X; U|W, Y) + I(X; V|W, Z), \tag{2a}$$
$$D_1 \geq \mathbb{E} d_1\big(X, \phi_1(U, W, Y)\big), \tag{2b}$$
$$D_2 \geq \mathbb{E} d_2\big(X, \phi_2(V, W, Z)\big). \tag{2c}$$

Theorem 1 *All triples in $\tilde{\mathcal{R}}(U, V, W, \phi_1, \phi_2)$ are achievable; equivalently*

$$\tilde{\mathcal{R}} \triangleq \bigcup_{\substack{U,V,W,\\ \phi_1,\phi_2}} \tilde{\mathcal{R}}(U, V, W, \phi_1, \phi_2) \subset \mathcal{R}.$$

The region $\tilde{\mathcal{R}}$ is convex. Furthermore, the union above can be restricted to those U, V, W triples for which $|\mathcal{W}| \leq |\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}| + 3$, $|\mathcal{U}| \leq |\mathcal{X} \times \mathcal{Y} \times \mathcal{W}| + 1$ and $|\mathcal{V}| \leq |\mathcal{X} \times \mathcal{Z} \times \mathcal{W}| + 1$.

Proof: The basic idea of the coding strategy is as follows. Since the decoders do *not* know each others' side-information, we cannot generate codebooks conditioned on the pair of side-information sources. Therefore, we establish a connection between the two decoders by designing an auxiliary sequence W^n which is decodable by both the decoders. Having established the 'common' auxiliary sequence, we now use conditional codebooks to give additional private information to each of the decoders in order to satisfy the individual distortion constraints.

For an arbitrary random variable S, we denote by $T_\delta^n(S)$, the sequences of length n that are strongly δ-typical with respect to the distribution of S.

Codebook generation: Given a distribution p_{XYZUVW}, and non-negative numbers A, B, C (to be determined later) draw 2^{nA} sequences $\mathbf{w}(1), \ldots, \mathbf{w}(2^{nA})$, independently with $\mathbf{w}(k) \sim p_W^n$, where p_W^n is the n-fold product distribution. Further, for each $\mathbf{w}(k)$ and $\mathbf{y} \in T_\delta^n(Y)$ draw 2^{nB} sequences $\mathbf{u}(1, k, \mathbf{y}), \ldots, \mathbf{u}(2^{nB}, k, \mathbf{y})$, independently from $p_{U|WY}^n(\cdot|\mathbf{w}(k), \mathbf{y})$. Analoguously, for every $\mathbf{w}(k)$ and $\mathbf{z} \in T_\delta^n(Z)$, generate 2^{nC} sequences $\mathbf{v}(1, k, \mathbf{z}), \ldots, \mathbf{v}(2^{nC}, k, \mathbf{z})$ independently from $p_{V|WZ}^n(\cdot|\mathbf{w}(k), \mathbf{z})$.

Encoder: Given a triple of n-length sequences $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, find k such that $(\mathbf{w}(k), \mathbf{x}, \mathbf{y}, \mathbf{z}) \in T_\delta^n(W, X, Y, Z)$. If there is no such k, then the encoding has failed. Now find i so that $(\mathbf{u}(i, k, \mathbf{y}), \mathbf{x}, \mathbf{y}, \mathbf{w}(k)) \in T_\delta^n(U, X, Y, W)$. Again, if no such i can be found, the encoding has failed. Similarly, find j so that $(\mathbf{v}(j, k, \mathbf{z}), \mathbf{x}, \mathbf{z}, \mathbf{w}(k)) \in T_\delta^n(V, X, Z, W)$. The message sent by the decoder consists of the triple $(k \mod 2^{nR_0}, i, j)$ that can be described in $nR = n(R_0 + B + C)$ bits. (Note that the encoder "bins" the index k by the modulo operation).

Decoders: We will only describe Decoder 1, *i.e.,* the decoder with side-information $\mathbf{y}$, and Decoder 2 is defined analogously. From the first part of the message the decoder knows $k \bmod 2^{nR_0}$. Among the $2^{n(A-R_0)}$ values of k' that are consistent with this information, the decoder finds a k' such that $(\mathbf{w}(k'), \mathbf{y}) \in T_\delta^n(W, Y)$. If there are several such sequences, then the decoding fails. From the second part of the message, the decoder finds $\mathbf{u}(i, k', \mathbf{y})$ and constructs $\hat{\mathbf{x}}_1$ by applying the function ϕ_1 component-wise to the triple $(\mathbf{u}(i, k', \mathbf{y}), \mathbf{w}(k'), \mathbf{y})$.

Analysis: Standard typicality arguments show that as long as $A > I(X,Y,Z;W)$ the encoder will succeed in finding a k with high probability, further, as long as $B > I(U;X|W,Y)$ it will find an i with high probability and as long as $C > I(V;X|W,Z)$ it will find a j with high probability. On the side of Decoder 1, the index k' will equal k with high probability as long as $A - R_0 < I(W;Y)$, similarly, Decoder 2 will find the index k correctly with high probability if $A - R_0 < I(W;Z)$. Note that since the encoder has ensured joint typicality of $(\mathbf{w}(k), \mathbf{x}, \mathbf{y}, \mathbf{z})$ by using knowledge of the side-information, we do not need the Markov condition (or lemma) as done in [8]. Consequently, provided that $R > I(X,Y,Z;W) - \min\{I(W;Y), I(W;Z)\} + I(U;X|W,Y) + I(V;X|W,Z)$ numbers A, B, C, R_0 can be found that satisfy all these inequalities and thus the encoder and the decoders will operate without failure with high probability.

But it is clear that whenever $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is typical and no encoding or decoding failures occur, then the sequences $(\mathbf{x}, \hat{\mathbf{x}}_1)$ and $(\mathbf{x}, \hat{\mathbf{x}}_2)$ are typical and thus

$$\begin{aligned} d_1(\mathbf{x}, \hat{\mathbf{x}}_1) \le n(E[d_1(X, \phi_1(W,U,Y))] + \delta) \\ d_2(\mathbf{x}, \hat{\mathbf{x}}_2) \le n(E[d_2(X, \phi_2(W,V,Z))] + \delta). \end{aligned}$$

Convexity: Suppose that

$$(R^{(1)}, D_1^{(1)}, D_2^{(1)}) \in \tilde{\mathcal{R}}$$

and

$$(R^{(2)}, D_1^{(2)}, D_2^{(2)}) \in \tilde{\mathcal{R}}.$$

We know that $\exists (U^{(j)}, V^{(j)}, W^{(j)}), \phi_1^{(j)}(\cdot)\phi_2^{(j)}(\cdot)$ for $j \in \{1,2\}$ s.t.

$$\begin{aligned} R^{(j)} \ &\ge\ I(X,Y,Z;W^{(j)}) - \min\{I(Y;W^{(j)}), I(Z;W^{(j)})\} \\ &\qquad + I(X;U^{(j)}|W^{(j)},Y) + I(X;V^{(j)}|W^{(j)},Z), \\ D_1^{(j)} \ &\ge\ \mathbb{E}\left[d_1(X, \phi_1^{(j)}(U^{(j)}, W^{(j)}, Y))\right], \\ D_2^{(j)} \ &\ge\ \mathbb{E}\left[d_2(X, \phi_2^{(j)}(V^{(j)}, W^{(j)}, Z))\right] \Big\}, \end{aligned}$$

for $j \in \{1,2\}$. Let $\lambda \in [0,1]$. We use the notation $\bar{\lambda} \triangleq 1 - \lambda$. We need to show that

$$(\lambda R^{(1)} + \bar{\lambda} R^{(2)}, \lambda D_1^{(1)} + \bar{\lambda} D_1^{(2)}, \lambda D_2^{(1)} + \bar{\lambda} D_2^{(2)}) \in \tilde{\mathcal{R}}.$$

To do so, we define the random variables (U, V, W, T) as follows:

$$T = \begin{cases} 1 & \text{w.p. } \lambda \\ 2 & \text{w.p. } \bar{\lambda} \end{cases}$$

and T is set to be independent of (XYZ). Further, given that T takes value j, (UVW) is distributed (jointly with (XYZ)) exactly as $(U^{(j)}V^{(j)}W^{(j)}$, for $j \in \{1, 2\}$. We also define

$$\phi_1(U, W, Y, T) = \phi_1^{(T)}(U^{(T)}, W^{(T)}, Y).$$

and

$$\phi_2(U, W, Y, T) = \phi_2^{(T)}(U^{(T)}, W^{(T)}, Y).$$

Now, we further define

$$W' = (W, T)$$

and

$$\phi_i'(U, W', Y) = \phi_i(U, W, Y, T),$$

for $i \in \{1, 2\}$. We have

$$\begin{aligned} \mathbb{E}\left[d_1\big(X, \phi_1'(U, W', Y)\big)\right] &= \mathbb{E}\left[d_1\big(X, \phi_1(U, W, Y, T)\big)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[d_1\big(X, \phi_1(U, W, Y, T)\big)|T\right]\right] \\ &= \lambda \mathbb{E}\left[d_1\big(X, \phi_1^{(1)}(U^{(1)}, W^{(1)}, Y)\big)\right] \\ &\qquad + \bar{\lambda} \mathbb{E}\left[d_1\big(X, \phi_1^{(2)}(U^{(2)}, W^{(2)}, Y)\big)\right] \\ &\leq \lambda D_1^{(1)} + \bar{\lambda} D_1^{(2)}. \end{aligned}$$

Similarly,

$$\mathbb{E}\left[d_2\big(X, \phi_2'(V, W', Y)\big)\right] \leq \lambda D_2^{(1)} + \bar{\lambda} D_2^{(2)}.$$

Finally,

$$\begin{aligned}
\lambda R^{(1)} + \bar{\lambda} R^{(2)} &\geq \lambda I(X,Y,Z;W^{(1)}) - \lambda \min\{I(Y;W^{(1)}), I(Z;W^{(1)})\} \\
&\quad + \lambda I(X;U^{(1)}|W^{(1)},Y) + \lambda I(X;V^{(1)}|W^{(1)},Z) \\
&\quad + \bar{\lambda} I(X,Y,Z;W^{(2)}) - \bar{\lambda} \min\{I(Y;W^{(2)}), I(Z;W^{(2)})\} \\
&\quad + \bar{\lambda} I(X;U^{(2)}|W^{(2)},Y) + \bar{\lambda} I(X;V^{(2)}|W^{(2)},Z) \\
&\overset{(a)}{\geq} \lambda I(X,Y,Z;W^{(1)}) + \bar{\lambda} I(X,Y,Z;W^{(2)}) \\
&\quad - \min\{\lambda I(Y;W^{(1)}) + \bar{\lambda} I(Y;W^{(2)}), \\
&\qquad\quad \lambda I(Z;W^{(1)}) + \bar{\lambda} I(Z;W^{(2)})\} \\
&\quad + \lambda I(X;U^{(1)}|W^{(1)},Y) + \bar{\lambda} I(X;U^{(2)}|W^{(2)},Y) \\
&\quad + \lambda I(X;V^{(1)}|W^{(1)},Z) \bar{\lambda} I(X;V^{(2)}|W^{(2)},Z) \\
&= I(X,Y,Z;W|T) - \min\{I(Y;W|T), I(Z;W|T)\} \\
&\quad + I(X;U|W,Y,T) + I(X;V|W,Z,T) \\
&\overset{(b)}{=} I(X,Y,Z;W') - \min\{I(Y;W'), I(Z;W')\} \\
&\quad + I(X;U|W',Y) + I(X;V|W',Z),
\end{aligned}$$

where (a) is the inequality

$$\min\{x_1 + x_2, y_1 + y_2\} \geq \min\{x_1, y_1\} + \min\{x_2, y_2\}.$$

In (b), we used the fact that T is independent of (XYZ). From the shown inequalities, it follows that

$$(\lambda R^{(1)} + \bar{\lambda} R^{(2)}, \lambda D_1^{(1)} + \bar{\lambda} D_1^{(2)}, \lambda D_2^{(1)} + \bar{\lambda} D_2^{(2)}) \in \tilde{\mathcal{R}}.$$

