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[1]. Indeed, the RD function is unknown for the seemingly simple case of deterministic distortion functions[2], 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 [2, 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. $X \in \mathcal{X}$. The n-fold Cartesian product of an alphabet is denoted by boldface font, e.g. $\boldsymbol{\mathcal{X}}$ is the n-fold product of $\mathcal{X}$. If a random vector (X, Y, Z) forms a Markov chain in the same order, then we write $X \leftrightarrow Y \leftrightarrow Z$. The symbol $\oplus$ 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

$$(\boldsymbol{R}, \boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}, \boldsymbol{L}) := (R_1, S_{1,1}, S_{2,1}, T_1, L_1), \\ (R_2, S_{1,2}, S_{2,2}, T_2, L_2), \dots, (R_n, S_{1,n}, S_{2,n}, T_n, L_n)$$

R. Timo is with the Institute for Telecommunications Research at the University of South Australia. E-mail: roy.timo@ieee.org. R. Timo is partly supported by the Australian Research Council Discovery Grant DP120102123.

Tobias J. Oechtering is with the School of Electrical Engineering and the ACCESS Linnaeus Center, KTH Royal Institute of Technology. E-mail: oech@kth.se. T. Oechtering is partly supported by the SRA program ICT-TNG of the Swedish Government and the Swedish Research Council (VR) under Grant C0406401.

Michèle Wigger is with the Communications and Electrical Department, Telecom ParisTech. E-mail: michele.wigger@telecom-paristech.fr. M. Wigger is partly supported by the city of Paris under the programme "Emergences."

Some of the material in this paper was presented at the IEEE Inform. Theory Workshop (ITW), Lausanne, Switzerland, Sep., 2012, and the IEEE Intl. Symp. Inform. Theory (ISIT), Istanbul, Turkey, Jul., 2013.

[1]Matsuta and Uyematsu [4] recently presented matching achievability and converse bounds for general sources and distortion functions using an information-spectrum approach; these bounds, however, are not computable.

[2]The Heegard-Berger problem with deterministic distortion functions also subsumes (an almost lossless version of) the popular *index coding* problem [5].

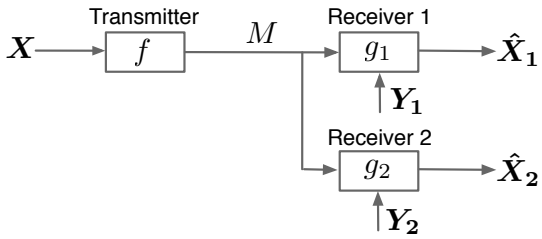

Fig. 1. The rate-distortion problem with side information at two receivers.

denote n i.i.d. copies of (R, S_1, S_2, T, L). Further, suppose that J is jointly distributed with $(\boldsymbol{R}, \boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}, \boldsymbol{L})$ and

$$J \leftrightarrow (\boldsymbol{R}, \boldsymbol{L}) \leftrightarrow (\boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T})$$

forms a Markov chain. Consider the difference

$$I(J; \boldsymbol{S_2}|\boldsymbol{L}) - I(J; \boldsymbol{S_1}|\boldsymbol{L}).$$

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 $(J, \boldsymbol{R}, \boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}, \boldsymbol{L})$ 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

$$I(J; \boldsymbol{S_2}|\boldsymbol{L}) - I(J; \boldsymbol{S_1}|\boldsymbol{L}) = n\big(I(W; S_2|L) - I(W; S_1|L)\big),$$

the cardinality of $\mathcal{W}$ satisfies $|\mathcal{W}| \leq |\mathcal{R}||\mathcal{L}|$, 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 $\mathcal{W}$ becomes $|\mathcal{W}| \leq |\mathcal{R}|$.

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 $(\boldsymbol{X}, \boldsymbol{Y_1}, \boldsymbol{Y_2})$ denote a string of n i.i.d. copies of (X, Y_1, Y_2). Let $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}_1}$ and $\boldsymbol{\mathcal{Y}_2}$ 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 $\boldsymbol{X}$, receiver 1 observes $\boldsymbol{Y_1}$ and receiver 2 observes $\boldsymbol{Y_2}$. The string $\boldsymbol{X}$ 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 $\boldsymbol{X}$ can be compressed while still allowing the receivers to reconstruct $\boldsymbol{X}$ to within specified average distortions, as described next.

A block code consists of three (possibly stochastic) mappings:

$$f : \boldsymbol{\mathcal{X}} \longrightarrow \mathcal{M}$$

and

$$g_j : \mathcal{M} \times \boldsymbol{\mathcal{Y}_j} \longrightarrow \hat{\boldsymbol{\mathcal{X}}}_{\boldsymbol{j}}, \quad j = 1, 2,$$

where $\mathcal{M}$ is an index set with finite cardinality $|\mathcal{M}|$ depending on n, $\hat{\mathcal{X}}_j$ is the reconstruction alphabet of receiver j and $\hat{\boldsymbol{\mathcal{X}}}_{\boldsymbol{j}}$ its n-fold Cartesian product. The transmitter sends $M := f(\boldsymbol{X})$ and receiver j reconstructs $\hat{\boldsymbol{X}}_{\boldsymbol{j}} := g_j(M, \boldsymbol{Y_j})$.

