Source Coding Problems with Conditionally Less Noisy Side Information

Roy Timo, Tobias J. Oechtering and Michèle Wigger

Abstract—A computable expression for Heegard and Berger’s rate-distortion (RD) function has eluded information theory for nearly three decades. Heegard and Berger’s single-letter achievability bound is well known to be optimal for physically degraded side information; however, it is not known whether the bound is optimal for arbitrarily correlated side information (general discrete memoryless sources). In this paper, we consider a new setup where the side information at one receiver is conditionally less noisy than that at the other. The new setup includes degraded side information as a special case, and it is motivated by the literature on degraded and less noisy broadcast channels. Our key contribution is a converse proving the optimality of Heegard and Berger’s achievability bound in a new setting, where the side information is conditionally less noisy and one distortion function is deterministic. The less noisy setup is also generalised to two different successive-refinement problems.

I. INTRODUCTION

Wyner and Ziv’s seminal 1976 paper [1] extended rate-distortion (RD) theory to include side information at the receiver. Nearly a decade later, Heegard and Berger [2] further extended the theory to include side information at multiple receivers: an example of which, and the principal subject of this paper, is shown in Fig. 1. Heegard and Berger’s RD function, however, has eluded complete characterisation in that matching computable [3, p. 259] achievability and converse bounds have yet to be obtained. Indeed, the RD function is unknown for the seemingly simple case of deterministic distortion functions [4], where each receiver needs to losslessly reconstruct a function of the source [6, 7].

The best single-letter achievability bound for two receivers is due to Heegard and Berger [4, Thm. 2], and the best bound for three or more receivers is due to Timo, Chan, and Grant [7, Thm. 2]. Both bounds hold for arbitrary discrete memoryless sources under average per-letter distortion constraints. Matching converses have been obtained only in some special cases, for example, see [2, 6, 8–12]. One such case is called physically degraded side information, and it refers to the situation where the side information at one receiver is a noisy version of that at the other. Degraded side information is essential to Heegard and Berger’s converse [2, pp. 733–734].

This paper considers a new setup where the side information at one receiver is conditionally less noisy than that at the other. Conditionally less noisy side information is a generalisation of physically degraded side information, and it is motivated by similar (but apparently unrelated) literature on broadcast channels [13, 14]. Our key contribution is a converse that proves the optimality of Heegard and Berger’s achievability bound when the side information is conditionally less noisy and one distortion function is deterministic.

Generalisations of Heegard and Berger’s RD problem include the successive-refinement work [15–19] and the joint source-channel coding work [20–22]. Other variations of the problem have been considered with causal side information [23, 24] and common reconstructions [25, 26]. The less noisy side information model may be useful in such problems; indeed, to conclude the paper, we apply our converse methods to obtain new results for two successive-refinement problems.

Paper Outline: Section II presents a single-letterization lemma that will be used throughout the paper. Sections III and IV present new converses for the Heegard-Berger problem and two successive-refinement problems with side information (degraded side information [15, 16] and scalable side information [17]). Longer proofs are given in the appendices.

Notation: All random variables in this paper are discrete and finite and denoted by uppercase letters, e.g., X. The alphabet of a random variable is identified by matching calligraphic font, e.g., X ∈ X. The n-fold Cartesian product of an alphabet is denoted by boldface font, e.g., X is the n-fold product of X. If a random vector (X, Y, Z) forms a Markov chain in the same order, then we write X ↔ Y ↔ Z. The symbol ⊕ denotes modulo-two addition.

II. A LEMMA

We start with a single-letterization (entropy characterisation) problem: Express the difference of two n-letter conditional mutual informations in a single-letter form.

Consider a tuple of random variables \((R, S_1, S_2, T, L)\) with an arbitrary joint distribution. Let

\[
(R, S_1, S_2, T, L) := (R_1, S_{1,1}, S_{2,1}, T_1, L_1), \quad (R_2, S_{1,2}, S_{2,2}, T_2, L_2), \ldots, (R_n, S_{1,n}, S_{2,n}, T_n, L_n)
\]
denote \( n \) i.i.d. copies of \((R, S_1, S_2, T, L)\). Further, suppose that \( J \) is jointly distributed with \((R, S_1, S_2, T, L)\) and
\[
J \leftrightarrow (R, L) \leftrightarrow (S_1, S_2, T)
\]
forms a Markov chain. Consider the difference
\[
I(J; S_2 | L) - I(J; S_1 | L).
\]
We wish to know whether this difference can be expressed in a single-letter form in the sense of Csiszár and Körner [3] p. 259]. The next lemma answers the question in the affirmative, and it is proved in Appendix A.

Lemma 1: Let \((J, R, S_1, S_2, T, L)\) be defined as above. There exists an auxiliary random variable \( W \), with alphabet \( \mathcal{W} \) and jointly distributed with \((R, S_1, S_2, T, L)\), such that
\[
I(J; S_2 | L) - I(J; S_1 | L) = n(I(W; S_2 | L) - I(W; S_1 | L))
\]
the cardinality of \( \mathcal{W} \) satisfies \(|\mathcal{W}| \leq |R| |L|\), and
\[
W \leftrightarrow (R, L) \leftrightarrow (S_1, S_2, T)
\]
forms a Markov chain. If, in addition, \( L \) is a function of \( R \), then the Markov chain can be replaced by \( W \leftrightarrow R \leftrightarrow (S_1, S_2, T) \) and the cardinality bound on \( W \) becomes \(|\mathcal{W}| \leq |R|\).

### III. The Heegard-Berger Problem

This section is devoted to Heegard and Berger’s RD problem for two receivers and is organised as follows: We recall the RD function’s operational definition in Section III-A we review some important results in Section III-B and we state our new results in Section III-C.

#### A. Operational Definition of the RD Function

Consider a tuple of random variables \((X, Y_1, Y_2)\) with an arbitrary joint distribution on \( \mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \). Let \((X, Y_1, Y_2)\) denote a string of \( n \) i.i.d. copies of \((X, Y_1, Y_2)\). Let \( \mathcal{X}, \mathcal{Y}_1 \) and \( \mathcal{Y}_2 \) denote the \( n \)-fold Cartesian products of \( \mathcal{X}, \mathcal{Y}_1 \) and \( \mathcal{Y}_2 \) respectively.

Consider the setup of Fig. 1. The transmitter observes \( X \), receiver 1 observes \( Y_1 \) and receiver 2 observes \( Y_2 \). The string \( X \) is to be compressed by the transmitter and reconstructed by both receivers using a block code. The RD function is the smallest rate at which \( X \) can be compressed while still allowing the receivers to reconstruct \( X \) to within specified average distortions, as described next.

A block code consists of three (possibly stochastic) mappings:
\[
f : \mathcal{X} \rightarrow \mathcal{M}
\]
and
\[
g_j : \mathcal{M} \times \mathcal{Y}_j \rightarrow \hat{X}_j, \quad j = 1, 2,
\]
where \( \mathcal{M} \) is an index set with finite cardinality \(|\mathcal{M}|\) depending on \( n \), \( \hat{X}_j \) is the reconstruction alphabet of receiver \( j \) and \( \hat{X}_j \) its \( n \)-fold Cartesian product. The transmitter sends \( M := f(X) \) and receiver \( j \) reconstructs \( \hat{X}_j := g_j(M, Y_j) \).

Let
\[
\delta_j : \mathcal{X} \times \hat{X}_j \rightarrow [0, \infty), \quad j = 1, 2,
\]
be bounded per-letter distortion functions. For simplicity, and without loss of generality, we assume that \( \delta_1 \) and \( \delta_2 \) are normal [27] p. 185]; that is, for each \( x \) in \( \mathcal{X}_j \) there exists some \( \hat{x} \) in \( \hat{X}_j \) such that \( \delta_j(x, \hat{x}) = 0 \).

**Definition 1:** A rate \( R \) is said to be \((D_1, D_2)\)-achievable if for any \( \epsilon > 0 \) there exists a block code \((f, g_1, g_2)\), with some sufficiently large blocklength \( n \), satisfying
\[
R + \epsilon \geq \frac{1}{n} \log |\mathcal{M}|
\]
and
\[
D_j + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} \delta_j(x_i, \hat{x}_{j,i}), \quad j = 1, 2.
\]

**Definition 2:** For distortions \( D_1 \geq 0 \) and \( D_2 \geq 0 \), Heegard and Berger’s RD function is
\[
R(D_1, D_2) := \min \left\{ R > 0 : R \text{ is } (D_1, D_2)\text{-achievable} \right\}.
\]

### B. Existing Results

Single-letter expressions for \( R(D_1, D_2) \) have been found in some special cases, for example, [2][9][10]. The achievability proofs of all these cases follow from the next simple, but surprisingly powerful, lemma. The converses, in contrast, are proved on a case-by-case basis using different approaches.

**Lemma 2 (Achievability):** The RD function \( R(D_1, D_2) \) is bounded from above by [2] Thm. 2
\[
R(D_1, D_2) \leq \min_{(A, B, C)} \left\{ \max \left\{ I(X; C|Y_1), I(X; C|Y_2) \right\} + I(X; A|C, Y_1) + I(X; B|C, Y_2) \right\},
\]
where minimisation is taken over all auxiliary random tuples \((A, B, C)\), jointly distributed with \((X, Y_1, Y_2)\), such that the following is true:

(i) The tuple \((A, B, C)\) is conditionally independent of the side information \((Y_1, Y_2)\) given \( X \),
\[
(A, B, C) \leftrightarrow X \leftrightarrow (Y_1, Y_2);
\]
(ii) The cardinalities of the alphabets of \( C, A \) and \( B \) are respectively bounded by
\[
|C| \leq |\mathcal{X}| + 3
\]
\[
|A| \leq |C||\mathcal{X}| + 1
\]
\[
|B| \leq |C||\mathcal{X}| + 1
\]
(these bounds are new and proved in Appendix B).
There exist deterministic maps
\[ \phi_1: A \times C \times Y_1 \rightarrow \hat{X}_1 \]
\[ \phi_2: B \times C \times Y_2 \rightarrow \hat{X}_2 \]
with
\[ D_1 \geq E_\delta_1(X, \phi_1(A, C, Y_1)) \]
\[ D_2 \geq E_\delta_2(X, \phi_2(B, C, Y_2)). \]

The next definition and theorem review a special case for which the upper bound of Lemma [2] is known to be tight.

**Definition 3:** The side information is said to be physically degraded if
\[ X \leftrightarrow Y_2 \leftrightarrow Y_1 \]
forms a Markov chain.