Cardinality Bounds: Consider the support lemma given in . [1, pp. 310]:

Lemma 2 *Let $\mathcal{P}(\mathbb{X})$ be the family of all probability distributions on the set $\mathbb{X}$, and let f_j, $j = 1, \ldots, k$ be real-valued, continuous functions on $\mathcal{P}(\mathbb{X})$. Then to any probability measure μ on the Borel σ-algebra of $\mathcal{P}(\mathbb{X})$ there exist k elements P_i of $\mathcal{P}(\mathbb{X})$ and non-negative numbers $\alpha_1, \ldots, \alpha_k$ with $\sum_{i=1}^k \alpha_i = 1$ such that for every $j = 1, \ldots, k$*

$$\int_{\mathcal{P}(\mathbb{X})} f_j(P) \mu(\mathrm{d}P) = \sum_{i=1}^{k} \alpha_i f_j(P_i). \tag{3}$$

The support lemma has the following interpretation: Assume that μ is assigning positive probability to a finite subset of $\mathcal{P}(\mathbb{X})$ only, *i.e.*, μ is a probability mass function (PMF). Assume that we are seeking a PMF μ^* on $\mathcal{P}(\mathbb{X})$ that assigns positive probability to less elements of $\mathcal{P}(\mathbb{X})$ (whose support in $\mathcal{P}(\mathbb{X})$ has smaller cardinality) than μ. However, we require μ^* to satisfy the following: For each of the k functions f_j, $j = 1, \ldots, k$, the expectation of $f_j(P)$ ($P \in \mathcal{P}(\mathbb{X})$) under μ^* has to be equals to the expectation of $f_j(P)$ under μ. In the lemma, $\mu^*(P_i) = \alpha_i$, and hence, the lemma says that we can find a μ^* that acts on a set of cardinality of size k.

Remark:

The proof of the support lemma is very simple in the case of μ being a PMF on a finite set. Let m be the size of the support set of μ and let $(\tilde{P}_1, \ldots, \tilde{P}_m) \subset \mathcal{P}(\mathbb{X})$ be this support set. Then (3) can be written as

$$\begin{bmatrix} f_1(\tilde{P}_1) \\ \vdots \\ f_k(\tilde{P}_1) \end{bmatrix} \mu(\tilde{P}_1) + \ldots + \begin{bmatrix} f_1(\tilde{P}_m) \\ \vdots \\ f_k(\tilde{P}_m) \end{bmatrix} \mu(\tilde{P}_m) = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix},$$

where $c_j = \int_{\mathcal{P}(\mathbb{X})} f_j(P)\mu(\mathrm{d}P) = \sum_{l=1}^{m} f_j(\tilde{P}_l)\mu(\tilde{P}_l)$.

Hence, we see that $[c_1, \ldots, c_k]^T$ is a vector in $\mathbb{R}^k$, that can be written as a convex combination of m vectors $[f_1(\tilde{P}_l), \ldots, f_k(\tilde{P}_l)]^T$, for $l = 1, \ldots, m$. Note that the f_j are continuous functions and that $\mathcal{P}(\mathbb{X})$ is a connected set. Hence, the set of all vectors

$$\{[f_1(P), \ldots, f_k(P)]^T : P \in \mathcal{P}(\mathbb{X})\}$$

is a connected subset of $\mathbb{R}^k$. To summarize, $[c_1, \ldots, c_k]^T$ is a vector in $\mathbb{R}^k$ that is spanned by m ($m > k$) vectors, all belonging to the same connected subset of $\mathbb{R}^k$. Hence, one can find k different vectors in the same subset of $\mathbb{R}^k$ that also span $[c_1, \ldots, c_k]^T$. We write this new convex combination as

$$\begin{bmatrix} f_1(P_1) \\ \vdots \\ f_k(P_1) \end{bmatrix} \mu^*(P_1) + \ldots + \begin{bmatrix} f_1(P_k) \\ \vdots \\ f_k(P_k) \end{bmatrix} \mu^*(P_k) = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix}.$$

End of Remark

Let us now use the support lemma to show the bounds on the cardinality of $\mathcal{W}$, $\mathcal{U}$ and $\mathcal{V}$. To this ends, assume that random variables (U, V, W) are given to us. We will replace them by different random variables (U', V', W') such that problem parameters and the performance of the solution remains unchanged, and such that $|\mathcal{W}| \leq |\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}| + 3$, $|\mathcal{U}| \leq |\mathcal{X} \times \mathcal{Y} \times \mathcal{W}| + 1$ and $|\mathcal{V}| \leq |\mathcal{X} \times \mathcal{Z} \times \mathcal{W}| + 1$. Let $\mathbb{X} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ and let $p_{XYZ|W}(\cdot|w)$ be an element of $\mathcal{P}(\mathbb{X})$ having μ-measure $p_W(w)$, for $w \in \mathcal{W}$. In words of the above interpretation, $m = |\mathcal{W}|$, $\tilde{P}_w = p_{XYZ|W}(\cdot|w)$, and $\mu(\tilde{P}_w) = p_W(w)$. We would like to find a new measure μ^* with a different support set $(P_1, \dots, P_k)$ of size $k < m$, leading to a new random variable W' having alphabet-size k and such that $p_{W'}(w) = \mu(P)$ for $j \in \mathcal{W}' = \{1, \dots, k\}$. In addition, the distribution of W' should be such that the rate, the two distortions as well as the distribution of (X, Y, Z) does not change. To this ends, we carefully define k functions on $\mathcal{P}(\mathbb{X})$:

- $f_{xyz}(P) = P(x, y, z)$ for all $(x, y, z) \in \mathbb{X}$ except for one. Note that

$$\int_{\mathcal{P}(\mathbb{X})} f_{xyz}(P)\mu(\mathrm{d}P) = \sum_{w \in \mathcal{W}} p_{XYZ|W}(x, y, z|w)p_W(w) = p_{XYZ}(x, y, z),$$

  and hence, preserving these values for all $(x, y, z) \in \mathbb{X}$ except for one will preserve the distribution of (X, Y, Z).

- To preserve the rate, we first consider the expressions

$$A = I(X, Y, Z; W) - I(Y; W) + I(X; U|W, Y) + I(X; V|W, Z)$$

  and

$$B = I(X, Y, Z; W) - I(Z; W) + I(X; U|W, Y) + I(X; V|W, Z).$$

  It is clear that by preserving A and B, we preserve the rate $R = \max\{A, B\}$. A can be expanded as

$$\begin{aligned} A &= I(X, Y, Z; W) - I(Y; W) + I(X; U|W, Y) + I(X; V|W, Z) \\ &= H(X, Y, Z) - H(X, Y, Z|W) - H(Y) + H(Y|W) + H(X|W, Y) \\ &\quad -H(X|U, W, Y) + H(X|W, Z) - H(X|V, W, Z). \end{aligned}$$

  Note that preservation of (X, Y, Z) will preserve the terms $H(X, Y, Z)$ and $H(Y)$. Hence, we are only concerned with

$$\begin{aligned} \tilde{A} &= -H(X, Y, Z|W) + H(Y|W) + H(X|W, Y) - H(X|U, W, Y) \\ &\quad +H(X|W, Z) - H(X|V, W, Z). \end{aligned}$$

The first term of $\tilde{A}$ can be written as

$$H(X,Y,Z|W) = \sum_{w\in\mathcal{W}} p_W(w) h\big(p_{XYZ|W}(\cdot|w)\big),$$

where $h(p_S(\cdot)) = -\sum_{s\in\mathcal{S}} p_S(s)\log(p_S(s))$.
The second term of $\tilde{A}$ can be written as

$$H(Y|W) = \sum_{w\in\mathcal{W}} p_W(w) h\big(\sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} p_{XYZ|W}(x,\cdot,z|w)\big).$$

The fourth term of $\tilde{A}$ can be written as

$$\begin{aligned} H(X|U,W,Y) &= \sum_{w\in\mathcal{W}} p_W(w) \sum_{u\in\mathcal{U}}\sum_{y\in\mathcal{Y}} p_{Y|W}(y|w) p_{U|WY}(u|w,y) \\ &\qquad\qquad h\big(p_{X|UWY}(\cdot|u,w,y)\big), \end{aligned}$$

where it is clear that $p_{Y|W}(y|w)$, $p_{U|WY}(u|w,y)$ and $p_{X|UWY}(\cdot|u,w,y)$ are continuous functions of $p_{XYZ|W}(x,y,z|w)$.
All remaining terms of $\tilde{A}$ decompose in a similar fashion.
Hence, we see that $\tilde{A}$ can be written as

$$\begin{aligned} \tilde{A} &= \sum_{w\in\mathcal{W}} p_W(w) f_A(p_{XYZ|W}(x,y,z|w) \\ &= \int_{\mathcal{P}(\mathbb{X})} \mu(P) f_A(P). \end{aligned}$$

Similarly, stripping already preserved terms from B yields a certain $\tilde{B}$, that can be written as

$$\begin{aligned} \tilde{B} &= \sum_{w\in\mathcal{W}} p_W(w) f_B(p_{XYZ|W}(x,y,z|w) \\ &= \int_{\mathcal{P}(\mathbb{X})} \mu(P) f_B(P). \end{aligned}$$

As long as $\tilde{A}$ and $\tilde{B}$ are preserved, the rate is preserved.

- Finally, note that

$$\begin{aligned} \mathbb{E}\left[d_1(X,\phi_1(U,W,Y))\right] &= \sum_{w\in\mathcal{W}} p_W(w) \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\sum_{u\in\mathcal{U}} p_{XYZ|W}(x,y,z|w) \\ &\quad p_{U|XYZW}(u|x,y,z,w) d_1(x,\phi_1(u,w,y)) \\ &= \sum_{w\in\mathcal{W}} p_W(w) f_{D_1}(p_{XYZ|W}(x,y,z|w)). \end{aligned}$$

Similarly, we can exhibit a function $f_{D_2}(P)$ such that

$$\mathbb{E}\left[d_2(X, \phi_2(V, W, Z))\right] = \sum_{w \in \mathcal{W}} p_W(w) f_{D_2}(p_{XYZ|W}(x, y, z|w)).$$

We therefore see that to successfully replace W by W', we need to preserve the expectation over W of $k = |\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}| - 1 + 4$ functions. Thus, by the support lemma, it is possible to find a suitable random variable W' whose alphabet-size is no larger than $|\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}| + 3$.

Note that due to replacing W by W', the distribution of U and V might have been modified. However, this is not a problem, because we preserved all the relevant quantities.

