LICOS-REPORT-2005-001

The Kaspi Rate-Distortion Problem with Encoder Side-Information: Gaussian Case

Etienne Perron[*], Suhas Diggavi, Emre Telatar

May 24, 2006

Abstract

We consider the problem of compressing a Gaussian source for two decoders, one of which has access to side-information. Kaspi has solved this problem for discrete memoryless sources for the two cases with and without encoder side-information. We focus on the case in which the encoder has access to side-information and we explicitly solve the optimization problem for Gaussian sources and squared error distortion measures. The achievability part of the proof provides an intuitive insight based on the idea of using innovations for the conditional rate-distortion problem. We compare our result to the case in which the encoder is not informed. Heegard and Berger have solved this problem for Gaussian sources. The comparison shows that in this multiple-decoder setup, side-information at the encoder improves the rate-distortion curve even in the Gaussian case.

1 Introduction

In [3], Kaspi has solved the rate-distortion problem for a setup where "side-information may be present at the decoder". He gives the rate-distortion

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

function for discrete memoryless sources for the two cases of an informed and uninformed encoder respectively. The case in which the encoder is informed, i.e., it has access to the side-information, is related to the conditional rate-distortion problem. If the sources (the source of interest and the side-information source) are Gaussian and the distortion measure is the squared error, then the conditional rate-distortion function can be achieved using an innovation scheme as explained in Appendix A. The same idea can be applied to find achievable schemes for the Kaspi problem with encoder side-information. In this paper, we do so and find that two cases can be distinguished. After outlining the problem, we treat the two cases in two different sections. For both cases, we provide an intuitive derivation of the achievability as well as a converse. In the last section, we compare our result to the rate-distortion function given by Heegard and Berger in [2] for an uninformed encoder. We conclude that encoder side-information is helpful in this Gaussian setup.

2 Problem Outline and Result

Consider the setup depicted in Figure 1, which corresponds to the case "switch A closed" in [3].

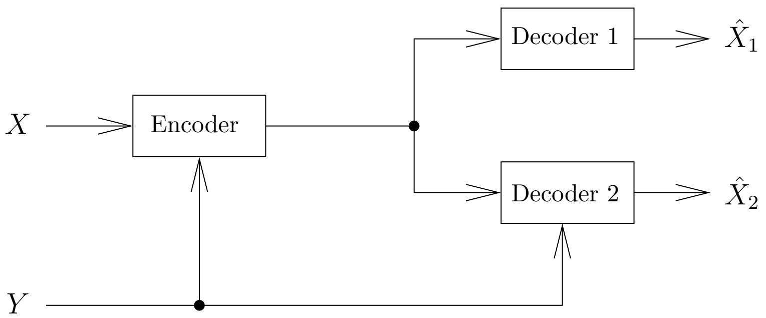

Figure 1: Kaspi Setup with Informed Encoder.

Given two distortion constraints D_1 and D_2 for Decoder 1 and 2, respectively, we are looking for the smallest encoding rate $R(D_1, D_2)$ that satisfies both distortion constraints simultaneously. In [3], Kaspi gives the precise rate-distortion function for discrete memoryless sources:

$$R(D_1, D_2) = \min \left[I(XY; \hat{X}_1) + I(X; \hat{X}_2 | Y \hat{X}_1) \right], \tag{1}$$

where the indicated minimization is with respect to all random variables $\hat{X}_1$ and $\hat{X}_2$ which take values in $\mathcal{X}$, which are jointly distributed with the given random variables $X \in \mathcal{X}$ and Y, and for which

$$\mathrm{E}\left[d(X, \hat{X}_i)^2 \right] \leq D_i$$

for $i = 1, 2$.

We focus on the Gaussian case, i.e.,

$$X \sim \mathcal{N}(0, \sigma_X^2)$$

and

$$Y = X + Z,$$

where $Z \sim \mathcal{N}(0, \sigma_Z^2)$ and independent of X. The distortion measure is the squared error distortion

$$d(X, \hat{X}_i) = (X - \hat{X}_i)^2.$$

Theorem 1 *For given distortion requirements D_1 and D_2, the smallest achievable rate for the above Gaussian setup is as follows:*

1. *if $\sigma_X^2 \leq D_1$ and $\frac{\sigma_X^2 \sigma_Z^2}{\sigma_X^2 + \sigma_Z^2} \leq D_2$, then*

$$R(D_1, D_2) = 0,$$

2. *if $\sigma_X^2 > D_1$ but $\frac{D_1 \sigma_Z^2}{D_1 + \sigma_Z^2} \leq D_2$ (includes the case $\frac{\sigma_X^2 \sigma_Z^2}{\sigma_X^2 + \sigma_Z^2} \leq D_2$), then*

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

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

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

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

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

where

$$\rho_0 = \frac{\sigma_X D_2}{\sigma_Z(\sigma_X^2 - D_2)D_1}\left[\sqrt{D_1(\sigma_X^2 - D_1)} - \sqrt{(\sigma_X^2\frac{D_1}{D_2}(\sigma_Z^2 - D_2) - \sigma_Z^2 D_1)(\frac{D_1}{D_2} - 1)}\right].$$

The regions given by above cases are shown in Figure 2.