**Theorem 3:** If the side information is physically degraded, then [2 Thm. 3]
\[ R(D_1, D_2) = \min_{(B, C)} \{ I(X; C|Y_1) + I(X; B|C, Y_2) \}, \]
where the minimisation is taken over all auxiliary random tuples \((B, C)\), jointly distributed with \((X, Y_1, Y_2)\), such that
(i) \((B, C) \leftrightarrow X \leftrightarrow (Y_1, Y_2)\) forms a Markov chain;
(ii) there exist deterministic maps
\[ \phi_1: C \times Y_1 \rightarrow \hat{X}_1 \]
\[ \phi_2: B \times C \times Y_2 \rightarrow \hat{X}_2 \]
with
\[ D_1 \geq E_\delta_1(X, \phi_1(C, Y_1)) \]
\[ D_2 \geq E_\delta_2(X, \phi_2(B, C, Y_2)). \]

The Markov chain \(X \leftrightarrow Y_2 \leftrightarrow Y_1\), which defines physically degraded side information, enables a crucial step in Heegard and Berger’s converse proof of Theorem [3]; see [2, pp. 733-734]. The aim of the next section is to broaden the scope of Theorem [3] by replacing \(X \leftrightarrow Y_2 \leftrightarrow Y_1\) with a more general condition. Our main results, however, will fail slightly short of this aim: We will need to restrict attention to the setting where receiver 1 requires an almost lossless copy of a function of \(X\). More specifically, we will require that \(D_1 = 0\) and \(\delta_1\) is deterministic in the following sense.

**Definition 4:** \(\delta_1\) is said to be deterministic [17, 28] if there is an alphabet \(\hat{X}\) with \(X_1 = \hat{X}\) and a deterministic map
\[ \psi: X \rightarrow \hat{X} \]
such that
\[ \delta_1(x, \hat{x}) := \begin{cases} 0 & \text{if } \hat{x} = \psi(x) \\ 1 & \text{otherwise.} \end{cases} \]

For later discussions, we need to specialise Theorem 3 to deterministic \(\delta_1\). Let
\[ \bar{X} := \psi(X). \]

Define
\[ S(D_2) := \min_B I(X; B|\bar{X}, Y_2), \quad D_2 \geq 0, \]
where the minimisation is taken over all auxiliary random variables \(B\), jointly distributed with \((X, Y_1, Y_2)\), such that
(i) \(B \leftrightarrow X \leftrightarrow (Y_1, Y_2)\) forms a Markov chain;
(ii) the cardinality of the alphabet of \(B\) is bounded by \(|B| \leq |X| + 1; \)
(iii) there exists a deterministic mapping
\[ \phi_2: B \times \bar{X} \times Y_2 \rightarrow \hat{X}_2 \]
with
\[ D_2 \geq E_\delta_2(X, \phi_2(B, \bar{X}, Y_2)). \]

The function \(S(D_2)\) is non-increasing, convex and continuous in \(D_2\) [1 Thm. A2]. The next corollary is proved in Appendix 5.

**Corollary 3.1:** If the side information is physically degraded and \(\delta_1\) is a deterministic distortion function, then
\[ R(0, D_2) = H(\bar{X}|Y_1) + S(D_2). \]

It will be useful to further specialise Corollary [3.1] to a “two-component” source model with Hamming distortion functions. The specialisation is central to our understanding of how Corollary [3.1] can be generalised to a less-noisy setup.

**Definition 5:** We say that \((X, Y_1, Y_2)\) is a two-source if
\[ X := X_1 \times X_2 \quad \text{and} \quad Y := (X_1, X_2), \]
where \(X_1\) and \(X_2\) are finite alphabets. In addition, we say that \(\delta_1\) and \(\delta_2\) are component Hamming distortion functions if
\[ \bar{X}_j = X_j \]
and for all \(x_j, \hat{x}_j \in X_j\)
\[ \delta_j(x_j, \hat{x}_j) = \begin{cases} 0 & \text{if } \hat{x}_j = x_j \\ 1 & \text{otherwise} \end{cases} \]
\[ j = 1, 2. \]

**Corollary 3.2:** Consider a two-source \((X_1, X_2, Y_1, Y_2)\) with component Hamming distortion functions. If the side information is physically degraded \((X_1, X_2) \leftrightarrow Y_2 \leftrightarrow Y_1\), then [2, 7]
\[ R(0, 0) = H(X_1|Y_1) + H(X_2|X_1, Y_2). \]

The corollary can be directly proved in a simple way that nicely motivates the possibility of a more general converse.

**Proof Outline (Converse):** If \(R\) is directly achievable, then for each \(\epsilon > 0\) there exists a block code \((f, g_1, g_2)\) for which
\[ R + \epsilon \geq \frac{1}{n} \log |\mathcal{M}| \geq \frac{1}{n} H(M) \geq \frac{1}{n} I(X_1, X_2, Y_1, Y_2; M) \]
\[ = \frac{1}{n} \left( I(X_1, Y_1; M) + I(X_2, Y_2; M|X_1, Y_1) \right) \]
\[ \geq \frac{1}{n} \left( I(X_1; M|Y_1) + I(X_2; M|X_1, Y_1, Y_2) \right) \]
\[ \geq \frac{1}{n} \left( H(X_1|Y_1) + H(X_2|X_1, Y_1, Y_2) - n\epsilon(\epsilon) \right) \]
\[ \geq H(X_1|Y_1) + H(X_2|X_1, Y_1, Y_2) - \epsilon(\epsilon) \]
\[ \geq H(X_1|Y_1) + H(X_2|X_1, Y_1, Y_2) - \epsilon. \]

The justifications for steps (a), (b), and (c) are as follows:
(a) \(\hat{X}_1\) and \(\hat{X}_2\) are determined by \((M, Y_1)\) and \((M, Y_2)\) respectively, so (a) follows by Fano’s inequality [14 Sec. 2.2]. Here \(\epsilon(\epsilon)\) can be chosen so that \(\epsilon(\epsilon) \rightarrow 0\) as \(\epsilon \rightarrow 0\).
(b) \((X_1, X_2, Y_1, Y_2)\) is i.i.d.
(c) The side information is physically degraded and consequently \(X_2 \leftrightarrow (X_1, Y_2) \leftrightarrow Y_1\).

**Proof Outline (Achievability):** Suppose that we use the Slepian-Wolf / Cover random-binning argument to send \(X_1\) losslessly to receiver 1 at rate \(R'\) close to \(H(X_1|Y_1)\). The side information is physically degraded, so we have

\[
R' \geq H(X_1|Y_1) \geq H(X_1|Y_2).
\]

A close inspection of the random binning proof, e.g. \[14\], reveals that \[3\] also suffices for receiver 2 to reliably decode \(X_1\). Assuming that \(X_1\) is successfully decoded by receiver 2, we can send \(X_2\) to receiver 2 at a rate \(R''\) close to \(H(X_2|X_1, Y_2)\) using \((X_1, Y_2)\) as side information. The total rate \(R = R' + R''\) is close to \(H(X_1|Y_1) + H(X_2|X_1, Y_2)\).

We notice that the Markov chain \((X_1, X_2) \leftrightarrow Y_2 \leftrightarrow Y_1\) is equivalent to

\[
X_1 \leftrightarrow Y_2 \leftrightarrow Y_1
\]

and

\[
X_2 \leftrightarrow (X_1, Y_2) \leftrightarrow Y_1.
\]

The chain \[(4a)\] is a sufficient, but not necessary, condition for the inequalities in \[(3)\] and hence the above achievability argument. In contrast, the chain \[(4b)\] is essential for equality in \[(3)\] and hence the converse argument. The generality of the achievability argument juxtaposed against the more restrictive converse argument suggests that Corollary \[3.2\] might hold for a broader class of two-sources. We show that this is indeed the case in the next subsection; specifically, we will see that the corollary holds when the Markov chain \[(4a)\] is replaced by \(H(X_1|Y_1) \geq H(X_1|Y_2)\) and the chain \[(4b)\] is replaced by a more general “conditionally less noisy” condition.

**Remark 1:**
(i) \(R(D_1, D_2)\) depends on the joint distribution of \((X, Y_1, Y_2)\) only via the distributions of \((X, Y_1)\) and \((X, Y_2)\).
(ii) The side information is said to be **stochastically degraded** if the joint distribution of \((X, Y_1, Y_2)\) is such that there exists some physically degraded side information \((X', Y_1', Y_2')\) with marginals \((X', Y_1')\) and \((X', Y_2')\) matching those of \((X, Y_1)\) and \((X, Y_2)\). By Remark 1 (i), Theorem \[3\] and Corollaries \[3.1\] and \[3.2\] also hold for stochastically degraded side information.
(iii) The function \(S(D_2)\), which is defined in \[1\], is the Wyner-Ziv RD function \[1\] Eqn. 15 for a source \(X\) with side information \((X, Y_2)\).
(iv) The asserted upper bound for \(R(D_1, D_2)\) in \[2\] Thm. 2] is incorrect for the case of three or more receivers \[7\].

**C. New Results**

Suppose that \(L\) is an auxiliary random variable that is jointly distributed with \((X, Y_1, Y_2)\).

**Definition 6:** We say that \(Y_2\) is **conditionally less noisy than** \(Y_1\) given \(L\), abbreviated as \((Y_2 \succeq Y_1 \mid L)\), if

\[
I(W; Y_2|L) \geq I(W; Y_1|L)
\]

holds for every auxiliary random variable \(W\), jointly distributed with \((X, Y_1, Y_2, L)\), for which

\[
W \leftrightarrow (X, L) \leftrightarrow (Y_1, Y_2)
\]

forms a Markov chain.

The next lemma and example collectively show that Definition \[6\] is broader than Definition \[5\]. The lemma is proved in Appendix \[C\].

**Lemma 4:**
(i) If the side information is physically degraded \(X \leftrightarrow Y_2 \leftrightarrow Y_1\) and

\[
L \leftrightarrow X \leftrightarrow (Y_1,Y_2),
\]

forms a Markov chain, then \((Y_2 \succeq Y_1 \mid L)\).

(ii) If a two-source \((X_1, X_2, Y_1, Y_2)\) satisfies

\[
X_2 \leftrightarrow X_1 \leftrightarrow Y_1
\]

and \(L = X_1\), then \((Y_2 \succeq Y_1 \mid X_1)\).

The next example describes a setup where the side information is not degraded, but \(X_2 \leftrightarrow X_1 \leftrightarrow Y_1\) is a Markov chain and therefore \((Y_2 \succeq Y_1 \mid X_1)\).

**Example 1:** Let \(X_2\), \(Y_2\), and \(Z\) be independent Bernoulli random variables with different, non-uniform, biases. Let

\[
X_1 = X_2 \oplus Y_2 \quad \text{and} \quad Y_1 = X_1 \oplus Z.
\]

We notice that

\[
X_2 \leftrightarrow X_1 \leftrightarrow Y_1
\]

forms a Markov chain, so assertion (ii) of Lemma \[4\] implies \((Y_2 \succeq Y_1 \mid X_1)\). In contrast, \((X_1, X_2)\) is not conditionally independent of \(Y_1\) given \(Y_2\).

The next lemma gives a converse for \(R(D_1, D_2)\). Its proof uses Lemma \[1\] and is the subject of Appendix \[E\]. Our main result in this section, Theorem \[6\] follows directly thereafter.