Next, we want to replace U by a new random variable U' with a smaller alphabet-size. First, note that in the rate expression as well as in the distortion, U appears together with (X, Y, W), but never with V or Z. Hence, we can replace U by a new variable $\tilde{U}$ such that $p_{XYZ\tilde{U}VW} = p_{XYZVW} p_{U|XYW}$, *i.e.*, $\tilde{U} \multimap (X, Y, W) \multimap (Z, V)$ forms a Markov chain, while the marginal $p_{\tilde{U}XYW} = p_{UXYW}$. Now, we set $\mathbb{X} = \mathcal{X} \times \mathcal{Y} \times \mathcal{W}$, and let $p_{XYW|\tilde{U}}(\cdot|u)$ be an element of $\mathcal{P}(\mathbb{X})$ having μ-measure $p_{\tilde{U}}(u)$. As previously, we define $|\mathcal{X} \times \mathcal{Y} \times \mathcal{W}| - 1$ functions on $\mathcal{P}(\mathbb{X})$ to preserve p_{XYW}. Thanks to the Markov-structure of $\tilde{U}$, preserving p_{XYW} also preserves p_{XYZVW}. In addition, we define a function to preserve the rate. Since $\tilde{U}$ affects only the first distortion, we only require one more function on $\mathcal{P}(\mathbb{X})$ to preserve the distortions. Again, the Markov lemma tells us that we can construct a random variable U' with alphabet-size no larger than $|\mathcal{X} \times \mathcal{Y} \times \mathcal{W}| - 1 + 2$. The proof for the alphabet-size of V is analoguous to the proof for U. $\square$

In the proof of the achievability we generate enough $\mathbf{w}$ sequences so as to guarantee that the encoder will find (with high probability) one for which $(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z})$ is typical. The reader may have observed that since the side-information sources are separately present at the decoders, this joint typicality of the quadruple is not required: it suffices to guarantee instead that both $(\mathbf{w}, \mathbf{x}, \mathbf{y})$ and $(\mathbf{w}, \mathbf{x}, \mathbf{z})$ are typical. This suggests that one can enlarge the achievability region by generating fewer $\mathbf{w}$ sequences to only make sure that the encoder finds with high probability one that satisfies this weaker requirement. Indeed we have the following

Lemma 3 *The size of the set*

$$A_\delta^n = \big\{(\mathbf{x},\mathbf{y},\mathbf{z}) : (\mathbf{x},\mathbf{y}) \in T_\delta(X,Y), \\ (\mathbf{x},\mathbf{z}) \in T_\delta(X,Z), (\mathbf{y},\mathbf{z}) \in T_\delta(Y,Z)\big\}$$

satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log |A_\delta^n| = \max H(\tilde{X}, Y, Z) \tag{4}$$

where the maximization is taken over all $\tilde{X}$ jointly distributed with (Y,Z) such that $(\tilde{X},Y) \sim (X,Y)$ and $(\tilde{X},Z) \sim (X,Z)$. (The notation $A \sim B$ means that A and B have the same probability distribution.)

Using the lemma with W in place of X, (X,Y) in place of Y and (X,Z) in place of Z, we see that we can replace the requirement $A > I(W;X,Y,Z)$ in the proof of Theorem 1 with $A > I_0(W;X,Y,Z)$, where

$$I_0(W;X,Y,Z) = \min_{\tilde{W}} I(\tilde{W};X,Y,Z) \tag{5}$$

and the minimization is over all $\tilde{W}$ jointly distributed with X,Y,Z such that $(\tilde{W},X,Y) \sim (W,X,Y)$ and $(\tilde{W},X,Z) \sim (W,X,Z)$, and prove the achievability of the region $\tilde{\mathcal{R}}_0(U,V,W,\phi_1,\phi_2)$ that differs from $\tilde{\mathcal{R}}(U,V,W,\phi_1,\phi_2)$ in only changing $I(W;X,Y,Z)$ to $I_0(W;X,Y,Z)$ in (2a). Since $I_0 \leq I$, $R_0(U,V,W,\phi_1,\phi_2) \supset \mathcal{R}(U,V,W,\phi_1,\phi_2)$. Unfortunately, this enlargement does not improve the union region $\tilde{\mathcal{R}}$ because

$$\tilde{\mathcal{R}}_0(U,V,W,\phi_1,\phi_2) = \tilde{\mathcal{R}}(U,V,W^*,\phi_1,\phi_2)$$

where W^* is the minimizer of (5). The reason for this equality is that replacing a given W by W^* in $\tilde{\mathcal{R}}(U,V,W,\phi_1,\phi_2)$ does not change any of the terms except for the first one, which becomes $I_0(X,Y,Z;W)$.

Note that if we consider a super source $(\mathbf{x},\mathbf{y},\mathbf{z})$ and define appropriate distortion measures, our problem reduces to the general question with uninformed encoder and decoders with side-information $\mathbf{y}$ and $\mathbf{z}$ respectively. Therefore, this means that having encoder side-information is a special case of the general problem studied in [2]. However, since that general problem is unsolved, this observation does not help us solve our problem. Note that even if we had degraded side-information, *i.e.*, $X \multimap Y \multimap Z$, the super-source construction implies that the new source with uninformed encoder would not have such a structure. The super-source argument also provides an alternate proof of Theorem 1.

4 Outer Bound

Assume that the side-information sequence Z^n is not only available at the encoder and at Decoder 2, but also at Decoder 1. The rate-distortion region of this setup obviously contains $\mathcal{R}$. One can think of this problem as having new sources $\tilde{X}, \tilde{Y}$ with distribution $p_{XY|Z}$, where $\tilde{X}$ is the source of interest and $\tilde{Y}$ is side-information at Decoder 1 (and at the encoder). This new setup can be thought of as a conditional version of the rate-distortion region established by Kaspi in [3] (conditioned on Z).

More precisely, for any random variables U, W taking values in $\mathcal{U}, \mathcal{W}$, jointly distributed with X, Y, Z, and any two functions $\phi_1' : \mathcal{U} \times \mathcal{W} \times \mathcal{Y} \times \mathcal{Z} \mapsto \hat{\mathcal{X}}_1$, and $\phi_2' : \mathcal{W} \times \mathcal{Z} \mapsto \hat{\mathcal{X}}_2$, define $\breve{\mathcal{R}}(U, W, \phi_1', \phi_2')$ as the set of triples (R, D_1, D_2) that satisfy

$$R \geq I(X, Y; W|Z) + I(X; U|W, Y, Z), \tag{6a}$$

$$D_1 \geq \mathbb{E}d_1\big(X, \phi_1'(U, W, Y, Z)\big), \tag{6b}$$

$$D_2 \geq \mathbb{E}d_2\big(X, \phi_2'(W, Z)\big). \tag{6c}$$

We have that

$$\breve{\mathcal{R}} \triangleq \bigcup_{\substack{U,W,\\ \phi_1',\phi_2'}} \breve{\mathcal{R}}(U, W, \phi_1', \phi_2') \supset \mathcal{R}.$$

5 Generalization of the Achievability

In this subsection, we present a generalization of our problem. Consider the same setup as in Figure 1, but with the side-information being available at the decoders only. The encoder has no access to Y and Z. In [2], Heegard and Berger study this setup, and find the following upper bound on the rate-distortion function:

$$\begin{aligned} R_{\mathrm{HB}}(D_1, D_2) \quad = \quad & \min_{P'(D_1,D_2)} \Big[\max\big\{I(X; W|Z), I(X; W|Y)\big\} \\ & \qquad + I(X; U|WY) + I(X; V|WZ)\Big], \end{aligned} \tag{7}$$

where $P'(D_1, D_2)$ is defined to be the set of all discrete random variables (U, V, W) jointly distributed with (X, Y, Z) such that:

- $(Y, Z) \multimap X \multimap (U, V, W)$.

- There exist fuctions $\hat{X}_1(U,W,Y)$ and $\hat{X}_2(V,W,Z)$ such that $\mathbb{E}\left[d_i(X,\hat{X}_i)\right] \leq D_i$, for $i \in \{1,2\}$.

To see how our problem is a special case of this setup, we apply a "super-source" argument. Consider a setup with no encoder side-information. Let (X,Y,Z) be the source of interest and let Y and Z be the side-information sources. Let the distortion measures be such that

$$d_i\big(\hat{x}_i,(x,y,z)\big) = d_i\big(\hat{x}_i,(x,y',z')\big)$$

for all $y,y' \in \mathcal{Y}$ and $z,z' \in \mathcal{Z}$, for $i \in \{1,2\}$. This "super-source" problem is the same as our original problem with encoder side-information. The achievable rate region $\tilde{\mathcal{R}}$ can be recovered using the Heegard-Berger achievable rate as follows. Replacing X by the "super-source" (X,Y,Z) in (7) yields

$$\begin{aligned} R_{\text{HB}}(D_1,D_2) = \min_{P'(D_1,D_2)} \Big[& \max\big\{I(XYZ;W|Z), I(XYZ;W|Y)\big\} \\ & + I(XYZ;U|WY) + I(XYZ;V|WZ)\Big]. \end{aligned}$$

First note that $(Y,Z) \multimap (X,Y,Z) \multimap (U,V,W)$ form a Markov chain for any (U,V,W), and the Markov-requirement can be dropped. The first term of $R_{\text{HB}}(D_1,D_2)$ can then be written as

$$\max\big\{I(XY;W|Z), I(XZ;W|Y)\big\}.$$

The second term becomes

$$\begin{aligned} I(XYZ;U|WY) &= I(XZ;U|WY) \\ &= I(X;U|WY) + I(Z;U|WXY) \\ &\geq I(X;U|WY), \end{aligned}$$

with equality if $U \multimap (X,Y,W) \multimap Z$. Note that distortion D_2 does not depend on U. Distortion D_1 is a function of X, U, W and Y, but not of Z. Therefore, imposing the Markov condition $U \multimap (X,Y,W) \multimap Z$ does not modify the distortions, nor does it alter any term of the rate other than $I(XZ;U|WY)$. Therefore, we can safely restrict our minimization to random variables U such that $U \multimap (X,Y,W) \multimap Z$. An analoguous argument holds for the random variable V in the last rate term, and we find that

$R_{\mathrm{HB}}(D_1, D_2) = \tilde{R}(D_1, D_2)$, which denotes the rate-distortion function associated with $\tilde{\mathcal{R}}$.

Heegard and Berger are able to prove optimality of R_{HB} only for degraded side-informations $X \multimap Y \multimap Z$. The "super-source" setup used in this section does not satisfy this condition.

6 Special Cases

6.1 Sgarro

If one requires the representation of X to be lossless, our setup reduces to the problem described in [5] by Sgarro. He shows that the smallest rate to perfectly describe X to both decoders is

$$R^* = \max\{H(X|Y), H(X|Z)\}.$$

Note that Sgarro considers the case where the side-information sources are not available at the encoder. It is, however, obvious that encoder side-information can not improve the above rate without violating the lower bound given by conditional lossless encoding of a source for a single decoder. One can verify that our achievable rate reduces to the Sgarro rate by setting $W = X$, $U = V = \text{const}$, $\hat{X}_1(U, W, Y) = \hat{X}_2(V, W, Z) = W$.

6.2 Gaussian Wyner-Ziv

Assume the following:

- The sources are Gaussian. More precisely, $X \sim \mathcal{N}(0, \sigma_X^2)$, $Y = X + N$ and $Z = X + N'$, where $N \sim \mathcal{N}(0, \sigma_N^2)$ and $N' \sim \mathcal{N}(0, \sigma_{N'}^2)$.
- The distortion measure is the squared error.
- The pair (X, Y) has the same distribution as (X, Z), i.e., $\sigma_N^2 = \sigma_{N'}^2$.
- $D_1 \le D_2$.