Let

$$\delta_j : \mathcal{X} \times \hat{\mathcal{X}}_j \longrightarrow [0, \infty), \quad j = 1, 2,$$

be bounded per-letter distortion functions. For simplicity, and without loss of generality, we assume that δ_1 and δ_2 are *normal* [27, p. 185]; that is, for each x in $\mathcal{X}_j$ there exists some $\hat{x}$ in $\hat{\mathcal{X}}_j$ 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

$$D_j + \epsilon \geq \mathbb{E}\, \frac{1}{n} \sum_{i=1}^{n} \delta_j(X_i, \hat{X}_{j,i}), \quad j = 1, 2.$$

Definition 2: For distortions $D_1 \geq 0$ and $D_2 \geq 0$, Heegard and Berger's *RD function* is

$$R(D_1, D_2) := \min \big\{R > 0 : R \text{ is } (D_1, D_2)\text{-achievable}\big\}.$$

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]

$$R(D_1, D_2) \leq \min_{(A,B,C)} \Big\{ \max\big\{I(X; C|Y_1), I(X; C|Y_2)\big\} + I(X; A|C, Y_1) + I(X; B|C, Y_2)\Big\},$$

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

$$|\mathcal{C}| \leq |\mathcal{X}| + 3$$
$$|\mathcal{A}| \leq |\mathcal{C}||\mathcal{X}| + 1$$
$$|\mathcal{B}| \leq |\mathcal{C}||\mathcal{X}| + 1$$

(these bounds are new and proved in Appendix B);

(iii) There exist deterministic maps

$$\phi_1 : \mathcal{A} \times \mathcal{C} \times \mathcal{Y}_1 \longrightarrow \hat{\mathcal{X}}_1$$
$$\phi_2 : \mathcal{B} \times \mathcal{C} \times \mathcal{Y}_2 \longrightarrow \hat{\mathcal{X}}_2$$

with

$$D_1 \geq \mathbb{E}\ \delta_1\big(X, \phi_1(A, C, Y_1)\big)$$
$$D_2 \geq \mathbb{E}\ \delta_2\big(X, \phi_2(B, C, Y_2)\big).$$

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]

$$R(D_1, D_2) = \min_{(B,C)} \Big\{ I(X; C|Y_1) + I(X; B|C, Y_2) \Big\},$$

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

$$\phi_1 : \mathcal{C} \times \mathcal{Y}_1 \longrightarrow \hat{\mathcal{X}}_1$$
$$\phi_2 : \mathcal{B} \times \mathcal{C} \times \mathcal{Y}_2 \longrightarrow \hat{\mathcal{X}}_2$$

with

$$D_1 \geq \mathbb{E}\ \delta_1\big(X, \phi_1(C, Y_1)\big)$$
$$D_2 \geq \mathbb{E}\ \delta_2\big(X, \phi_2(B, C, Y_2)\big).$$

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 fall 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 δ_1 is deterministic in the following sense.

Definition 4: δ_1 is said to be *deterministic* [17, 28] if there is an alphabet $\tilde{\mathcal{X}}$ with $\hat{\mathcal{X}}_1 = \tilde{\mathcal{X}}$ and a deterministic map

$$\psi : \mathcal{X} \longrightarrow \tilde{\mathcal{X}}$$

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 δ_1. Let

$$\tilde{X} := \psi(X).$$

Define

$$S(D_2) := \min_{B}\ I(X; B|\tilde{X}, Y_2), \quad D_2 \geq 0, \tag{1}$$

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

$$|\mathcal{B}| \leq |\mathcal{X}| + 1;$$

(iii) there exists a deterministic mapping

$$\phi_2 : \mathcal{B} \times \tilde{\mathcal{X}} \times \mathcal{Y}_2 \longrightarrow \hat{\mathcal{X}}_2$$

with

$$D_2 \geq \mathbb{E}\ \delta_2\big(X, \phi_2(B, \tilde{X}, Y_2)\big).$$

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

Corollary 3.1: If the side information is physically degraded and δ_1 is a deterministic distortion function, then

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

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

$$\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \quad \text{and} \quad X := (X_1, X_2),$$

where $\mathcal{X}_1$ and $\mathcal{X}_2$ are finite alphabets. In addition, we say that δ_1 and δ_2 are *component Hamming distortion functions* if

$$\hat{\mathcal{X}}_j = \mathcal{X}_j$$

and for all $x_j, \hat{x}_j \in \mathcal{X}_j$

$$\delta_j(x_j, \hat{x}_j) = \begin{cases} 0 & \text{if } \hat{x}_j = x_j \\ 1 & \text{otherwise} \end{cases} \qquad j = 1, 2.$$

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 achievable, then for each $\epsilon > 0$ there exists a block code (f, g_1, g_2) for which

$$\begin{aligned} R + \epsilon &\geq \frac{1}{n} \log |\mathcal{M}| \geq \frac{1}{n} H(M) \geq \frac{1}{n} I(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{Y_1}, \boldsymbol{Y_2}; M) \\ &= \frac{1}{n} \Big(I(\boldsymbol{X_1}, \boldsymbol{Y_1}; M) + I(\boldsymbol{X_2}, \boldsymbol{Y_2}; M|\boldsymbol{X_1}, \boldsymbol{Y_1}) \Big) \\ &\geq \frac{1}{n} \Big(I(\boldsymbol{X_1}; M|\boldsymbol{Y_1}) + I(\boldsymbol{X_2}; M|\boldsymbol{X_1}, \boldsymbol{Y_1}, \boldsymbol{Y_2}) \Big) \\ &\overset{\text{(a)}}{\geq} \frac{1}{n} \Big(H(\boldsymbol{X_1}|\boldsymbol{Y_1}) + H(\boldsymbol{X_2}|\boldsymbol{X_1}, \boldsymbol{Y_1}, \boldsymbol{Y_2}) - n\varepsilon(\epsilon) \Big) \\ &\overset{\text{(b)}}{=} H(X_1|Y_1) + H(X_2|X_1, Y_1, Y_2) - \varepsilon(\epsilon) \\ &\overset{\text{(c)}}{=} H(X_1|Y_1) + H(X_2|X_1, Y_2) - \varepsilon(\epsilon). \end{aligned} \tag{2}$$

The justifications for steps (a), (b) and (c) are as follows:

(a) $\hat{\boldsymbol{X}}_{\boldsymbol{1}}$ and $\hat{\boldsymbol{X}}_{\boldsymbol{2}}$ are determined by $(M, \boldsymbol{Y_1})$ and $(M, \boldsymbol{Y_2})$ respectively, so (a) follows by Fano's inequality [14, Sec. 2.2]. Here $\varepsilon(\epsilon)$ can be chosen so that $\varepsilon(\epsilon) \to 0$ as $\epsilon \to 0$.

