LICOS-REPORT-2006-004

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

Etienne Perron[*], Suhas Diggavi, Emre Telatar

April 2, 2007

Abstract

We consider the problem of compressing a binary symmetric i.i.d. source stream for two decoders, one of which has access to side-information. Kaspi found the rate-distortion tradeoff for this problem for general 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. We assume that the source has a symmetric Bernoulli distribution and that the side-information sequence is the output of an erasure channel whose input is the source stream. We explicitly solve the optimization problem for this setup. We also show that the Wyner-Ziv rate distortion function and the conditional rate-distortion function coincide for general discrete memoryless sources when the side-information is of the erasure type.

1 Introduction

In [2], 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 for this setup and 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. In this report, we evaluate the Kaspi rate-distortion function with encoder side-information for the special case of a binary symmetric source and an erasure side-information. The distortion metric is the Hamming distance. In Section 2, we formally define a binary source and an erasure side-information. For this pair of sources, we first state the classical rate-distortion result in Section 3. In Section 4, we compute the conditional rate-distortion function for such a pair of sources, followed by the computation of the Wyner-Ziv rate-distortion function in Section 5, where we show that for a binary source with erasure side-information, making the side-information available at the encoder is not beneficial. This result is extended to general discrete memoryless sources in Appendix B. Section 6 contains our main result, namely the computation of the Kaspi rate-distortion function for binary sources and erasure side-information.

2 Definition of Source and Side-Information

In this writeup, we consider a binary source stream, given by a sequence $\{X_i\}_{i\in\mathbb{N}}$, where the X_i are independent and identically distributed (i.i.d.) with

$$X_i = \begin{cases} 0 & \text{with probability} & \frac{1}{2} \\ 1 & \text{with probability} & \frac{1}{2}. \end{cases} \tag{1}$$

We assume the existence of a second information stream, called the side-information, given by a sequence $\{Y_i\}_{i\in\mathbb{N}}$, where the Y_i are i.i.d. with

$$Y_i = \begin{cases} X_i & \text{with probability} & 1-p \\ \epsilon & \text{with probability} & p. \end{cases} \tag{2}$$

The symbol ϵ is called an erasure, and whenever $Y_i = \epsilon$, we say that the source symbol has been erased, meaning that all information about the value X_i is lost.

In the following, we propose several rate-distortion settings, where a decoder wishes to obtain a representation $\{\hat{X}_i\}_{i\in\mathbb{N}} \in \{0,1\}$ of $\{\hat{X}_i\}$ such that the average distortion between the two is no larger than a certain limit D. The distortion measure is the Hamming distance, *i.e.*,

$$d(x,\hat{x}) = x \oplus \hat{x},$$

where $\oplus$ denotes addition modulo-2.

3 Rate-Distortion Setup without Side-Information

First we assume that there is an encoder and only one decoder, neither of which has access to the side-information. It is well known (*e.g.* [3]) that the rate-distortion function for this setup is

$$R_X(D) = \begin{cases} 1 - h(D) & \text{if } D < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The optimal description $\hat{X}$ for this setup should be such that $\hat{X}$ is the input to a binary symmetric channel (BSC) whose output is X and that has crossover probability D. Since X is a symmetric source, one can also view $\hat{X}$ as being the output of the same BSC when X is the input, *i.e.*,

$$\hat{X} = \begin{cases} \bar{X} & \text{w.p. } D \\ X & \text{w.p. } 1 - D, \end{cases}$$

where, for a binary symbol x, we define

$$\bar{x} = \begin{cases} 0 & \text{if } x = 1 \\ 1 & \text{if } x = 0. \end{cases}$$

4 Conditional Rate-Distortion Setup

Assume now that the side-information source Y is available at the encoder and at the decoder. For general discrete memoryless sources, it is known ([1]) that

$$R_{X|Y}(D) = \min_{p_{\hat{X}|XY}} I(X; \hat{X}|Y),$$

where the minimization is over all distributions $p_{\hat{X}|XY}$ such that the distortion requirement is satisfied.

Theorem 1 *For a binary source and an erasure side-information as defined in* (1) *and* (2)*, and a Hamming distortion measure, the conditional rate-distortion function is*

$$R_{X|Y}(D) = \begin{cases} p\big(1 - h(\frac{D}{p})\big) & \textit{when } D < \frac{p}{2} \\ 0 & \textit{otherwise,} \end{cases}$$

where $h(\cdot)$ is the binary entropy function.

Appendix B contains a more general version of this theorem and its proof. Nevertheless, we provide a proof for the binary special case in this section.
Proof: First, note that whenever $D \geq \frac{p}{2}$, then distortion D can be achieved by a decoder that declares

$$\hat{X} = \begin{cases} Y & \text{if } Y \neq \epsilon \\ 0 & \text{otherwise.} \end{cases}$$

The expected distortion is indeed

$$\mathrm{E}\left[X \oplus \hat{X}\right] = \Pr\left\{X \neq \hat{X}\right\} = (1-p)0 + p\frac{1}{2} = \frac{p}{2} \leq D.$$

This scheme requires no transmission of information from the encoder, and thus, $R = 0$. Now, we treat the case $D < \frac{p}{2}$.

Converse: For any given $p_{\hat{X}|XY}$ that satisfies the distortion constraint, we find the following lower bound on $I(X;\hat{X}|Y)$:

$$\begin{aligned} I(X;\hat{X}|Y) &= H(X|Y) - H(X|\hat{X},Y) \\ &= \sum_{y\in\{0,1\}} \Pr\{Y=y\} \underbrace{H(X|Y=y)}_{=0} + p\underbrace{H(X|Y=\epsilon)}_{=1} - H(X|\hat{X},Y) \\ &= p - H(X|\hat{X},Y). \end{aligned}$$

We have that

$$\begin{aligned} H(X|\hat{X},Y) &= pH(X|\hat{X},Y=\epsilon) + (1-p)\frac{1}{2}H(X|\hat{X},Y=0) \\ &\quad + (1-p)\frac{1}{2}H(X|\hat{X},Y=1) \\ &\overset{(a)}{=} pH(X|\hat{X},Y=\epsilon) + (1-p)\frac{1}{2}H(X|\hat{X},X=0,Y=0) \\ &\quad + (1-p)\frac{1}{2}H(X|\hat{X},X=1,Y=1) \\ &= pH(X|\hat{X},Y=\epsilon) \\ &= pH(X\oplus\hat{X}|\hat{X},Y=\epsilon) \\ &\leq pH(X\oplus\hat{X}|Y=\epsilon) \\ &= ph\big(\Pr\left\{X\neq\hat{X}|Y=\epsilon\right\}\big), \end{aligned}$$

where (a) is true because when $Y \neq \epsilon$, then $Y = X$. Now, by assumption, $\hat{X}$ satisfies the distortion requirement, *i.e.*,

$$\begin{aligned} D &\geq \Pr\left\{X \neq \hat{X}\right\} \\ &= p\Pr\left\{X \neq \hat{X}|Y = \epsilon\right\} + (1-p)\Pr\left\{X \neq \hat{X}|Y \neq \epsilon\right\} \\ &\geq p\Pr\left\{X \neq \hat{X}|Y = \epsilon\right\}. \end{aligned}$$

Hence,

$$\Pr\left\{X \neq \hat{X}|Y = \epsilon\right\} \leq \frac{D}{p} < \frac{1}{2},$$

and we get the lower bound

$$\begin{aligned} I(X;\hat{X}|Y) &\geq p - ph\big(\Pr\left\{X \neq \hat{X}|Y = \epsilon\right\}\big) \\ &\geq p - ph(\frac{D}{p}), \end{aligned}$$

because for $x \in [0, \frac{1}{2}]$, $h(x)$ is an increasing function.

Achievability 1: The most intuitive optimal coding strategy is as follows. Given a pair of sequences $(\mathbf{x}, \mathbf{y})$ of length n, the encoder extracts a sequence $\tilde{\mathbf{x}}$, which is the subsequence of $\mathbf{x}$ indicated by the positions in $\mathbf{y}$ for which $y_i = \epsilon$. More precisely,

$$\tilde{x}_i = \begin{cases} x_i & \text{if } y_i = \epsilon \\ \phi & \text{otherwise,} \end{cases}$$

where $\tilde{x}_i = \phi$ means that position i is nonexistent. The subsequence $\tilde{\mathbf{x}}$ has an average length of pn. Now, the encoder describes the subsequence $\tilde{\mathbf{x}}$ at an average distortion $\frac{D}{p}$. The required rate is

$$\tilde{R} = 1 - h(\frac{D}{p}).$$

However, since the subsequence is only of length pn, the per-symbol rate for the n-length block is

$$R = p\big(1 - h(\frac{D}{p})\big).$$

The decoder, that observes $\mathbf{y}$, replaces all the erasures in $\mathbf{y}$ by the corresponding symbol in the decoded subsequence and declares the result to be $\hat{\mathbf{x}}$. The expected distortion is

$$\Pr\left\{X \neq \hat{X}\right\} = (1-p)0 + p\frac{D}{p} = D.$$

Achievability 2: We also present an alternative optimal coding strategy, which will be useful for the Kaspi setup in Section 6. Consider the binary auxiliary random variable W defined as follows:

$$W = \begin{cases} 0 & \text{w.p. } \frac{1}{2} & \text{if } Y \neq \epsilon \\ 1 & \text{w.p. } \frac{1}{2} & \text{if } Y \neq \epsilon \\ \bar{X} & \text{w.p. } \frac{D}{p} & \text{if } Y = \epsilon \\ X & \text{w.p. } 1-\frac{D}{p} & \text{if } Y = \epsilon. \end{cases}$$

For such a p_{XYW}, we generate $2^{nR-\epsilon}$ $\mathbf{w}$ sequences randomly from the marginal distribution p_W^n. The encoder, observing $(\mathbf{x}, \mathbf{y})$, sends the index of a $\mathbf{w}$ that is jointly typical with $(\mathbf{x}, \mathbf{y})$. It is well known that such a random code works with high probability if

$$\begin{aligned} R &= I(X,Y;W) \\ &= H(W) - H(W|X,Y) \\ &= h\big(\Pr\{W=1\}\big) - H(W|X,Y). \end{aligned}$$

We compute

$$\begin{aligned} \Pr\{W=1\} &= p\Pr\{W=1|Y=\epsilon\} + (1-p)\Pr\{W=1|Y\neq\epsilon\} \\ &= p\frac{1}{2} + (1-p)\frac{1}{2} \\ &= \frac{1}{2}, \end{aligned}$$

and hence,

$$H(W) = h(\frac{1}{2}) = 1.$$

Now, the second term of the rate is

$$
\begin{aligned}
H(W|X,Y) &= \frac{1}{2}p\big[H(W|Y=\epsilon, X=0) + H(W|Y=\epsilon, X=1)\big] \\
&\quad + \frac{1}{2}(1-p)\big[H(W|Y\neq\epsilon, X=0) + H(W|Y\neq\epsilon, X=1)\big] \\
&= \frac{1}{2}p\big[h(\frac{D}{p}) + h(1-\frac{D}{p})\big] \\
&\quad + \frac{1}{2}(1-p)\big[h(\frac{1}{2}) + h(\frac{1}{2})\big] \\
&= ph(\frac{D}{p}) + 1 - p.
\end{aligned}
$$

From this, it follows that

$$R = H(W) - H(W|X,Y) = p - ph(\frac{D}{p}).$$

The decoder declares the reconstruction

$$\hat{X} = \begin{cases} W & \text{if } Y = \epsilon \\ Y & \text{if } Y \neq \epsilon, \end{cases}$$

which again yields an expected distortion of

$$\Pr\left\{X \neq \hat{X}\right\} = (1-p)0 + p\frac{D}{p} = D.$$

This concludes the proof. □

5 Wyner-Ziv Setup

Now, we assume that there is one encoder and one decoder, but that the side-information Y is only available at the decoder. In [4], Wyner and Ziv have shown that in this case, the rate-distortion function for general discrete memoryless sources is

$$R(D) = \min_{p_{W|XY}} I(X;W|Y),$$

where the minimization is over all distributions of an auxiliary variable W such that

- $W -\!\!\circ\!\!- X -\!\!\circ\!\!- Y$ is a Markov chain,
- and there exists a decoding function $f : \mathcal{Y} \times \mathcal{W} \mapsto \hat{\mathcal{X}}$ such that the distortion requirement is satisfied.