In this case, the rate-distortion function becomes

$$R(D_1, D_2) = \log \frac{\mathrm{Var}\,(X|Y)}{D_1}, \tag{8}$$

where $\mathrm{Var}\,(X|Y) = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}$. We know from [8] that the right hand side of (8) is the rate-distortion function for describing the given Gaussian source to a single decoder using side-information. Therefore, $R(D_1, D_2)$ can be no smaller than the expression in (8). To see how this rate is achievable, note that Wyner and Ziv show in [8] that for Gaussian sources and squared error distortion, the side-information can be ignored at the encoder. Let the encoder ignore Y and Z, and use the binning scheme proposed by Wyner and Ziv to describe X to Decoder 1. The side-information Z available at Decoder 2 is statistically equivalent to Y, and thus, Decoder 2 can use the message produced by the encoder to represent X with a distortion no larger than D_1.

We will now show that under the assumptions made in this section, the achievable rate $\tilde{R}(D_1, D_2)$ reduces to (8). Define W' to be Gaussian, zero-mean and such that

$$(X, Y, Z, W') \sim \mathcal{N}(\mathbf{0}, K),$$

with

$$K = \begin{bmatrix} \sigma_X^2 & \sigma_X^2 & \sigma_X^2 & \tau\sigma_X\sigma_{W'} \\ \sigma_X^2 & \sigma_X^2 + \sigma_N^2 & \sigma_X^2 + \xi\sigma_N^2 & \tau\sigma_X\sigma_{W'} \\ \sigma_X^2 & \sigma_X^2 + \xi\sigma_N^2 & \sigma_X^2 + \sigma_N^2 & \tau\sigma_X\sigma_{W'} \\ \tau\sigma_X\sigma_{W'} & \tau\sigma_X\sigma_{W'} & \tau\sigma_X\sigma_{W'} & \sigma_{W'}^2 \end{bmatrix}.$$

In the above matrix, the correlation of W' with Y has been chosen to be the same as with Z. Additionally, W' has been chosen to be independent of N. Finally, we set the distribution of W' to be such that $\mathrm{Var}\,(X|ZW') = D_1$. From (2) we know that

$$\begin{aligned}
\tilde{R}(D_1, D_2) &= \min_{P(D_1,D_2)} \Big[\max\big\{ I(XY;W|Z), I(XZ;W|Y) \big\} \\
&\qquad\qquad + I(X;U|WY) + I(X;V|WZ) \Big] \\
&\overset{(a)}{\leq} I(XY;W'|Z) \\
&= h(XYZ) - h(Z) - h(XYW'Z) + h(W'Z) \\
&= -h(Z) - h(W'|XYZ) + h(W'Z) \\
&\overset{(b)}{=} -h(Z) - h(W'|X) + h(W'Z) \\
&= h(W'Z) - h(W'X) + h(X) - h(Z) \\
&= \frac{1}{2}\log \frac{\sigma_N^2\sigma_{W'}^2 + \sigma_X^2\sigma_{W'}^2(1-\tau^2)}{\sigma_X^2\sigma_{W'}^2(1-\tau^2)} + \frac{1}{2}\log\frac{\sigma_X^2}{\sigma_X^2+\sigma_N^2}.
\end{aligned}$$

In inequality (a), we set $U = V = \text{const}$ and $W = W'$. In (b), we use the fact that $W' \perp N$, which implies that $W \multimap X \multimap (YZ)$ is a Markov chain. On the other hand, the rate in (8) is

$$\begin{aligned} R(D_1, D_2) &= \frac{1}{2} \log \frac{\sigma_X^2 \sigma_N^2}{(\sigma_X^2 + \sigma_N^2) D_1} \\ &= \frac{1}{2} \log \frac{\sigma_N^2}{D_1} + \frac{1}{2} \log \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2}. \end{aligned}$$

We have picked W' such that

$$\begin{aligned} D_1 &= \operatorname{Var}(X|ZW') \\ &= \sigma_X^2 - \mathbb{E}\left[X \begin{bmatrix} Z & W' \end{bmatrix}\right] K_{ZW'}^{-1} \mathbb{E}\left[X \begin{bmatrix} Z \\ W' \end{bmatrix}\right] \\ &= \frac{\sigma_N^2 \sigma_X^2 \sigma_{W'}^2 (1 - \tau^2)}{\sigma_N^2 \sigma_{W'}^2 + \sigma_X^2 \sigma_{W'}^2 (1 - \tau^2)}. \end{aligned}$$

Hence, $\tilde{R}(D_1, D_2) \leq R(D_1, D_2)$. This together with the optimality of $R(D_1, D_2)$ implies that $R(D_1, D_2) = \tilde{R}(D_1, D_2)$.

6.3 Kaspi

In [3], Kaspi solves a special case of our problem. He considers a setup where only one of the decoders has access to side-information. The second decoder bases its reconstruction of X solely on the message that the encoder outputs. In our setup, this corresponds to the second side-information source Z being constant.

Kaspi's region is described in equations (1). Observe that by setting $V = \text{constant}$ in equations (2) we recover Kaspi's equations, showing that in this setting the region $\tilde{\mathcal{R}}$ is tight.

It turns out that one does not shrink $\tilde{\mathcal{R}}$ by taking $\hat{X}_1 = \phi_1'(U, W, Y) = U$, $\hat{X}_2 = \phi_2'(W) = W$, showing that rate distortion region is given by the union over $\hat{X}_1, \hat{X}_2$ of regions satisfying

$$\begin{aligned} R &\geq I(XY; \hat{X}_2) + I(X; \hat{X}_1 | \hat{X}_2 Y), \\ D_1 &\geq \mathbb{E}[d_1(X, \hat{X}_1)], \\ D_2 &\geq \mathbb{E}[d_2(X, \hat{X}_2)]. \end{aligned}$$

7 The Gaussian Kaspi Problem

As discussed the previous Section, in [3], Kaspi gives a complete solution for a special case of our problem where only one of the decoders has access to side-information, i.e., the case when Z is constant.

In this section we give an explicit characterization of Kaspi's rate-distortion function when we have a jointly Gaussian source and side-information. This will allow us to quantify the benefit of encoder side-information in terms of compression rate. We will consider $X \sim \mathcal{N}(0, \sigma_X^2)$ which is jointly Gaussian with side-information $Y = X + N$, where $N \sim \mathcal{N}(0, \sigma_N^2)$ and independent of X. The distortion measure is the squared error distortion $d_i(X, \hat{X}_i) = (X - \hat{X}_i)^2$, for $i \in \{1, 2\}$. The main result of this section is the following explicit characterization of the rate-distortion function.

Theorem 4 *For given distortion requirements D_1 and D_2, the smallest achievable rate for the above Gaussian setup is as follows:*

1. *if $\sigma_X^2 \leq D_2$ and $\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2} \leq D_1$, then $R(D_1, D_2) = 0$.*

2. *if $\sigma_X^2 > D_2$ but $\frac{D_2 \sigma_N^2}{D_2 + \sigma_N^2} \leq D_1$ (includes the case $\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2} \leq D_1$), $R(D_1, D_2) = \frac{1}{2} \log \frac{\sigma_X^2}{D_2}$.*

3. *if $\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2} > D_1$ and $D_1 + \frac{\sigma_X^4}{\sigma_X^2 + \sigma_N^2} \leq D_2$, then $R(D_1, D_2) = \frac{1}{2} \log \frac{\sigma_X^2 \sigma_N^2}{D_1(\sigma_X^2 + \sigma_N^2)}$.*

4. *if $\sigma_X^2 > D_2$, $\frac{\sigma_N^2 D_2}{\sigma_N^2 + D_2} > D_1$ and $D_1 + \frac{\sigma_X^4}{\sigma_X^2 + \sigma_N^2} > D_2$, then*

$$R(D_1, D_2) = \frac{1}{2} \log \frac{\sigma_X^2}{D_2(1 - \rho_0^2)},$$

where $\rho_0 = \frac{\sigma_X D_1}{\sigma_N (\sigma_X^2 - D_1) D_2} \mathcal{A}$ and

$$\mathcal{A} = \sqrt{D_2(\sigma_X^2 - D_2)} - \sqrt{(\sigma_X^2 \frac{D_2}{D_1}(\sigma_N^2 - D_1) - \sigma_N^2 D_2)(\frac{D_2}{D_1} - 1)}.$$

The regions given by the above regimes are shown in Figure 2.

Sketch of Proof. We only give an intuitive interpretation of the different strategies of Theorem 4. These explanations are at the same time a sketch of the proof. A more detailed proof can be found in [4].

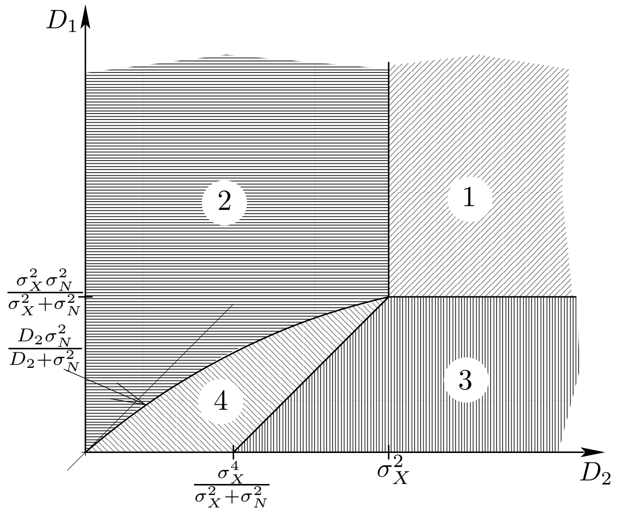

Figure 2: Graphical representation of the different cases of Theorem 4.

In regime 1), clearly, if $\mathrm{Var}\left(()\, X\right) \leq D_2$ and $\mathrm{Var}\left(()\, X|Y\right) \leq D_1$, we need rate 0.

The regime 2) correponds to when given a description $\hat{X}_2$ to Decoder 2, Decoder 1, along with its side-information Y is able to reconstruct $\hat{X}_1$ within the required fidelity, *i.e.,* $D_1 \geq \mathrm{Var}\left(()\, X|Y\hat{X}_2\right) = \frac{D_2\sigma_N^2}{D_2+\sigma_N^2}$. Clearly in this case, a rate of $\frac{1}{2}\log\frac{\sigma_X^2}{D_2}$ is adequate.

Regime 3) corresponds to a conditional rate-distortion for Decoder 1. Conditional rate-distortion compression can be interpreted as describing a new source, $F = \mathbb{E}[X|Y] - X$, *i.e.,* the innovation sequence, within distortion D_1. If, by receiving $\hat{F}$, the approximation to the innovation sequence, Decoder 2 is able to reconstruct $\hat{X}_2$ within distortion D_2, then clearly the conditional rate-distortion rate is adequate. We can show that this can happen if $D_2 \geq \mathrm{Var}\left(()\, X|\hat{F}\right) = D_1 + \frac{\sigma_X^4}{\sigma_X^2+\sigma_N^2}$, and this resolves regime 3).

Regime 4) is perhaps the most interesting case, since it requires more than a single-terminal rate as is the case in the other regimes. The encoder starts by describing X in a single-terminal fashion for Decoder 2. Then, to add information useful to Decoder 1, we correlate $\hat{X}_2$ with the side-information Y as shown below. The representation $\hat{X}_2$ can be expressed using the forward test channel

$$\hat{X}_2 = \alpha X + \Psi,$$

where $\alpha = \frac{\sigma_X^2 - D_2}{\sigma_X^2}$ and $\Psi \sim \mathcal{N}(0, \alpha D_2)$ is independent of X. In order to correlate $\hat{X}_2$ with Y we correlate Ψ with N. This does not affect the performance of Decoder 2 in terms of distortion. However, the rate $I(X; \hat{X}_2)$ increases by $I(Y; \hat{X}_2|X)$ to $I(XY; \hat{X}_2)$. The best correlation factor ρ_0 is negative and its absolute value is just large enough to make $\mathrm{Var}\left(()\, X|\hat{X}_2 Y\right) = D_1$. One can show that the rate achieved by this strategy is optimal under the conditions given in Theorem 4. $\square$

It is remarkable that in all four strategies of Theorem 4, a single message is produced by the encoder, even though the achievability given in [3] suggests two messages.