**Lemma 5 (Converse):** If \(\delta_1\) is a deterministic distortion function specified by \(\bar{X} = \psi(X)\), then the following statements are true.
(i) For arbitrarily distributed \((X, Y_1, Y_2)\), we have

\[
R(0, D_2) \geq H(\bar{X}|Y_1) + S(D_2)
\]

\[+ \min \left\{ I(W; Y_2|\bar{X}) - I(W; Y_1|\bar{X}) \right\}, \]

where the minimisation is taken over all auxiliary \(W\), jointly distributed with \((X, Y_1, Y_2)\), such that

\[
W \leftrightarrow X \leftrightarrow (Y_1, Y_2)
\]

forms a Markov chain and \(|W| \leq |X'|\).
(ii) If \((X, Y_1, Y_2)\) satisfies \((Y_2 \succeq Y_1 \mid \bar{X})\), then

\[
R(0, D_2) \geq H(\bar{X}|Y_1) + S(D_2).
\]

It is worth highlighting that

\[
\min \left\{ I(W; Y_2|\bar{X}) - I(W; Y_1|\bar{X}) \right\}
\]

is non-positive because, for example, we can always choose \(W\) to be a constant. Assertion (ii) of the lemma follows from assertion (i) upon invoking Definition \[6\] with \(L = \bar{X}\).
The next theorem gives a single-letter expression for $R(D_1, D_2)$ in a new setting. The theorem is a consequence of the achievability of Lemma 2 and the converse of Lemma 5 (ii).

**Theorem 6:** If $\delta_1$ is a deterministic distortion function specified by $\tilde{X} = \psi(X)$, $(Y_2 \geq Y_1 | \tilde{X})$ and

$$H(\tilde{X}|Y_1) \geq H(\tilde{X}|Y_2),$$

then

$$R(0, D_2) = H(\tilde{X}|Y_1) + S(D_2).$$

**Proof:** The achievability of Theorem 6 follows from Lemma 2 with $C = \tilde{X}$ and $A = $ constant. The converse follows from Lemma 2.

The next corollary generalises Corollary 3.2 from physically degraded to the conditionally less noisy setting.

**Corollary 6.1:** Consider a two-source $(X_1, X_2, Y_1, Y_2)$ with component Hamming distortion functions. If

$$(Y_2 \geq Y_1 | X_1) \quad \text{and} \quad H(X_1|Y_1) \geq H(X_1|Y_2),$$

then

$$R(0, 0) = H(X_1|Y_1) + H(X_2|X_1, Y_2).$$

**Proof:** The proof follows from Theorem 6 upon noting $\tilde{X} = X_1$ and $S(0) = H(X_2|X_1, Y_2)$.

**Example 2:** Suppose that $X_1$ and $Z$ are independent Bernoulli random variables with

$$P[X_1 = 0] = P[X_1 = 1] = \frac{1}{2}$$

and

$$P[Z = 0] = 1 - P[Z = 1] = \frac{1}{3}.$$

Let

$$X_2 = X_1 \oplus Z.$$

Furthermore, let $Y_2$ and $Y_1$ be the outcomes of passing $X_1$ through two independent channels: A BEC(2/3) and a BSC(1/4) respectively, see Fig. 2.

We have $(Y_2 \geq Y_1 | X_1)$ from condition (ii) of Lemma 4. Moreover,

$$H(X_1|Y_2) = 2/3$$

is smaller than

$$H(X_1|Y_1) = H_b(1/4) \approx 0.8113,$$

where

$$H_b(\alpha) := -\alpha \log_2 \alpha - (1 - \alpha) \log_2 (1 - \alpha)$$

is the binary entropy function. From Corollary 6.1 we have

$$R(0, 0) = H_b(1/4) + H_b(1/3).$$

Finally, we notice that the side information $(Y_1, Y_2)$ is not physically or stochastically degraded with respect to $X_1$ [14, p. 121], [29], and hence with respect to $X = (X_1, X_2)$.

**Remark 2:**

(i) Theorem 6 includes Corollary 3.1 for physically degraded side information as a special case, since $X \leftrightarrow Y_2 \leftrightarrow Y_1$

(ii) It appears that our approach to proving Lemma 5 does not readily generalise to an arbitrary distortion function $\delta_1$. An apparent difficulty follows from the use of the Wyner-Ziv style converse argument to construct the $S(D_2)$ term using $(\tilde{X}, Y_1)$ as side information. The argument needs $(\tilde{X}, Y_1)$ to be i.i.d., and this need not be the case when $\delta_1$ is arbitrary.

(iii) Theorem 6 employs the conditionally less noisy definition for the special case where $L$ is a deterministic function of the source $X$. In this case, we can remove $L$ from the Markov chain in Definition 6.

(iv) If $L = \emptyset$, then Definition 6 reduces to the less noisy concept for information-theoretic security for source coding recently introduced by Villard and Piantanida [30]. In fact, recall Example 1 with $Pr[X_2 = 0] = p$ and $Pr[Z = 0] = r$. If $r$ is sufficiently small (or large) compared to $p$ so that

$$H(X_1|Y_1) < H(X_2),$$

the side information $Y_2$ is conditionally less noisy than $Y_1$ given $X_2$, but $Y_2$ is not less noisy than $Y_1$. To see the latter, select $W = X_1$. We have

$$I(W; Y_1) = H(X_1) - H(X_1|Y_1)$$

and

$$I(W; Y_2) = H(X_1) - H(X_1|Y_2) = H(X_1) - H(X_2),$$

and thus $I(W; Y_1) > I(W; Y_2)$.

**IV. SUCCESSIVE REFINEMENT WITH SIDE INFORMATION**

The method used in Appendix 4 to prove Lemma 5 can, with appropriate modification, yield useful converses for various generalisations of Heegard and Berger’s RD problem. In this section, we extend the setup of Fig. 1 to two different successive-refinement problems with receiver side information.
A. Problem Formulation

Consider a tuple of random variables \((X, Y_1, Y_2, Y_3)\) with an arbitrary joint distribution. Let \((X, Y_1, Y_2, Y_3)\) denote a string of \(n\) i.i.d. copies of \((X, Y_1, Y_2, Y_3)\). A successive-refinement block code for the setup shown in Fig. 3 consists of four (possibly stochastic) maps

\[
f : \mathcal{X} \rightarrow \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3
\]

and

\[
g_1 : \mathcal{M}_1 \times \mathcal{Y}_1 \rightarrow \hat{X}_1
\]

\[
g_2 : \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{Y}_2 \rightarrow \hat{X}_2
\]

\[
g_3 : \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3 \times \mathcal{Y}_3 \rightarrow \hat{X}_3
\]

where \(\mathcal{M}_1, \mathcal{M}_2\) and \(\mathcal{M}_3\) are finite index sets. The transmitter sends \((\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3) := f(X)\) over the noiseless channels, as shown in Fig. 3. Receiver 1 reconstructs \(\hat{X}_1 := g_1(\mathcal{M}_1, Y_1)\), receiver 2 reconstructs \(\hat{X}_2 := g_2(\mathcal{M}_1, \mathcal{M}_2, Y_2)\) and receiver 3 reconstructs \(\hat{X}_3 := g_3(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, Y_3)\).

Definition 7: A rate tuple \((R_1, R_2, R_3)\) is said to be achievable with distortions \((D_1, D_2, D_3)\) if for any \(\epsilon > 0\) there exists a block code \((f, g_1, g_2, g_3)\), with some sufficiently large blocklength \(n\), satisfying

\[
R_j + \epsilon \geq \frac{1}{n} \log |\mathcal{M}_j|
\]

and

\[
D_j + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} \delta_j(X_j, \hat{X}_j,i)
\]

for \(j = 1, 2, 3\).

Definition 8: The RD region \(\mathcal{R}(D_1, D_2, D_3)\) is the set of all rates \((R_1, R_2, R_3)\) that are achievable with distortions \((D_1, D_2, D_3)\).

B. Three Stages with \(Y_3\) better than \(Y_2\) better than \(Y_1\) (starting from \(X \leftrightarrow Y_3 \leftrightarrow Y_2 \leftrightarrow Y_1\))

Let us now assume that Receiver 3 obtains the best side information and Receiver 1 the worst. Tian and Diggavi [16] modelled such a relation with physically degraded side information, that is, \(X \leftrightarrow Y_3 \leftrightarrow Y_2 \leftrightarrow Y_1\), and they derived the corresponding RD region. The goal here is to broaden their result to a conditionally less noisy setup.

We will need the following achievable RD region that holds for arbitrarily distributed side information. The region is distilled from a more general achievability result in [7], see Appendix F.

Let \(\mathcal{R}_{in}(D_1, D_2, D_3)\) denote the set of all rate tuples \((R_1, R_2, R_3)\) for which there exists an auxiliary tuple \((A_1, A_2, A_3)\), jointly distributed with \((X, Y_1, Y_2, Y_3)\), such that

(i) \((A_1, A_2, A_3) \leftrightarrow X \leftrightarrow (Y_1, Y_2, Y_3)\) forms a Markov chain;

(ii) The auxiliary alphabet cardinalities are bounded by \(\Phi \)

\[
|A_1| \leq |X| + 6
\]

\[
|A_2| \leq |X| |A_1| + 4
\]

\[
|A_3| \leq |X| |A_1| |A_2| + 1.
\]

(iii) There exist (deterministic) maps for each \(j = 1, 2, 3\)

\[
\phi_j : A_j \times Y_j \rightarrow \hat{X}_j
\]

with

\[
D_j \geq \mathbb{E} \delta_j(X, \phi_j(A_j, Y_j)).
\]

(iv) The rate tuple \((R_1, R_2, R_3)\) satisfies

\[
R_1 \geq I(X; A_1|Y_1),
\]

\[
R_1 + R_2 \geq \max_{j=1,2} I(X; A_1|Y_j) + I(X; A_2|A_1, Y_2)
\]

and

\[
R_1 + R_2 + R_3 \geq \max_{j=1,2,3} I(X; A_1|Y_j)
\]

\[
+ \max_{j=2,3} I(X; A_2|A_1, Y_j)
\]

\[
+ I(X; A_3|A_1, A_2, Y_3).
\]

Lemma 7:

\[
\mathcal{R}_{in}(D_1, D_2, D_3) \subseteq \mathcal{R}(D_1, D_2, D_3).
\]

The next theorem, which is due to Tian and Diggavi [16], shows that the entire RD region is subsumed by \(\mathcal{R}_{in}(D_1, D_2, D_3)\) whenever the side information is physically degraded.

Theorem 8: If the side information is physically degraded \(X \leftrightarrow Y_3 \leftrightarrow Y_2 \leftrightarrow Y_1\), then [16] Thm. 1

\[
\mathcal{R}_{in}(D_1, D_2, D_3) = \mathcal{R}(D_1, D_2, D_3).
\]

Moreover, the rate constraints defining \(\mathcal{R}_{in}(D_1, D_2, D_3)\) simplify to

\[
R_1 \geq I(X; A_1|Y_1)
\]

\[
R_1 + R_2 \geq I(X; A_1|Y_1) + I(X; A_2|A_1, Y_2)
\]

\[
R_1 + R_2 + R_3 \geq I(X; A_1|Y_1) + I(X; A_2|A_1, Y_2)
\]

\[
+ I(X; A_3|A_1, A_2, Y_3),
\]

where \(A_1, A_2\) and \(A_3\) obey the same cardinality constraints as those for \(\mathcal{R}_{in}(D_1, D_2, D_3)\), see also [16] Thm. 1.

The achievability part of Theorem 8 is given by Lemma 7 and the simplified rate constraints follow from degraded side

Reference [7] does not provide cardinality constraints, and these bounds follow by the standard convex cover method.
information (the Markov chain $X \leftrightarrow Y_3 \leftrightarrow Y_2 \leftrightarrow Y_1$). The converse assertion was proved by Tian and Diggavi in [16].