It is known ([4]) that for jointly Gaussian sources and the squared error distortion measure, the rate-distortion functions for the Wyner-Ziv setup and the conditional setup are the same. The following theorem establishes this equivalence for the present setup with discrete sources.

Theorem 2 *For a binary source and an erasure side-information as defined in* (1) *and* (2)*, and a Hamming distortion measure, the Wyner-Ziv rate-distortion function is*

$$R(D) = \begin{cases} p\big(1 - h(\frac{D}{p})\big) & \text{when } D < \frac{p}{2} \\ 0 & \text{otherwise,} \end{cases}$$

and is therefore equal to the conditional rate-distortion function.

A more general version of this second theorem can be found in Appendix B.
Proof: For $D \geq \frac{p}{2}$, the argument in the proof of Theorem 1 is also valid here. Now, assume that $D < \frac{p}{2}$. Note that any scheme that achieves a certain (R, D) pair in the Wyner-Ziv setup can also be applied to the conditional setup (by ignoring Y at the encoder). Hence, the conditional rate-distortion function is certainly a lower bound to the rate-distortion tradeoff achievable by any scheme here. This establishes the converse proof.

Achievability: Choose the auxiliary variable W to be binary and distributed as the input of a binary symmetric channel (BSC) whose output is X and whose crossover probability is $\frac{D}{p}$. Since X is a symmetric source, W can also be considered as the output of the same BSC when X is the input, *i.e.*,

$$W = \begin{cases} \bar{X} & \text{w.p. } \frac{D}{p} \\ X & \text{w.p. } 1 - \frac{D}{p}. \end{cases}$$

In other words, W is a good lossy representation of X at distortion $\frac{D}{p}$. To reveal W to the decoder, we use random coding and binning as proposed by Wyner and Ziv. The rate of the scheme is then

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

where

$$H(X|Y) = p$$

and

$$\begin{aligned}
H(X|W,Y) &= pH(X|W,Y=\epsilon) + (1-p)\frac{1}{2}H(X|W,Y=0) \\
&\quad + (1-p)\frac{1}{2}H(X|W,Y=1) \\
&\stackrel{(a)}{=} pH(X|W,Y=\epsilon) + (1-p)\frac{1}{2}H(X|W,X=0,Y=0) \\
&\quad + (1-p)\frac{1}{2}H(X|W,X=1,Y=1) \\
&= pH(X|W,Y=\epsilon) \\
&\stackrel{(b)}{=} pH(X|W) \\
&= pH(X \oplus W|W) \\
&\stackrel{(c)}{=} pH(X \oplus W) \\
&= ph(\frac{D}{p}),
\end{aligned}$$

where

(a) holds because when $Y \neq \epsilon$, then $Y = X$,

(b) holds because $Y \multimap X \multimap W$, and thus,

$$\begin{aligned}
p_{X|WY}(x|w,\epsilon) &= \frac{p_{XWY}(x,w,\epsilon)}{p_{XWY}(0,w,\epsilon) + p_{XWY}(1,w,\epsilon)} \\
&= \frac{p_Y(\epsilon)p_{X|Y}(x|\epsilon)p_{W|X}(w|x)}{p_Y(\epsilon)\big(p_{X|Y}(0|\epsilon)p_{W|X}(w|0) + p_{X|Y}(1|\epsilon)p_{W|X}(w|1)\big)} \\
&= \frac{p_X(x)p_{W|X}(w|x)}{p_X(0)p_{W|X}(w|0) + p_X(1)p_{W|X}(w|1)} \\
&= p_{X|W}(x|w),
\end{aligned}$$

where the second-last equality is true because $p_{X|Y}(x|\epsilon) = p_X(x)$.

(c), finally, holds because $X \oplus W$ is independent of W by construction (X being a noisy version of W).

It follows that for our choice of W,

$$R = H(X|Y) - H(X|W,Y) = p - ph(\frac{D}{p}).$$

The decoder declares

$$\hat{X} = \begin{cases} W & \text{if } Y = \epsilon \\ Y & \text{otherwise.} \end{cases}$$

The expected distortion is then once more given by

$$\Pr\left\{X \neq \hat{X}\right\} = (1-p)0 + p\frac{D}{p} = D.$$

Hence, the proposed choice of W achieves a rate-distortion tradeoff that corresponds to the lower bound given by the conditional rate-distortion function. This proves the theorem. □

6 Kaspi Setup with Encoder Side-Information

Consider the setup depicted in Figure 1, which corresponds to the case "switch A closed" in [2]. In this scheme, Decoders 1 and 2 require an approximation of X at distortions no larger than D_1 and D_2, respectively. The encoder and Decoder 1 have access to the side-information source Y.

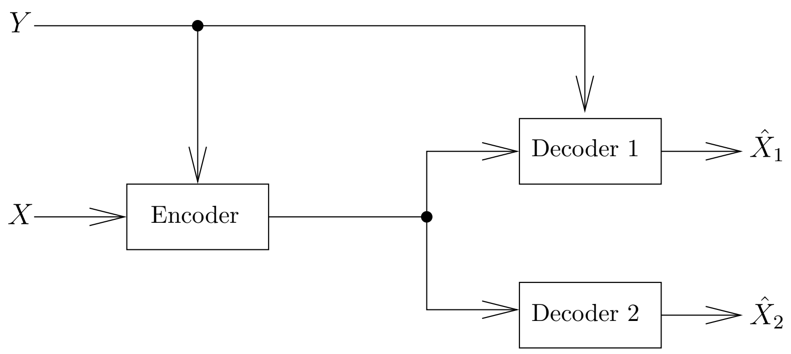

Figure 1: Kaspi Setup with Informed Encoder.

In [2], Kaspi found the rate-distortion function for this setup with general discrete memoryless sources. The smallest achievable rate in terms of the two distortion requirements is

$$R(D_1, D_2) = \min_{p_{\hat{X}_1 \hat{X}_2 | XY}} \big[I(X, Y; \hat{X}_2) + I(X; \hat{X}_1 | Y, \hat{X}_2) \big], \tag{4}$$

where the indicated minimization is over all distributions of $(\hat{X}_1, \hat{X}_2)$, jointly with (X, Y), such that the two distortion requirements are satisfied. We obtain the following result.

Theorem 3 *For a binary source and an erasure side-information as defined in* (1) *and* (2)*, and a Hamming distortion measure, the rate-distortion function for the Kaspi setup is given by the following four regimes.*

1) if $D_1 \geq \frac{p}{2}$ *and* $D_2 \geq \frac{1}{2}$*, then*

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

2) if $D_1 \geq pD_2$ *and* $D_2 < \frac{1}{2}$*, then*

$$R(D_1, D_2) = 1 - h(D_2),$$

3) if $D_1 < \frac{p}{2}$ *and* $D_2 \geq D_1 + \frac{1-p}{2}$*, then*

$$R(D_1, D_2) = p\big(1 - h(\frac{D_1}{p})\big),$$

4) otherwise

$$R(D_1, D_2) = 1 - h\big(\frac{D_2 - D_1}{1-p}\big) + p\Big(h\big(\frac{D_2 - D_1}{1-p}\big) - h\big(\frac{D_1}{p}\big)\Big).$$

The four regimes of Theorem 3 are depicted in Figure 2.

Proof: **Regime 1):** Here, no converse proof is needed since $R = 0$ is a natural lower bound on the rate. This rate is for instance achievable by the following decoding rules:

$$\hat{X}_1 = \begin{cases} Y & \text{if } Y \neq \epsilon \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\hat{X}_2 = 0.$$

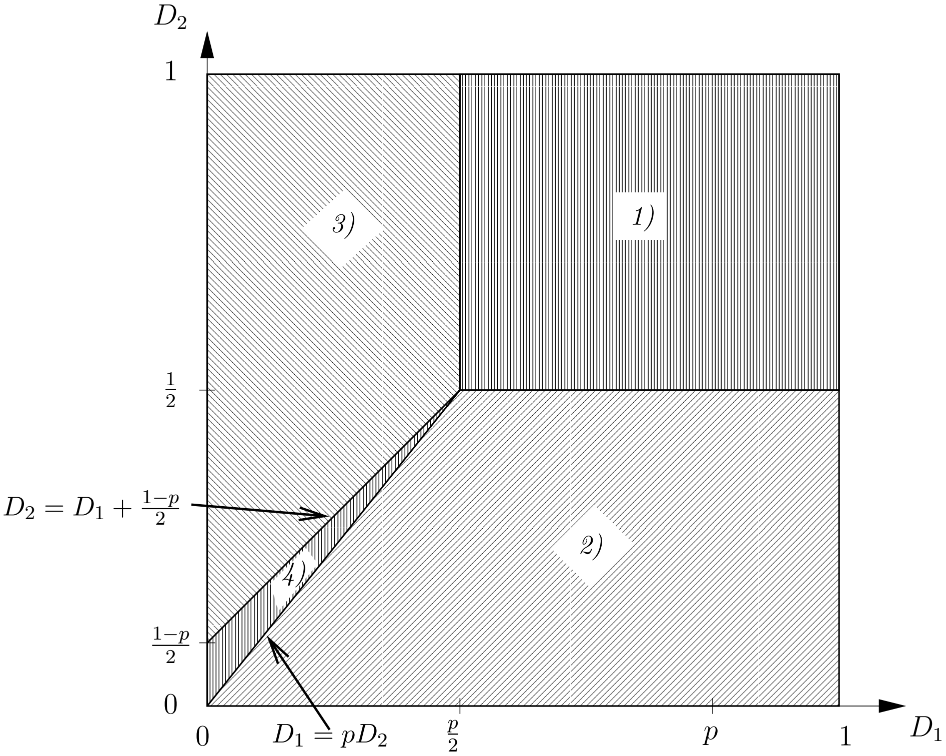

Figure 2: The regimes of Theorem 3.