(b) $(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{Y_1}, \boldsymbol{Y_2})$ 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 $\boldsymbol{X_1}$ losslessly to receiver 1 at rate R' close to $H(X_1|Y_1)$. The side information is physically degraded, so we have

$$R' \geq H(X_1|Y_1) \geq H(X_1|Y_2). \tag{3}$$

A close inspection of the random binning proof, e.g. [14], reveals that (3) also suffices for receiver 2 to reliably decode $\boldsymbol{X_1}$. Assuming that $\boldsymbol{X_1}$ is successfully decoded by receiver 2, we can send $\boldsymbol{X}_2$ to receiver 2 at a rate R'' close to $H(X_2|X_1, Y_2)$ using $(\boldsymbol{X_1}, \boldsymbol{Y_2})$ 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

$$X_1 \leftrightarrow Y_2 \leftrightarrow Y_1 \tag{4a}$$

and

$$X_2 \leftrightarrow (X_1, Y_2) \leftrightarrow Y_1. \tag{4b}$$

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 (c) in (2) 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 $\boldsymbol{X}$ with side information $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_2})$.
(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 3. 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 D. Our main result in this section, Theorem 6, follows directly thereafter.

Lemma 5 (Converse): If δ_1 is a deterministic distortion function specified by $\tilde{X} = \psi(X)$, then the following statements are true.

(i) For arbitrarily distributed (X, Y_1, Y_2), we have

$$R(0, D_2) \geq H(\tilde{X}|Y_1) + S(D_2) + \min\left\{I(W; Y_2|\tilde{X}) - I(W; Y_1|\tilde{X})\right\},$$

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 $|\mathcal{W}| \leq |\mathcal{X}|$.
(ii) If (X, Y_1, Y_2) satisfies $(Y_2 \succeq Y_1 \mid \tilde{X})$, then

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

It is worth highlighting that

$$\min\left\{I(W; Y_2|\tilde{X}) - I(W; Y_1|\tilde{X})\right\}$$

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 $L = \tilde{X}$.

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 δ_1 is a deterministic distortion function specified by $\tilde{X} = \psi(X)$, $(Y_2 \succeq Y_1 \mid \tilde{X})$ 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 5. ∎

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

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

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

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

and

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

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 $(Y_2 \succeq Y_1 \mid X_1)$ 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

$$H_\mathrm{b}(\alpha) := -\alpha \log_2 \alpha - (1 - \alpha) \log_2 (1 - \alpha)$$

is the binary entropy function. From Corollary 6.1, we have

$$R(0, 0) = H_\mathrm{b}(1/4) + H_\mathrm{b}(1/3).$$

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

$$X \leftrightarrow Y_2 \leftrightarrow Y_1$$

and

$$\tilde{X} \leftrightarrow X \leftrightarrow (Y_1, Y_2)$$

together imply $(Y_2 \succeq Y_1 \mid \tilde{X})$ and $H(\tilde{X}|Y_1) \geq H(\tilde{X}|Y_2)$ by the data processing lemma.

(ii) It appears that our approach to proving Lemma 5 does not readily generalise to an arbitrary distortion function δ_1. An apparent difficulty follows from the use of a Wyner-Ziv style converse argument to construct the $S(D_2)$ term using $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})$ as side information. The argument needs $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})$ to be i.i.d., and this need not be the case when δ_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

$$\begin{aligned} I(W; Y_2) &= H(X_1) - H(X_1|Y_2) \\ &= H(X_1) - H(X_2), \end{aligned}$$

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

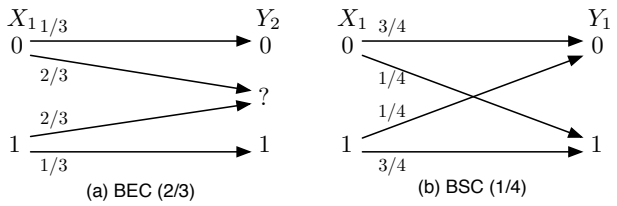

Fig. 2. Binary channels defining the side information in Example 2: (a) Binary Erasure Channel (BEC) with erasure probability $2/3$; and (b) Binary Symmetric Channel (BSC) with crossover probability $1/4$.

IV. Successive Refinement with Side Information

The method used in Appendix D 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.

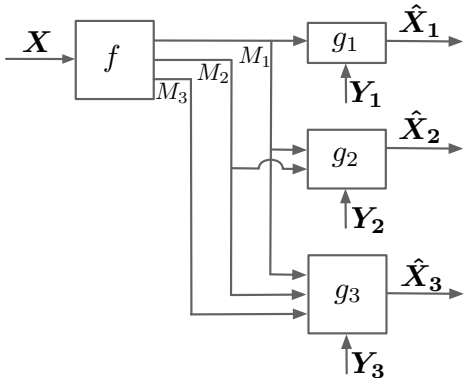

Fig. 3. The successive refinement of information problem with three stages and side information at the receivers.