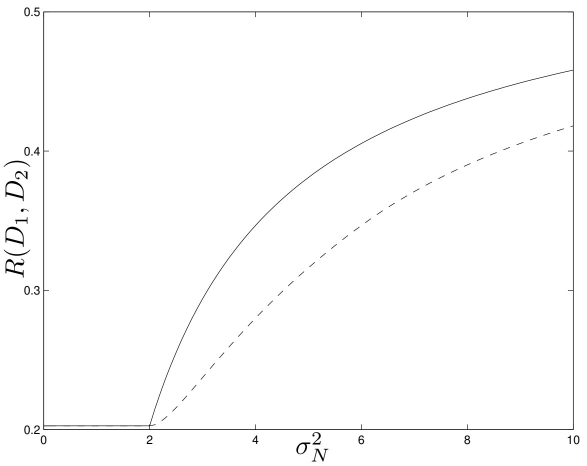

Figure 3: The rate in nats as a function of σ_N^2 for informed (dashed line) and uninformed (solid line) encoder.

Given this explicit characterization of the Gaussian Kaspi rate-distortion function, we will use it to quantify the advantage of encoder side-information in terms of compression rate. We do so by comparing it to the rate-distortion function for the uninformed encoder computed in [2]. Note that in the Gaussian single-decoder case, Wyner and Ziv [8] have shown that encoder side-information is useless in terms of rate-distortion. In Figure 3 we have plotted the rate from Theorem 4 along with the Heegard-Berger rate for a Gaussian source with $\sigma_X^2 = 3$ and distortion requirements $D_2 = 2$ and $D_1 = 1$. We plot the optimal rate as a function of the variance of the side-information noise. We see that encoder side-information can improve the rate-distortion function as compared to the uninformed case.

8 The Gaussian Case with Degraded Side-Information

In this section, we present a special case of the problem stated in Section 2 for which the inner and the outer bound match. The sources considered here take real values. Let $X \sim \mathcal{N}(0, \sigma_X^2)$, and let the side-information sources be $Y = X + N_1$ and $Z = X + N_1 + N_2$, where the variables N_1 and N_2 are jointly distributed with X, Gaussian, zero-mean, independent of X and of each other, and they have respective variances σ_1^2 and σ_2^2. Note that the side-information Y, Z is physically degraded. Also, when we let σ_2^2 go to infinity, then we recover the result of Section 7. We also assume that $D_1 < D_2$. If this is not the case, we can achieve the Wyner-Ziv rate for decoder 2 in a trivial manner.

First, consider the achievable region given in (2). Let U and V be constants. Define

$$\tilde{R}_W = I(X, Y, Z; W) - \min\big\{I(Y; W), I(Z; W)\big\}.$$

The rate-distortion function given by (2) is then

$$\tilde{R}(D_1, D_2) \quad = \quad \min \tilde{R}_W,$$

where the indicated minimization is taken over all Gaussian random variables W, jointly distributed with X, Y, Z and such that

$$D_1 \geq \mathrm{Var}\left(X|W, Y\right), \quad D_2 \geq \mathrm{Var}\left(X|W, Z\right).$$

For Gaussian codebooks, this minimization can be done explicitly, yielding an upper bound on $R(D_1, D_2)$.

Now, we consider the lower bound to $R(D_1, D_2)$ provided in (6). We define $\tilde{X}, \tilde{Y}$ to be as defined in Section 4, but normalize $\tilde{X}$ to have zero mean, *i.e.*,

$$\tilde{X} \sim \mathcal{N}\left(0, \frac{\sigma_X^2(\sigma_1^2 + \sigma_2^2)}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2}\right).$$

Similarly, we scale and normalize $\tilde{Y}$ to be of the form $\tilde{Y} = \tilde{X} + \tilde{N}$, where

$$\tilde{N} \sim \mathcal{N}\left(0, \frac{\sigma_1^2}{\sigma_2^2}(\sigma_1^2 + \sigma_2^2)\right).$$

The above expressions can be found by computing the covariance matrix of (X, Y) conditioned on Z. We then scale the random variable corresponding to Y conditioned on Z until the off-diagonal elements of that covariance matrix equals the $(1, 1)$ element. Note that scaling by or subtracting a constant does not affect mutual informations and conditional variances, that are the only mathematical objects involved in the outer bound. The source pair $\tilde{X}, \tilde{Y}$ is now of the same form as the pair of Gaussian sources in the Kaspi setup of Section 7. Hence, the result of Theorem 4, applied to $\tilde{X}, \tilde{Y}$, yields a lower bound $\breve{R}(D_1, D_2)$ on the rate-distortion function.

Theorem 5 *In the Gaussian setup with physically degraded side-information,*

$$\tilde{R}(D_1, D_2) = \breve{R}(D_1, D_2) = R(D_1, D_2).$$

Proof: To prove this theorem, we evaluate $\breve{R}(D_1, D_2)$ for the sources $\tilde{X}, \tilde{Y}$. This yields a lower bound on $R(D_1, D_2)$, given by different expressions for the four different regimes of Theorem 4. Now, for each of these regimes, we exhibit a Gaussian random variable W for which the achievable rate $\tilde{R}(D_1, D_2)$ equals $\breve{R}(D_1, D_2)$. This leads to the conclusion that the Gaussian coding strategy given by $\tilde{R}(D_1, D_2)$ is optimal.

We can assume that

$$D_1 \leq D_2. \tag{9}$$

If this is not the case, then Decoder 2, observing a weaker side-information, has a tighter distortion requirement, and the optimal solution is Wyner-Ziv encoding for Decoder 2. We also assume that

$$D_1 \leq \mathrm{Var}\,(X|Y) \tag{10}$$

and

$$D_2 \leq \mathrm{Var}\,(X|Z)\,. \tag{11}$$

Again, if any of these conditions is violated, then the problem is trivially solvable.

1. This regime is specified by

$$D_2 \geq \sigma_{\tilde{X}}^2 = \frac{\sigma_X^2(\sigma_1^2 + \sigma_2^2)}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2}$$

and

$$D_1 \geq \frac{\sigma^2_{\tilde{X}}\sigma^2_{\tilde{N}}}{\sigma^2_{\tilde{X}}+\sigma^2_{\tilde{N}}} = \cdots = \frac{\sigma^2_X\sigma^2_1}{\sigma^2_X+\sigma^2_1}.$$

In this regime, the lower bound is $\breve{R}(D_1, D_2) = 0$. The bound is achievable under the above conditions by setting $W =$ constant.

2. In this second regime, we assume that

$$D_2 < \sigma^2_{\tilde{X}} = \frac{\sigma^2_X(\sigma^2_1+\sigma^2_2)}{\sigma^2_X+\sigma^2_1+\sigma^2_2}$$

and

$$D_1 \geq \frac{D_2\sigma^2_{\tilde{N}}}{D_2+\sigma^2_{\tilde{N}}} = \frac{D_2\sigma^2_1(\sigma^2_1+\sigma^2_2)}{D_2\sigma^2_2+\sigma^2_1(\sigma^2_1+\sigma^2_2)}.$$

In this regime, the lower bound is

$$\breve{R}(D_1, D_2) = \frac{1}{2}\log\frac{\sigma^2_{\tilde{X}}}{D_2} = \frac{1}{2}\log\frac{\sigma^2_X(\sigma^2_1+\sigma^2_2)}{D_2(\sigma^2_X+\sigma^2_1+\sigma^2_2)}$$

To see how this rate can be achieved, we set

$$W = X + \Phi,$$

where $\Phi \sim \mathcal{N}(0, \sigma^2_\Phi)$, $\Phi \perp (X, N_1, N_2)$. We choose the variance σ^2_Φ such that

$$\mathrm{Var}\,(X|Z, W) = D_2,$$

which yields

$$\sigma^2_\Phi = \frac{D_2\sigma^2_X(\sigma^2_1+\sigma^2_2)}{\sigma^2_X(\sigma^2_1+\sigma+2^2) - D_2(\sigma^2_X+\sigma^2_1+\sigma^2_2)}.$$

Since W satisfies the Markov chain $W \multimap X \multimap Y \multimap Z$, we have that $I(Z;W) \leq I(Y;W)$. The rate is then

$$\begin{aligned}
\tilde{R}_W &= I(X,Y,Z;W) - I(Z;W) \\
&= I(X,Z;W) - I(Z;W) \\
&= I(X;W|Z) \\
&= h(X|Z) - h(X|W,Z) \\
&= \frac{1}{2}\log 2\pi e \frac{\sigma_X^2(\sigma_1^2+\sigma_2^2)}{\sigma_X^2+\sigma_1^2+\sigma_2^2} - \frac{1}{2}\log 2\pi e \mathrm{Var}\left(X|W,Z\right) \\
&= \frac{1}{2}\log \frac{\sigma_X^2(\sigma_1^2+\sigma_2^2)}{D_2(\sigma_X^2+\sigma_1^2+\sigma_2^2)}.
\end{aligned}$$

This rate is achievable only if

$$D_1 \geq \mathrm{Var}\left(X|Y,W\right) = \frac{\sigma_X^2\sigma_\Phi^2\sigma_1^2}{\sigma_X^2\sigma_1^2+\sigma_\Phi^2(\sigma_X^2+\sigma_1^2)}.$$

Using the expression obtained for σ_Φ^2, this condition can be written as

$$\begin{aligned}
\sigma_\Phi^2\big(\sigma_X^2\sigma_1^2 - D_1(\sigma_X^2+\sigma_1^2)\big) &\leq D_1\sigma_X^2\sigma_1^2 \\
\frac{D_2\sigma_X^2\sigma_1^2(\sigma_1^2+\sigma_2^2) - D_1D_2(\sigma_1^2+\sigma_2^2)(\sigma_X^2+\sigma_1^2)}{\sigma_X^2(\sigma_1^2+\sigma_2^2) - D_2(\sigma_X^2+\sigma_1^2+\sigma_2^2)} &\leq D_1\sigma_1^2.
\end{aligned}$$

We know from a condition of regime 2) that the denominator of the left hand side is positive. Multiplying both sides by this denominator and simplifying yields the second condition of regime 2), *i.e.*,

$$D_1 \geq \frac{D_2\sigma_1^2(\sigma_1^2+\sigma_2^2)}{D_2\sigma_2^2+\sigma_1^2(\sigma_1^2+\sigma_2^2)}.$$

Hence, under the conditions of regime 2), $\breve{R}(D_1,D_2)$ is achievable.