We now consider Theorem 8 with deterministic distortion functions at receivers 1 and 2. In particular, receivers 1 and 2 wish to reconstruct almost losslessly

$$\hat{X}_1 := \psi_1(X) \quad \text{and} \quad \hat{X}_2 := \psi_2(X),$$

respectively, where $\psi_1$ and $\psi_2$ are functions of the form

$$\psi_j : \mathcal{X} \rightarrow \mathcal{X}_j, \quad j = 1, 2.$$

Theorem 8 with deterministic $\delta_1$ and $\delta_2$, simplifies as follows. Define

$$S'(D_3) := \min I(X; A_3 \mid \hat{X}_1, \hat{X}_2, Y_3), \quad D_3 \geq 0,$$

where the minimisation is taken over all auxiliary $A_3$, jointly distributed with $(X, Y_1, Y_2, Y_3)$, such that

(i) $A_3 \leftrightarrow X \leftrightarrow (Y_1, Y_2, Y_3)$ forms a Markov chain;
(ii) $|A_3| \leq |\mathcal{X}| + 1$;
(iii) there exists a (deterministic) map

$$\phi_3 : A_3 \times \hat{X}_1 \times \hat{X}_2 \times Y_3 \rightarrow \hat{X}_3$$

with

$$D_3 \geq \delta_3 (X, \phi_3(A_3, \hat{X}_1, \hat{X}_2, Y_3)).$$

**Corollary 8.1:** If the side information is physically degraded $X \leftrightarrow Y_3 \leftrightarrow Y_2 \leftrightarrow Y_1$ and the distortion functions $\delta_1$ and $\delta_2$ are deterministic, then $R(0, 0, D_3)$ is equal to the set of all rate tuples $(R_1, R_2, R_3)$ satisfying

$$R_1 \geq H(\hat{X}_1 | Y_1)$$

$$R_1 + R_2 \geq H(\hat{X}_1 | Y_1) + H(\hat{X}_2 | \hat{X}_1, Y_2)$$

$$R_1 + R_2 + R_3 \geq H(\hat{X}_1 | Y_1) + H(\hat{X}_2 | \hat{X}_1, Y_2) + S'(D_3).$$

**Proof:** The achievability part follows directly from Theorem 8 upon selecting the auxiliary random variables as $A_1 = \hat{X}_1$ and $A_2 = \hat{X}_2$ as well as recalling the definition of $S'(D_3)$. The converse can be proved following arguments similar to those used in Appendix E and is omitted.

The next lemma is a converse for deterministic distortion functions $\delta_1$ and $\delta_2$ and arbitrarily distributed side information; it is a successive-refinement version of Lemma 5. Let $R_{\text{out}}(D_3)$ denote the set of all rate tuples $(R_1, R_2, R_3)$ for which

$$R_1 \geq H(\hat{X}_1 | Y_1)$$

$$R_1 + R_2 \geq H(\hat{X}_1 | Y_1) + H(\hat{X}_2 | \hat{X}_1, Y_2)$$

$$+ \min_W \left\{ I(W; Y_2 | \hat{X}_1) - I(W; Y_1 | \hat{X}_1) \right\}$$

and

$$R_1 + R_2 + R_3 \geq H(\hat{X}_1 | Y_1) + H(\hat{X}_2 | \hat{X}_1, Y_2) + S'(D_3)$$

$$+ \min_W \left\{ I(W; Y_2 | \hat{X}_1) - I(W; Y_1 | \hat{X}_1) \right\}$$

where each minimisation is independently taken over an auxiliary random variable $W$, jointly distributed with $(X, Y_1, Y_2, Y_3)$, such that $|W| \leq |\mathcal{X}|$ and $W \leftrightarrow X \leftrightarrow (Y_1, Y_2, Y_3)$.

**Lemma 9 (Converse):** If $\delta_1$ and $\delta_2$ are deterministic distortion functions, then

$$R_{\text{out}}(D_3) \supseteq R(0, 0, D_3).$$

Our proof of Lemma 9 is quite similar to that of Lemma 5 and it is given in Appendix C.

The next theorem shows that the outer bound (converse) of Lemma 9 matches the inner bound (achievability) of Lemma 7 for a certain conditionally less noisy setting.

**Theorem 10:** If $\delta_1$ and $\delta_2$ are deterministic distortion functions, $(Y_2 \geq Y_1 \mid \hat{X}_1)$, $(Y_3 \geq Y_2 \mid \hat{X}_1, \hat{X}_2)$, and

$$H(\hat{X}_1 | Y_1) \geq \max \{ H(\hat{X}_1 | Y_2), H(\hat{X}_1 | Y_3) \},$$

then $R(0, 0, D_3)$ is equal to the set of all rate tuples $(R_1, R_2, R_3)$ satisfying

$$R_1 \geq H(\hat{X}_1 | Y_1)$$

$$R_1 + R_2 \geq H(\hat{X}_1 | Y_1) + H(\hat{X}_2 | \hat{X}_1, Y_2)$$

$$R_1 + R_2 + R_3 \geq H(\hat{X}_1 | Y_1) + H(\hat{X}_2 | \hat{X}_1, Y_2) + S'(D_3).$$

**Proof:** The converse follows directly by Lemma 9 and uses the conditionally less noisy assumptions $(Y_2 \geq Y_1 \mid \hat{X}_1)$ and $(Y_3 \geq Y_2 \mid \hat{X}_1, \hat{X}_2)$. The achievability follows by Lemma 7 with $A_1 = \hat{X}_1$ and $A_2 = \hat{X}_2$.

**Remark 3:** Theorem 10 includes Corollary 8.1. To see this: The Markov chain $X \leftrightarrow Y_3 \leftrightarrow Y_2 \leftrightarrow Y_1$ implies, by the data processing lemma, that

$$H(\hat{X}_1 | Y_1) \geq H(\hat{X}_1 | Y_2) \geq H(\hat{X}_1 | Y_3).$$

Moreover, we also have $\hat{X}_2 \leftrightarrow (\hat{X}_1, Y_3) \leftrightarrow Y_2$ and therefore

$$H(\hat{X}_2 | \hat{X}_1, Y_3) = H(\hat{X}_2 | \hat{X}_1, Y_2) \leq H(\hat{X}_2 | \hat{X}_1, Y_2).$$

Physical degradedness implies conditionally less noisy: For every auxiliary random variable $W$ satisfying $W \leftrightarrow (X, \hat{X}_1) \leftrightarrow (Y_1, Y_2, Y_3)$ we have $W \leftrightarrow (\hat{X}_1, Y_2) \leftrightarrow Y_1$ and thus

$$I(W; Y_2 | \hat{X}_1) = H(W | \hat{X}_1) - H(W | Y_1, \hat{X}_1, Y_2)$$

$$\geq I(W; Y_1 | \hat{X}_1).$$

The less noisy condition $(Y_3 \geq Y_2 \mid \hat{X}_1, \hat{X}_2)$ follows by a similar argument.

**Remark 4:** Steinberg and Merhav [15] were the first to consider and solve the two-stage successive refinement problem with physically degraded side information. Tian and Diggavi’s work [16] generalised Steinberg and Merhav’s result to three or more stages with physically degraded side information.

### C. Two Stages with $Y_1$ better than $Y_2$ (starting from $X \leftrightarrow Y_1 \leftrightarrow Y_2$)

Reconsider the successive-refinement problem in Fig. 3 but now with only two receivers, receiver 1 and 2. Suppose that the side information at receiver 1 is better than the side
forms a Markov chain. Here we notice that the roles of $Y_1$ and $Y_2$ in the Markov chain (5) are reversed with respect to Definition 3 and Theorem 8. In contrast to Theorem 8, however, there is no known computable expression for the RD region under (6). Tian and Diggavi gave achievability and converse bounds in [17], and they show that these bounds match for degraded deterministic distortion measures. In this section, we relax the Markov chain in (5) to a conditionally less noisy setting.

The next lemma gives an achievable rate region for arbitrarily distributed side information. The rate constraints can be distilled from those in [7], see Appendix F and the cardinality bounds can be derived by the standard convex cover method [13]. The lemma includes Tian and Diggavi’s bound [17] Cor. 1 for arbitrarily distributed side information as a special case.

Let $R^*_m(D_1, D_2)$ denote the set of all rate pairs $(R_1, R_2)$ for which there exists a tuple of auxiliary random variables $(A_{12}, A_1, A_2)$, jointly distributed with $(X, Y_1, Y_2)$, such that

(i) $(A_{12}, A_1, A_2) \leftrightarrow X \leftrightarrow (Y_1, Y_2)$ forms a Markov chain;

(ii) the auxiliary alphabet cardinalities satisfy

$$|A_{12}| \leq |X| + 3$$
$$|A_1| \leq |X| |A_{12}| + 1$$
$$|A_2| \leq |X| |A_{12}| + 1$$;

(iii) there exist deterministic maps for $j = 1, 2$

$$\phi_j : A_j \times Y_j \rightarrow \hat{X}_j,$$

with

$$D_j \geq E \delta_j(X, \phi_j(A_j, Y_j));$$

(iv) the rate pair $(R_1, R_2)$ satisfies

$$R_1 \geq I(X; A_{12}, A_1|Y_1)$$

$$R_1 + R_2 \geq \max \left\{ I(X; A_{12}|Y_1), I(X; A_1|Y_2), I(X; A_2|Y_2) \right\}$$

$$R_1 + R_2 \geq \max \left\{ I(X; A_{12}|Y_1), I(X; A_1|Y_2), I(X; A_2|Y_2) \right\}$$

Lemma 11:

$$R^*_m(D_1, D_2) \subseteq R(D_1, D_2).$$

The next and final result of the paper generalises Tian and Diggavi’s result [17] Thm. 4, which holds under the Markov chain in (5), to a conditionally less noisy setting. Suppose that $\delta_1$ and $\delta_2$ are deterministic distortion functions, with $\tilde{X}_1 = \psi_1(X)$ and $\tilde{X}_2 = \psi_2(X)$. It is said that $\delta_2$ is a degraded version of $\delta_1$ if

$$\psi_2 = \psi' \circ \psi_1$$

for some deterministic map $\psi'$. The next theorem is proved in Appendix H.

**Theorem 12**: Suppose that $\delta_1$ and $\delta_2$ are deterministic distortion functions.

(i) If $\delta_2$ is a degraded version of $\delta_1$,

$$H(\tilde{X}_2|Y_1) \leq H(\tilde{X}_2|Y_2)$$

and $(Y_1 \geq Y_2 \mid \tilde{X}_2)$,

then $R^*_m(0, 0) = R(0, 0)$ and the rate constraints of 9 simplify to

$$R_1 \geq H(\tilde{X}_1|Y_1)$$

$$R_1 + R_2 \geq H(\tilde{X}_2|Y_2) + H(\tilde{X}_1|\tilde{X}_2, Y_1).$$

(ii) If $\delta_1$ is a degraded version of $\delta_2$ and

$$H(\tilde{X}_1|Y_1) \leq H(\tilde{X}_1|Y_2)$$

then $R^*_m(0, 0) = R(0, 0)$ and the rate constraints (4) simplify to

$$R_1 \geq H(\tilde{X}_1|Y_1)$$

$$R_1 + R_2 \geq H(\tilde{X}_2|Y_2).$$

Remark 5: Theorem 12 applies to the reverse degraded side information case, since by Lemma 4 (i) the Markov chain $X \leftrightarrow Y_1 \leftrightarrow Y_2$ implies $(Y_1 \geq Y_2 \mid \tilde{X}_2)$ and by the data processing lemma it also implies $H(X_j|Y_1) \leq H(\tilde{X}_j|Y_2)$ for $j = 1, 2$.