As computed in Section 4, the first decoder can achieve a distortion of at best $\frac{D_1}{p}$. At the same time, the second decoder achieves an expected distortion of $\frac{1}{2}$.

Regime 2): No converse is needed since $R = 1 - h(D_2)$ would be the smallest rate if only Decoder 2 were to be satisfied. In this regime, it is indeed sufficient to use an encoder as in Section 3, *i.e.*,

$$\hat{X}_2 = \begin{cases} \bar{X} & \text{w.p. } D_2 \\ X & \text{w.p. } 1 - D_2, \end{cases}$$

The rate is then $R = 1 - h(D_2)$ and distortion D_2 is achieved at Decoder 2. Decoder 1, on the other hand, is also able to decode $\hat{X}_2$ and combines it with Y in the following way:

$$\hat{X}_1 = \begin{cases} Y & \text{if } Y \neq \epsilon \\ \hat{X}_2 & \text{if } Y = \epsilon. \end{cases}$$

The expected distortion at Decoder 1 is then

$$\Pr\left\{X \neq \hat{X}_1\right\} = (1-p)0 + pD_2 = pD_2.$$

Thus, the encoder need not send any additional information as long as $D_1 \geq pD_2$.

Regime 3): No converse is needed here, since $R = p(1 - h(\frac{D_1}{p}))$ would be the smallest required rate if only Decoder 1 were to be satisfied. In this regime, a conditional rate-distortion strategy as describe in Section 4 for Decoder 1 is sufficient. To increase the benefit of the encoded information for Decoder 2, we use the scheme described in **Achievability 2** in the proof of Theorem 1. In this scheme, the encoder describes the following auxiliary object to both decoders:

$$W = \begin{cases} 0 & \text{w.p. } \frac{1}{2} & \text{if } Y \neq \epsilon \\ 1 & \text{w.p. } \frac{1}{2} & \text{if } Y \neq \epsilon \\ \bar{X} & \text{w.p. } \frac{D_1}{p} & \text{if } Y = \epsilon \\ X & \text{w.p. } 1 - \frac{D_1}{p} & \text{if } Y = \epsilon. \end{cases}$$

We know from Section 4 that this strategy yields a distortion of D_1 at Decoder 1 and that it uses rate $p(1 - h(\frac{D_1}{p}))$. Decoder 2 declares

$$\hat{X}_2 = W.$$

The average distortion that it achieves by doing so is

$$\Pr\left\{X \neq \hat{X}_2\right\} = (1-p)\frac{1}{2} + p\frac{D_1}{p} = D_1 + \frac{1-p}{2}.$$

Thus, as long as $D_2 \geq D_1 + \frac{1-p}{2}$, this scheme is sufficient for Decoder 2.

Regime 4): To prove that in this regime, the specified rate-distortion function is optimal, we simply solve the minimization (4) explicitly. First note that whenever Decoder 1 observes $Y \neq \epsilon$, it should declare $\hat{X}_1 = Y$ with probability 1. In fact, assume that we are given a scheme without this property. Then we construct a new scheme by using the same encoder and decoders, however, before declaring $\hat{X}_1$, Decoder 1 replaces it by Y in case $Y \neq \epsilon$. The pair (R, D_2) of the scheme is unchanged. The distortion D_1, however, is lowered, because when $Y \neq \epsilon$, then $Y = X$.

Keeping this in mind, we parametrize the joint distribution of $(X, Y, \hat{X}_1, \hat{X}_2)$ in the following way:

x	y	$p_{\hat{X}_2\|XY}(1\|x,y)$
0	0	q_{00}
0	ϵ	$q_{0\epsilon}$
1	1	q_{11}
1	ϵ	$q_{1\epsilon}$

and

x	y	$\hat{x}_2$	$p_{\hat{X}_1\|XY\hat{X}_2}(1\|x,y,\hat{x}_2)$
0	0	0	0
0	0	1	0
0	ϵ	0	p_1
0	ϵ	1	p_2
1	1	0	1
1	1	1	1
1	ϵ	0	p_3
1	ϵ	1	p_4.

Claim 1: Without loss of generality, we can restrict the optimization (4) to distributions $p_{\hat{X}_1\hat{X}_2XY}$ that are symmetric in the sense that conditioned on a value of Y, $\hat{X}_2$ is the output of a BSC whose input is X, and that conditioned on a value of $(Y, X \oplus \hat{X}_2)$, $\hat{X}_1$ is the output of a different BSC (whose input is X). More precisely, we claim that we can restrict ourselves to distributions of the form

x	y	$p_{\hat{X}_2\|XY}(1\|x,y)$
0	0	a
0	ϵ	b
1	1	$1-a$
1	ϵ	$1-b$

and

x	y	$\hat{x}_2$	$p_{\hat{X}_1\|XY\hat{X}_2}(1\|x,y,\hat{x}_2)$
0	0	0	0
0	0	1	0
0	ϵ	0	c
0	ϵ	1	d
1	1	0	1
1	1	1	1
1	ϵ	0	$1-d$
1	ϵ	1	$1-c$.

Proof of Claim 1: We start off with a choice of $(q_{00}, q_{0\epsilon}, q_{11}, q_{1\epsilon}, p_1, p_2, p_3, p_4)$ and let $(\hat{X}_1, \hat{X}_2)$ denote the corresponding random variables. The rate-distortion triple achieved by this choice is

$$I(X,Y;\hat{X}_2) + I(X;\hat{X}_1|\hat{X}_2,Y),$$

$$\mathrm{E}\left[X \oplus \hat{X}_1\right]$$

and

$$\mathrm{E}\left[X \oplus \hat{X}_2\right].$$

Consider now new random variables $(\tilde{X}_1, \tilde{X}_2)$ given by the parameters

$$\begin{aligned}
\tilde{q}_{00} &= 1 - q_{11} \\
\tilde{q}_{0\epsilon} &= 1 - q_{1\epsilon} \\
\tilde{q}_{11} &= 1 - q_{00} \\
\tilde{q}_{1\epsilon} &= 1 - q_{0\epsilon} \\
\tilde{p}_1 &= 1 - p_4 \\
\tilde{p}_2 &= 1 - p_3 \\
\tilde{p}_3 &= 1 - p_2 \\
\tilde{p}_4 &= 1 - p_1.
\end{aligned}$$

We show now that the rate-distortion triple achieved by $(\tilde{X}_1, \tilde{X}_2)$ is equals to the triple achieved by $(\hat{X}_1, \hat{X}_2)$. We first compute the two mutual information terms in the rate:

$$I(X,Y;\tilde{X}_2) = h\big(p_{\tilde{X}_2}(1)\big) - H(\tilde{X}_2|X,Y),$$

where

$$\begin{aligned}
p_{\tilde{X}_2}(1) &= \frac{1}{2}\big((1-p)(\tilde{q}_{00} + \tilde{q}_{11}) + p(\tilde{q}_{0\epsilon} + \tilde{q}_{1\epsilon})\big) \\
&= \frac{1}{2}\big((1-p)(2 - q_{11} - q_{00}) + p(2 - q_{1\epsilon} - q_{0\epsilon})\big) \\
&= 1 - p_{\hat{X}_2}(1)
\end{aligned}$$

and

$$
\begin{aligned}
H(\tilde{X}_2|X,Y) &= \frac{1}{2}\Big((1-p)\big(h(\tilde{q}_{00})+h(\tilde{q}_{11})\big) \\
&\quad + p\big(h(\tilde{q}_{0\epsilon})+h(\tilde{q}_{1\epsilon})\big)\Big) \\
&= \frac{1}{2}\Big((1-p)\big(h(q_{11})+h(q_{00})\big) \\
&\quad + p\big(h(q_{1\epsilon})+h(q_{0\epsilon})\big)\Big) \\
&= H(\hat{X}_2|X,Y).
\end{aligned}
$$

Hence,

$$
I(X,Y;\tilde{X}_2) = I(X,Y;\hat{X}_2).
$$

Further,

$$
I(X;\tilde{X}_1|\tilde{X}_2,Y) = H(\tilde{X}_1|\tilde{X}_2,Y) - H(\tilde{X}_1|X,Y,\tilde{X}_2),
$$

where

$$
\begin{aligned}
H(\tilde{X}_1|X,Y,\tilde{X}_2) &= \frac{1}{2}p\big((1-\tilde{q}_{0\epsilon})h(\tilde{p}_1)+\tilde{q}_{0\epsilon}h(\tilde{p}_2) \\
&\quad + (1-\tilde{q}_{1\epsilon})h(\tilde{p}_3)+\tilde{q}_{1\epsilon}h(\tilde{p}_4)\big) \\
&= \frac{1}{2}p\big(q_{1\epsilon}h(p_4)+(1-q_{1\epsilon})h(p_3) \\
&\quad + q_{0\epsilon}h(p_2)+(1-q_{0\epsilon})h(p_1)\big) \\
&= H(\hat{X}_1|X,Y,\hat{X}_2).
\end{aligned}
$$