A. Problem Formulation

Consider a tuple of random variables (X, Y_1, Y_2, Y_3) with an arbitrary joint distribution. Let $(\boldsymbol{X}, \boldsymbol{Y_1}, \boldsymbol{Y_2}, \boldsymbol{Y_3})$ 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

$$f : \boldsymbol{\mathcal{X}} \longrightarrow \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3$$

and

$$\begin{aligned} g_1 &: \mathcal{M}_1 \times \boldsymbol{\mathcal{Y}_1} \longrightarrow \boldsymbol{\hat{\mathcal{X}}_1} \\ g_2 &: \mathcal{M}_1 \times \mathcal{M}_2 \times \boldsymbol{\mathcal{Y}_2} \longrightarrow \boldsymbol{\hat{\mathcal{X}}_2} \\ g_3 &: \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3 \times \boldsymbol{\mathcal{Y}_3} \longrightarrow \boldsymbol{\hat{\mathcal{X}}_3}, \end{aligned}$$

where $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are finite index sets. The transmitter sends $(M_1, M_2, M_3) := f(\boldsymbol{X})$ over the noiseless channels, as shown in Fig. 3. Receiver 1 reconstructs $\boldsymbol{\hat{X}_1} := g_1(M_1, \boldsymbol{Y_1})$, receiver 2 reconstructs $\boldsymbol{\hat{X}_2} := g_2(M_1, M_2, \boldsymbol{Y_2})$ and receiver 3 reconstructs $\boldsymbol{\hat{X}_3} := g_3(M_1, M_2, M_3, \boldsymbol{Y_3})$.

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

$$D_j + \epsilon \geq \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \delta_j(X_j, \hat{X}_{j,i})$$

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 $\mathcal{R}_{\text{in}}(D_1, D_2, D_3)$ 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[3]

$$\begin{aligned} |\mathcal{A}_1| &\leq |\mathcal{X}| + 6 \\ |\mathcal{A}_2| &\leq |\mathcal{X}|\ |\mathcal{A}_1| + 4 \\ |\mathcal{A}_3| &\leq |\mathcal{X}|\ |\mathcal{A}_1|\ |\mathcal{A}_2| + 1. \end{aligned}$$

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

$$\phi_j : \mathcal{A}_j \times \mathcal{Y}_j \longrightarrow \hat{\mathcal{X}}_j$$

with

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

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

$$\begin{aligned} R_1 &\geq I(X; A_1|Y_1), \\ R_1 + R_2 &\geq \max_{j=1,2} I(X; A_1|Y_j) + I(X; A_2|A_1, Y_2) \end{aligned}$$

and

$$\begin{aligned} R_1 + R_2 + R_3 \geq \max_{j=1,2,3} I(X; A_1|Y_j) \\ + \max_{j=2,3} I(X; A_2|A_1, Y_j) \\ + I(X; A_3|A_1, A_2, Y_3). \end{aligned}$$

Lemma 7:

$$\mathcal{R}_{\text{in}}(D_1, D_2, D_3) \subseteq \mathcal{R}(D_1, D_2, D_3).$$

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

$$\mathcal{R}_{\text{in}}(D_1, D_2, D_3) = \mathcal{R}(D_1, D_2, D_3).$$

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

$$\begin{aligned} R_1 &\geq I(X; A_1|Y_1) \\ R_1 + R_2 &\geq I(X; A_1|Y_1) + I(X; A_2|A_1, Y_2) \\ R_1 + R_2 + R_3 &\geq I(X; A_1|Y_1) + I(X; A_2|A_1, Y_2) \\ &\quad + I(X; A_3|A_1, A_2, Y_3), \end{aligned}$$

where A_1, A_2 and A_3 obey the same cardinality constraints as those for $\mathcal{R}_{\text{in}}(D_1, D_2, D_3)$, 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

[3]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, App. I] and there, again, degraded side information played a crucial role.

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

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

respectively, where ψ_1 and ψ_2 are functions of the form

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

Theorem 8, with deterministic δ_1 and δ_2, simplifies as follows. Define

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

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) $|\mathcal{A}_3| \leq |\mathcal{X}| + 1$;
(iii) there exists a (deterministic) map

$$\phi_3 : \mathcal{A}_3 \times \tilde{\mathcal{X}}_1 \times \tilde{\mathcal{X}}_2 \times \mathcal{Y}_3 \longrightarrow \hat{\mathcal{X}}_3$$

with

$$D_3 \geq \mathbb{E}\ \delta_3\big(X, \phi_3(A_3, \tilde{X}_1, \tilde{X}_2, Y_3)\big).$$

Corollary 8.1: If the side information is physically degraded $X \leftrightarrow Y_3 \leftrightarrow Y_2 \leftrightarrow Y_1$ and the distortion functions δ_1 and δ_2 are deterministic, then $\mathcal{R}(0, 0, D_3)$ is equal to the set of all rate tuples (R_1, R_2, R_3) satisfying

$$\begin{aligned} R_1 &\geq H(\tilde{X}_1|Y_1) \\ R_1 + R_2 &\geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) \\ R_1 + R_2 + R_3 &\geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) + S'(D_3). \end{aligned}$$

Proof: The achievability part follows directly from Theorem 8 upon selecting the auxiliary random variables as $A_1 = \tilde{X}_1$ and $A_2 = \tilde{X}_2$ 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 δ_1 and δ_2 and arbitrarily distributed side information; it is a successive-refinement version of Lemma 5. Let $\mathcal{R}_{\text{out}}(D_3)$ denote the set of all rate tuples (R_1, R_2, R_3) for which

$$\begin{aligned} R_1 &\geq H(\tilde{X}_1|Y_1) \\ R_1 + R_2 &\geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) \\ &\quad + \min_W \Big\{ I(W; Y_2|\tilde{X}_1) - I(W; Y_1|\tilde{X}_1) \Big\} \end{aligned}$$

and

$$\begin{aligned} R_1 + R_2 + R_3 &\geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) + S'(D_3) \\ &\quad + \min_W \Big\{ I(W; Y_2|\tilde{X}_1) - I(W; Y_1|\tilde{X}_1) \Big\} \\ &\quad + \min_W \Big\{ I(W; Y_3|\tilde{X}_1, \tilde{X}_2) - I(W; Y_2|\tilde{X}_1, \tilde{X}_2) \Big\}, \end{aligned}$$