**APPENDIX A**

**PROOF OF LEMMA 11**

We first notice that

$$I(J; S_2|L) - I(J; S_1|L) = I(J; S_2, L) - I(J; S_1, L).$$

(7)

by the chain rule for mutual information. Expand the first mutual information on the right hand side of (7) as follows:

$$I(J; S_2, L)$$

$$\begin{array}{c}
\overset{(a)}{=} \sum_{i=1}^n I(J; S_{2,i}, L_i; S_{2,i}^{i-1}, L_1^{i-1}) \\
\overset{(b)}{=} \sum_{i=1}^n I(J; S_{2,i}^{i-1}, L_1^{i-1}; S_{2,i}, L_i) \\
\overset{(c)}{=} \sum_{i=1}^n \left[ I(J; S_{1,i+1}, S_{2,i}^{i-1}, L_1^{i-1}, L_{i+1}^{i-1}; S_{2,i}, L_i) \\
- I(S_{1,i+1}, L_{i+1}^{i-1}; S_{2,i}, L_i|J, S_{2,i}^{i-1}, L_1^{i-1}) \right] \\
\overset{(d)}{=} \sum_{i=1}^n \left[ I(W_i; S_{2,i}, L_i) \\
- I(S_{1,i+1}, L_{i+1}^{i-1}; S_{2,i}, L_i|J, S_{2,i}^{i-1}, L_1^{i-1}) \right] \\
\end{array}$$

(8)

where (a) and (c) follow from the chain rule for mutual information; (b) exploits the fact that the source is i.i.d. and

$$H(S_{2,i}, L_i; S_{2,i}^{i-1}, L_1^{i-1}) = H(S_{2,i}, L_i);$$

and, finally, in (d) we define and substitute the random variable

$$W_i := (J, S_{2,i}^{i-1}, S_{2,i}^{i-1}, L_1^{i-1}, L_{i+1}^{i-1}).$$

(9)

Expand the second mutual information on the right hand side of (7) as follows:

$$I(J; S_1, L)$$
\[
\sum_{i=1}^{n} \left( I(J, S_{2,1}^n, L_{i-1}; S_{1,i}, L_i^n) - I(J, S_{2,1}^n, L_{i-1}; S_{1,i+1}, L_{i+1}^n) \right)
\]

\[
\sum_{i=1}^{n} \left( I(J, S_{2,1}^n, L_{i-1}; S_{1,i}, L_i^n) - I(S_{2,i}, L_i; S_{1,i+1}^n, L_{i+1}^n) \right)
\]

\[
\sum_{i=1}^{n} \left( I(J, S_{2,1}^n, L_{i-1}; S_{1,i}, L_i^n) - I(S_{2,i}, L_i; S_{1,i+1}^n, L_{i+1}^n) \right)
\]

\[
\sum_{i=1}^{n} \left( W_i; S_{1,i}, L_i \right)
\]

\[
I(J, S_2, L) - I(J, S_1, L)
\]

\[
= n \left( I(\bar{W}; S_2, L) - I(\bar{W}; S_1, L) \right)
\]

We also notice that

\[
W_i \leftrightarrow (R_i, L_i) \leftrightarrow (S_{1,i}, S_{2,i}, T_i),
\]

forms a Markov chain for all \( i = 1, 2, \ldots, n \). Each of the \( n \) Markov chains in (14) follows from the definition of \( W_i \); the \( n \)-letter Markov chain

\[
J \leftrightarrow (R, L) \leftrightarrow (S_1, S_2, T),
\]

and the fact that \((R, S_1, S_2, T, L)\) is i.i.d. Now define

\[
R = R_Q \quad \text{and} \quad T = T_Q.
\]

Using the independence of \( Q \) from \((R, T, S_1, S_2, L)\), we have the desired Markov chain,

\[
\bar{W} \leftrightarrow (R, L) \leftrightarrow (S_1, S_2, T).
\]

It remains to show that the auxiliary random variable \( \bar{W} \), whose alphabet cardinality is unbounded in \( n \), can be replaced by some \( W \) with an alphabet satisfying \(|W| \leq |R||L|\). We now prove the existence of such using the convex cover method, for example, [14] App. C.

For each and every \( \tilde{w} \) in the support set of \( \bar{W} \), let \( q_{\tilde{w}} \) denote the conditional distribution of \((R, S_1, S_2, T, L)\) given \( \bar{W} = \tilde{w} \). Let \( P \) denote the set of all joint distributions on \( R \times S_1 \times S_2 \times T \times L \).

For each and every pair \((r, l)\) in \( R \times L \) but one — the omitted pair, say \((r^*, l^*)\), can be chosen arbitrarily — define the functional \( g_{r,l} : P \rightarrow [0, 1], \)

\[
ge_{r,l}(q) := \sum_{s_1 \in S_1} \sum_{s_2 \in S_2} \sum_{t \in T} q(r, s_1, s_2, t, l).
\]

The \(|P||L| - 1\)-functionals defined in (16) will be used to preserve the joint distribution of \((R, S_1, S_2, T, L)\) when the Support Lemma [14] Sec. App. C) is invoked shortly. Indeed, we notice that for each such pair \((r, l)\) the expectation

\[
\mathbb{E}_{\bar{W}} \left\{ g_{r,l}(q_{\bar{w}}) \right\} = \sum_{\tilde{w} \in \bar{W}} P[\bar{W} = \tilde{w}] g_{r,l}(q_{\tilde{w}})
\]

is equal to the true probability \( P[(R, L) = (r, l)] \). Moreover, this agreement extends over \( R \times S_1 \times S_2 \times T \times L \) because

\[
\mathbb{E} \left\{ g_{r,l}(q_{\bar{w}}) \right\} \cdot P[S_1 = s_1, S_2 = s_2, T = t | R = r, L = l] = 1
\]

is equal to the true joint probability \( P[R = r, S_1 = s_1, S_2 = s_2, T = t, L = l] \).

If the joint distribution of \((R, L, S_1, S_2, T)\) is preserved, we can additionally preserve the difference

\[
I(\bar{W}; S_2, L) - I(\bar{W}; S_1, L)
\]

by simply preserving \( H(S_2, L|\bar{W}) - H(S_1, L|\bar{W}) \). To this end, define the functional \( g : P \rightarrow [-|S_1|, |S_2|], \)

\[
g(q) := H(S_2, L) - H(S_1, L),
\]
where the joint distribution of \((R, S_1, S_2, T, L)\) is understood to be given by \(q\). We also notice that
\[
E_W\{g(q_{W})\} = \sum_{\tilde{w} \in \tilde{W}} P[W = \tilde{w}] g(q_{\tilde{w}}) = H(S_2, L | \tilde{W}) - H(S_1, L | \tilde{W}).
\]

The Support Lemma asserts that there exists an auxiliary random variable \(W\) defined on an alphabet \(\mathcal{W}\) with cardinality
\[
|\mathcal{W}| \leq |\mathcal{R}| |\mathcal{L}|
\]
and a collection of (conditional) joint distributions \(\{q_w\}\) from \(\mathcal{P}\), indexed by the elements \(w\) of \(\mathcal{W}\), such that
\begin{enumerate}[(i)]  
  
  \item for all \((r, t)\) in \(\mathcal{R} \times \mathcal{L}\) — excluding the omitted pair \((r^*, t^*)\) — we have
  \[
  E_W\{g_{r,t}(q_{W})\} = E_W\{g_{r,t}(q_{W(t)})\},
  \]
  (20)
  
  \item and
  \[
  E_W\{g(q_{W})\} = E_{\tilde{W}}\{g(q_{\tilde{W}})\}.
  \]
  (21)
\end{enumerate}

The new auxiliary random variable \(W\) and the distributions \(\{q_w\}\) induce a joint distribution on \(\mathcal{W} \times \mathcal{R} \times \mathcal{L}\). The equality (20) ensures that the \((R, L)\)-marginal of this new distribution is equal to the true distribution of \((R, L)\). This agreement extends to the full joint distribution via (17); that is, we impose the Markov chain
\[
W \leftrightarrow (R, L) \leftrightarrow (S_1, S_2, T).
\]
Finally, the equalities (20) and (21) imply
\[
I(W; S_2, L) - I(W; S_1, L) = I(\tilde{W}; S_2, L) - I(\tilde{W}; S_1, L).
\]
(23)

For the case when \(L\) is a function of \(R\): The tighter cardinality bound \(|\mathcal{W}| \leq |\mathcal{R}|\) can be proved using the above method with the following modifications. If \(L\) is a function of \(R\), then \(L \leftrightarrow R \leftrightarrow \tilde{W}\) from (15) we have
\[
\tilde{W} \leftrightarrow R \leftrightarrow (L, S_1, S_2, T).
\]
(24)

Replace the \((|\mathcal{R}| |\mathcal{L}| - 1)\) functionals \(\{g_{r,t}\}\) in (16) by
\[
g_r(q) := \sum_{s_1 \in \mathcal{S}_1} \sum_{s_2 \in \mathcal{S}_2} \sum_{t \in \mathcal{L}} q(r, s_1, s_2, t, l)
\]
(25)

for all \(r\) in \(\mathcal{R}\) but one. The \((|\mathcal{R}| - 1)\)-functionals in (25) combined with the Markov chain (24) are sufficient to preserve the joint distribution of \((R, S_1, S_2, T, L)\) using the Support Lemma. The remainder of the proof remains unchanged except that \(g_r\) replaces \(g_{r,t}\) in (20) and \(W \leftrightarrow R \leftrightarrow (L, S_1, S_2, T)\) replaces (22).

\begin{remark}
\begin{enumerate}[(i)]  
  
  \item The proof of Lemma [1] can be manipulated so as to replace the telescoping sum step (10) with a Csiszár sum identity [14] Sec. 2.4 step. We feel that the telescoping approach gives a cleaner proof.
  