We also have

$$
\begin{aligned}
p_{\tilde{X}_1|Y\tilde{X}_2}(1|\epsilon,0) &= \frac{p_{\tilde{X}_1\tilde{X}_2|Y}(1,0|\epsilon)}{p_{\tilde{X}_2|Y}(0|\epsilon)} \\
&= \frac{\frac{1}{2}\big(p_{\tilde{X}_1\tilde{X}_2|XY}(1,0|0,\epsilon)+p_{\tilde{X}_1\tilde{X}_2|XY}(1,0|1,\epsilon)\big)}{\frac{1}{2}\big(p_{\tilde{X}_2|XY}(0|0,\epsilon)+p_{\tilde{X}_2|XY}(0|1,\epsilon)\big)} \\
&= \frac{(1-\tilde{q}_{0\epsilon})\tilde{p}_1+(1-\tilde{q}_{1\epsilon})\tilde{p}_3}{1-\tilde{q}_{0\epsilon}+1-\tilde{q}_{1\epsilon}} \\
&= \frac{q_{1\epsilon}(1-p_4)+q_{0\epsilon}(1-p_2)}{q_{1\epsilon}+q_{0\epsilon}} \\
&= \frac{q_{1\epsilon}+q_{0\epsilon}-q_{1\epsilon}p_4-q_{0\epsilon}p_2}{q_{1\epsilon}+q_{0\epsilon}} \\
&= 1-p_{\hat{X}_1|Y\hat{X}_2}(1|\epsilon,1) \\
&= p_{\hat{X}_1|Y\hat{X}_2}(0|\epsilon,1)
\end{aligned}
$$

and

$$
\begin{aligned}
p_{\tilde{X}_1|Y\tilde{X}_2}(1|\epsilon,1) &= \frac{\tilde{q}_{0\epsilon}\tilde{p}_2+\tilde{q}_{1\epsilon}\tilde{p}_4}{\tilde{q}_{0\epsilon}+\tilde{q}_{1\epsilon}} \\
&= \frac{(1-q_{1\epsilon})(1-p_3)+(1-q_{0\epsilon})(1-p_1)}{1-q_{1\epsilon}+1-q_{0\epsilon}} \\
&= 1-p_{\hat{X}_1|Y\hat{X}_2}(1|\epsilon,0) \\
&= p_{\hat{X}_1|Y\hat{X}_2}(0|\epsilon,0),
\end{aligned}
$$

and therefore,

$$
\begin{aligned}
H(\tilde{X}_1|Y,\tilde{X}_2) &= \frac{1}{2}p(2-\tilde{q}_{0\epsilon}-\tilde{q}_{1\epsilon})h\big(p_{\tilde{X}_1|Y\tilde{X}_2}(1|\epsilon,0)\big) \\
&\quad+\frac{1}{2}p(\tilde{q}_{0\epsilon}+\tilde{q}_{1\epsilon})h\big(p_{\tilde{X}_1|Y\tilde{X}_2}(1|\epsilon,1)\big) \\
&= \frac{1}{2}p(q_{0\epsilon}+q_{1\epsilon})h\big(p_{\hat{X}_1|Y\hat{X}_2}(0|\epsilon,1)\big) \\
&\quad+\frac{1}{2}p(2-q_{0\epsilon}-q_{1\epsilon})h\big(p_{\hat{X}_1|Y\hat{X}_2}(1|\epsilon,0)\big) \\
&= H(\hat{X}_1|Y,\hat{X}_2).
\end{aligned}
$$

This proves that the rate required for describing $(\tilde{X}_1,\tilde{X}_2)$ matches the rate for $(\hat{X}_1, \hat{X}_2)$. For the distortions achieved by $(\tilde{X}_1, \tilde{X}_2)$, we have

$$\begin{aligned}
\mathrm{E}\left[X \oplus \tilde{X}_1\right] &= \frac{1}{2}p\big((1-\tilde{q}_{0\epsilon})\tilde{p}_1 + \tilde{q}_{0\epsilon}\tilde{p}_2 \\
&\quad + (1-\tilde{q}_{1\epsilon})(1-\tilde{p}_3) + \tilde{q}_{1\epsilon}(1-\tilde{p}_4)\big) \\
&\quad + \frac{1}{2}(1-p)0 \\
&= \frac{1}{2}p\big(q_{1\epsilon}(1-p_4) + (1-q_{1\epsilon})(1-p_3) \\
&\quad + q_{0\epsilon}p_2 + (1-q_{0\epsilon})p_1\big) \\
&= \mathrm{E}\left[X \oplus \hat{X}_1\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[X \oplus \tilde{X}_2\right] &= \frac{1}{2}(1-p)\big(\tilde{q}_{00} + (1-\tilde{q}_{11})\big) \\
&\quad + \frac{1}{2}p\big(\tilde{q}_{0\epsilon} + (1-\tilde{q}_{1\epsilon})\big) \\
&= \frac{1}{2}(1-p)\big((1-q_{11}) + q_{00}\big) \\
&\quad + \frac{1}{2}p\big((1-q_{1\epsilon}) + q_{0\epsilon}\big) \\
&= \mathrm{E}\left[X \oplus \hat{X}_2\right].
\end{aligned}$$

From all the above, it follows that $(\tilde{X}_1, \tilde{X}_2)$ yield exactly the same rate-distortion triple as $(\hat{X}_1, \hat{X}_2)$. Now, we define new random variables $(\breve{X}_1, \breve{X}_2)$ such that

$$p_{\breve{X}_1\breve{X}_2|XY} = \begin{cases} p_{\hat{X}_1\hat{X}_2|XY} & \text{w.p. } \frac{1}{2} \\ p_{\tilde{X}_1\tilde{X}_2|XY} & \text{w.p. } \frac{1}{2}. \end{cases}$$

Kaspi shows in [2] that

$$I(X,Y;\hat{X}_2) + I(X;\hat{X}_1|\hat{X}_2,Y) = I(\hat{X}_2;Y) + I(X;\hat{X}_1\hat{X}_2|Y).$$

For random variables (A, B), we know that $I(A;B)$ is convex in $p_{A|B}$. This implies that $I(\hat{X}_2;Y)$ is convex in $p_{\hat{X}_2|Y}$ and that $I(X;\hat{X}_1\hat{X}_2|Y)$ is convex in $p_{\hat{X}_1\hat{X}_2|XY}$. In addition, the two expected distortions are linear in $p_{\hat{X}_1\hat{X}_2|XY}$. Hence, $(\breve{X}_1, \breve{X}_2)$ achieve a rate which is at least as small as the rate achieved by $(\hat{X}_1, \hat{X}_2)$, and distortions which are unchanged. The parameters corresponding to $(\breve{X}_1, \breve{X}_2)$ are

$$
\begin{aligned}
\breve{q}_{00} &= \frac{1}{2}(q_{00} + 1 - q_{11}) \\
\breve{q}_{0\epsilon} &= \frac{1}{2}(q_{0\epsilon} + 1 - q_{1\epsilon}) \\
\breve{q}_{11} &= \frac{1}{2}(q_{11} + 1 - q_{00}) \\
\breve{q}_{1\epsilon} &= \frac{1}{2}(q_{1\epsilon} + 1 - q_{0\epsilon}) \\
\breve{p}_1 &= \frac{1}{2}(p_1 + 1 - p_4) \\
\breve{p}_2 &= \frac{1}{2}(p_2 + 1 - p_3) \\
\breve{p}_3 &= \frac{1}{2}(p_3 + 1 - p_2) \\
\breve{p}_4 &= \frac{1}{2}(p_4 + 1 - p_1).
\end{aligned}
$$

It can be verified that

$$
\begin{aligned}
a &\triangleq \breve{q}_{00} = 1 - \breve{q}_{11}, \\
b &\triangleq \breve{q}_{0\epsilon} = 1 - \breve{q}_{1\epsilon}, \\
c &\triangleq \breve{p}_1 = 1 - \breve{p}_4
\end{aligned}
$$

and

$$d \triangleq \breve{p}_2 = 1 - \breve{p}_3.$$

Thus, the scheme specified by $(\breve{X}_1, \breve{X}_2)$ is symmetric in the sense described earlier, and it can only achieve a better rate-distortion triple than $(\hat{X}_1, \hat{X}_2)$. **This completes the proof of Claim 1.**

We are now facing the constrained minimization problem

$$\min_{a,b,c,d \in [0,1]} \rho(a, b, c, d) \tag{5}$$

subject to

$$
\begin{aligned}
\delta_1(a, b, c, d) &= D_1, &(6)\\
\delta_2(a, b, c, d) &= D_2, &(7)
\end{aligned}
$$

where $\rho(\cdot)$, $\delta_1(\cdot)$ and $\delta_2(\cdot)$ are the rate and the two distortions as functions of a, b, c, and d. Note that we enforce equality in the distortion constraints. This can be done without loss of generality. In fact, assume that for a given triple (R, D_1, D_2), the optimal scheme has an expected distortion at Decoder 1 that is strictly less than D_1. We derive a new scheme that uses the same code as the previous scheme, but Decoder 1 introduces a small amount of noise before declaring $\breve{X}_1$. Introducing noise corresponds to flipping the bits of $\breve{X}_1$ with a small probability, in a symmetric manner. If this probability is chosen correctly, the distortion at Decoder 1 can be raised to be equals to D_1. The resulting new description $\breve{X}_1$ still satisfies all symmetry properties and preserves R and D_2. An analoguous argument can be made for Decoder 2.

Since the distortion is zero whenever $y \neq \epsilon$, the expected distortions can be computed as

$$\begin{aligned}
\delta_1(a,b,c,d) &= \mathrm{E}\left[X \oplus \breve{X}_1\right] \\
&= \sum_{x=0,1} \sum_{\breve{x}_2=0,1} p_{XY}(x,\epsilon) p_{\breve{X}_2|XY}(\breve{x}_2|x,y) p_{\breve{X}_1|XY\breve{X}_2}(\bar{x}|x,\epsilon,\breve{x}_2) \\
&= \frac{1}{2} p\big((1-b)c + bd + bd + (1-b)c\big) \\
&= p\big((1-b)c + bd\big)
\end{aligned}$$

and

$$\begin{aligned}
\delta_2(a,b,c,d) &= \mathrm{E}\left[X \oplus \breve{X}_2\right] \\
&= \sum_{x=0,1} \sum_{y=x,\epsilon} p_{XY}(x,y) p_{\breve{X}_2|XY}(\bar{x}|x,y) \\
&= \frac{1}{2}(1-p)(a+a) + \frac{1}{2} p(b+b) \\
&= (1-p)a + pb.
\end{aligned}$$

To compute the rate $\rho(\cdot)$, we need

$$H(\breve{X}_2) = 1$$

(since $\breve{X}_2$ is the output of a BSC with input X),

$$\begin{aligned} H(\breve{X}_2|X,Y) &= \sum_{x=0,1}\sum_{y=x,\epsilon} p_{XY}(x,y)h\big(p_{\breve{X}_2|XY}(1|x,y)\big) \\ &= \frac{1}{2}(1-p)\big(h(a)+h(1-a)\big)+\frac{1}{2}p\big(h(b)+h(1-b)\big) \\ &= (1-p)h(a)+ph(b), \end{aligned}$$

$$\begin{aligned} H(\breve{X}_1|X,Y,\breve{X}_2) &= \sum_{x=0,1}\sum_{\breve{x}_2=0,1} p_{XY}(x,\epsilon)p_{\breve{X}_2|XY}(\breve{x}_2|x,\epsilon)h\big(p_{\breve{X}_1|XY\breve{X}_2}(1|x,\epsilon,\breve{x}_2)\big) \\ &= \frac{1}{2}p\big((1-b)h(c)+bh(d)+bh(1-d)+(1-b)h(1-c)\big) \\ &= p\big((1-b)h(c)+bh(d)\big) \end{aligned}$$

(since $H(\breve{X}_1|X,Y\neq\epsilon,\breve{X}_2)=0$). Further, we need

$$\begin{aligned} p_{\breve{X}_1|Y\breve{X}_2}(1|\epsilon,0) &= \frac{p_{XY\breve{X}_2\breve{X}_1}(0,\epsilon,0,1)+p_{XY\breve{X}_2\breve{X}_1}(1,\epsilon,0,1)}{p_{XY\breve{X}_2}(0,\epsilon,0)+p_{XY\breve{X}_2}(1,\epsilon,0)} \\ &= \frac{\frac{1}{2}p\big((1-b)c+b(1-d)\big)}{\frac{1}{2}p\big(1-b+b\big)} \\ &= (1-b)c+b(1-d), \end{aligned}$$

$$\begin{aligned} p_{\breve{X}_1|Y\breve{X}_2}(1|\epsilon,1) &= bd+(1-b)(1-c) \\ &= 1-(1-b)c-b(1-d) \end{aligned}$$