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

Lemma 9 (Converse): If δ_1 and δ_2 are deterministic distortion functions, then

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

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

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 δ_1 and δ_2 are deterministic distortion functions, $(Y_2 \succeq Y_1 \mid \tilde{X}_1)$, $(Y_3 \succeq Y_2 \mid \tilde{X}_1, \tilde{X}_2)$, and

$$\begin{aligned} H(\tilde{X}_1|Y_1) &\geq \max\big\{H(\tilde{X}_1|Y_2), H(\tilde{X}_1|Y_3)\big\}, \\ H(\tilde{X}_2|\tilde{X}_1, Y_2) &\geq H(\tilde{X}_2|\tilde{X}_1, Y_3), \end{aligned}$$

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

$$\begin{aligned} R_1 &\geq H(\tilde{X}_1|Y_1) \\ R_1 + R_2 &\geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) \\ R_1 + R_2 + R_3 &\geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) + S'(D_3). \end{aligned}$$

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

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

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

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

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

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

$$\begin{aligned} I(W; Y_2|\tilde{X}_1) &= H(W|\tilde{X}_1) - H(W|\tilde{X}_1, Y_1, Y_2) \\ &\geq I(W; Y_1|\tilde{X}_1). \end{aligned}$$

The less noisy condition $(Y_3 \succeq Y_2 \mid \tilde{X}_1, \tilde{X}_2)$ 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

information at receiver 2. *Side information scalable source coding* refers to the special case where

$$X \leftrightarrow Y_1 \leftrightarrow Y_2 \tag{5}$$

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 (5). 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 [14]. The lemma includes Tian and Diggavi's bound [17, Cor. 1] for arbitrarily distributed side information as a special case.

Let $\mathcal{R}_{\text{in}}^*(D_1, D_2)$ 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

$$\begin{aligned} |\mathcal{A}_{12}| &\leq |\mathcal{X}| + 3 \\ |\mathcal{A}_1| &\leq |\mathcal{X}|\ |\mathcal{A}_{12}| + 1 \\ |\mathcal{A}_2| &\leq |\mathcal{X}|\ |\mathcal{A}_{12}| + 1; \end{aligned}$$

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

$$\phi_j : \mathcal{A}_j \times \mathcal{Y}_j \longrightarrow \hat{\mathcal{X}}_j,$$

with

$$D_j \geq \mathbb{E}\ \delta_j\big(X, \phi_j(A_j, Y_j)\big);$$

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

$$R_1 \geq I(X; A_{12}, A_1 | Y_1) \tag{6a}$$

$$\begin{aligned} R_1 + R_2 \geq \max\Big\{ & I(X; A_{12}|Y_1), I(X; A_{12}|Y_2) \Big\} \\ & + I(X; A_1 | A_{12}, Y_1) \\ & + I(X; A_2 | A_{12}, Y_2). \end{aligned} \tag{6b}$$

Lemma 11:

$$\mathcal{R}_{\text{in}}^*(D_1, D_2) \subseteq \mathcal{R}(D_1, D_2).$$

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 δ_1 and δ_2 are deterministic distortion functions, with $\tilde{X}_1 = \psi_1(X)$ and $\tilde{X}_2 = \psi_2(X)$. It is said that δ_2 is a *degraded version* of δ_1 if

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

for some deterministic map ψ'. The next theorem is proved in Appendix H.

Theorem 12: Suppose that δ_1 and δ_2 are deterministic distortion functions.

(i) If δ_2 is a degraded version of δ_1,

$$H(\tilde{X}_2|Y_1) \leq H(\tilde{X}_2|Y_2) \quad \text{and} \quad (Y_1 \succeq Y_2 \mid \tilde{X}_2),$$

then $\mathcal{R}_{\text{in}}^*(0,0) = \mathcal{R}(0,0)$ and the rate constraints of (6) simplify to

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

(ii) If δ_1 is a degraded version of δ_2 and

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

then $\mathcal{R}_{\text{in}}^*(0,0) = \mathcal{R}(0,0)$ and the rate constraints (6) simplify to

$$\begin{aligned} R_1 &\geq H(\tilde{X}_1|Y_1) \\ R_1 + R_2 &\geq H(\tilde{X}_2|Y_2). \end{aligned}$$

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 $(Y_1 \succeq Y_2 \mid \tilde{X}_2)$ and by the data processing lemma it also implies $H(\tilde{X}_j|Y_1) \leq H(\tilde{X}_j|Y_2)$ for $j = 1, 2$.

Appendix A
Proof of Lemma 1

We first notice that

$$I(J; \boldsymbol{S}_2|\boldsymbol{L}) - I(J; \boldsymbol{S}_1|\boldsymbol{L}) = I(J; \boldsymbol{S}_2, \boldsymbol{L}) - I(J; \boldsymbol{S}_1, \boldsymbol{L}), \tag{7}$$

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

$$\begin{aligned} & I(J; \boldsymbol{S_2}, \boldsymbol{L}) \\ & \overset{\text{(a)}}{=} \sum_{i=1}^{n} I(J; S_{2,i}, L_i | S_{2,1}^{i-1}, L_1^{i-1}) \\ & \overset{\text{(b)}}{=} \sum_{i=1}^{n} I(J, S_{2,1}^{i-1}, L_1^{i-1}; S_{2,i}, L_i) \\ & \overset{\text{(c)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{1,i+1}^{n}, S_{2,1}^{i-1}, L_1^{i-1}, L_{i+1}^{n}; S_{2,i}, L_i) \\ & \qquad\qquad - I(S_{1,i+1}^{n}, L_{i+1}^{n}; S_{2,i}, L_i | J, S_{2,1}^{i-1}, L_1^{i-1}) \Big) \\ & \overset{\text{(d)}}{=} \sum_{i=1}^{n} \Big(I(W_i; S_{2,i}, L_i) \\ & \qquad\qquad - I(S_{1,i+1}^{n}, L_{i+1}^{n}; S_{2,i}, L_i | J, S_{2,1}^{i-1}, L_1^{i-1}) \Big) \end{aligned} \tag{8}$$