3. The third regime is specified by

$$\begin{aligned}
D_1 &< \frac{\sigma_{\tilde{X}}^2\sigma_{\tilde{N}}^2}{\sigma_{\tilde{X}}^2+\sigma_{\tilde{N}}^2} \\
&= \mathrm{Var}\left(\tilde{X}|\tilde{Y}\right) \\
&= \mathrm{Var}\left(X|Y,Z\right)
\end{aligned}$$

and

$$\begin{aligned} D_2 &\geq D_1 + \frac{\sigma_{\tilde{X}}^4}{\sigma_{\tilde{X}}^2 + \sigma_{\tilde{N}}^2} \\ &= D_1 + \frac{\sigma_X^4 \sigma_2^2}{(\sigma_X^2 + \sigma_1^2)(\sigma_X^2 + \sigma_1^2 + \sigma_2^2)}. \end{aligned}$$

Our first concern is to express $\mathrm{Var}\,(X|Y, Z)$. To this ends, observe that

$$\begin{aligned} h(X|Y, Z) &= h(X|Y, Z - Y) \\ &= h(X|Y, N_2) \\ &= h(X|Y), \end{aligned}$$

because N_2 is independent of (X, Y). Hence,

$$\mathrm{Var}\,(X|Y, Z) = \mathrm{Var}\,(X|Y) = \frac{\sigma_X^2 \sigma_1^2}{\sigma_X^2 + \sigma_1^2}.$$

The lower bound for this regime is

$$\begin{aligned} \breve{R}(D_1, D_2) &= \frac{1}{2} \log \frac{\mathrm{Var}\left(\tilde{X}|\tilde{Y}\right)}{D_1} \\ &= \frac{1}{2} \log \frac{\sigma_X^2 \sigma_1^2}{D_1(\sigma_X^2 + \sigma_1^2)}. \end{aligned}$$

To achieve this lower bound, we first construct a so-called innovation

$$\begin{aligned} F &= \mathbb{E}[X|Y] - X \\ &= qY - X \\ &= (q - 1)X + qN_1 \end{aligned}$$

(innovation with respect to Y), where $q = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_1^2}$. The auxiliary W shall then be an approximation of F at distortion D_1, *i.e.*,

$$W = \beta F + \Psi,$$

where $\beta = \frac{\sigma_F^2 - D_1}{\sigma_F^2}$ and $\Psi \sim \mathcal{N}(0, \beta D_1)$ is independent of (F, X, Y, Z). Note that since F is the estimation error when estimating X from

Y, F is independent of Y. Furthermore, F is chosen such that $F \multimap (X,Y) \multimap Z$ is a Markov chain. Hence, $Z \perp (Y, Z)$. All of this implies that $W \perp (Y, Z)$ and $W \multimap (X,Y) \multimap Z$ is Markov. With such a choice of W, the rate is

$$
\begin{aligned}
\tilde{R}_W &= I(X, Y, Z; W) \\
&= I(Y; W) + I(X; W|Y) + I(Z; W|X, Y) \\
&= I(X; W|Y) \\
&= I(qY - X; W|Y) \\
&= I(F; W|Y) \\
&= I(FY; W) - \underbrace{I(Y; W)}_{=0} \\
&= I(F; W) + I(Y; W|F) \\
&\overset{(a)}{=} I(F; W) \\
&= h(F) - h(F|W) \\
&= \frac{1}{2}\log 2\pi e \mathrm{Var}\,(F) - \frac{1}{2}\log 2\pi e \mathrm{Var}\,(F|W) \\
&\overset{(b)}{=} \frac{1}{2}\log 2\pi e \frac{\sigma_X^2\sigma_1^2}{\sigma_X^2+\sigma_1^2} - \frac{1}{2}\log 2\pi e D_1,
\end{aligned}
$$

where (a) is true because W is a noisy version of F that satisfies the Markov chain $W \multimap F \multimap Y$ and (b) follows from the fact that F is the innovation of X with respect to Y and that W is a D_1-good description of F. Hence, $\tilde{R} = \breve{R}$. Now, we check whether W satisfies the distortion constraints under the conditions of this regime. Note that

$$
\begin{aligned}
h(X|W, Y) &= h(qY - X|W, Y) \\
&= h(F|W, Y) \\
&= h(F|W) \\
&= \frac{1}{2}\log 2\pi e D_1,
\end{aligned}
$$

which implies that $\mathrm{Var}\,(X|W, Y) = D_1$. For the distortion at Decoder 2, first note that

$$
\begin{aligned}
W &= \beta F + \Psi \\
&= \beta(q-1)X + \beta q N_1 + \Psi.
\end{aligned}
$$

The covariance matrix of (X, W, Z) is given by

$$\left[\begin{array}{ccc} \sigma_X^2 & \beta(q-1)\sigma_X^2 & \sigma_X^2 \\ \beta(q-1)\sigma_X^2 & \beta^2(q-1)^2\sigma_X^2 + \beta^2 q^2 \sigma_1^2 + \beta D_1 & \beta(q-1)\sigma_X^2 + \beta q \sigma_1^2 \\ \sigma_X^2 & \beta(q-1)\sigma_X^2 + \beta q \sigma_1^2 & \sigma_X^2 + \sigma_1^2 + \sigma_2^2 \end{array}\right].$$

We compute

$$\mathrm{Var}\,(X|W,Z) = \sigma_X^2 - \sigma_X^4 \left[\begin{array}{cc} \beta(q-1) & 1 \end{array}\right] K_{WZ}^{-1} \left[\begin{array}{c} \beta(q-1) \\ 1 \end{array}\right],$$

where K_{WZ} is the covariance matrix of (W, Z). After replacing $q = \frac{\sigma_X^2}{\sigma_X^2+\sigma_1^2}$ and $\beta = \frac{(q-1)^2\sigma_X^2+q^2\sigma_1^2-D_1}{(q-1)^2\sigma_X^2+q^2\sigma_1^2}$, and after lengthy simplifications, we obtain

$$\mathrm{Var}\,(X|W,Z) = D_1 + \frac{\sigma_X^4\sigma_2^2}{(\sigma_X^2+\sigma_1^2)(\sigma_X^2+\sigma_1^2+\sigma_2^2)}.$$

Thus, the condition $D_2 \geq \mathrm{Var}\,(X|W,Z)$ matches one of the conditions of the current regime 3), which shows that the chosen W achieves the distortion requirements.

4. The conditions for this regime can be summarized as “all values of the problem parameters that do not fall into regimes 1), 2) or 3)”. The computations for this fourth regime lead to very long analytical expressions. For this reason, we refer to the attached files with Mathematica code for the details. The lower bound on the rate is computed in the

file *conditional_Kaspi_rate.nb* and it yields

$$
\begin{aligned}
2^{2\breve{R}(D_1,D_2)} \;=\; & \Big[D_2\sigma_1^2(\sigma_1^2+\sigma_2^2)\sigma_X^2((\sigma_1^2+\sigma_2^2)\sigma_X^2 - D_1(\sigma_1^2+\sigma_2^2+\sigma_X^2))^2\Big] / \\
& \Bigg[D_1(\sigma_1^2+\sigma_2^2+\sigma_X^2) \\
& (D_2(\sigma_1^2+\sigma_2^2)\sigma_X^2(-(D_2\sigma_1^2(\sigma_1^2+\sigma_2^2)) + \\
& (-(D_2\sigma_1^2)+\sigma_1^4+D_2\sigma_2^2+\sigma_1^2\sigma_2^2)\sigma_X^2) + \\
& D_1({D_2}^2(\sigma_1^2+\sigma_2^2)(\sigma_1^2+\sigma_X^2)(\sigma_1^2+\sigma_2^2+\sigma_X^2) - \\
& D_2(\sigma_1^2+\sigma_2^2)\sigma_X^2(\sigma_1^4+2\sigma_2^2\sigma_X^2+\sigma_1^2(\sigma_2^2+\sigma_X^2)) + \\
& 2\sigma_2^2\sigma_X^2(\sigma_1^2+\sigma_2^2+\sigma_X^2) \\
& \sqrt{\frac{(D_1-D_2)D_2(\sigma_1^2+\sigma_2^2)^2(D_1\sigma_1^2+(D_1-\sigma_1^2)\sigma_X^2)}{{D_1}^2\sigma_2^2(\sigma_1^2+\sigma_2^2+\sigma_X^2)}} \\
& \sqrt{\frac{D_2((\sigma_1^2+\sigma_2^2)\sigma_X^2 - D_2(\sigma_1^2+\sigma_2^2+\sigma_X^2))}{\sigma_1^2+\sigma_2^2+\sigma_X^2}}))\Bigg].
\end{aligned}
$$

Unfortunately, we do not have an intuitive explanation of how to achieve this lower bound. However, we are able to exhibit an auxiliary random variable W that achieves the distortion constraints and for which $\tilde{R}(D_1, D_2)$ matches $\breve{R}(D_1, D_2)$. First, we arbitrarily impose that W should be such that $I(W; Z) \leq I(W; Y)$, and hence, the achievable rate becomes

$$
\begin{aligned}
\tilde{R}_W &= I(X, Y, Z; W) - I(W; Z) \\
&= I(X, Y; W|Z).
\end{aligned}
$$

We further choose W to be jointly Gaussian with (X, Y, Z), more precisely,

$$W = aX + bN_1 + cN_2 + \xi,$$

where $\xi \sim \mathcal{N}(0, 1)$ and a, b, c are real numbers. The rate can be written as

$$
\begin{aligned}
\tilde{R}_W &= I(X, Y; W|Z) \\
&= h(W|Z) - h(W|X, Y, Z) \\
&= h(W - \alpha Z|Z) - h(aX + bN_1 + cN_2 + \xi|X, N_1, N_2) \\
&= h(W - \alpha Z|Z) - h(\xi). \qquad (12)
\end{aligned}
$$

We choose α such that $W - \alpha Z$ is independent of Z, *i.e.*,

$$\alpha = \frac{a\sigma_X^2 + b\sigma_1^2 + c\sigma_2^2}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2}.$$

Then,

$$\begin{aligned}
\tilde{R}_W &= h((a-\alpha)X + (b-\alpha)N_1 + (c-\alpha)N_2 + \xi) - h(\xi) \\
&= \frac{1}{2}\log\frac{(a-\alpha)^2\sigma_X^2 + (b-\alpha)^2\sigma_1^2 + (c-\alpha)^2\sigma_2^2 + 1}{1} \qquad (13) \\
&= \frac{1}{2}\log\Big(1 + \frac{(a-b)^2\sigma_X^2\sigma_1^2 + (a-c)^2\sigma_X^2\sigma_2^2 + (b-c)^2\sigma_1^2\sigma_2^2}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2}\Big).
\end{aligned}$$

We want to find the values of a, b, c that minimize $\tilde{R}_W$, while satisfying the distortion constraints and the constraint $I(W;Z) \leq I(W;Y)$. This can be done by minimizing the objective function

$$\mathcal{O}(a,b,c) = (a-\alpha)^2\sigma_X^2 + (b-\alpha)^2\sigma_1^2 + (c-\alpha)^2\sigma_2^2$$

subject to the constraints

$$\begin{aligned}
\mathrm{Var}\,(X|W,Y) &\leq D_1 \\
\mathrm{Var}\,(X|W,Z) &\leq D_2 \\
I(W;Z) &\leq I(W;Y).
\end{aligned}$$

Computing the conditional variances of X yields

$$\mathrm{Var}\,(X|W,Y) = \frac{\sigma_X^2\sigma_1^2(1 + c^2\sigma_2^2)}{(a-b)^2\sigma_X^2\sigma_1^2 + (\sigma_X^2 + \sigma_1^2)(1 + c^2\sigma_2^2)}$$

and

$$\mathrm{Var}\,(X|W,Z) = \frac{\sigma_X^2\big(\sigma_1^2 + \sigma_2^2 + (b-c)^2\sigma_1^2\sigma_2^2\big)}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2 + \sigma_X^2\sigma_1^2(a-b)^2 + \sigma_X^2\sigma_2^2(a-c)^2 + \sigma_1^2\sigma_2^2(b-c)^2}.$$