(through a similar computation), and finally,

$$\begin{aligned} H(\breve{X}_1|Y,\breve{X}_2) &= p_{Y,\breve{X}_2}(\epsilon,0)h\big(p_{\breve{X}_1|Y,\breve{X}_2}(1|\epsilon,0)\big)+p_{Y,\breve{X}_2}(\epsilon,1)h\big(p_{\breve{X}_1|Y,\breve{X}_2}(1|\epsilon,1)\big) \\ &= \frac{1}{2}p(1-b+b)h\big(1-bd-(1-b)(1-c)\big) \\ &\quad+\frac{1}{2}p(b+1-b)h\big(bd+(1-b)(1-c)\big) \\ &= ph\big(bd+(1-b)(1-c)\big). \end{aligned}$$

Putting the above four conditional entropies together yields

$$\begin{aligned} \rho(a,b,c,d) &= I(X,Y;\breve{X}_2)+I(X;\breve{X}_1|Y,\breve{X}_2) \\ &= H(\breve{X}_2)-H(\breve{X}_2|X,Y)+H(\breve{X}_1|Y,\breve{X}_2)-H(\breve{X}_1|X,Y,\breve{X}_2) \\ &= 1-(1-p)h(a)-ph(b) \\ &\quad+p\Bigg(h\big(bd+(1-b)(1-c)\big)-(1-b)h(c)-bh(d)\Bigg). \end{aligned}$$

Now, we solve the constrained minimization problem (5), (6), (7). Using the two distortion constraints, we express a and d in terms of b and c:

$$\begin{aligned}\delta_2(a,b,c,d) = (1-p)a + pb &= D_2 \\ a &= \frac{D_2 - pb}{1-p},\end{aligned} \tag{8}$$

$$\begin{aligned}\delta_1(a,b,c,d) = p\big((1-b)c + bd\big) &= D_1 \\ d &= \frac{\frac{D_1}{p} - (1-b)c}{b}.\end{aligned} \tag{9}$$

Substituting a and d into $\rho(a,b,c,d)$ yields

$$\begin{aligned}\rho(b,c) &= 1 - (1-p)h\Big(\frac{D_2 - pb}{1-p}\Big) - ph(b) \\ &\quad + p\bigg(h\Big(\frac{D_1}{p} + (1-b)(1-2c)\Big) - (1-b)h(c) - bh\Big(\frac{1}{b}\Big(\frac{D_1}{p} - (1-b)c\Big)\Big)\bigg),\end{aligned}$$

which we minimize now over b and c, under the constraints that b and c should be in $[0,1]$. In addition, the arguments of all occurences of $h(\cdot)$ in $\rho(b,c)$ should lie in $[0,1]$, because the binary entropy function is not defined outside that interval. This implies the following requirements:

$$D_2 - pb \geq 0$$

$$\Longrightarrow b \leq \frac{D_2}{p},$$

$$\frac{D_1}{p} + (1-b)(1-2c) \geq 0$$

$$\Longrightarrow c \leq \frac{1}{2}\Big(\frac{D_1}{p(1-b)} + 1\Big), \tag{10}$$

$$\frac{D_1}{p} - (1-b)c \geq 0$$

$$\Longrightarrow c \leq \frac{D_1}{p(1-b)}. \tag{11}$$

Since we also require that $c \leq 1$, the bounds (10) (11) are only active when

$$\frac{D_1}{p(1-b)} < 1.$$

However, if this is the case, then the bound (11) is tighter than (10). Hence, we obtain the following upper bounds on b and c:

I)

$$b \leq b_{\max} \triangleq \min\{\frac{D_2}{p}, 1\}$$

II)

$$c \leq c_{\max}(b) \triangleq \min\{\frac{D_1}{p(1-b)}, 1\}$$

We also require

$$\frac{D_2 - pb}{1-p} \leq 1$$
$$\implies b \geq \frac{D_2 - 1 + p}{p},$$

$$\begin{aligned} \frac{D_1}{p} + (1-b)(1-2c) &\leq 1 \\ 1 - 2c &\leq \frac{1 - \frac{D_1}{p}}{1-b} \\ c &\geq \frac{1 - b - 1 + \frac{D_1}{p}}{2(1-b)} \end{aligned}$$

$$\implies c \geq \frac{\frac{D_1}{p} - b}{2(1-b)}, \tag{12}$$

$$\frac{D_1}{p} - (1-b)c \leq b$$
$$\implies c \geq \frac{\frac{D_1}{p} - b}{1-b}. \tag{13}$$

Note that if $\frac{\frac{D_1}{p} - b}{1-b} \geq 0$, then the bound in (13) is always tighter than the one in (12) and thus, we obtain the following lower bounds on b and c:

III)

$$b \geq b_{\min} \triangleq \max\{0, \frac{D_1 - 1 + p}{p}\}$$

IV)

$$c \geq c_{\min}(b) \triangleq \max\{0, \frac{\frac{D_1}{p} - b}{1 - b}\}.$$

Constraints I) through IV) specify an admissible region in the (b, c)-plane. To minimize $\rho(b, c)$ over that region, we proceed in the following way. First, for any fixed admissible b, we take the derivative of $\rho(b, c)$ in the c-direction and set it to zero. This yields two extremal points $c_1(b)$ and $c_2(b)$, which are expressed in terms of b. We then realize that if $b \leq \frac{D_1}{p}$ or if $b \geq 1 - \frac{D_1}{p}$, then $c_1(b)$ is admissible and $c_2(b)$ is not. On the other hand, if $\frac{D_1}{p} \leq b \leq 1 - \frac{D_1}{p}$, then $c_2(b)$ is admissible and $c_1(b)$ is not. We define

$$c^*(b) = \begin{cases} c_1(b) & \text{if } b \in \left[b_{\min}, \frac{D_1}{b}\right] \cup \left[1 - \frac{D_1}{p}, b_{\max}\right] \\ c_2(b) & \text{if } b \in \left[\frac{D_1}{b}, 1 - \frac{D_1}{b}\right]. \end{cases}$$

We further verify that the second derivative of $\rho(b, c)$ in the c-direction is positive at $c^*(b)$, confirming that the minimizer of $\rho(b, c)$ has to lie on the curve $c = c^*(b)$. Hence, we substitute this expression for c into $\rho(b, c)$ to obtain a function $\rho^*(b)$. We then show that for $b \leq \frac{D_1}{p}$, the derivative of $\rho^*(b)$ is ≤ 0 ($\rho^*(b)$ is decreasing) and that for $b \geq \frac{D_1}{p}$, the derivative is ≥ 0 ($\rho^*(b)$ is increasing). This allows us to conclude that the minimum should lie at

$$b^* = \frac{D_1}{p}$$
$$c^* = c^*\left(\frac{D_1}{p}\right).$$

Now we do the computations that we have just outlined. First, note that

$$\frac{\partial h(x)}{\partial x} = \log(1 - x) - \log x.$$

We first regard b as a parameter.

$$\begin{aligned}
\frac{\partial \rho(b,c)}{\partial c} &= p\Bigg(\Big(\log\big(1-\frac{D_1}{p}-(1-b)(1-2c)\big)\\
&\quad-\log\big(\frac{D_1}{p}+(1-b)(1-2c)\big)\Big)(-2(1-b))\\
&\quad-(1-b)\big(\log(1-c)-\log c\big)\\
&\quad-b\Big(\log\frac{b-\frac{D_1}{p}+(1-b)c}{b}-\log\frac{\frac{D_1}{p}-(1-b)c}{b}\Big)\Big(\frac{-(1-b)}{b}\Big)\Bigg)\\
&= p(1-b)\log\frac{\big(\frac{D_1}{p}+(1-b)(1-2c)\big)^2 c\big(b-\frac{D_1}{p}+(1-b)c\big)}{\big(1-\frac{D_1}{p}-(1-b)(1-2c)\big)^2(1-c)\big(\frac{D_1}{p}-(1-b)c\big)}.
\end{aligned}$$

We find that if $b=1$, this expression is always 0 (for all values of c). Hence, there is an extremal edge at $b=1$ (c arbitrary). However, it turns out that this extremal edge does not minimize $\rho(b,c)$. We only show this at the very end of the proof. For now, assume that $b<1$. Hence, setting the above expression to zero means that the numerator and the denominator of the logarithm argument should be equal. This results in a quadratic equation, whose solutions are (see mathematica-file *quadratic_eqn.nb*)

$$c_1(b)=\frac{b-\frac{D_1}{p}}{2b-1}$$

and

$$c_2(b)=\frac{\big(b-\frac{D_1}{p}\big)\frac{D_1}{p}}{(b-1)\big(2\frac{D_1}{p}-1\big)}.$$

Claim 2: When $b\le\frac{D_1}{p}$, then $c_1(b)$ is admissible and $c_2(b)$ is not.

Proof of Claim 2: Note that in Regime 4), $D_1<\frac{p}{2}$, and therefore, $\frac{D_1}{p}<\frac{1}{2}$. When $b\le\frac{D_1}{p}$, then the numerator of $c_1(b)$ is non-positive, and so is the denominator, because $2b\le 2\frac{D_1}{p}<2\frac{1}{2}=1$. Hence, $c_1(b)$ is non-negative whenever $b\le\frac{D_1}{p}$. Moreover,

$$c_1(b)=\frac{b-\frac{D_1}{p}}{2b-1}=\frac{\frac{D_1}{p}-b}{\underbrace{1-2b}_{\in[0,1]}}\ge\frac{\frac{D_1}{p}-b}{1-b}$$

and hence, $c_1(b)$ satisfies condition IV). For condition II) to hold, we need to check whether

$$\begin{aligned}
\frac{\frac{D_1}{p} - b}{1 - 2b} &\overset{?}{\leq} \frac{\frac{D_1}{p}}{1 - b} \\
\big(\frac{D_1}{p} - b\big)(1 - b) &\overset{?}{\leq} \frac{D_1}{p}(1 - 2b) \\
\frac{D_1}{p}(1 - b) - b(1 - b) &\overset{?}{\leq} \frac{D_1}{p}(1 - b) - b\frac{D_1}{p} \\
\frac{D_1}{p} &\overset{?(*)}{\leq} 1 - b \\
\frac{D_1}{p} + b &\overset{?}{\leq} 1,
\end{aligned}$$

which is true since

$$\frac{D_1}{p} \leq \frac{1}{2}$$

and

$$b \leq \frac{D_1}{p} \leq \frac{1}{2}.$$

At the same time, $(*)$ proves that

$$\frac{D_1}{p(1 - b)} \leq 1$$

and hence, condition II) is satisfied, and $c_1(b)$ is admissible. On the other hand, b being smaller than $\frac{D_1}{p}$ also implies that

$$(b - \frac{D_1}{p}) \leq 0.$$

In addition,

$$(b - 1) < 0$$

(assuming that $b < 1$) and

$$(2\frac{D_1}{p} - 1) \leq 0.$$

This implies that

$$c_2(b) = \frac{\big(b - \frac{D_1}{p}\big)\frac{D_1}{p}}{(b - 1)\big(2\frac{D_1}{p} - 1\big)} < 0,$$

and is therefore not admissible. **End of proof of Claim 2.**

Claim 3: When $\frac{D_1}{p} \leq b \leq 1 - \frac{D_1}{p}$, then $c_2(b)$ is admissible and $c_1(b)$ is not.