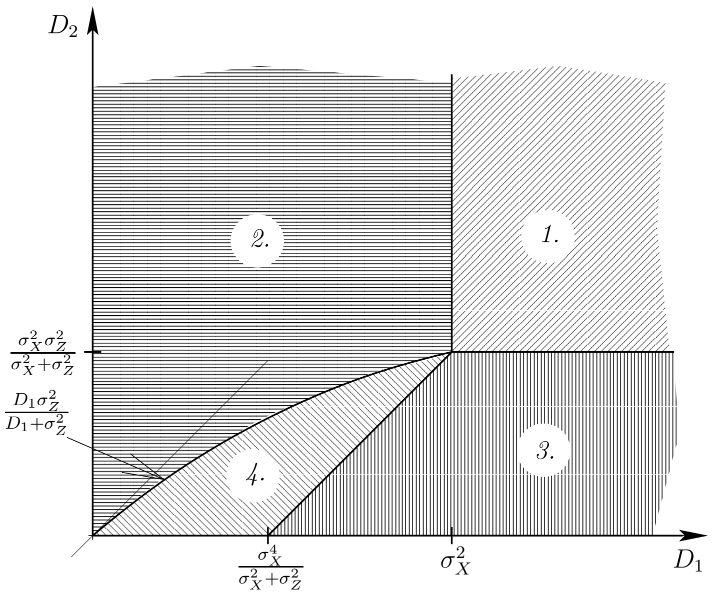

Figure 2: Graphical representation of the different cases of Theorem 1.

The proof for any of the cases

$$\sigma_X^2 \le D_1,$$

or

$$\frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2+\sigma_Z^2} \le D_2$$

is trivial. Therefore, we omit it and assume henceforth that neither of the above inequalities holds.

Depending on the values of D_1, D_2, σ_X^2 and σ_Z^2, the optimal rate is achieved by one of two different approaches. In the following two sections we outline the approaches and derive the corresponding rate-distortion functions. In the first approach, the encoder only produces a description for Decoder 2 (the first term of the Kaspi rate (1) is zero), in the second approach it produces a description only for Decoder 1 (the second term of the Kaspi rate is zero).

3 First Approach

A trivial lower bound is

$$R(D_1, D_2) \geq \frac{1}{2} \log \frac{\sigma_X^2 \sigma_Z^2}{D_2(\sigma_X^2 + \sigma_Z^2)}. \tag{2}$$

To see when this lower bound can be achieved, consider the following scheme: To satisfy the second decoder, the encoder computes the innovation

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

where

$$q = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}.$$

The encoder then describes F to Decoder 2 such that the expected distortion between the representation $\hat{F}$ and F is equals to D_2. The innovation F is Gaussian, and we know from single-user rate-distortion theory that the optimal representation $\hat{F}$ can be written using the forward test channel:

$$\hat{F} = \beta F + V,$$

where

$$\beta = \frac{\sigma_F^2 - D_2}{\sigma_F^2},$$

V is independent of F and $V \sim \mathcal{N}(0, \beta D_2)$. The output of the decoder is computed as $\hat{X}_2 = qY - \hat{F}$ and achieves an average distortion of D_2. It can be shown that this scheme requires exactly the rate given in (2). For more details see Appendix A.

Observe now that

$$\begin{aligned}\hat{F} &= \beta F + V \\ &= \beta(q-1)X + \beta q Z + V \\ &= \beta(q-1)[X + \frac{1}{\beta(q-1)}(\beta q Z + V)] \\ &= \beta(q-1)[X+U],\end{aligned}$$

where

$$U \triangleq \frac{1}{\beta(q-1)}(\beta q Z + V)$$

is Gaussian, independent of X, and has variance

$$\sigma_U^2 = \frac{1}{\beta^2(q-1)^2}(\beta^2 q^2 \sigma_Z^2 + \beta D_2).$$

One can show (e.g. using geometrical arguments) that in this innovation scheme for the conditional rate distortion problem, V should not only be independent of F, but also of Z. This fact has been used in the above expression for σ_U^2.

The representation of the innovation, $\hat{F}$, can also be recovered by Decoder 1. Since $\hat{F}$ carries information about X, it can be viewed as a side-information, helping Decoder 1 to estimate X.

To compute the variance of U in terms of the problem parameters, we first express β:

$$\begin{aligned}\beta &= 1 - \frac{D_2}{\sigma_F^2} \\ &= 1 - \frac{D_2}{(q-1)^2\sigma_X^2 + q^2\sigma_Z^2} \\ &= 1 - \frac{D_2}{\left(\frac{-\sigma_Z^2}{\sigma_X^2+\sigma_Z^2}\right)^2\sigma_X^2 + \left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}\right)^2\sigma_Z^2} \\ &= \frac{\sigma_Z^2\sigma_X^2 - D_2(\sigma_X^2+\sigma_Z^2)}{\sigma_X^2\sigma_Z^2}.\end{aligned}$$

After a few simplifications, we see that

$$\sigma_U^2 = \frac{1}{\sigma_Z^2}\left(\frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2+\sigma_Z^2} - D_2\right)^2 + \left(1 - D_2\frac{\sigma_X^2+\sigma_Z^2}{\sigma_X^2\sigma_Z^2}\right)D_2.$$

We can now express the variance of X given $\hat{F}$:

$$\begin{aligned}\mathrm{Var}\left(X|\hat{F}\right) &= \frac{\sigma_X^2\sigma_U^2}{\sigma_X^2+\sigma_U^2}\\ &= \ldots\\ &= D_2+\frac{\sigma_X^4}{\sigma_X^2+\sigma_Z^2}.\end{aligned}$$

If we want to achieve the lower bound (2), then the encoder must not send any additional message, i.e., Decoder 1 has to base its estimation of X solely on the "side-information" $\hat{F}$. This can be done if and only if

$$D_1 \geq \mathrm{Var}\left(X|\hat{F}\right) = D_2+\frac{\sigma_X^4}{\sigma_X^2+\sigma_Z^2}.$$

If this condition is satisfied, then we can achieve the rate

$$R(D_1,D_2)=\frac{1}{2}\log\frac{\sigma_X^2\sigma_Z^2}{D_2(\sigma_X^2+\sigma_Z^2)}. \tag{3}$$

4 Second Approach

4.1 Achievability

Assume now that

$$D_1 < \mathrm{Var}\left(X|\hat{F}_2\right) = D_2+\frac{\sigma_X^4}{\sigma_X^2+\sigma_Z^2} \tag{4}$$

and thus that the rate (3) of Section 3 can not be achieved.