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

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

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

$$W_i := (J, S_{1,i+1}^{n}, S_{2,1}^{i-1}, L_1^{i-1}, L_{i+1}^{n}). \tag{9}$$

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

$$I(J; \boldsymbol{S_1}, \boldsymbol{L})$$

$$\stackrel{\text{(a)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{2,1}^{i-1}, L_1^{i-1}; S_{1,i}^n, L_i^n) - I(J, S_{2,1}^{i}, L_1^{i}; S_{1,i+1}^n, L_{i+1}^n) \Big)$$

$$\stackrel{\text{(b)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{2,1}^{i-1}, L_1^{i-1}; S_{1,i}, L_i | S_{1,i+1}^n, L_{i+1}^n) - I(S_{2,i}, L_i; S_{1,i+1}^n, L_{i+1}^n | J, S_{2,1}^{i-1}, L_1^{i-1}) \Big)$$

$$\stackrel{\text{(c)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{1,i+1}^n, S_{2,1}^{i-1}, L_1^{i-1}, L_{i+1}^n; S_{1,i}, L_i) - I(S_{2,i}, L_i; S_{1,i+1}^n, L_{i+1}^n | J, S_{2,1}^{i-1}, L_1^{i-1}) \Big)$$

$$\stackrel{\text{(d)}}{=} \sum_{i=1}^{n} \Big(I(W_i; S_{1,i}, L_i) - I(S_{2,i}, L_i; S_{1,i+1}^n, L_{i+1}^n | J, S_{2,1}^{i-1}, L_1^{i-1}) \Big), \quad (10)$$

where (a) is a telescoping sum and we understand $S_{2,1}^0$ and L_1^0, for $i = 1$, and $S_{1,n+1}^n$ and L_{n+1}^n, for $i = n$, to be degenerate random variables (constants); (b) again uses the chain rule for mutual information; (c) exploits the i.i.d. source and hence

$$H(S_{1,i}, L_i | S_{1,i+1}^n, L_{i+1}^n) = H(S_{1,i}, L_i);$$

and, finally, in (d) we substitute

$$W_i \equiv (J, S_{1,i+1}^n, S_{2,1}^{i-1}, L_1^{i-1}, L_{i+1}^n).$$

Subtract (10) from (8) to obtain

$$I(J; \boldsymbol{S_2}, \boldsymbol{L}) - I(J; \boldsymbol{S_1}, \boldsymbol{L}) = \sum_{i=1}^{n} \Big(I(W_i; S_{2,i}, L_i) - I(W_i; S_{1,i}, L_i) \Big). \quad (11)$$

We now *single-letterize* the quantity on the right hand side of (11). To this end, let us introduce a time-sharing random variable: Let Q be uniform on $\{1, 2, \ldots, n\}$ and independent of the tuple $(\boldsymbol{R}, \boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}, \boldsymbol{L})$. Dividing (11) by n, we have

$$\frac{1}{n} \sum_{i=1}^{n} \Big(I(W_i; S_{2,i}, L_i) - I(W_i; S_{1,i}, L_i) \Big)$$
$$\stackrel{\text{(a)}}{=} \frac{1}{n} \sum_{i=1}^{n} \Big(I(W_i; S_{2,i}, L_i | Q = i) - I(W_i; S_{1,i}, L_i | Q = i) \Big)$$
$$\stackrel{\text{(b)}}{=} I(W_Q; S_{2,Q}, L_Q | Q) - I(W_Q; S_{1,Q}, L_Q | Q)$$
$$\stackrel{\text{(c)}}{=} I(W_Q, Q; S_{2,Q}, L_Q) - I(W_Q, Q; S_{1,Q}, L_Q)$$
$$\stackrel{\text{(d)}}{=} I(\tilde{W}; S_2, L) - I(\tilde{W}; S_1, L), \quad (12)$$

where in (a) we use that Q is independent of $(S_{1,i}, S_{2,i}, L_i, W_i)$; in (b) that Q is uniformly distributed; in (c) that $(\boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{L})$ is i.i.d. and independent of Q, and therefore

$$H(S_{j,Q}, L_Q | Q) = H(S_{j,Q}, L_Q), \quad j = 1, 2;$$

and, finally, in (d) we define and substitute

$$\tilde{W} = (W_Q, Q), \ S_1 = S_{1,Q}, \ S_2 = S_{2,Q}, \text{ and } L = L_Q.$$

From (11) and (12), we have

$$I(J; \boldsymbol{S_2}, \boldsymbol{L}) - I(J; \boldsymbol{S_1}, \boldsymbol{L}) = n\big(I(\tilde{W}; S_2, L) - I(\tilde{W}; S_1, L) \big). \quad (13)$$

We also notice that

$$W_i \leftrightarrow (R_i, L_i) \leftrightarrow (S_{1,i}, S_{2,i}, T_i), \quad (14)$$

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

$$J \leftrightarrow (\boldsymbol{R}, \boldsymbol{L}) \leftrightarrow (\boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}),$$

and the fact that $(\boldsymbol{R}, \boldsymbol{S}_1, \boldsymbol{S_2}, \boldsymbol{T}, \boldsymbol{L})$ is i.i.d. Now define

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

Using the independence of Q from $(\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{S_1}, \boldsymbol{S}_2, \boldsymbol{L})$, we have the desired Markov chain,

$$\tilde{W} \leftrightarrow (R, L) \leftrightarrow (S_1, S_2, T). \quad (15)$$

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

For each and every $\tilde{w}$ in the support set of $\tilde{W}$, let $q_{\tilde{w}}$ denote the conditional distribution of (R, S_1, S_2, T, L) given $\tilde{W} = \tilde{w}$. Let $\mathcal{P}$ denote the set of all joint distributions on $\mathcal{R} \times \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{T} \times \mathcal{L}$.