Proof of Claim 3: When $b \geq \frac{D_1}{p}$, then

$$c_2(b) = \frac{\overbrace{\left(b - \frac{D_1}{p}\right)}^{\geq 0} \frac{D_1}{p}}{\underbrace{(b-1)}_{\leq 0} \underbrace{\left(2\frac{D_1}{p} - 1\right)}_{\leq 0}} \geq 0.$$

At the same time,

$$\frac{\frac{D_1}{p} - b}{1-b} \leq 0$$

and therefore, $c_2(b)$ satisfies condition IV). For condition II) to hold, we verify whether

$$\begin{aligned} c_2(b) = \frac{\frac{D_1}{p}}{1-b} \frac{\frac{D_1}{p} - b}{2\frac{D_1}{p} - 1} &\overset{?}{\leq} \frac{\frac{D_1}{p}}{1-b} \\ \frac{D_1}{p} - b &\overset{?}{\geq} 2\frac{D_1}{p} - 1 \\ b &\overset{?}{\leq} 1 - \frac{D_1}{p}. \end{aligned}$$

This is true by the assumption of Claim 3. By the same assumption, $\frac{D_1}{p} \leq 1 - b$, and hence,

$$\frac{\frac{D_1}{p}}{1-b} \leq 1,$$

Therefore, condition II) holds.

At the same time, b being in $(\frac{D_1}{p}, 1 - \frac{D_1}{p})$ implies that $c_1(b)$ is not admissible. This can be shown by splitting the interval into two parts. For $\frac{D_1}{p} \leq b \leq \frac{1}{2}$, we have

$$c_1(b) = \frac{\overbrace{b - \frac{D_1}{p}}^{\geq 0}}{\underbrace{2b-1}_{\leq 0}} \leq 0.$$

On the other hand, for $\frac{1}{2} \leq b \leq 1 - \frac{D_1}{p}$, we find that the following sequence of statements is true:

$$\begin{aligned} b &\leq 1 - \frac{D_1}{p} \\ \underbrace{2b-1}_{\geq 0} &\leq b - \frac{D_1}{p} \\ 1 &\leq \frac{b - \frac{D_1}{p}}{2b-1} = c_1(b). \end{aligned}$$

This shows that $c_1(b) = \frac{b - \frac{D_1}{p}}{2b-1}$ is not admissible for all b in $(\frac{D_1}{p}, 1 - \frac{D_1}{p})$. **End of Proof of Claim 3.**

Claim 4: When $1 - \frac{D_1}{p} \leq b \leq \min\{\frac{D_2}{p}, 1\}$, then $c_1(b)$ is admissible and $c_2(b)$ is not.

Proof of Claim 4: We first consider condition IV). Note that since $\frac{D_1}{p} \leq \frac{1}{2}$, $b \geq 1 - \frac{D_1}{p}$ implies that $b \geq \frac{D_1}{p}$ and $b \geq \frac{1}{2}$. Thus,

$$\frac{\overbrace{\frac{D_1}{p} - b}^{\leq 0}}{\underbrace{1-b}_{\geq 0}} \leq 0,$$

and condition IV) becomes $c_1(b) \geq 0$. This is the case, since

$$c_1(b) = \frac{\overbrace{b - \frac{D_1}{p}}^{\geq 0}}{\underbrace{2b-1}_{\geq 0}} \geq 0.$$

Next, we consider condition II). Note that since $\frac{D_1}{p} \geq 1 - b$,

$$\frac{D_1}{p(1-b)} \geq 1,$$

and hence, condition II) becomes $c_1(b) \leq 1$. This is satisfied, since

$$\begin{aligned} b &\geq 1 - \frac{D_1}{p} \\ \underbrace{2b-1}_{\geq 0} &\geq b - \frac{D_1}{p} \\ 1 &\geq \frac{b - \frac{D_1}{p}}{2b-1} = c_1(b), \end{aligned}$$

and thus, $c_1(b)$ is admissible. To show that $c_2(b)$ is not admissible, we verify whether

$$\begin{aligned} c_2(b) = \frac{(b - \frac{D_1}{p})\frac{D_1}{p}}{(\underbrace{1-b}_{\geq 0})\underbrace{(1 - 2\frac{D_1}{p})}_{\geq 0}} &\overset{?}{\geq} 1 \\ b\frac{D_1}{p} - \frac{D_1^2}{p^2} &\overset{?}{\geq} (1-b)(1 - 2\frac{D_1}{p}) \\ b\big(1 - \frac{D_1}{p}\big) &\overset{?}{\geq} \frac{D_1^2}{p^2} - 2\frac{D_1}{p} + 1 \\ b\underbrace{\big(1 - \frac{D_1}{p}\big)}_{\geq 0} &\overset{?}{\geq} -\frac{D_1}{p}\big(1 - \frac{D_1}{p}\big) + 1 - \frac{D_1}{p} \\ b &\overset{?}{\geq} 1 - \frac{D_1}{p}, \end{aligned}$$

which is true by assumption. Hence, $c_2(b)$ is indeed not admissible. **End of Proof of Claim 4.**

Now, we have, for every admissible b, a unique admissible $c^*(b)$ that is an extremal point of $\rho(b,c)$ in the c-direction. The function $c^*(b)$ is defined as follows:

$$c^*(b) = \begin{cases} \frac{b - \frac{D_1}{p}}{2b-1} & \text{if } b \in \big[b_{\min}, \frac{D_1}{b}\big] \cup \big[1 - \frac{D_1}{p}, b_{\max}\big] \\ \frac{\left(b - \frac{D_1}{p}\right)\frac{D_1}{p}}{(b-1)\left(2\frac{D_1}{p} - 1\right)} & \text{if } b \in \big[\frac{D_1}{b}, 1 - \frac{D_1}{b}\big]. \end{cases}$$

We verify that $c^*(b)$ is continuous for $b \in [b_{\min}, b_{\max}]$. In fact, the only potential points of discontinuity are $b = \frac{D_1}{p}$ and $b = 1 - \frac{D_1}{p}$. However, we have

$$c_1(\frac{D_1}{p}) = 0 = c_2(\frac{D_1}{p}),$$

which is true since $\frac{D_1}{p}$ is strictly less than $\frac{1}{2}$. At the second critical value of b, we have

$$c_2(1-\frac{D_1}{p}) = \frac{1-2\frac{D_1}{p}}{-1\left(2\frac{D_1}{p}-1\right)} = 1 = c_1(1-\frac{D_1}{p}).$$

An example of the curve $c = c^*(b)$ is depicted in Figure 3, consisting of the union of the admissible pieces of $c_1(b)$ and $c_2(b)$. The admissible region is the area bounded by all the solid curves and the two lines $b = 0$ and $b = 1$.

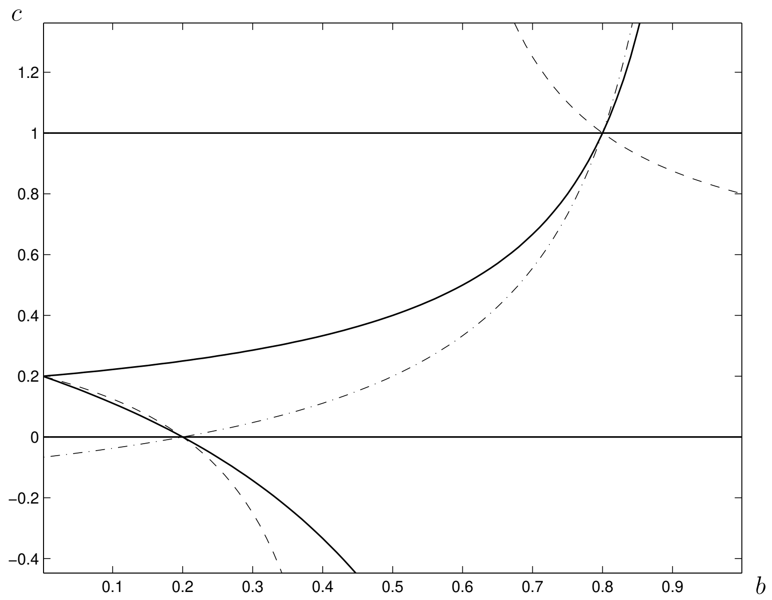

Figure 3: Admissible region in the b-c plane, with $c_1(b)$ (dashed line) and $c_2(b)$ (dash-dotted line). In this example, $\frac{D_1}{p} = 0.2$ (and hence $1-\frac{D_1}{p} = 0.8$). We assume that $b_{\min} \leq 0$ and $b_{\max} \geq 0$.

Next, we verify that for every admissible b, the extremal point (in the c-direction) given by $c^*(b)$ is actually a minimum. To this ends, we show that for all admissible b (actually, for all $b \in [0, 1]$),

$$\frac{\partial^2 \rho(b,c)}{\partial c^2}\Big|_{c=c^*(b)} \geq 0.$$

We split the proof into two claims.

Claim 5: For $b \in \left[b_{\min}, \frac{D_1}{b}\right] \cup \left[1 - \frac{D_1}{p}, b_{\max}\right]$, we have

$$\frac{\partial^2 \rho(b,c)}{\partial c^2}\Big|_{c=c_1(b)} \geq 0.$$

Proof of Claim 5: After a longer computation (done in *second_derivative.nb*), we find that

$$\frac{\partial^2 \rho(b,c)}{\partial c^2}\Big|_{c=c_1(b)} = \frac{(1-2b)^4(1-b)p}{b\left(b^2 - b + \frac{D_1}{p} - \frac{D_1^2}{p^2}\right)}.$$

It is obvious that $(1-2b)^4$, $(1-b)$, p and b are all non-negative. It also follows from Lemma 4 in Appendix A that if $b \leq \frac{D_1}{p}$ (and hence $b, \frac{D_1}{p} \in [0, \frac{1}{2})$, and therefore $b + \frac{D_1}{p} \leq 1$) then

$$b^2 - b + \frac{D_1}{p} - \frac{D_1^2}{p^2} = \frac{D_1}{p} - b - \left(\frac{D_1^2}{p^2} - b^2\right) \geq 0.$$

Also note that if $b \geq 1 - \frac{D_1}{p}$, then $\frac{D_1}{p} \leq b$ and $b + \frac{D_1}{p} \geq 1$, and thus, it follows from Lemma 5 in the Appendix that

$$b - \frac{D_1}{p} - \left(b^2 - \frac{D_1^2}{p^2}\right) \leq 0,$$

and therefore,

$$b^2 - b + \frac{D_1}{p} - \frac{D_1^2}{p^2} = -\Big(b - \frac{D_1}{p} - \left(b^2 - \frac{D_1^2}{p^2}\right)\Big) \geq 0.$$

End of Proof of Claim 5.