Consider the following achievable scheme. The encoder first describes X to Decoder 1, which is optimally done through a representation

$$\hat{X}_1=\alpha X+W,$$

where

$$\alpha=\frac{\sigma_X^2-D_1}{\sigma_X^2}$$

and W is Gaussian, independent of X and has variance $\sigma_W^2=\alpha D_1$.

The representation $\hat{X}_1$ can also be recovered by Decoder 2, and can serve there as an additional side-information. If

$$\mathrm{Var}\left(X|Y\hat{X}_1\right) \leq D_2,$$

then we are done and the rate

$$R(D_1, D_2) = \frac{1}{2}\log\frac{\sigma_X^2}{D_1} \tag{5}$$

can be achieved (for details on how to derive this rate, see page 344 in [4]). On the other hand, if the variance of X given Y and $\hat{X}_1$ is too large, then the encoder can augment the usability of $\hat{X}_1$ for Decoder 2 by correlating it with Y. More precisely, the variable W can be correlated with the side-information noise Z. We quantify this using a correlation factor ρ defined as

$$\rho = \frac{1}{\sigma_W \sigma_Z}\mathrm{E}\left[WZ\right],$$

where $\rho \in [-1, 1]$. By setting $\rho = 0$, one obtains that

$$\mathrm{Var}\left(X|Y\hat{X}_1\right)\Big|_{\rho=0} = \frac{\sigma_Z^2 D_1}{\sigma_Z^2 + D_1}.$$

($\mathrm{Var}\left(X|Y\hat{X}_1\right)$ as a function of ρ will be given later in (7).) We conclude that if

$$\frac{\sigma_Z^2 D_1}{\sigma_Z^2 + D_1} \leq D_2,$$

then the rate (5) can be achieved and is trivially optimal. Hence, we assume from now on that $\frac{\sigma_Z^2 D_1}{\sigma_Z^2 + D_1} > D_2$.

From the point of view of Decoder 1, the statistics of $\hat{X}_1$ are the same for any value of ρ. This could lead to the wrong assumption that the rate required to produce $\hat{X}_1$ is also the same for any ρ. This is wrong, because the choice of W depends now on a given realization of $Y = X + Z$, and the encoder has less freedom when choosing the representation of X. It is obvious that using such a correlation technique, an arbitrarily small distortion can be achieved at Decoder 2 without requiring an additional message. In the extreme case, when $\rho = -1$, $D_2 = 0$ can be achieved. (This would require an infinite rate.)

We will now derive the rate required by this scheme. For a given $\hat{X}_1$ (specified by a given correlation factor ρ),

$$\begin{aligned} R(D_1, D_2) &= I(XY; \hat{X}_1) \\ &= I(X; \hat{X}_1) + I(Y; \hat{X}_1 | X). \end{aligned}$$

The first term is obviously

$$I(X; \hat{X}_1) = \frac{1}{2} \log \frac{\sigma_X^2}{D_1}.$$

The second term can be further decomposed as

$$I(Y; \hat{X}_1 | X) = H(Y|X) - H(Y|X\hat{X}_1),$$

and $H(Y|X) = H(Z) = \frac{1}{2}\log(2\pi e \sigma_Z^2)$. To compute $H(Y|X\hat{X}_1)$, we write

$$Z = tW + T,$$

where W is the additive variable in $\hat{X}_1 = \alpha X + W$, t is a real number, and where T is Gaussian and independent of W. To express the variance of T, start by expressing t:

$$\mathrm{E}\left[ZW\right] = t\sigma_W^2 = \rho \sigma_Z \sigma_W,$$

end hence,

$$t = \rho \frac{\sigma_Z}{\sigma_W}.$$

The variance of Z is

$$\begin{aligned} \sigma_Z^2 &= t^2 \sigma_W^2 + \sigma_T^2 \\ &= \rho^2 \sigma_Z^2 + \sigma_T^2 \end{aligned}$$

and thus,

$$\sigma_T^2 = (1 - \rho^2)\sigma_Z^2.$$

Now we can see that

$$\begin{aligned}
H(Y|X\hat{X}_1) &= H(Y-X|X,\hat{X}_1-\alpha X)\\
&= H(Z|X,\hat{X}_1-\alpha X)\\
&= H(Z|X,W)\\
&= H(Z|W)\\
&= H(Z-tW|W)\\
&= H(T|W)\\
&= H(T)\\
&= \frac{1}{2}\log(2\pi e(1-\rho^2)\sigma_Z^2).
\end{aligned}$$

Hence, the rate required to produce the given $\hat{X}_1$ is

$$\begin{aligned}
R(D_1,D_2) &= \frac{1}{2}\log\frac{\sigma_X^2}{D_1}+\frac{1}{2}\log(2\pi e\sigma_Z^2)-\frac{1}{2}\log(2\pi e(1-\rho^2)\sigma_Z^2)\\
&= \frac{1}{2}\log\frac{\sigma_X^2}{D_1(1-\rho^2)}. \qquad (6)
\end{aligned}$$