  \item We note that steps (a) and (b) of (10) are reminiscent of those used in Kramer’s converse for the Gelfand-Pinsker problem (coding for channels with state), see [31] Sec. F or [32] Sec. 6.6. It is not clear, as yet, whether there is a deeper relationship between the two problems.
\end{enumerate}
\end{remark}
The Support Lemma asserts that there exists a new auxiliary random variable $C^\dagger$ defined on an alphabet $C^\dagger$ with cardinality $|C^\dagger| \leq |X| + 3$


together with a collection of $|C^\dagger|$ distributions $\{q^\dagger_i\}$ from $\mathcal{P}_1$— indexed by the elements $c$ of $C^\dagger$— such that

$$
\mathbb{E}_C\{g_x(q_C)\} = \mathbb{E}_{C^\dagger}\{g_x(q^\dagger_{C^\dagger})\}, \quad \forall x \in \mathcal{X} \text{ except } x^* \tag{29}
$$

and

$$
\mathbb{E}_C\{g_j(q_C)\} = \mathbb{E}_{C^\dagger}\{g_j(q^\dagger_{C^\dagger})\}, \quad \forall j = 1, 2, 3, 4. \tag{30}
$$

The new variable $C^\dagger$, the distributions $\{q^\dagger_i\}$, and the true state information channel come together via the Markov chain

$$(A^\dagger, B^\dagger, C^\dagger) \leftrightarrow X \leftrightarrow (Y^\dagger_1, Y^\dagger_2) \tag{31}$$

to specify a tuple $(A^\dagger, B^\dagger, C^\dagger, X^\dagger_1, Y^\dagger_1, Y^\dagger_2)$ on $A \times B \times C \times X \times Y_1 \times Y_2$. The equality (29) ensures that $(X^\dagger_1, Y^\dagger_1, Y^\dagger_2)$ and $(X, Y_1, Y_2)$ have the same distribution, which also implies

$$H(X^\dagger_1|Y^\dagger_1) = H(X|Y^\dagger_1) \quad \text{and} \quad H(X^\dagger_1|Y^\dagger_2) = H(X|Y^\dagger_2). \tag{32}$$

Similarly, (30) ensures

$$I(X^\dagger_1, B^\dagger|Y^\dagger_1, C^\dagger) - H(X^\dagger_1|A^\dagger, C^\dagger, Y^\dagger_1) = I(X; B|Y_2, C) - H(X|A, C, Y_1); \tag{33a}$$

$$I(X^\dagger_1; A^\dagger|Y^\dagger_1, C^\dagger) - H(X^\dagger_1|B^\dagger, C^\dagger, Y^\dagger_2) = I(X; A|Y_1, C) - H(X|B, C, Y_2); \tag{33b}$$

and

$$\min_{\phi_i: A \times C \times X \rightarrow Y_1} \mathbb{E}_i \delta_1\left(X^\dagger, \phi_i(A^\dagger, C^\dagger, Y^\dagger_1)\right) = \min_{\phi_i: A \times C \times X \rightarrow Y_1} \mathbb{E}_i \delta_1\left(X, \phi_i(A, C, Y_1)\right) \tag{34a}$$

$$\min_{\phi_j: B \times C \times X \rightarrow Y_2} \mathbb{E}_j \delta_2\left(X^\dagger, \phi_j(B^\dagger, C^\dagger, Y^\dagger_2)\right) = \min_{\phi_j: B \times C \times X \rightarrow Y_2} \mathbb{E}_j \delta_2\left(X, \phi_j(B, C, Y_2)\right). \tag{34b}$$

Finally, the equalities (32) and (33) together give

$$I(X^\dagger_1, C^\dagger|Y^\dagger_1) + I(X^\dagger_1; A^\dagger|C^\dagger, Y^\dagger_1) + I(X^\dagger_1; B^\dagger|C^\dagger, Y^\dagger_2) = I(X; C|Y_1) + I(X; A|C, Y_1) + I(X; B|C, Y_2)$$

and

$$I(X^\dagger_1, C^\dagger|Y^\dagger_2) + I(X^\dagger_1; A^\dagger|C^\dagger, Y^\dagger_1) + I(X^\dagger_1; B^\dagger|C^\dagger, Y^\dagger_2) = I(X; C|Y_2) + I(X; A|C, Y_1) + I(X; B|C, Y_2)$$

and therefore

$$\max_{j=1,2} I(X; C|Y^\dagger_j) + I(X; A|C, Y_1) + I(X; B|C, Y_2)$$

$$= \max_{j=1,2} I(X^\dagger_1; C^\dagger|Y^\dagger_j) + I(X^\dagger_1; A^\dagger|C^\dagger, Y^\dagger_1)$$

$$+ I(X^\dagger_1; B^\dagger|C^\dagger, Y^\dagger_2). \tag{35}$$

Consider the tuple $(A^\dagger, B^\dagger, C^\dagger, X^\dagger_1, Y^\dagger_1, Y^\dagger_2)$. We have the Markov chain (31) by construction, and we notice that $A^\dagger$ and $B^\dagger$ always appear separately in (33) and (34). We may therefore replace the joint distribution of $(A^\dagger, B^\dagger, C^\dagger, X^\dagger_1, Y^\dagger_1, Y^\dagger_2)$ with another that shares the same Markov chain (31) and marginals $(A^\dagger, C^\dagger, X^\dagger_1), (B^\dagger, C^\dagger, X^\dagger_1)$ and $(X^\dagger_1, Y^\dagger_1, Y^\dagger_2)$, but imposes the new chain

$$A^\dagger \leftrightarrow (C^\dagger, X^\dagger_1) \leftrightarrow B^\dagger. \tag{36}$$

Or put another way, the Markov chain (36) does not alter the left hand sides of (33) or (34). The chain (36) will be important in the sequel because it allows the cardinalities of $A$ and $B$ to be bounded independently. With a slight abuse of notation, we retain the same notation $(A^\dagger, B^\dagger, C^\dagger, X^\dagger_1, Y^\dagger_1, Y^\dagger_2)$ for this new distribution.

Consider the variable $A^\dagger$. For each and every $a$ in the support set of $A^\dagger$, let $q_a$ denote the conditional distribution of $(C, X^\dagger_1)$ given $A^\dagger = a$. Let $\mathcal{P}_2$ denote the set of all joint distributions on $C \times X$. For each and every $(c, x)$ in $C \times X$ but one, define $g_{c,x}: \mathcal{P}_2 \rightarrow [0, 1]$ by setting

$$g_{c,x}(q) := q(c, x).$$

Here

$$E_{A^\dagger}\{g_{c,x}(q_{A^\dagger})\} = P[(C, X^\dagger_1) = (c, x)]$$

returns the desired probability for all $(c, x)$ in $C \times X$ but one. In addition, define

$$g_5(q) := H(X|C, Y_1)$$

and

$$g_6(q) := \sum_{c \in C} \sum_{x \in X}\min_{\hat{x}}\sum_{y_1, y_2} q(c, x)p(y_1, y_2|x)\delta_1(\hat{x}, x),$$

where the joint distribution of $(C, X, Y_1, Y_2)$ is understood as follows: $(C, X)$ is distributed according to $q$, and $(Y_1, Y_2)$ conditionally depends on $X$ via the true state information channel. We have

$$E_{A^\dagger}\{g_5(q_{A^\dagger})\} = H(X^\dagger_1|A^\dagger, C^\dagger, Y^\dagger_1),$$

and

$$E_{A^\dagger}\{g_6(q_{A^\dagger})\} = \min_{\phi_i: C \times X \rightarrow Y^\dagger_i} \mathbb{E}_i \delta_1\left(X, \phi_i(A^\dagger, C^\dagger, Y^\dagger_i)\right).$$

The Support Lemma asserts that there exists a random variable $A\dagger$ defined on an alphabet $A\dagger$ with cardinality

$$|A\dagger| \leq |C\dagger| \times |X| + 1$$

together with a collection of $|A\dagger|$ distributions $\{q^\dagger_j\}$ from $\mathcal{P}_2$— indexed by the elements $a$ of $A\dagger$— such that

$$E_{A\dagger}\{g_{c,x}(q_{A\dagger})\} = E_{A\dagger}\{g_{c,x}(q_{A\dagger})\}, \quad j = 5, 6. \tag{37}$$

The new variable $A\dagger$, the distributions $\{q^\dagger_j\}$, the true state information channel, the conditional distribution $P(B^\dagger|X^\dagger_1, C^\dagger)$, and the Markov chains (31) and (36) come together to specify a tuple $(A\dagger, B\dagger, C\dagger, X^\dagger_1, Y^\dagger_1, Y^\dagger_2)$ on $A \times B \times C \times X \times Y_1 \times Y_2$. 
The equalities in \((37)\) ensure that \((C^t, X^t)\) and \((C^t, X^t)\) have the same distribution. By construction, we also have that \((B^t, C^t, X^t, Y^t_1, Y^t_2)\) and \((B^t, C^t, X^t, Y^t_1, Y^t_2)\) have the same distribution, and therefore

\[
\max \left\{ I(X^t; C^t|Y^t_1), I(X^t; C^t|Y^t_2) \right\} + H(X^t; B^t|C^t, Y^t_2) = \max \left\{ I(X^t; C^t|Y^t_1), I(X^t; C^t|Y^t_2) \right\} + H(X^t; B^t|C^t, Y^t_2).
\]

Combining \((35), (34), (39), (40)\) and \((41)\) gives

\[
H(X^t|A^t, C^t, Y^t_1) = H(X^t|A^t, C^t, Y^t_1).
\]

and

\[
\min \phi^t_{1: A^t \times C^t \times Y^t_1 \rightarrow X^t} \mathbb{E} \delta_1(X^t, \phi^t_1(A^t, C^t, Y^t_1)) = \min \phi^t_{1: A^t \times C^t \times Y^t_1 \rightarrow X^t} \mathbb{E} \delta_1(X^t, \phi^t_1(A^t, C^t, Y^t_1)).
\]

Combining \((35), (34), (39), (40)\) and \((41)\) gives

\[
\max \left\{ I(X^t; C^t|Y^t_1), I(X^t; C^t|Y^t_2) \right\} + I(X^t; A^t|C^t, Y^t_1) + I(X^t; B^t|C^t, Y^t_2) = \max \left\{ I(X^t; C|Y^t_1), I(X^t; C|Y^t_2) \right\} + I(X^t; A|C, Y^t_1) + I(X^t; B|C, Y^t_2).
\]

as well as

\[
\min \phi^t_{1: A^t \times C^t \times Y^t_1 \rightarrow X^t} \mathbb{E} \delta_1(X^t, \phi^t_1(A^t, C^t, Y^t_1)) = \min \phi^t_{1: A^t \times C^t \times Y^t_1 \rightarrow X^t} \mathbb{E} \delta_1(X, \phi^t_1(A, C, Y^t_1))
\]

and

\[
\min \phi^t_{2: B^t \times Y^t_2 \rightarrow X^t} \mathbb{E} \delta_2(X^t, \phi^t_2(B^t, C^t, Y^t_2)) = \min \phi^t_{2: B^t \times Y^t_2 \rightarrow X^t} \mathbb{E} \delta_2(X, \phi^t_2(B, C, Y^t_2)),
\]

as desired.

Using analogous arguments as above, we can find a random vector \((A', B', C', X', Y'_1, Y'_2)\) over \(A^t \times B^t \times C^t \times X' \times Y'_1 \times Y'_2\), where the cardinality of the alphabet \(B'\) satisfies

\[
|B'| \leq |C^t||X'| + 1,
\]

and such that \((42)\) and \((43)\) are satisfied when the tuple \((A^t, B^t, C^t, X^t, Y^t_1, Y^t_2)\) is replaced by the new tuple \((A', B', C', X', Y'_1, Y'_2)\). This concludes the proof of the cardinality bounds.