Claim 6: For $b \in \left[\frac{D_1}{b}, 1 - \frac{D_1}{p}\right]$, we have

$$\frac{\partial^2 \rho(b,c)}{\partial c^2}\Big|_{c=c_2(b)} \geq 0.$$

Proof of Claim 6: After another longer computation (also done in *second_derivative.nb*), we find that

$$\frac{\partial^2 \rho(b,c)}{\partial c^2}\Big|_{c=c_2(b)} = \frac{(1-b)^2(1-2\frac{D_1}{p})^4 p}{(1-\frac{D_1}{p})\frac{D_1}{p}(b - b^2 - \frac{D_1}{p} + \frac{D_1^2}{p^2})}.$$

Since $0 \leq \frac{D_1}{p} < \frac{1}{2}$, we know that $(1-b)^2$, $(1-2\frac{D_1}{p})$, p, $(1-\frac{D_1}{p})$ and $\frac{D_1}{p}$ are all non-negative. We know that $b \geq \frac{D_1}{p}$ and that $b + \frac{D_1}{p} \leq 1$. Hence, it follows from Lemma 4 that

$$b - b^2 - \frac{D_1}{p} + \frac{D_1^2}{p^2} = b - \frac{D_1}{p} - \big(b^2 - \frac{D_1^2}{p^2}\big) \geq 0$$

as well. **End of Proof of Claim 6.**

Now, we know that the curve $c = c^*(b)$ is a unique "valley" in the "landscape" given by $\rho(b,c)$, restricted to admissible (b,c) points. Hence, the constrained minimizer of $\rho(b,c)$ must lie on the curve $c = c^*(b)$. To find the optimal value for b, we substitute $c = c^*(b)$ into $\rho(b,c)$ to obtain a function $\rho^*(b)$. We find that for $b \leq \frac{D_1}{p}$, $\rho^*(b)$ is decreasing, while it is increasing for $b \geq \frac{D_1}{p}$. Again, we prove this using two claims.

Claim 7: We define

$$\rho_1(b) = \rho(b,c)\big|_{c=c_1(b)}.$$

The claim is that for $b \in [b_{\min}, \frac{D_1}{p}]$, $\rho_1(b)$ is decreasing, while it is increasing for $b \in [1-\frac{D_1}{p}, b_{\max}]$.

Proof of Claim 7: We first compute

$$\begin{aligned} \rho_1(b) &= \rho\Big(b, \frac{b-\frac{D_1}{p}}{2b-1}\Big) \\ &= 1-(1-p)h\big(\frac{D_2-pb}{1-p}\big) - ph(b) \\ &\quad + p\Big(h\big(\underbrace{\frac{D_1}{p} + \frac{(1-b)(2b-1-2b+2\frac{D_1}{p})}{2b-1}}_{\triangleq \mathcal{A}}\big) \\ &\quad - (1-b)h\big(\underbrace{\frac{b-\frac{D_1}{p}}{2b-1}}_{\triangleq \mathcal{B}}\big) \\ &\quad - bh\big(\underbrace{\frac{2b\frac{D_1}{p} - \frac{D_1}{p} - (1-b)(b-\frac{D_1}{p})}{b(2b-1)}}_{\triangleq \mathcal{C}}\big)\Big). \end{aligned}$$

Note that

$$\begin{aligned}
\mathcal{A} &= \frac{2b\frac{D_1}{p} - \frac{D_1}{p} - 1 + 2\frac{D_1}{p} + b - 2b\frac{D_1}{p}}{2b-1} \\
&= \frac{\frac{D_1}{p} - 1 + b}{2b-1}
\end{aligned}$$

and

$$\mathcal{C} = \frac{\frac{D_1}{p} - 1 + b}{2b-1} = \mathcal{A}.$$

In addition, we have

$$1 - \mathcal{B} = \frac{2b - 1 - b + \frac{D_1}{p}}{2b-1} = \mathcal{A}.$$

Hence,

$$\begin{aligned}
\rho_1(b) &= 1 - (1-p)h\big(\frac{D_2 - pb}{1-p}\big) - ph(b) \\
&\quad + p(1-b)\big(h(\mathcal{A}) - h(\mathcal{B})\big) \\
&= 1 - (1-p)h\big(\frac{D_2 - pb}{1-p}\big) - ph(b) \\
&\quad + p(1-b)\big(h(\mathcal{A}) - h(\mathcal{A})\big) \\
&= 1 - (1-p)h\big(\frac{D_2 - pb}{1-p}\big) - ph(b).
\end{aligned}$$

The derivative of $\rho_1(b)$ with respect to b is

$$\begin{aligned}
\frac{\partial \rho_1(b)}{\partial b} &= -(1-p)\Big(\log\frac{1-p-D_2+pb}{1-p} - \log\frac{D_2 - pb}{1-p}\Big)\big(-\frac{p}{1-p}\big) \\
&\quad - p\big(\log(1-b) - \log b\big) \\
&= p\log\frac{(1-p-D_2+pb)b}{(D_2-pb)(1-b)}.
\end{aligned}$$

Since we know that $b \le b_{\max} \le \frac{D_2}{p}$, we obtain the following chain of implications.

$$\begin{aligned}
p \log \frac{(1-p-D_2+pb)b}{(D_2-pb)(1-b)} &\overset{?}{\leq} 0 \\
\frac{(1-p-D_2+pb)b}{(\underbrace{D_2-pb}_{\geq 0})(\underbrace{1-b}_{\geq 0})} &\overset{?}{\leq} 1 \\
(1-p-D_2+pb)b &\overset{?}{\leq} (D_2-pb)(1-b) \\
b-pb-D_2b+pb^2 &\overset{?}{\leq} D_2-bD_2-pb+pb^2 \\
b &\overset{?}{\leq} D_2.
\end{aligned}$$

This is indeed the case whenever

$$b \leq \frac{D_1}{p} < D_2,$$

where the last inequality is a condition of Regime 4). Hence, for $b \in [b_{\min}, \frac{D_1}{p}]$, the function $\rho_1(b)$ is decreasing. Similarly, we find that if

$$b \geq D_2,$$

then

$$\frac{\partial \rho_1(b)}{\partial b} \geq 0,$$

i.e., $\rho_1(b)$ is increasing. This is the case whenever

$$b \geq 1 - \frac{D_1}{p} > 1 - D_2 \geq D_2,$$

because in Regime 4), $D_1 < pD_2$ and $D_2 < \frac{1}{2}$. Hence, for $b \in [1-\frac{D_1}{p}, b_{\max}]$, the function $\rho_1(b)$ is increasing. **End of Proof of Claim 7.**

Claim 8: We define

$$\rho_2(b) = \rho(b,c)\big|_{c=c_2(b)}.$$

The claim is that if $b \in [\frac{D_1}{p}, 1-\frac{D_1}{p}]$, then $\rho_1(b)$ is increasing.

Proof of Claim 8: We refer to the Mathematica file *derivative_for_b.nb* for the computation of $\rho_2(b)$ and its derivate, which turns out to be

$$\frac{\partial \rho_2(b)}{\partial b} = p \log \frac{D_2-1+p-pb}{pb-D_2}.$$

To verify that this derivative is positive, we check whether

$$\begin{aligned}
\frac{D_2 - 1 + p - pb}{\underbrace{pb - D_2}_{\leq 0 \text{ since } b \leq \frac{D_2}{p}}} &\overset{?}{\geq} 1 \\
D_2 - 1 + p - pb &\overset{?}{\leq} pb - D_2 \\
2D_2 + p - 1 &\overset{?}{\leq} 2bp \\
\frac{D_2 - \frac{1-p}{2}}{p} &\overset{?}{\leq} b.
\end{aligned}$$

The last inequality is true because by assumption of Claim 8,

$$b \geq \frac{D_1}{p},$$

and by a condition of Regime 4), we have

$$D_1 \geq D_2 - \frac{1-p}{2}.$$

Hence, $\rho_2(b)$ is increasing for $b \in [\frac{D_1}{p}, 1 - \frac{D_1}{p}]$, which **ends the proof of Claim 8.**

From Claims 7 and 8 it follows that $\rho^*(b)$ is decreasing for $b \leq \frac{D_1}{p}$ and increasing for $b \geq \frac{D_1}{p}$. Therefore, $\rho(b, c)$ is smallest for

$$b^* = \frac{D_1}{p},$$

which yields an optimal

$$c^* = c^*(b^*) = 0.$$

Plugging these values back into (8) and (9) yields

$$a^* = \frac{D_2 - D_1}{1 - p}$$

and

$$d^* = 1.$$

The corresponding rate is

$$\begin{aligned}\rho^*(a^*,b^*,c^*,d^*) &= 1-(1-p)h\big(\frac{D_2-D_1}{1-p}\big)-ph\big(\frac{D_1}{p}\big)\\ &\quad + p\Big(h\big(\frac{D_1}{p}+1-\frac{D_1}{p}\big)-(1-\frac{D_1}{p})h(0)-\frac{D_1}{p}h(1)\Big)\\ &= 1-h\big(\frac{D_2-D_1}{1-p}\big)+p\Big(h\big(\frac{D_2-D_1}{1-p}\big)-h\big(\frac{D_1}{p}\big)\Big).\end{aligned}$$

Recall that we have also found an extremal edge at $b=1$. The optimal a^*, b^*, c^* and d^* given above have been shown to be the minimizer under the assumption $b<1$. To show that they are actually minimal, we evaluate $\rho(b,c)$ at $b=1$:

$$\rho(1,c)=1-(1-p)h\big(\frac{D_2-p}{1-p}\big)+p\Big(h\big(\frac{D_1}{p}\big)-h\big(\frac{D_1}{p}\big)\Big).$$

Note that $\rho(1,c)$ does not depend on c. Hence, $(b=1, c_{\min}\le c\le c_{\max}$ is actually a “flat” extremal edge. However, we find that

$$\begin{aligned}\rho^*(a^*,b^*,c^*,d^*) &= 1-(1-p)h\big(\frac{D_2-D_1}{1-p}\big)-ph\big(\frac{D_1}{p}\big)\\ &\le 1-(1-p)h\big(\underbrace{\frac{D_2-D_1}{1-p}}_{\ge\frac{D_2-p}{1-p}}\big).\end{aligned}$$