To achieve this rate, ρ should be such that

$$\mathrm{Var}\left(X|Y\hat{X}_1\right)\le D_2,$$

i.e., $\hat{X}_2$ shoud be a function of $\hat{X}_1$ and Y and hence, no additional message should be required for Decoder 2. We derive the conditional variance

$$\mathrm{Var}\left(X|Y\hat{X}_1\right)=\mathrm{E}\left[(X-\tilde{X})^2\right],$$

where $\tilde{X}$ is the best estimate of X given Y and $\hat{X}_1$:

$$\tilde{X}=\mathrm{E}\left[X\left[\begin{array}{cc} Y & \hat{X}_1\end{array}\right]\right]\quad \mathrm{E}\left[\left[\begin{array}{c} Y\\ \hat{X}_1\end{array}\right]\left[\begin{array}{cc} Y & \hat{X}_1\end{array}\right]\right]^{-1}\quad\left[\begin{array}{c} Y\\ \hat{X}_1\end{array}\right]$$

Since $X-\tilde{X}$ is independent of $\tilde{X}$,

$$\begin{aligned}
\mathrm{Var}\left(X|Y\hat{X}_1\right) &= \mathrm{E}\left[X(X-\tilde{X})\right]\\
&= \sigma_X^2-\mathrm{E}\left[X\left[\begin{array}{cc} Y & \hat{X}_1\end{array}\right]\right]\quad K^{-1}\quad \mathrm{E}\left[X\left[\begin{array}{c} Y\\ \hat{X}_1\end{array}\right]\right],
\end{aligned}$$

where

$$\begin{aligned} K &= \mathrm{E}\left[\left[\begin{array}{c} Y \\ \hat{X}_1 \end{array}\right]\left[\begin{array}{cc} Y & \hat{X}_1 \end{array}\right]\right] \\ &= \left[\begin{array}{cc} \sigma_X^2+\sigma_Z^2 & \alpha\sigma_X^2+\rho\sigma_Z\sigma_W \\ \alpha\sigma_X^2+\rho\sigma_Z\sigma_W & \alpha^2\sigma_X^2+\sigma_W^2 \end{array}\right] \end{aligned}$$

and

$$\mathrm{E}\left[X\left[\begin{array}{c} Y \\ \hat{X}_1 \end{array}\right]\right]=\sigma_X^2\left[\begin{array}{c} 1 \\ \alpha \end{array}\right].$$

This yields, after some algebra,

$$\mathrm{Var}\left(X|Y\hat{X}_1\right)=\frac{\sigma_X^2(1-\rho^2)\sigma_Z^2\sigma_W^2}{\det K}, \tag{7}$$

where

$$\det K=(1-\rho^2)\sigma_Z^2\sigma_W^2-2\alpha\sigma_Z\sigma_W\sigma_X^2\rho+\sigma_X^2(\sigma_W^2+\alpha^2\sigma_Z^2).$$

Let ρ_0 be the correlation factor that is best in terms of rate and for which

$$\mathrm{Var}\left(X|Y\hat{X}_1\right)\le D_2.$$

While satisfying the above inequality, ρ_0 should minimize the rate (6). One can observe that the conditional variance $\mathrm{Var}\left(X|Y\hat{X}_1\right)$ decreases with decreasing ρ. Since we assume that for $\rho=0$, $\mathrm{Var}\left(X|Y\hat{X}_1\right)>D_2$, we conclude that ρ_0 has to be negative. However, as $\rho<0$ decreases, its absolute value increases, and so does the rate given in (6). Therefore, ρ_0 should be the largest ("least negative") ρ such that $\mathrm{Var}\left(X|Y\hat{X}_1\right)\le D_2$, in other words, ρ_0 is the largest negative solution to

$$\mathrm{Var}\left(X|Y\hat{X}_1\right)=D_2. \tag{8}$$

The rate achieved by this scheme is then

$$R(D_1,D_2)=\frac{1}{2}\log\frac{\sigma_X^2}{D_1(1-\rho_0^2)}. \tag{9}$$

To find a closed form expression for ρ_0, one has to solve (8), which simplifies to the following quadratic equation:

$$\begin{aligned} f(\rho) &= \rho^2\sigma_Z^2(D_2-\sigma_X^2)D_1+\rho 2\sigma_X D_2\sigma_Z\sqrt{D_1(\sigma_X^2-D_1)} \\ &\quad +\sigma_Z^2(\sigma_X^2-D_2)D_1-\sigma_X^2 D_1 D_2-\sigma_Z^2 D_2(\sigma_X^2-D_1) \\ &= 0. \end{aligned}$$

Since $\sigma_Z^2(D_2-\sigma_X^2)D_1 \leq 0$, the function $f(\rho)$ has a global maximum at

$$\begin{aligned} \rho &= \frac{\sigma_X D_2\sqrt{D_1(\sigma_X^2-D_1)}}{\sigma_Z D_1(\sigma_X^2-D_2)} \\ &\geq 0. \end{aligned}$$

Since the maximum of the function $f(\rho)$ occurs for positive ρ, there can be at most one negative solution to $f(\rho)=0$. However, we assume that for $\rho=0$, $\mathrm{Var}\left(X|Y\hat{X}_1\right) > D_2$, and therefore that $f(0)>0$. Thus, there is exactly one $\rho<0$ for which $f(\rho)=0$ (and exactly one $\rho>0$ for which $f(\rho)=0$). For reasons mentioned earlier, the solution we pick for ρ_0 should be negative, and we obtain

$$\rho_0=\frac{\sigma_X D_2}{\sigma_Z(\sigma_X^2-D_2)D_1}\left[\sqrt{D_1(\sigma_X^2-D_1)}-\sqrt{(\sigma_X^2\frac{D_1}{D_2}(\sigma_Z^2-D_2)-\sigma_Z^2 D_1)(\frac{D_1}{D_2}-1)}\right],$$

that is the unique negative solution to $f(\rho)=0$.