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

$$g_{r,l}(q) := \sum_{s_1 \in \mathcal{S}_1} \sum_{s_2 \in \mathcal{S}_2} \sum_{t \in \mathcal{T}} q(r, s_1, s_2, t, l). \quad (16)$$

The $(|\mathcal{R}||\mathcal{L}| - 1)$-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

$$\mathbb{E}_{\tilde{W}}\big\{ g_{r,l}\big(q_{\tilde{W}}\big) \big\} \equiv \sum_{\tilde{w} \in \tilde{\mathcal{W}}} \mathbb{P}[\tilde{W} = \tilde{w}] \ g_{r,l}(q_{\tilde{w}})$$

is equal to the true probability $\mathbb{P}[(R, L) = (r, l)]$. Moreover, this agreement extends over $\mathcal{R} \times \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{T} \times \mathcal{L}$ because

$$\mathbb{E}\big\{ g_{r,l}(q_{\tilde{W}}) \big\} \cdot \mathbb{P}\big[S_1 = s_1, S_2 = s_2, T = t | R = r, L = l \big] \quad (17)$$

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

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

$$I(\tilde{W}; S_2, L) - I(\tilde{W}; S_1, L) \quad (18)$$

by simply preserving $H(S_2, L|\tilde{W}) - H(S_1, L|\tilde{W})$. To this end, define the functional $g \colon \mathcal{P} \mapsto [-|\mathcal{S}_1|, |\mathcal{S}_2|]$,

$$g(q) := H(\mathsf{S}_2, \mathsf{L}) - H(\mathsf{S}_1, \mathsf{L}), \quad (19)$$

where the joint distribution[4] of $(\mathsf{R}, \mathsf{S}_1, \mathsf{S}_2, \mathsf{T}, \mathsf{L})$ is understood to be given by q. We also notice that

$$\begin{aligned}\mathbb{E}_{\tilde{W}}\big\{g(q_{\tilde{W}})\big\} &\equiv \sum_{\tilde{w}\in\tilde{\mathcal{W}}} \mathbb{P}[\tilde{W}=\tilde{w}]g(q_{\tilde{w}}) \\ &= H(S_2, L|\tilde{W}) - H(S_1, L|\tilde{W}).\end{aligned}$$

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

(i) for all (r, l) in $\mathcal{R} \times \mathcal{L}$ — excluding the omitted pair (r^*, l^*) — we have

$$\mathbb{E}_W\big\{g_{r,l}(q_W)\big\} = \mathbb{E}_{\tilde{W}}\big\{g_{r,l}(q_{\tilde{W}})\big\}, \tag{20}$$

(ii) and

$$\mathbb{E}_W\big\{g(q_W)\big\} = \mathbb{E}_{\tilde{W}}\big\{g(q_{\tilde{W}})\big\}. \tag{21}$$

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

$$W \leftrightarrow (R, L) \leftrightarrow (S_1, S_2, T). \tag{22}$$

Finally, the equalities (20) and (21) imply

$$\begin{aligned}&I(W; S_2, L) - I(W; S_1, L) \\ &\qquad = I(\tilde{W}; S_2, L) - I(\tilde{W}; S_1, L).\end{aligned} \tag{23}$$

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 $L \leftrightarrow R \leftrightarrow \tilde{W}$ and from (15) we have

$$\tilde{W} \leftrightarrow R \leftrightarrow (L, S_1, S_2, T)\,. \tag{24}$$

Replace the $(|\mathcal{R}||\mathcal{L}| - 1)$ functionals $\{g_{r,l}\}$ in (16) by

$$g_r(q) := \sum_{s_1\in\mathcal{S}_1}\sum_{s_2\in\mathcal{S}_2}\sum_{t\in\mathcal{T}}\sum_{l\in\mathcal{L}} q(r, s_1, s_2, t, l) \tag{25}$$

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 $g_{r,l}$ in (20) and $W \leftrightarrow R \leftrightarrow (L, S_1, S_2, T)$ replaces (22). ■

Remark 6:

(i) 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.

(ii) 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.

[4]We use sans serif font to emphasise that this joint distribution differs to that of (R, S_1, S_2, T, L).

Appendix B
Proof of Cardinality Bounds in Lemma 2

Suppose that we have auxiliary random variables (A, B, C) as well as functions ϕ_1 and ϕ_2 such that $(A, B, C) \leftrightarrow X \leftrightarrow (Y_1, Y_2)$ and

$$\begin{aligned}D_1 &\geq \mathbb{E}\ \delta_1\big(X, \phi_1(A, C, Y_1)\big) \\ D_2 &\geq \mathbb{E}\ \delta_2\big(X, \phi_2(B, C, Y_2)\big),\end{aligned}$$

but without the cardinality bounds in Lemma 2; that is, the alphabets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are finite but otherwise arbitrary.

Consider the variable C. For each and every c in the support set of C, let q_c denote the conditional distribution of (A, B, X) given $C = c$. Let $\mathcal{P}_1$ denote the set of all joint distributions on $\mathcal{A} \times \mathcal{B} \times \mathcal{X}$.