One can show that

$$\frac{D_2-p}{1-p}\le\frac{1}{2}$$

which, since $h(x)$ is increasing for $0\le x\le\frac{1}{2}$, implies that

$$\rho^*(a^*,b^*,c^*,d^*)\le\rho(1,c).$$

This completes the proof. □

The most interesting part of Theorem 3 is of course Regime 4). An intutive explanation of the achievability for Regime 4) is the following. The encoder creates a description $\hat{X}_2$ for Decoder 2 in a single-decoder fashion, *i.e.*, $\hat{X}_2$ is the output of a BSC with input X. The average distortion achieved by $\hat{X}_2$ should be D_2 (corresponding to an average crossover probability D_2 of the BSC). Decoder 1 can use $\hat{X}_2$ in the following way to create a description $\hat{X}_1$:

$$\hat{X}_1 = \begin{cases} Y & \text{if } Y \neq \epsilon \\ \hat{X}_2 & \text{if } Y = \epsilon. \end{cases}$$

To help Decoder 1, the encoder should design $\hat{X}_2$ such that it has a lower distortion (crossover probability of the BSC) when $Y = \epsilon$ and a higher distortion when $Y \neq \epsilon$. This imbalance should be just high enough such that Decoder 1 can achieve a distortion of D_1. Let the crossover probability when $Y = \epsilon$ be b. The corresponding distortion at Decoder 1 is

$$\mathrm{E}\left[X \oplus \hat{X}_1\right] = pb + (1-p)0 = pb = D_1,$$

which implies that

$$b = \frac{D_1}{p}.$$

Let the crossover probability when $Y \neq \epsilon$ be a. This probability should now be such that the average crossover probability (averaged over Y) is still D_2:

$$\begin{aligned} \mathrm{E}\left[X \oplus \hat{X}_2\right] &= pb + (1-p)a \\ &= D_1 + (1-p)a = D_2, \end{aligned}$$

which implies that

$$a = \frac{D_2 - D_1}{1-p}.$$

These values of a and b are optimal, as we have shown in the proof of Theorem 3. The strategy of deciding $\hat{X}_1 = \hat{X}_2$ whenever $Y = \epsilon$ is expressed by

$$c = 0$$

and

$$d = 1,$$

which has also been shown to be optimal.

One can also write the optimal Kaspi rate as

$$\rho^*(a^*, b^*, c^*, d^*) = p\Big(1 - ph\big(\frac{D_1}{p}\big)\Big) + (1-p)\Big(1 - h\big(\frac{D_2 - D_1}{1-p}\big)\Big),$$

which resembles the addition of two binary-erasure conditional rate-distortion functions as given in Theorem 1. This suggests that we can also view the intuitive achievability given above in a different way: If only Decoder 1 were to be satisfied, the encoder would generate an auxiliary variable

$$W = \begin{cases} 0 & \text{w.p. } \frac{1}{2} & \text{if } Y \neq \epsilon \\ 1 & \text{w.p. } \frac{1}{2} & \text{if } Y \neq \epsilon \\ \bar{X} & \text{w.p. } \frac{D_1}{p} & \text{if } Y = \epsilon \\ X & \text{w.p. } 1 - \frac{D_1}{p} & \text{if } Y = \epsilon, \end{cases}$$

which leads to the first term of the rate. Since Decoder 2 is not satisfied with this (in Regime 4)), the encoder instead uses the following auxiliary variable:

$$W = \begin{cases} \bar{X} & \text{w.p. } \frac{D_2 - D_1}{1-p} & \text{if } Y \neq \epsilon \\ X & \text{w.p. } 1 - \frac{D_2 - D_1}{1-p} & \text{if } Y \neq \epsilon \\ \bar{X} & \text{w.p. } \frac{D_1}{p} & \text{if } Y = \epsilon \\ X & \text{w.p. } 1 - \frac{D_1}{p} & \text{if } Y = \epsilon. \end{cases}$$

Note that this W has the same statistics as the $\hat{X}_2$ used in the above intuitive explanation of Regime 4).

A Two Lemmata

Lemma 4 *If $r \leq s$ and $r + s \leq 1$, then*

$$s - r \geq s^2 - r^2.$$

Proof:

$$s^2 - r^2 = (s - r)(\underbrace{s + r}_{\leq 1}) \leq s - r.$$

□

Lemma 5 *If $r \leq s$ and $r + s \geq 1$, then*

$$s - r \leq s^2 - r^2.$$

Proof:

$$s^2 - r^2 = (s - r)(\underbrace{s + r}_{\geq 1}) \geq s - r.$$

□

B Generalization to Discrete Memoryless Sources

This appendix contains a more general version of Theorems 1 and 2. The content of this appendix has been presented at the *2007 TUD/EPFL Joint Workshop on Information Theory and Signal Processing in Communications* in Flims, Switzerland, March 2007.

Recall that the rate-distortion function of any source X is

$$R_X(D) = \min_{p_{\hat{X}|X} \in \mathcal{A}(D)} I(X;\hat{X}), \tag{14}$$

where $\mathcal{A}(D)$ is the set of all distributions $p_{\hat{X}|X}$ such that $\mathrm{E}\left[d(X,\hat{X})\right] \leq D$. The rate-distortion function for the conditional rate-distortion problem is in general

$$R_{X|Y}(D) = \min_{p_{\hat{X}|XY} \in \mathcal{B}(D)} I(X;\hat{X}|Y), \tag{15}$$

where $\mathcal{B}(D)$ is the set of all distributions $p_{\hat{X}|XY}$ such that $\mathrm{E}\left[d(X,\hat{X})\right] \leq D$. The rate-distortion function for the Wyner-Ziv problem is

$$R^{\mathrm{WZ}}(D) = \min_{p_{W|X} \in \mathcal{C}(D)} I(X;W|Y), \tag{16}$$

where $\mathcal{C}(D)$ is the set of all distributions of auxiliary variables W such that $W \multimap X \multimap Y$ forms a Markov chain and such that there exists a function $f(W,Y) = \hat{X}$ for which $\mathrm{E}\left[d(X,\hat{X})\right] \leq D$.

Let X be a memoryless source with values in some discrete alphabet $\mathcal{X}$, having probability mass function $p_X(x)$, $x \in \mathcal{X}$. Let Y be a second source, jointly distributed with X such that

$$Y = \begin{cases} X & \text{w.p.} \quad 1-p \\ \epsilon & \text{w.p.} \quad p. \end{cases}$$

The reconstruction alphabet is $\mathcal{X}$, and we assume that the distortion measure is such that $d(x,x) = 0$ for $x \in \mathcal{X}$. We refer to this setup as "discrete source with erased side-information".

Theorem 6 *For a discrete source with erased side-information, we have*

$$R^{WZ}(D) = R_{X|Y}(D) = pR_X(\frac{D}{p}).$$

Proof: We first show that $R_{X|Y}(D) \geq pR_X(\frac{D}{p})$. Let $p_{\hat{X}|XY} \in \mathcal{B}(D)$. Then, the corresponding rate is

$$\begin{aligned} R &= I(X;\hat{X}|Y) \\ &= pI(X;\hat{X}|Y=\epsilon) + \sum_{x \neq \epsilon} \underbrace{I(X;\hat{X}|Y=x)}_{=0} p_Y(x) \\ &= pI(X;\tilde{X}) \\ &\geq p \min_{p_{\breve{X}|X} \in \mathcal{A}(\frac{D}{p})} I(X;\breve{X}) \\ &= pR_X(\frac{D}{p}), \end{aligned} \tag{17}$$

where $\tilde{X}$ is defined such that $p_{\tilde{X}|X}(\hat{x}|x) \triangleq p_{\hat{X}|XY}(\hat{x}|x,\epsilon)$ and the last inequality is true because $\tilde{X}$ is a member of $\mathcal{A}(\frac{D}{p})$. To realize this fact, note that

$$\begin{aligned} D &\geq \mathrm{E}\left[d(X,\hat{X})\right] \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X} \cup \{\epsilon\}} \sum_{\hat{x} \in \mathcal{X}} p_{XY\hat{X}}(x,y,\hat{x}) d(x,\hat{x}) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}} p_{XY\hat{X}}(x,y,\hat{x}) d(x,\hat{x}) \\ &\quad + p \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}} p_{X\hat{X}|Y}(x,\hat{x}|\epsilon) d(x,\hat{x}) \\ &\geq p \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}} p_X(x) p_{\tilde{X}|X}(\hat{x}|x) d(x,\hat{x}) \\ &= p\mathrm{E}\left[d(X,\tilde{X})\right]. \end{aligned}$$

From this, it follows that $\mathrm{E}\left[d(X,\tilde{X})\right] \leq \frac{D}{p}$, and hence, $\tilde{X}$ is in $\mathcal{A}(\frac{D}{p})$. The lower bound (17) holds for any $\hat{X}$ that is a member of $\mathcal{B}(D)$, and we get $R_{X|Y}(D) \geq pR_X(\frac{D}{p})$.

Further, it is easy to see that $R^{\mathrm{WZ}}(D) \geq R_{X|Y}(D)$. This is so because informing the encoder can only help.

Finally, we show that $pR_X(\frac{D}{p}) \geq R^{\mathrm{WZ}}(D)$. We choose the auxiliary W and the function f in $\mathcal{C}(D)$ as $p_{W|X} = \arg\min_{p_{\hat{X}|X} \in \mathcal{A}(\frac{D}{p})} I(X;\hat{X})$ and

$f(w,y) = w$ if $y = \epsilon$ and $f(w,y) = y$ if $y \neq \epsilon$. The Wyner-Ziv rate for this particular W is

$$\begin{aligned} I(X;W|Y) &= pI(X;W|Y=\epsilon) + \sum_{x \in \mathcal{X}} p_Y(x) \underbrace{I(X;W|Y=x)}_{=0} \\ &\stackrel{(a)}{=} pI(X;W) \\ &= pR_X(\frac{D}{p}), \end{aligned}$$

where the last step is true by the choice of W, and (a) holds because $p_{XW|Y}(x,w|\epsilon) = p_{XW}(x,w)$ for any (x,w) in $\mathcal{X}^2$. The distortion achieved by this scheme is

$$\mathrm{E}\left[d\big(X, f(W,Y)\big)\right] = p\frac{D}{p} + (1-p)0 = D.$$

Hence, the indicated W and f achieve distortion D while using a rate of $pR_X(\frac{D}{p})$. The optimal Wyner-Ziv scheme can only do better, and therefore,

$$pR_X(\frac{D}{p}) \geq R^{\mathrm{WZ}}(D).$$

But now, we have shown that

$$R^{\mathrm{WZ}}(D) \geq R_{X|Y}(D) \geq pR_X(\frac{D}{p}) \geq R^{\mathrm{WZ}}(D),$$

which implies that all three quantities are equal. □

References

[1] T. Berger, *Multiterminal Source Coding*, Springer-Verlag, Vienna/New York, 1978.

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

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

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