We incorporate two of the above constraints into a new objective function using Lagrange multipliers λ and μ:

$$F(a,b,c) = \mathcal{O}(a,b,c) + \lambda\mathrm{Var}\,(X|W,Y) + \mu\mathrm{Var}\,(X|W,Z)\,.$$

We ignore the third constraint $I(W;Z) \leq I(W;Y)$ for now, and we will later show that the W that optimizes the Lagrangian problem satisfies it. In the Mathematica file *derivatives.nb*, we compute the values $\frac{\partial F}{\partial a}$, $\frac{\partial F}{\partial b}$ and $\frac{\partial F}{\partial c}$. Since at the optimum, all three partial derivatives vanish, we know that the optimal solution has to satisfy

$$\frac{\partial F}{\partial a} + \frac{\partial F}{\partial b} + \frac{\partial F}{\partial c} = 0. \tag{14}$$

The Mathematica file shows that one solution to this equation is $a = b$. However, in that case, we would have

$$W = aX + aN_1 + cN_2 + \xi = aY + cN_2 + \xi.$$

Thus, from the view of Decoder 1, W would be a noisy version of Y, and hence of no use at all. One can indeed verify that if $a = b$, then $\mathrm{Var}\,(X|W,Y) = \mathrm{Var}\,(X|Y)$. An alternate solution to equation (14) is $c = 0$, and we will retain this value of c in our final choice of W. At this point, one can already verify whether

$$\begin{aligned}
I(W;Z) &\overset{?}{\leq} I(W;Y) \\
-\frac{1}{2}\log(1-\rho_{WZ}^2) &\overset{?}{\leq} -\frac{1}{2}\log(1-\rho_{WY}^2) \\
\rho_{WZ}^2 &\overset{?}{\leq} \rho_{WY}^2 \\
\frac{\mathbb{E}[WZ]^2}{\mathbb{E}[W^2]\mathbb{E}[Z^2]} &\overset{?}{\leq} \frac{\mathbb{E}[WY]^2}{\mathbb{E}[W^2]\mathbb{E}[Y^2]} \\
\frac{\mathbb{E}[WZ]^2}{\mathbb{E}[Z^2]} &\overset{?}{\leq} \frac{\mathbb{E}[WY]^2}{\mathbb{E}[Y^2]} \\
\frac{(a\sigma_X^2 + b\sigma_1^2 + c\sigma_2^2)^2}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2} &\overset{?}{\leq} \frac{(a\sigma_X^2 + b\sigma_1^2)^2}{\sigma_X^2 + \sigma_1^2} \\
\frac{(a\sigma_X^2 + b\sigma_1^2)^2}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2} &\overset{?}{\leq} \frac{(a\sigma_X^2 + b\sigma_1^2)^2}{\sigma_X^2 + \sigma_1^2},
\end{aligned}$$

and we see that the inequality holds for all values of a and b. In the last step, we have substituted $c = 0$.

From here, we drop the Lagrangian formulation and use the two equations

$$\begin{aligned}
\mathrm{Var}\,(X|W,Y) &= D_1 \\
\mathrm{Var}\,(X|W,Z) &= D_2
\end{aligned}$$

to determine a and b. Substituting $c = 0$ into the first equation yields

$$\begin{aligned}
D_1 &= \frac{\sigma_X^2\sigma_1^2}{\sigma_X^2 + \sigma_1^2 + (a-b)^2\sigma_X^2\sigma_1^2} \\
(a-b)^2 &= \frac{\sigma_X^2\sigma_1^2 - D_1(\sigma_X^2 + \sigma_1^2)}{D_1\sigma_X^2\sigma_1^2} \\
a &= b \pm \sqrt{\frac{\sigma_X^2\sigma_1^2 - D_1(\sigma_X^2 + \sigma_1^2)}{D_1\sigma_X^2\sigma_1^2}} \\
&\triangleq b \pm \sqrt{\mathcal{B}}.
\end{aligned}$$

We arbitrarily choose $a = b - \sqrt{\mathcal{B}}$, and we substitute this together with $c = 0$ into $\mathrm{Var}\,(X|W,Z) = D_2$:

$$D_2 = \frac{\sigma_X^2(\sigma_1^2 + \sigma_2^2 + b^2\sigma_1^2\sigma_2^2)}{\sigma_X^2 + \sigma_1^2 + \sigma_2^2 + \sigma_X^2\sigma_1^2\mathcal{B} + \sigma_X^2\sigma_2^2(b - \sqrt{\mathcal{B}})^2 + \sigma_1^2\sigma_2^2 b^2}.$$

This equation is quadratic in b:

$$\begin{gathered}
b^2(D_2\sigma_X^2\sigma_2^2 + D_2\sigma_1^2\sigma_2^2 - \sigma_X^2\sigma_1^2\sigma_2^2) \\
-bD_2 2\sqrt{\mathcal{B}}\sigma_X^2\sigma_2^2 \\
+D_2\big(\sigma_X^2 + \sigma_1^2 + \sigma_2^2 + \sigma_X^2\sigma_1^2\mathcal{B} + \sigma_X^2\sigma_2^2\mathcal{B}\big) - \sigma_X^2(\sigma_1^2 + \sigma_2^2) = 0.
\end{gathered}$$

In the Mathematica file *achievable_rate.nb*, we solve this equation for b. Again, we arbitrarily retain one of the two choices for b, namely

$$b = \frac{2D_1D_2\sigma_X^2\sigma_1^2\sigma_2^2\sqrt{\mathcal{C}} - \sqrt{\mathcal{D}}}{2(D_1D_2\sigma_1^4\sigma_2^2 + D_1D_2\sigma_X^2\sigma_1^2\sigma_2^2 - D_1\sigma_X^2\sigma_1^4\sigma_2^2)},$$

where

$$\mathcal{C} = \frac{\sigma_X^2\sigma_1^2 - D_1(\sigma_X^2 + \sigma_1^2)}{D_1\sigma_X^2\sigma_1^2}$$

and

$$\begin{aligned}
\mathcal{D} = {} & 4D_1D_2^2\sigma_X^2\sigma_1^2\sigma_2^4\big(\sigma_X^2\sigma_1^2 - D_1(\sigma_X^2 + \sigma_1^2)\big) \\
& -4(D_2\sigma_X^2\sigma_1^4 - D_1\sigma_X^2\sigma_1^4 - D_1D_2\sigma_X^2\sigma_2^2 - D_1\sigma_X^2\sigma_1^2\sigma_2^2 + D_2\sigma_X^2\sigma_1^2\sigma_2^2) \\
& (D_1D_2\sigma_1^4\sigma_2^2 + D_1D_2\sigma_X^2\sigma_1^2\sigma_2^2 - D_1\sigma_X^2\sigma_1^4\sigma_2^2).
\end{aligned}$$

For b to be a real number, $\mathcal{C}$ and $\mathcal{D}$ must both be non-negative. It is immediate to see that $\mathcal{C}$ is non-negative whenever

$$D_1 \leq \frac{\sigma_X^2 \sigma_1^2}{\sigma_X^2 + \sigma_1^2} = \mathrm{Var}\,(X|Y)\,,$$

which holds by assumption (10). After some simplifications, the second square-root expression yields

$$\begin{aligned}\mathcal{D} &= 4D_1\sigma_1^6\sigma_2^2\sigma_X^2(D_2 - D_1)\\ &\quad \Big(\sigma_X^2(\sigma_1^2 + \sigma_2^2) - D_2(\sigma_X^2 + \sigma_1^2 + \sigma_2^2)\Big).\end{aligned}$$

The non-negativity of this expression is directly implied by assumptions (9) and (11). To summarize, by fixing a, b and c as given, we fix a distribution for

$$W = aX + bN_1 + cN_2 + \xi$$

which satisfies both distortion requirements. In addition, we have $I(W;Z) \leq I(W;Y)$, and therefore, the rate given in (12) is achievable. Now, we plug the given expressions for a, b and c into (13) to compute $\tilde{R}_W$. This is done in the Mathematica file *achievable_rate.nb*, where the output called "rate" corresponds to $2^{2\tilde{R}_W}$, namely to the argument of the logarithm in $\tilde{R}_W$:

$$2^{2\tilde{R}_W} = G\frac{A - 2\sigma_X^2\sqrt{B}}{C},$$

where

$$\begin{aligned}A &= \sigma_1^2\Bigg(D_1\Big(D_2(\sigma_X^2 + \sigma_1^2)(\sigma_X^2 + \sigma_1^2 + \sigma_2^2)\\ &\quad -\sigma_X^2\big(\sigma_1^4 + 2\sigma_X^2\sigma_2^2 + \sigma_1^2(\sigma_X^2 + \sigma_2^2)\big)\Big)\\ &\quad +\sigma_X^2\Big(\sigma_X^2\sigma_1^4 + \sigma_X^2\sigma_1^2\sigma_2^2\\ &\quad -D_2\big(\sigma_1^4 - \sigma_X^2\sigma_2^2 + \sigma_1^2(\sigma_X^2 + \sigma_2^2)\big)\Big)\Bigg),\end{aligned}$$

$$\begin{aligned}B &= (\sigma_X^2\sigma_1^2 - D_1\sigma_X^2 - D_1\sigma_1^2)\\ &\quad \Big((D_2 - D_1)\sigma_1^4\sigma_2^2\big(\sigma_X^2(\sigma_1^2 + \sigma_2^2) - D_2(\sigma_X^2 + \sigma_1^2 + \sigma_2^2)\big)\Big),\end{aligned}$$

$$C = \left(\sigma_X^2\sigma_1^2 - D_2(\sigma_X^2+\sigma_1^2)\right)^2$$

and

$$G = \frac{\sigma_X^2}{D_1(\sigma_X^2+\sigma_1^2+\sigma_2^2)}.$$

Now, it remains to show that $\tilde{R}_W = \breve{R}(D_1, D_2)$. To this ends, we rewrite the lower bound as

$$2^{2\breve{R}(D_1,D_2)} = G\frac{\sigma_1^4(\sigma_X^2+\sigma_1^2+\sigma_2^2)^2(\mathrm{Var}\,(X|Z)-D_1)^2}{A+2\sigma_X^2\sqrt{B}}.$$

We also write $\tilde{R}_W$ differently:

$$\begin{aligned} 2^{2\tilde{R}_W} &= G\frac{(A-2\sigma_X^2\sqrt{B})(A+2\sigma_X^2\sqrt{B})}{C(A+2\sigma_X^2\sqrt{B})} \\ &= G\frac{(A^2-4\sigma_X^4 B)}{C(A+2\sigma_X^2\sqrt{B})}. \end{aligned}$$

One can see that the only identity that still needs verification is

$$\frac{A^2-4\sigma_X^4 B}{C} = \sigma_1^4(\sigma_X^2+\sigma_1^2+\sigma_2^2)(\mathrm{Var}\,(X|Z)-D_1)^2.$$

In the Mathematica file *last_step.nb*, we compute

$$A^2 - 4\sigma_X^4 B - C\sigma_1^4(\sigma_X^2+\sigma_1^2+\sigma_2^2)(\mathrm{Var}\,(X|Z)-D_1)^2$$

and verify that it is indeed zero. This concludes the proof of Theorem 5.

□

Observe that in the achievability of Theorem 5, only the auxiliary variable W, destined for both decoders, remains. The private messages U and V in (2) disappear. The same is the case in the Gaussian Kaspi setup of Theorem 4. Also note that in Theorem 5, the optimal auxiliary random variable W satisfies the Markov chain $W \multimap (X,Y) \multimap Z$. This will be useful in Section 8.2.