For completeness, it can be shown that if

$$D_2+\frac{\sigma_X^4}{\sigma_X^2+\sigma_Z^2}=D_1$$

(condition (4) is satisfied with equality), then

$$\frac{\sigma_X^2}{D_1(1-\rho_0^2)}=\frac{\sigma_X^2\sigma_Z^2}{D_2(\sigma_X^2+\sigma_Z^2)}.$$

In other words, at the boundary between the first (Section 3) and the second approach (this Section), the two rate distortion functions (3) and (9) coincide.

4.2 Converse

In the following, we prove optimality of (9) under the conditions $D_1 < D_2 + \frac{\sigma_X^4}{\sigma_X^2+\sigma_Z^2}$ and $D_2 < \frac{\sigma_Z^2 D_1}{\sigma_Z^2+D_1}$.

4.2.1 Gaussianity

Lemma 2 *The optimal distribution of* $(\hat{X}_1\hat{X}_2XY)$ *is Gaussian.*

Proof: Assume that a distribution $p(\hat{X}_1\hat{X}_2XY)$ of a certain correlation matrix Σ is given to us. This distribution achieves a certain rate-distortion-triple (R, D_1, D_2). Note that the distortions

$$D_i = \mathrm{E}\left[(\hat{X}_i - X)^2\right], \qquad i = 1, 2,$$

depend only on the correlation Σ (but not on the actual distribution of the $\hat{X}_i$). Call $\gamma(\hat{X}_1\hat{X}_2XY)$ the jointly Gaussian distribution with zero mean and correlation Σ. If we replace p by γ, the distortions do not change.

The rate can be written as

$$\begin{aligned} R &= I(XY;\hat{X}_1) + I(X;\hat{X}_2|\hat{X}_1Y) \\ &= h(XY) - h(XY|\hat{X}_1) + h(X|\hat{X}_1Y) - h(X|\hat{X}_1\hat{X}_2Y) \\ &= h(Y) + h(X|Y) - h(Y|\hat{X}_1) - h(X|\hat{X}_1\hat{X}_2Y) \\ &= I(\hat{X}_1;Y) + I(X;\hat{X}_1\hat{X}_2|Y). \end{aligned}$$

We will investigate the effect of replacing p by γ on each of the above terms.

$$\begin{aligned} I(\hat{X}_1;Y) &= h(Y) - h(Y|\hat{X}_1) \\ &\overset{(a)}{=} h(Y) - h(Y - s\hat{X}_1|\hat{X}_1) \\ &\geq h(Y) - h(Y - s\hat{X}_1) \\ &\geq h(Y) - \frac{1}{2}\log(2\pi e\mathrm{Var}\left(Y - s\hat{X}_1\right)), \end{aligned}$$

where (a) is true for any value of $s \in \mathbb{R}$. We pick s to be the best linear estimator of Y given the Gaussian version of $\hat{X}_1$. In this case, both inequalities become equalities if and only if $\hat{X}_1$ is Gaussian. Hence, replacing p by γ decreases the mutual information. Similarly,

$$\begin{aligned} I(X;\hat{X}_1\hat{X}_2|Y) &= h(X|Y) - h(X|\hat{X}_1\hat{X}_2Y) \\ &= h(X|Y) - h(X - u\hat{X}_1 - v\hat{X}_2 - wY|\hat{X}_1\hat{X}_2Y) \\ &\geq h(X|Y) - h(X - u\hat{X}_1 - v\hat{X}_2 - wY) \\ &\geq h(X|Y) - \frac{1}{2}\log\left(2\pi e\mathrm{Var}\left(X - u\hat{X}_1 - v\hat{X}_2 - wY\right)\right), \end{aligned}$$

where (u, v, w) are the best estimators of X given the Gaussian version of $(\hat{X}_1, \hat{X}_2, Y)$. Again the inequalities become equalities if and only if we replace p by γ.

Therefore, replacing p by γ decreases both terms of R, and hence, the optimal distribution of $(\hat{X}_1 \hat{X}_2 X Y)$ must be Gaussian. □

4.2.2 Optimal Form of $\hat{X}_2$

Lemma 3 *Assume that $\hat{X}_1$ has been fixed to have a given joint Gaussian distribution with X and Y. Then, for a given distortion requirement D_2, the optimal distribution of $\hat{X}_2$ is given by solving the conditional rate-distortion problem for Decoder 2 considering $(Y, \hat{X}_1)$ as side information. The rate that corresponds to such an optimal $\hat{X}_2$ is*

$$R(D_1, D_2) = I(XY; \hat{X}_1) + \frac{1}{2} \log \frac{\operatorname{Var}\left(X|Y\hat{X}_1\right)}{D_2}.$$

Proof: For a fixed $\hat{X}_1$, the optimization that remains to carry out is

$$\begin{aligned} R(D_1, D_2) &= \min \left[I(XY; \hat{X}_1) + I(X; \hat{X}_2 | Y\hat{X}_1) \right] \\ &= I(XY; \hat{X}_1) + \min I(X; \hat{X}_2 | Y\hat{X}_1), \end{aligned}$$

where the minimization is over all joint distributions of $\hat{X}_2$ (with respect to $(\hat{X}_1, X, Y)$) that satisfy the distortion constraint

$$\mathrm{E}\left[(\hat{X}_2 - X)^2 \right] \leq D_2.$$

The second term of $R(D_1, D_2)$, together with the above condition, is the same optimization problem as the conditional rate distortion problem described in Appendix A, and thus, the optimal distribution must be of the type stated in the lemma. □