\section{APPENDIX C}

\textbf{Proof of Lemma 5}

\begin{itemize}
  \item \textbf{Assertion (i)}
  
  Consider any auxiliary random variable \(W\) for which
  \[
  W \leftrightarrow (X, L) \leftrightarrow (Y_1, Y_2)
  \]
  is a Markov chain. We have
  \[
  I(W; Y_2|L) = H(W|L) - H(W|L, Y_2)
  \]
  \[
  = \left[ H(W|L) - H(W|L, Y_2) \right] \geq H(W|L) - H(W|L, Y_1)
  \]
  \[
  = I(W; Y_1|L),
  \]
  where (a) uses the fact that
  \[
  W \leftrightarrow (Y_2, L) \leftrightarrow Y_1,
  \]
  which follows from \((44)\), the Markov chain \(L \leftrightarrow X \leftrightarrow (Y_1, Y_2)\), and the physically degraded side information.

\item \textbf{B. Assertion (ii)}

  Take any auxiliary random variable \(W\) for which
  \[
  W \leftrightarrow (X_1, X_2) \leftrightarrow (Y_1, Y_2).
  \]
  Consider Definition 5 with \(L = X_1\). We have
  \[
  0 \leq I(W; Y_1|X_1) = H(Y_1|X_1) - H(Y_1|W, X_1)
  \]
  \[
  \geq H(Y_1|X_1, X_2) - H(Y_1|W, X_1)
  \]
  \[
  = H(Y_1|X_1, X_2) - H(Y_1|W, X_1, X_2)
  \]
  \[
  = I(W; Y_1|X_1, X_2),
  \]
  where the indicated steps apply the following Markov chains:
  \[
  (a) \quad X_2 \leftrightarrow X_1 \leftrightarrow Y_1
  \]
  \[
  (b) \quad X_2 \leftrightarrow (W, X_1) \leftrightarrow Y_1
  \]
  \[
  (c) \quad W \leftrightarrow (X_1, X_2) \leftrightarrow (Y_1, Y_2).
  \]
  Thus, we have that
  \[
  I(W; Y_1|X_1) = 0
  \]
  and therefore \(I(W; Y_1|X_1)\) is no larger than \(I(W; Y_2|X_1)\). \hfill \blacksquare
\end{itemize}

\section{APPENDIX D}

\textbf{Proof of Lemma 5}

Fix a distortion \(D_2 \geq 0\) and an \((0, D_2)\)-achievable rate \(R > 0\). By definition, for each \(\epsilon > 0\) we can find a block code \((f, g_1, g_2)\) with sufficiently large blocklength \(n\) such that

\[
R + \epsilon \geq \frac{1}{n} \log |\mathcal{M}|, \tag{45}
\]

\[
\epsilon \geq \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \delta_1(X_1, \tilde{X}_{1,i}), \tag{46}
\]

and

\[
D_2 + \epsilon \geq \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \delta_2(X_1, \tilde{X}_{2,i}). \tag{47}
\]

Fix \(\epsilon > 0\) and consider such a block code. Define \(P_{e,i}\) as the probability that the \(i\)-th symbol \(\tilde{X}_i \equiv \psi(X_i)\) is reconstructed in error at Receiver 1,

\[
P_{e,i} := P[\tilde{X}_{1,i} \neq \tilde{X}_i].
\]
The probability $P_{e,i}$ can be expressed as $P_{e,i} = \mathbb{E} \delta_1(X_i, \tilde{X}_{1,i})$, so by (46)
\[
\frac{1}{n} \sum_{i=1}^{n} P_{e,i} \leq \epsilon. \tag{48}
\]

Consider next the conditional entropy $H(\tilde{X}|M, Y_1)$. Starting from the fact that $\tilde{X}_1$ is determined by $(M, Y_1)$, we have
\[
H(\tilde{X}|M, Y_1) \overset{(a)}{=} H(\tilde{X}|M, Y_1, \tilde{X}_1) \leq H(\tilde{X}|\tilde{X}_1)
\]
\[
\leq \sum_{i=1}^{n} \left( h(P_{e,i}) + P_{e,i} \log |\tilde{X}| \right)
\leq n h\left( \frac{1}{n} \sum_{i=1}^{n} P_{e,i} \right) + \left( \sum_{i=1}^{n} P_{e,i} \right) \log |\tilde{X}|
\leq nh(\epsilon) + n\epsilon \log |\tilde{X}|
\overset{(0)}{=} n\epsilon(\epsilon), \tag{49}
\]
where (a) applies the Markov chain
\[
\tilde{X} \leftrightarrow (M, Y_1) \leftrightarrow \tilde{X}_1;
\]
(b) invokes the chain rule for entropy and the fact that conditioning cannot increase entropy; (c) applies Fano’s inequality; (d) combines the concavity of the binary entropy function with Jensen’s inequality; (e) invokes (48); and (f) substitutes
\[
\epsilon(\epsilon) := h(\epsilon) + \epsilon \log |\tilde{X}|.
\]
Finally, we notice that $\epsilon(\epsilon) \to 0$ as $\epsilon \to 0$.

Now consider the rate condition (43). We have
\[
R + \epsilon \geq \frac{1}{n} \log |M|
\geq \frac{1}{n} H(M)
\geq \frac{1}{n} I(X; M|Y_1)
\geq \frac{1}{n} \left( I(\tilde{X}; M|Y_1) + I(X; M|\tilde{X}, Y_1) \right)
\overset{(a)}{=} \frac{1}{n} \left( H(\tilde{X}|Y_1) - n\epsilon(\epsilon) + I(X; M|\tilde{X}, Y_1) \right)
\overset{(b)}{=} H(\tilde{X}|Y_1) - \epsilon(\epsilon) + \frac{1}{n} I(X; M|\tilde{X}, Y_1), \tag{50}
\]
where (a) substitutes (49) and (b) invokes the fact that $(\tilde{X}, \tilde{X}_1, Y_1)$ is i.i.d.

Consider the conditional mutual information term on the right hand side of (50). Rearranging this term, with the intent of conditioning on $(\tilde{X}, Y_2)$ instead of $(\tilde{X}, Y_1)$, we obtain
\[
I(X; M|\tilde{X}, Y_1)
\overset{(a)}{=} I(X; M|\tilde{X}, Y_2) - H(M|\tilde{X}, Y_2) + H(M|\tilde{X}, Y_1)
= I(X; M|\tilde{X}, Y_2) + I(M; Y_2|\tilde{X}) - I(M; Y_1|\tilde{X}) \tag{51}
\]
where (a) invokes that $M$ is a function of $X$ or, in the more general case of stochastic encoders, that
\[
M \leftrightarrow X \leftrightarrow (\tilde{X}, Y_1, Y_2).
\]
Consider the first conditional mutual information on the right hand side of (51). Expand this term using the method of Wyner and Ziv [1, Eqn. (52)] as follows:
\[
I(X; M|\tilde{X}, Y_2)
= \sum_{i=1}^{n} I(X_i; M|\tilde{X}, Y_2, X_{1:i-1})
\overset{(a)}{=} \sum_{i=1}^{n} I(X_i; M, \tilde{X}_{i-1}, \tilde{X}_{i+1}, Y_{i-1}, Y_{i+1}, X_{1:i-1}|\tilde{X}_i, Y_2, X_{1:i-1})
\leq \sum_{i=1}^{n} I(X_i; M, Y_{i-1}, Y_{i+1}|\tilde{X}_i, Y_2, X_{1:i-1})
\leq \sum_{i=1}^{n} I(X_i; B_i|\tilde{X}_i, Y_2), \tag{52}
\]
where (a) follows because $(X, Y_2, \tilde{X})$ i.i.d. and therefore
\[
H(X_i|\tilde{X}, Y_2, X_{1:i-1}) = H(X_i|\tilde{X}_i, Y_2, X_{1:i-1}),
\]
and in (b) we define
\[
B_i := (M, Y_{i-1}, Y_{i+1}).
\]
Continuing on from (52), we have
\[
\frac{1}{n} I(X; M|\tilde{X}, Y_2) \geq \frac{1}{n} \sum_{i=1}^{n} I(X_i; B_i|\tilde{X}_i, Y_2, X_{1:i-1})
\overset{(a)}{=} \frac{1}{n} \sum_{i=1}^{n} S(\varepsilon(\tilde{X}_i, Y_{2:i}))
\overset{(b)}{=} S\left( E \left( \frac{1}{n} \sum_{i=1}^{n} \delta_2(X_i, \tilde{X}_{2:i}) \right) \right)
\overset{(c)}{=} S(D_2 + \epsilon), \tag{53}
\]
where
(a) follows from the definition of $S(D_2)$ upon noticing that the $i$-th reconstructed symbol, $\tilde{X}_{2:i}$, can be expressed as a deterministic function of $(B_i, Y_{2:i})$ and
\[
B_i \leftrightarrow X_i \leftrightarrow (Y_{1:i}, Y_{2:i});
\]
(b) combines the convexity of $S(D_2)$ in $D_2$ with Jensen’s inequality; and
(c) $S(D_2)$ is non-increasing in $D_2$ and
\[
D_2 + \epsilon \geq E \frac{1}{n} \sum_{i=1}^{n} \delta_2(X_i, \tilde{X}_{2:i}).
\]
Consider (50), (51) and (53). We have
\[
R + \epsilon \geq H(\tilde{X}|Y_1) - \epsilon(\epsilon) + S(D_2 + \epsilon)
+ \frac{1}{n} I(M; Y_2|\tilde{X}) - I(M; Y_1|\tilde{X}) \tag{54}
\]
We now apply Lemma 1 with
\[
R = X, S_1 = Y_1, S_2 = Y_2, T = \emptyset, L = \tilde{X} \text{ and } J = M.
\]
There exists $W$, jointly distributed with $(X, Y_1, Y_2, \tilde{X})$, such that
\[ W \leftrightarrow X \leftrightarrow (Y_1, Y_2), \]
$|W| \leq |X|$, and
\[ R + \epsilon \geq H(\tilde{X}|Y_1) - \epsilon(\epsilon) + S(D_2 + \epsilon) + I(W; Y_2|\tilde{X}) - I(W; Y_1|\tilde{X}). \]
The converse proof is completed by letting $\epsilon \to 0$ and invoking the continuity of $S(D_2)$ in $D_2$.

**APPENDIX E**

**Proof of Corollary 3.1**
Choose $C = \tilde{X}$ in Theorem 13 and apply the definition of $S(D_2)$ to obtain
\[ R(0, D_2) \leq H(\tilde{X}|Y_1) + S(D_2). \]
The reverse inequality can be proved using a short converse; specifically, we have
\[ H(M) \geq I(X, \tilde{X}, Y_1, Y_2; M) \]
\[ \geq I(\tilde{X}; M|Y_1) + I(X; M|\tilde{X}, Y_1, Y_2) \]
\[ \geq n \left( H(\tilde{X}|Y_1) - \epsilon(\epsilon) + S(D_2 + \epsilon) \right), \]
where (a) applies $M \leftrightarrow (\tilde{X}, Y_2) \leftrightarrow Y_1$ and (b) repeats the steps in (49), (53), where $\epsilon(\epsilon)$ can be chosen so that $\epsilon(\epsilon) \to 0$ as $\epsilon \to 0$.