8.1 Uninformed Encoder

In [2], Heegard and Berger find the rate-distortion function for several decoders with degraded side-information when the encoder is not informed. In particular, when there is only one side-information source, they explicitly solve the quadratic Gaussian case. We simply state their result here. Let $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y = X + N_1$, where $N_1 \perp X$ and $N_1 \sim \mathcal{N}(0, \sigma_1^2)$. The side-information Y is only available at Decoder 1, *i.e.*, we are considering the setup of Figure 4 with Z being absent. The distortion measures are the squared error. Then, the rate-distortion function is given by

$$R_{\mathrm{HB},Y} = \begin{cases} \frac{1}{2}\log\frac{\sigma_X^2\sigma_1^2}{D_1(D_2+\sigma_1^2)}, \text{ if } D_1 \leq \frac{D_2\sigma_1^2}{D_2+\sigma_1^2} \text{ and } D_2 \leq \sigma_X^2 \\ \frac{1}{2}\log\frac{\sigma_X^2}{D_2}, \text{ if } D_1 > \frac{D_2\sigma_1^2}{D_2+\sigma_1^2} \text{ and } D_2 \leq \sigma_X^2 \\ \frac{1}{2}\log\frac{\sigma_X^2\sigma_1^2}{D_1(\sigma_X^2+\sigma_1^2)}, \text{ if } D_1 \leq \frac{\sigma_X^2\sigma_1^2}{\sigma_X^2+\sigma_1^2} \text{ and } D_2 > \sigma_X^2 \\ 0, \text{ if } D_1 > \frac{\sigma_X^2\sigma_1^2}{\sigma_X^2+\sigma_1^2} \text{ and } D_2 > \sigma_X^2. \end{cases} \tag{15}$$

In [7], Tian and Diggavi explicitly compute the rate-distortion function for the case when both Y and Z are present and physically degraded, and the encoder is not informed (Figure 4). We assume that X and Y are as above, and $Z = Y + N_2$, *i.e.*, the same setup as in the previous parts of Section 8. Let

$$\gamma = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

The rate-distortion function is given by

$$R_{\mathrm{HB}} = \begin{cases} \frac{1}{2}\log\frac{\sigma_X^2\sigma_1^2\sigma_2^2}{D_1(\sigma_X^2+\sigma_1^2+\sigma_2^2)((1-\gamma)^2D_2+\gamma\sigma_2^2)} \text{ if } D_1 \leq \frac{D_2\gamma\sigma_1^2}{\gamma\sigma_1^2+(1-\gamma)^2D_1} \text{ and } D_2 \leq \frac{\sigma_X^2(\sigma_1^2+\sigma_2^2)}{\sigma_X^2+\sigma_1^2+\sigma_2^2} \\ \frac{1}{2}\log\frac{\sigma_X^2(\sigma_1^2+\sigma_2^2)}{D_2(\sigma_X^2+\sigma_1^2+\sigma_2^2)}, \text{ if } D_1 > \frac{D_2\gamma\sigma_1^2}{\gamma\sigma_1^2+(1-\gamma)^2D_1} \text{ and } D_2 \leq \frac{\sigma_X^2(\sigma_1^2+\sigma_2^2)}{\sigma_X^2+\sigma_1^2+\sigma_2^2} \\ \frac{1}{2}\log\frac{\sigma_X^2\sigma_1^2}{D_1(\sigma_X^2+\sigma_1^2)}, \text{ if } D_1 \leq \frac{\sigma_X^2\sigma_1^2}{\sigma_X^2+\sigma_1^2} \text{ and } D_2 > \frac{\sigma_X^2(\sigma_1^2+\sigma_2^2)}{\sigma_X^2+\sigma_1^2+\sigma_2^2} \\ 0, \text{ if } D_1 > \frac{\sigma_X^2\sigma_1^2}{\sigma_X^2+\sigma_1^2} \text{ and } D_2 > \frac{\sigma_X^2(\sigma_1^2+\sigma_2^2)}{\sigma_X^2+\sigma_1^2+\sigma_2^2}. \end{cases} \tag{16}$$

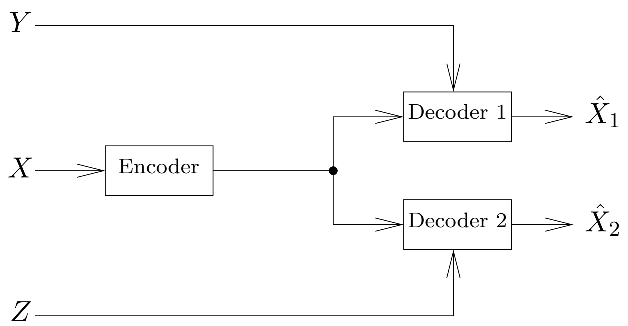

Figure 4: Uninformed Encoder.

8.2 Partially Informed Encoder

Let us consider the setup depicted in Figure 5. A lower bound to this setup can be established by making Z also available to Decoder 1. Now, all parties observe Z, and hence, the problem becomes a Heegard-Berger setup (only one side-information) with sources $(\tilde{X}, \tilde{Y})$, that correspond to X and Y conditioned on Z (as defined in the beginning of Section 8. One can verify that by plugging $(\tilde{X}, \tilde{Y})$ into (15), we obtain the Heegard-Berger rate-distortion function (16) for two side-informations. This means that the lower bound given by the conditional setup is achievable using the strategy described in [7]. In other words, making only Z available to the encoder is of no benefit in terms on rate-distortion.

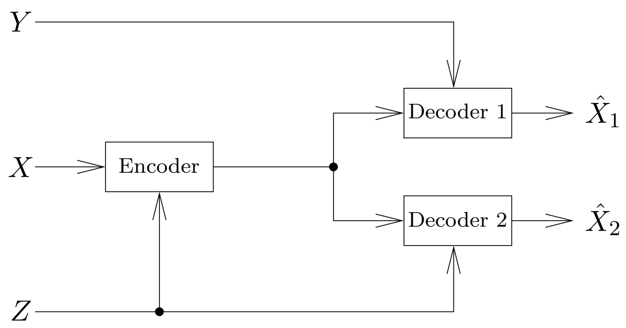

Figure 5: Only Z is available at the encoder.

Now, we consider the case when only Y is available at the encoder. The rate-distortion function found in Theorem 5 is obviously a lower bound on the rate-distortion of this new setup. We remarked earlier that the achievable strategy in Theorem 5 is such that the auxiliary W satisfies $W -\!\circ\!- (X, Y) -\!\circ\!- Z$. Hence, the same achievability can be applied to this new setup. This means that in this Gaussian setup with physically degraded side information, it does not matter whether Z is available at the encoder or not.

9 Discussion

The main challenge appears to be in understanding how to make use of side-information at the encoder. Since the side-information sources are separately

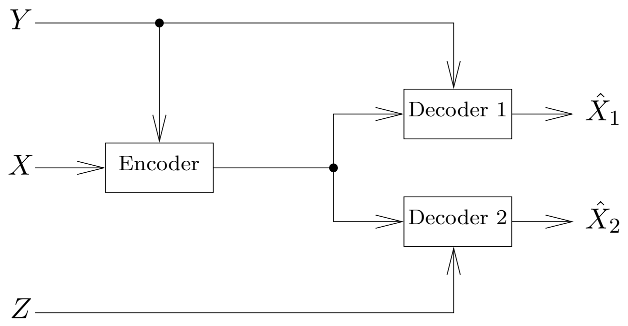

Figure 6: Only Y is available at the encoder.

available to the decoders, we need to use the encoder knowledge in a non-trivial manner. The rate-distortion functions given in the two Gaussian cases reveal that the description to a decoder shall be designed to maximally benefit the other decoder as well. This idea of establishing a connection between the two decoders through a common auxiliary sequence is the crux of our achievability. The result established by Kaspi in [3] shows that additional private auxiliary sequences are necessary. The content of this paper is work in progress. Some of us believe that the region $\tilde{\mathcal{R}}$ is optimal and we are hoping to find some tight characterizations for discrete sources.

10 Acknowledgment

We would like to thank Dinkar Vasudevan for the fruitful discussions, and in particular for his contribution to Section 8.2.

A Maximum Entropy

Consider three discrete random variables A,B and C. Let R be the type of (AB), S the type of AC and T the type of BC. Define $\mathcal{P}$ to be the set of all distributions $p(\tilde{C}|AB)$ such that $(A\tilde{C}) \sim (AC)$ and $(B\tilde{C}) \sim (BC)$. We are interested in the asymptotic behaviour of the following number N:

$$
\begin{aligned}
N &= \#\{(\mathbf{a},\mathbf{b},\mathbf{c}) : (\mathbf{a},\mathbf{b}) \in R, (\mathbf{a},\mathbf{c}) \in S, (\mathbf{b},\mathbf{c}) \in T\} \\
&= \#\big\{ \cup_{p(\tilde{C}|AB)\in\mathcal{P}} \mathrm{Type}(A,B,\tilde{C})\big\} \\
&\le \sum_{p(\tilde{C}|AB)\in\mathcal{P}} \#\mathrm{Type}(A,B,\tilde{C}) \\
&= \sum_{p(\tilde{C}|AB)\in\mathcal{P}} 2^{nH(AB\tilde{C})} \\
&\doteq 2^{n\big(\max_{p(\tilde{C}|AB)\in\mathcal{P}} H(AB\tilde{C})\big)}.
\end{aligned}
$$

We also have

$$
\begin{aligned}
N &= \#\big\{ \cup_{p(\tilde{C}|AB)\in\mathcal{P}} \mathrm{Type}(A,B,\tilde{C})\big\} \\
&\ge \max_{p(\tilde{C}|AB)\in\mathcal{P}} \#\mathrm{Type}(A,B,\tilde{C})\big\} \\
&= 2^{n\big(\max_{p(\tilde{C}|AB)\in\mathcal{P}} H(AB\tilde{C})\big)},
\end{aligned}
$$

and thus,

$$
N \doteq 2^{n\big(\max_{p(\tilde{C}|AB)\in\mathcal{P}} H(AB\tilde{C})\big)}.
$$

B Attached Mathematica Files

derivatives.nb
conditional_Kaspi_rate.nb
achievable_rate.nb
last_step.nb

References

[1] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[2] C. Heegard and T. Berger, *Rate distortion when side information may be absent*, IEEE Trans. Inform. Theory, 31 (1985), pp. 727–734.

[3] A. Kaspi, *Rate-distortion function when side-information may be present at the decoder*, IEEE Trans. Inform. Theory, 40 (1994), pp. 2031–2034.

[4] E. Perron, S. Diggavi, and E. Telatar, *The kaspi rate-distortion problem with encoder side-information: Gaussian case*, Tech. Rep. LICOS-REPORT-2005-001, École Polytechnique Fédérale de Lausanne, Switzerland, Nov. 2005.

[5] A. Sgarro, *Source coding with side information at several decoders*, IEEE Trans. Inform. Theory, 23 (1977), pp. 179–182.

[6] D. Slepian and J. Wolf, *Noiseless coding of correlated information sources*, IEEE Trans. Inform. Theory, 19 (1973), pp. 471–480.

[7] C. Tian and S. Diggavi, *On multistage successive refinement for wyner-ziv source coding with degraded side-information*, in Proc. of the IEEE Int. Symposium on Inform. Theory, Seattle, USA, 2006.

[8] A. Wyner and J. Ziv, *The rate-distortion function for source coding with side information at the decoder*, IEEE Trans. Inform. Theory, 22 (1976).