4.2.3 Optimal Form of $\hat{X}_1$

Lemma 4 *The optimal distribution of $\hat{X}_1$ is such that*

$$(\hat{X}_1 - X) \perp \hat{X}_1.$$

Proof: Assume that some distribution on $\hat{X}_1$ is given and that that $(\hat{X}_1 - X)$ is not independent of $\hat{X}_1$. The given distribution achieves a certain rate-distortion triple (R, D_1, D_2). Scaling $\hat{X}_1$ by some factor a will not change either of the mutual informations in the rate expression. Furthermore, whatever Decoder 2 can do using $\hat{X}_1$, it can also do using $a\hat{X}_1$. However, scaling $\hat{X}_1$ could influence the distortion D_1. In particular,

$$\begin{aligned} \frac{\partial}{\partial a} \mathrm{E}\left[(a\hat{X}_1 - X)^2\right]\Big|_{a=1} &= 2\mathrm{E}\left[\hat{X}_1(a\hat{X}_1 - X)\right]\Big|_{a=1} \\ &= 2\mathrm{E}\left[\hat{X}_1(\hat{X}_1 - X)\right], \end{aligned}$$

which is non-zero by assumption. Therefore, the distortion D_1 can be lowered through scaling of $\hat{X}_1$ and thus, the given distribution is not optimal.

□

4.2.4 Optimal Correlation

We know now that the optimal representation at Decoder 1 is such that $(\hat{X}_1 - X) \perp \hat{X}_1$ and hence satisfies the forward test channel relationship

$$\hat{X}_1 = \alpha X + W,$$

where $\alpha = \frac{\sigma_X^2 - D_1}{\sigma_X^2}$. We also know that the variance of the independent term W is equals to $\sigma_W^2 = D_1 \alpha$. Thus, the only choice that is left is how strongly W is correlated with Z. This choice can be expressed using the correlation factor ρ, that we already introduced in Section 4.1.

Depending on the correlation factor ρ, the variance of X given $(Y, \hat{X}_1)$ might or might not be smaller than D_2. If $\mathrm{Var}\left(X|Y\hat{X}_1\right) > D_2$, an additional message needs to be produced by the encoder, in order to enable Decoder 2 to estimate X within the required distortion bound. As shown in Lemma 3, the additional rate needed is

$$\frac{1}{2} \log \frac{\mathrm{Var}\left(X|Y\hat{X}_1\right)}{D_2}.$$

This variance as well as the expression of the first rate-term have been given in Section 4.1. The total rate is

$$\begin{aligned} R(D_1, D_2) &= \frac{1}{2}\log\frac{\sigma_X^2}{D_1(1-\rho^2)} + \frac{1}{2}\log\frac{\sigma_X^2(1-\rho^2)\sigma_Z^2\sigma_W^2}{D_2 \det K} \\ &= \frac{1}{2}\log\frac{\sigma_X^4\sigma_Z^2\sigma_W^2}{D_1 D_2 \det K}. \end{aligned} \tag{10}$$

The only free parameter is ρ. In the above rate expression, only the determinant of the covariance matrix K depends on ρ. Hence, to minimize the rate, we should maximize the determinant

$$(\det K)(\rho) = (1-\rho^2)\sigma_Z^2\sigma_W^2 - 2\alpha\sigma_Z\sigma_W\sigma_X^2\rho + \sigma_X^2(\sigma_W^2 + \alpha^2\sigma_Z^2).$$

One finds that $(\det K)(\rho)$ has a local maximum at

$$\rho^* = -\frac{\sigma_X\sqrt{\sigma_X^2 - D_1}}{\sigma_Z\sqrt{D_1}}.$$

One can easily verify that the slope of $(\det K)(\rho)$ is negative at $\rho = 0$, and therefore $\rho^* < 0$. Figure 3 shows a possible shape of $(\det K)$ as a function of ρ.

One might be tempted to think that ρ^* is the optimal correlation factor. However, we suggest in Section 4.1 that the optimal correlation factor be ρ_0 (provided that $D_2 < \frac{\sigma_Z^2 D_1}{\sigma_Z^2 + D_1}$), which is the ρ for which $\mathrm{Var}\left(X|Y\hat{X}_1\right) = D_2$. Note that if $\rho < \rho_0$, then the second rate term is 0 (it can not become negative), and the objective function is no longer (10) but only the first term of (10). Thus, the optimal value of ρ is certainly no smaller than ρ_0, and we find that the optimum is

$$\rho_{\text{opt}} = \max\{\rho^*, \rho_0\}.$$

The question is now whether ρ_0 can be smaller than ρ^*. To answer this question, we evaluate $\mathrm{Var}\left(X|Y\hat{X}_1\right)$ at ρ^*:

$$\begin{aligned} \mathrm{Var}\left(X|Y\hat{X}_1\right)\Big|_{\rho=\rho^*} &= \frac{\sigma_X^2(1-(\rho^*)^2)\sigma_Z^2\sigma_W^2}{(\det K)(\rho^*)} \\ &= \dots \\ &= \frac{\left(\sigma_Z^2 D_1 - \sigma_X^2(\sigma_X^2 - D_1)\right)\sigma_X^2}{\sigma_Z^2 D_1 - \sigma_X^2(\sigma_X^2 - D_1) + \sigma_X^2(2\sigma_X^2 - D_1) + \sigma_Z^2(\sigma_X^2 - D_1)}. \end{aligned}$$