**APPENDIX F**

**Proof of Lemmas 7 and 11**
Lemmas 7 and 11 are both special cases of the next theorem.

**Theorem 13 (Thm. 1, [7]):** Let $(U_{123}, U_{12}, U_{13}, U_{23}, U_1, U_2, U_3)$ be any tuple of auxiliary random variables, jointly distributed with $(X, Y_1, Y_2, Y_3)$, such that
\[ (Y_1, Y_2, Y_3) \leftrightarrow X \leftrightarrow (U_{123}, U_{12}, U_{13}, U_{23}, U_1, U_2, U_3); \]
forms a Markov chain, and there exist three deterministic mappings
\[ \phi_j : U_j \times Y_j \to \tilde{X}_j, \quad j = 1, 2, 3, \]
with
\[ D_j \geq E \delta_j(X, \phi_j(U_j, Y_j)). \]
Then, for each such tuple of auxiliary random variables, any rate tuple $(R_1, R_2, R_3)$ satisfying (57) is achievable with distortions $(D_1, D_2, D_3)$.

**A. Proof of Lemma 7**
Suppose that the auxiliary random variables $(A_1, A_2, A_3)$ meet the conditions of Lemma 7. Consider Theorem 13 with $U_{12}$ and $U_{13}$ being constants and $U_{123} = U_1 = A_1$
\[ U_{23} = U_2 = A_2 \]
\[ U_3 = A_3. \]
The rate constraints of (57) now simplify to those of Lemma 7.

**B. Proof of Lemma 11**
Suppose that the auxiliary random variables $(A_{12}, A_1, A_2)$ meet the conditions of Lemma 11. Consider Theorem 13 with infinite $D_3$, set $U_{123}, U_{13}, U_{23}$ and $U_3$ to be constants, and $U_{12} = A_{12}, U_1 = A_1$ and $U_2 = A_2$. The rate constraints of (57) now simplify to those of Lemma 7.

**APPENDIX G**

**Proof of Lemma 9**
We have
\[ R_1 + \epsilon \geq \frac{1}{n} H(M_1) \]
\[ \geq \frac{1}{n} I(\tilde{X}_1; M_1|Y_1) \]
\[ \geq \frac{1}{n} \left( H(\tilde{X}_1|Y_1) - \epsilon_1(\epsilon) \right) \]
\[ \geq \frac{1}{n} I(\tilde{X}_1; M_1|Y_1) - \epsilon_2(\epsilon), \]
where (a) applies Fano’s inequality in the same way as (49), where $\epsilon_1(\epsilon)$ can be chosen so that $\epsilon_1(\epsilon) \to 0$ as $\epsilon \to 0$; and (b) follows because the pair $(\tilde{X}_1, Y_1)$ is i.i.d. The sum rate $R_1 + R_2$ is bounded in (60). The justification for the steps leading to (60) is:
(a) The Markov chain $(M_1, M_2) \leftrightarrow (\tilde{X}_1, X) \leftrightarrow (Y_1, Y_2)$;
(b) $\tilde{X}_2$ is determined by $X$;
(c) exploits the fact that $(\tilde{X}_1, \tilde{X}_2, Y_1, Y_2)$ is i.i.d. and applies Fano’s inequality twice, in a manner similar to (49), where $\epsilon_1(\epsilon)$ and $\epsilon_2(\epsilon)$ can be chosen so that they tend to 0 as $\epsilon \to 0$; and
(d) the nonnegativity of conditional mutual information.

We now bound the sum rate $R_1 + R_2 + R_3$. Notice that the steps leading to (59) remain valid if we replace $R_1 + R_2$ by $R_1 + R_2 + R_3$ and the pair of messages $(M_1, M_2)$ by the triple $(M_1, M_2, M_3)$. Indeed, we have (62), where (a) invokes the Markov chain
\[ (M_1, M_2, M_3) \leftrightarrow (\tilde{X}_1, \tilde{X}_2, X) \leftrightarrow (Y_2, Y_3). \]
Consider the first conditional mutual information on the right hand side of (62). We have
\[ \frac{1}{n} I(X; M_1, M_2, M_3|\tilde{X}_1, \tilde{X}_2, Y_3) \]
\[ \geq \frac{1}{n^3} \sum_{i=1}^n I(X_i; M_1, M_2, M_3, Y_{3,i}^{i-1}, Y_{3,i+1}^n, \tilde{X}_1, \tilde{X}_2, Y_{3,i}) \]
\[ = \frac{1}{n} \sum_{i=1}^n I(X_i; C_i|\tilde{X}_1, \tilde{X}_2, Y_{3,i}) \]
\[ \geq S' \left( \frac{1}{n^2} \sum_{i=1}^n \delta_3(X_i, \tilde{X}_{3,i}) \right) \]
\[ \geq S'(D_3 + \epsilon), \]
where (a) follows from the same reasoning as step (a) of (52); in (b), we define
\[ C_i := (M_1, M_2, M_3, Y_{3,i}^{i-1}, Y_{3,i+1}^n); \]
\[ R_1 \geq I(X; U_{123}) - I(U_{123}; Y_1) + I(X; U_{12}|U_{123}) - I(U_{12}; Y_1|U_{123}) + I(X; U_{12}; U_{13}|U_{123}) - I(U_{13}; U_{12}Y_1|U_{123}) + I(X; U_1|U_{123}, U_{12}, U_{13}) - I(U_1; Y_1|U_{123}, U_{12}, U_{13}) \] (57a)

\[ R_1 + R_2 \geq I(X; U_{123}) - \min \left\{ I(U_{123}; Y_1), I(U_{123}; Y_2) \right\} + I(X; U_{12}|U_{123}) - \min \left\{ I(U_{12}; Y_1|U_{123}), I(U_{12}; Y_2|U_{123}) \right\} + I(X; U_{12}; U_{13}|U_{123}) - \min \left\{ I(U_{13}; U_{12}, Y_1|U_{123}), I(U_{13}; Y_3|U_{123}) \right\} + I(X; U_1|U_{123}, U_{12}, U_{13}) - I(U_1; Y_1|U_{123}, U_{12}, U_{13}) + I(X; U_2|U_{123}, U_{12}, U_{23}) - I(U_2; Y_2|U_{123}, U_{12}, U_{23}) + I(X; U_3|U_{123}, U_{13}, U_{23}) - I(U_3; Y_3|U_{123}, U_{13}, U_{23}) \] (57b)

\[ R_1 + R_2 + R_3 \geq I(X; U_{123}) - \min \left\{ I(U_{123}; Y_1), I(U_{123}; Y_2), I(U_{123}; Y_3) \right\} + I(X; U_{12}|U_{123}) - \min \left\{ I(U_{12}; Y_1|U_{123}), I(U_{12}; Y_2|U_{123}) \right\} + I(X; U_{12}; U_{13}|U_{123}) - \min \left\{ I(U_{13}; U_{12}, Y_1|U_{123}), I(U_{13}; Y_3|U_{123}) \right\} + I(X; U_1|U_{123}, U_{12}, U_{13}) - I(U_1; Y_1|U_{123}, U_{12}, U_{13}) + I(X; U_2|U_{123}, U_{12}, U_{23}) - I(U_2; Y_2|U_{123}, U_{12}, U_{23}) + I(X; U_3|U_{123}, U_{13}, U_{23}) - I(U_3; Y_3|U_{123}, U_{13}, U_{23}). \] (57c)

\[ R_1 + R_2 + \epsilon \geq \frac{1}{n}H(M_1, M_2) \geq \frac{1}{n}H(\tilde{X}_1; X; M_1, M_2|Y_1) \]

\[ = \frac{1}{n} \left( I(\tilde{X}_1; M_1, M_2|Y_1) + I(X; M_1, M_2|\tilde{X}_1, Y_1) \right) \]

\[ (a) \frac{1}{n} \left( I(\tilde{X}_1; M_1, M_2|Y_1) + I(X; M_1, M_2|\tilde{X}_1, Y_2) + I(Y_2; M_1, M_2|\tilde{X}_1) - I(Y_1; M_1, M_2|\tilde{X}_1) \right) \]

\[ (b) \frac{1}{n} \left( I(\tilde{X}_1; M_1, M_2|Y_1) + I(\tilde{X}_2; M_1, M_2|\tilde{X}_1, Y_2) + I(X; M_1, M_2|\tilde{X}_1, \tilde{X}_2, Y_2) + I(Y_2; M_1, M_2|\tilde{X}_1) - I(Y_1; M_1, M_2|\tilde{X}_1) \right) \]

\[ (c) \geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) - \epsilon_1(\epsilon) - \epsilon_2(\epsilon) + \frac{1}{n} \left( I(X; M_1, M_2|\tilde{X}_1, \tilde{X}_2, Y_2) \right) \]

\[ (d) \geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) - \epsilon_1(\epsilon) - \epsilon_2(\epsilon) + \frac{1}{n} \left( I(Y_2; M_1, M_2|\tilde{X}_1) - I(Y_1; M_1, M_2|\tilde{X}_1) \right). \] (59)

and (c), (d) and (e) each follow the same reasoning as steps (a), (b) and (c) of (53) respectively. From (62) and (63) we obtain (64).

Consider (60) and (64), and apply Lemma 1 twice: once for

\[ R = X, \ S_1 = Y_1, \ S_2 = Y_2, \ T = Y_3 \] and \( L = \tilde{X}_1 \),

and once for

\[ R = X, \ S_1 = Y_2, \ S_2 = Y_3, \ T = Y_1 \] and \( L = (\tilde{X}_1, \tilde{X}_2) \). We conclude that there exist auxiliary random variables \( W_1, W_2 \) and \( W_3 \) with

\[ |W_1|, |W_2|, |W_3| \leq |X|, \]

\[ W_j \leftrightarrow X \leftrightarrow (Y_1, Y_2, Y_3), \quad j = 1, 2, 3, \]

such that the rate tuple \((R_1, R_2, R_3)\) satisfies

\[ R_1 + R_2 + \epsilon \geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) + I(W_1; Y_2|\tilde{X}_1) - I(W_1; Y_1|\tilde{X}_1) - \epsilon_1(\epsilon) - \epsilon_2(\epsilon) \] (65)

and

\[ R_1 + R_2 + R_3 + \epsilon \geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) + S'(D_3 + \epsilon) - \epsilon_2(\epsilon) - \epsilon_1(\epsilon) \]

\[ + I(W_3; Y_3|\tilde{X}_1, \tilde{X}_2) - I(W_3; Y_2|\tilde{X}_1, \tilde{X}_2) + I(W_2; Y_2|\tilde{X}_1) - I(W_2; Y_1|\tilde{X}_1). \] (66)
\[ R_1 + R_2 + R_3 + \epsilon \geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) - \varepsilon_1(\epsilon) - \varepsilon_2(\epsilon) + \frac{1}{n} \left( I(\mathbf{X}; M_1, M_2, M_3|\tilde{X}_1, \tilde{X}_2, Y_2) - I(Y_2; M_1, M_2, M_3|\tilde{X}_1) \right) \]
\[
+ I(M_1, M_2, M_3; Y_2|\tilde{X}_1, \tilde{X}_2) - I(M_1, M_2, M_3; Y_2|\tilde{X}_1, \tilde{X}_2) 
+ I(Y_2; M_1, M_2, M_3|\tilde{X}_1) - I(Y_1; M_1, M_2, M_3|\tilde{X}_1) \right) \]
REFERENCES