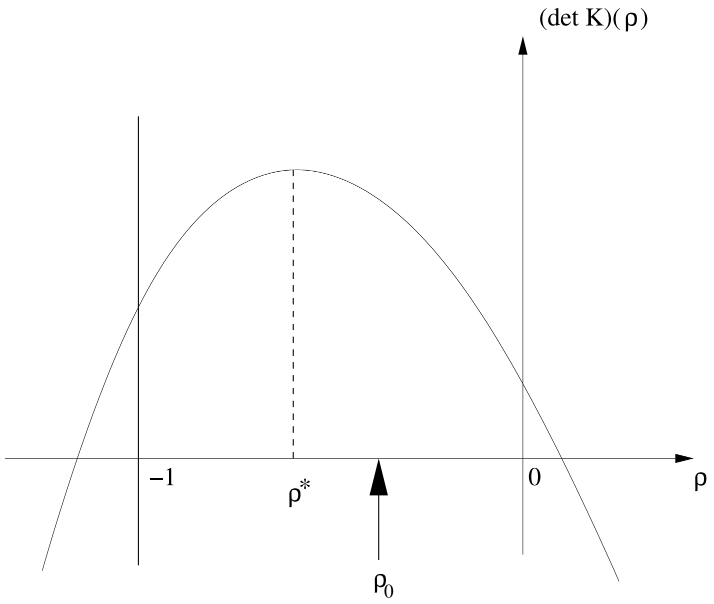

Figure 3: The determinant of K as a function of the correlation factor ρ.

We now derive a condition for $\mathrm{Var}\left(X|Y\hat{X}_1\right)\Big|_{\rho=\rho^*} < D_2$, i.e., a condition for negativity of the second rate term at $\rho = \rho^*$. This condition will tell us in what case $\rho_{\mathrm{opt}} = \rho_0$.

$$\begin{aligned}
\frac{\left(\sigma_Z^2 D_1 - \sigma_X^2(\sigma_X^2 - D_1)\right)\sigma_X^2}{\sigma_Z^2 D_1 - \sigma_X^2(\sigma_X^2 - D_1) + \sigma_X^2(2\sigma_X^2 - D_1) + \sigma_Z^2(\sigma_X^2 - D_1)} &< D_2 \\
1 + \frac{\sigma_X^2(2\sigma_X^2 - D_1) + \sigma_Z^2(\sigma_X^2 - D_1)}{\sigma_Z^2 D_1 - \sigma_X^2(\sigma_X^2 - D_1)} &> \frac{\sigma_X^2}{D_2} \\
\frac{\sigma_X^2(\sigma_X^2 + \sigma_Z^2)}{\sigma_Z^2 D_1 - \sigma_X^4 + \sigma_X^2 D_1} &> \frac{\sigma_X^2}{D_2} \\
D_2 &> D_1 - \frac{\sigma_X^4}{\sigma_X^2 + \sigma_Z^2}.
\end{aligned}$$

Surprisingly, we obtain the same condition as in (4). This means that, provided that the approach of Section 3 is not feasible, ρ_0 is always larger than

ρ^*.

Therefore,

$$\rho_{\text{opt}} = \rho_0,$$

which means that at the optimum, the second rate term is always zero. Hence, the rate (9) is optimal.

For completeness, we mention the following property. One can easily verify that whenever

$$\sigma_Z^2 \leq \frac{\sigma_X^2(\sigma_X^2 - D_1)}{D_1},$$

then the slope of $(\det K)(\rho)$ is non-positive at $\rho = -1$ and therefore, $\rho^* \leq -1$. Since in this case ρ^* is not a valid correlation, we can directly argue that $\rho_{\text{opt}} = \rho_0$. If σ_Z^2 satisfies the above inequality, we also see that the condition

$$D_1 < \text{Var}\left(X|\hat{F}_2\right) = D_2 + \frac{\sigma_X^4}{\sigma_X^2 + \sigma_Z^2}$$

becomes

$$D_1 < D_2 + D_1$$

and is satisfied for any value of D_2.

5 The Benefit of Encoder Side-Information

In [3], Kaspi also gives the rate-distortion function for what he calls the case "switch A open", i.e., the case where the encoder has no access to the side-information Y. In [2], Heegard and Berger independently derive the same result (Theorem 2 in [3]), and they explicitly solve the minimization for the Gaussian case. Figures 4 and 5 show a numerical comparison of the Heegard-Berger Gaussian rate to the rate of Theorem 1. The numerical values used are $\sigma_X^2 = 3$, $D_1 = 2$ and $D_2 = 1$. The variance of the side-information noise varies. In Figure 4, one can see that for $\sigma_Z^2 \leq 2$, the two rates coincide. This corresponds to the regime where

$$\text{Var}\left(X|\hat{X}_1 Y\right) \leq D_2$$

for a description $\hat{X}_1$ that is independent of Y ($\rho = 0$), and therefore, $R(D_1, D_2) = \frac{1}{2}\log\frac{\sigma_X^2}{D_1}$. This rate can be achieved by an encoder that does not use Y and is therefore achievable in both cases (informed and uninformed encoder).

However, for all other values of σ_Z^2, the informed encoder achieves a strictly lower rate than the uninformed encoder. This result is somewhat unexpected since for several known source-coding problems, e.g. the Wyner-Ziv setup, it is known that in the Gaussian case, encoder side-information does not improve the rate-distortion curve.

Figure 5 displays the same two curves for a higher range of the noise-variance. Naturally, as the noise variance σ_Z^2 goes to infinity, the side-information becomes useless. Therefore, the two rate curves converge to the same limit, that is the single-user rate-distortion function for the decoder that has the more severe constraint, i.e., $R(D_1, D_2) = \frac{1}{2} \log \frac{\sigma_X^2}{D_2}$.

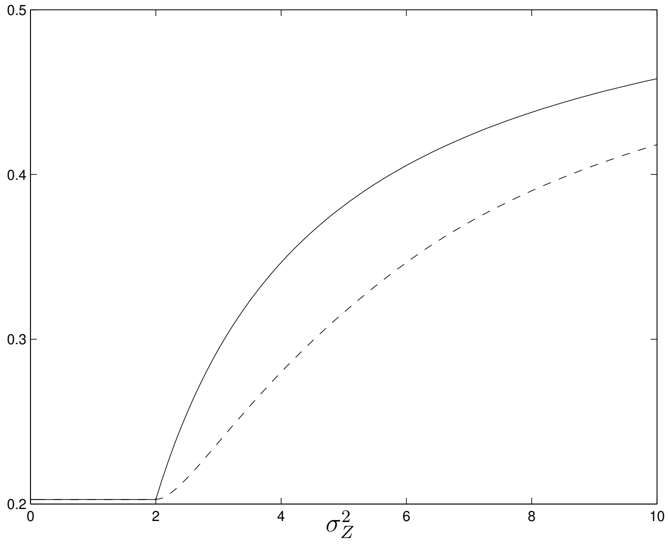

Figure 4: The rate in nats as a function of σ_Z^2 for informed (dashed line) and uninformed (solid line) encoder.

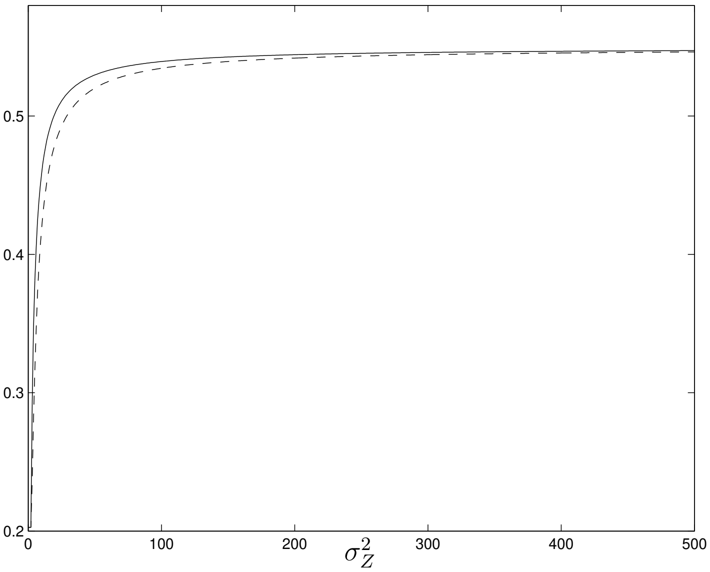

Figure 5: Asymptotic behaviour of the rate (in nats) for informed (dashed line) and uninformed (solid line) encoder.

A Gaussian Conditional Rate Distortion

It is known (cf. page 183 in [1]) that the rate distortion problem of a source X with side information Y available at the encoder and at the decoder is given by

$$R(D) = \min I(X; \hat{X}|Y),$$

where the minimization is over all distributions $p(X, Y, \hat{X})$ such that

$$\mathrm{E}\left[d(X, \hat{X})\right] \leq D.$$

We will carry out this minimization for the Gaussian case, where $X \sim \mathcal{N}(0, \sigma_X^2)$, $Y = X + Z$, $Z \sim \mathcal{N}(0, \sigma_Z^2)$ and independent of X. The result is of course well known and we derive it here only for the sake of completeness.

For any $\hat{X}$ that achieves a distortion D, the following lower bound holds.

$$\begin{aligned}
I(X; \hat{X}|Y) &= h(X|Y) - h(X|\hat{X}Y) \\
&= \frac{1}{2}\log(2\pi e \mathrm{Var}\,(X|Y)) - h(X|\hat{X}Y) \\
&= \frac{1}{2}\log(2\pi e \mathrm{Var}\,(X|Y)) - h(X - \hat{X}|\hat{X}Y) \\
&\geq \frac{1}{2}\log(2\pi e \mathrm{Var}\,(X|Y)) - h(X - \hat{X}) \\
&\geq \frac{1}{2}\log(2\pi e \mathrm{Var}\,(X|Y)) - \frac{1}{2}\log(2\pi e \mathrm{Var}\left(X - \hat{X}\right)) \\
&\geq \frac{1}{2}\log(2\pi e \mathrm{Var}\,(X|Y)) - \frac{1}{2}\log(2\pi e D) \\
&= \frac{1}{2}\log\frac{\mathrm{Var}\,(X|Y)}{D}
\end{aligned}$$

This lower bound can be achieved as follows: The encoder constructs the so-called innovation

$$F = \mathrm{E}\,[X|Y] - X,$$

where $\mathrm{E}\,[X|Y]$ is the best estimate of X given Y. The innovation F is Gaussian with zero mean and variance $\mathrm{Var}\,(X|Y)$. It can be viewed as a new source, and the encoder describes this source to the decoder in an optimal way at distortion D, i.e., the decoder will end up with a representation $\hat{F}$ s.t.

$$\mathrm{E}\left[(F - \hat{F})^2\right] \leq D.$$

The rate required to do so is

$$\begin{aligned} R &= \frac{1}{2}\log\frac{\mathrm{Var}\,(F)}{D} \\ &= \frac{1}{2}\log\frac{\mathrm{Var}\,(X|Y)}{D}. \end{aligned}$$

The decoder then recovers a representation of X:

$$\hat{X} = \mathrm{E}\,[X|Y] - \hat{F}.$$

The distortion achieved by this scheme is

$$\begin{aligned} \mathrm{E}\left[(X-\hat{X})^2\right] &= \mathrm{E}\left[\left((\mathrm{E}\,[X|Y]-F)-(\mathrm{E}\,[X|Y]-\hat{F})\right)^2\right] \\ &= \mathrm{E}\left[(F-\hat{F})^2\right] \\ &\leq D. \end{aligned}$$

Hence, this innovation-scheme achieves the lower bound that we gave earlier, which lets us conclude that the Gaussian conditional rate distortion function is

$$R(D) = \frac{1}{2}\log\frac{\mathrm{Var}\,(X|Y)}{D}.$$

References

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

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