For each and every x in $\mathcal{X}$ but one, say x^*, define $g_x : \mathcal{P}_1 \longrightarrow [0, 1]$ by setting

$$g_x(q) := \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} q(a, b, x).$$

We notice that, for all x except x^*,

$$\mathbb{E}_C\big\{g_x(q_C)\big\} = \mathbb{P}[X = x] \tag{26}$$

gives the true marginal distribution of X. Now define the following functionals — each mapping $\mathcal{P}_1$ to $[-|\mathcal{X}|, |\mathcal{X}|]$ — by setting

$$g_1(q) := I(\mathsf{X}; \mathsf{B}|\mathsf{Y}_2) - H(\mathsf{X}|\mathsf{A}, \mathsf{Y}_1) \tag{27}$$

$$g_2(q) := I(\mathsf{X}; \mathsf{A}|\mathsf{Y}_1) - H(\mathsf{X}|\mathsf{B}, \mathsf{Y}_2) \tag{28}$$

and

$$\begin{aligned}g_3(q) := &\sum_{a\in\mathcal{A}}\sum_{y_1\in\mathcal{Y}_1} \\ &\min_{\hat{x}\in\hat{\mathcal{X}}_1}\sum_{b\in\mathcal{B}}\sum_{x\in\mathcal{X}}\sum_{y_2\in\mathcal{Y}_2} q(a, b, x)p(y_1, y_2|x)\delta_1(\hat{x}, x)\end{aligned}$$

$$\begin{aligned}g_4(q) := &\sum_{b\in\mathcal{B}}\sum_{y_2\in\mathcal{Y}_2} \\ &\min_{\hat{x}\in\hat{\mathcal{X}}_2}\sum_{a\in\mathcal{A}}\sum_{x\in\mathcal{X}}\sum_{y_1\in\mathcal{Y}_1} q(a, b, x)p(y_1, y_2|x)\delta_2(\hat{x}, x).\end{aligned}$$

The joint distribution of $(\mathsf{A}, \mathsf{B}, \mathsf{X}, \mathsf{Y}_1, \mathsf{Y}_2)$ in (27) and (28) is understood as follows: $(\mathsf{A}, \mathsf{B}, \mathsf{X})$ is distributed according to q and $(\mathsf{Y}_1, \mathsf{Y}_2)$ conditionally depends on X via the true side information channel (the conditional distribution $\mathbb{P}[Y_1 = y_1, Y_2 = y_2|X = x]$); in particular, we have imposed the Markov chain $(\mathsf{A}, \mathsf{B}) \leftrightarrow \mathsf{X} \leftrightarrow (\mathsf{Y}_1, \mathsf{Y}_2)$. We also notice that

$$\begin{aligned}\mathbb{E}_C\big\{g_1(q_C)\big\} &= I(X; B|Y_2, C) - H(X|A, C, Y_1) \\ \mathbb{E}_C\big\{g_2(q_C)\big\} &= I(X; A|Y_1, C) - H(X|B, C, Y_2)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_C\big\{g_3(q_C)\big\} &= \min_{\phi_1:\mathcal{A}\times\mathcal{C}\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\,\delta_1\big(X, \phi_1(A, C, Y_1)\big) \\ \mathbb{E}_C\big\{g_4(q_C)\big\} &= \min_{\phi_2:\mathcal{B}\times\mathcal{C}\times\mathcal{Y}_2\to\hat{\mathcal{X}}_2} \mathbb{E}\,\delta_2\big(X, \phi_2(B, C, Y_2)\big).\end{aligned}$$

The Support Lemma asserts that there exists a new auxiliary random variable $C^\dagger$ defined on an alphabet $\mathcal{C}^\dagger$ with cardinality

$$|\mathcal{C}^\dagger| \leq |\mathcal{X}| + 3$$

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

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

and

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

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

$$(A^\dagger, B^\dagger, C^\dagger) \leftrightarrow X^\dagger \leftrightarrow (Y_1^\dagger, Y_2^\dagger) \quad (31)$$

to specify a tuple $(A^\dagger, B^\dagger, C^\dagger, X^\dagger, Y_1^\dagger, Y_2^\dagger)$ on $\mathcal{A}\times\mathcal{B}\times\mathcal{C}^\dagger\times\mathcal{X}\times\mathcal{Y}_1\times\mathcal{Y}_2$. The equality (29) ensures that $(X^\dagger, Y_1^\dagger, Y_2^\dagger)$ and (X, Y_1, Y_2) have the same distribution, which also implies

$$H(X^\dagger|Y_1^\dagger) = H(X|Y_1) \quad \text{and} \quad H(X^\dagger|Y_2^\dagger) = H(X|Y_2). \quad (32)$$

Similarly, (30) ensures

$$I(X^\dagger; B^\dagger|Y_2^\dagger, C^\dagger) - H(X^\dagger|A^\dagger, C^\dagger, Y_1^\dagger) = I(X; B|Y_2, C) - H(X|A, C, Y_1); \quad (33a)$$

$$I(X^\dagger; A^\dagger|Y_1^\dagger, C^\dagger) - H(X^\dagger|B^\dagger, C^\dagger, Y_2^\dagger) = I(X; A|Y_1, C) - H(X|B, C, Y_2); \quad (33b)$$

and

$$\min_{\phi_1^\dagger:\mathcal{A}\times\mathcal{C}^\dagger\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\,\delta_1\big(X^\dagger, \phi_1^\dagger(A^\dagger, C^\dagger, Y_1^\dagger)\big) = \min_{\phi_1:\mathcal{A}\times\mathcal{C}\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\,\delta_1\big(X, \phi_1(A, C, Y_1)\big) \quad (34a)$$

$$\min_{\phi_2^\dagger:\mathcal{B}\times\mathcal{C}^\dagger\times\mathcal{Y}_2\to\hat{\mathcal{X}}_2} \mathbb{E}\,\delta_2\big(X^\dagger, \phi_2^\dagger(B^\dagger, C^\dagger, Y_2^\dagger)\big) = \min_{\phi_2:\mathcal{B}\times\mathcal{C}\times\mathcal{Y}_2\to\hat{\mathcal{X}}_2} \mathbb{E}\,\delta_2\big(X, \phi_2(B, C, Y_2)\big). \quad (34b)$$

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

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

and

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

and therefore

$$\max_{j=1,2} I(X; C|Y_j) + I(X; A|C, Y_1) + I(X; B|C, Y_2) = \max_{j=1,2} I(X^\dagger; C^\dagger|Y_j^\dagger) + I(X^\dagger; A^\dagger|C^\dagger, Y_1^\dagger) + I(X^\dagger; B^\dagger|C^\dagger, Y_2^\dagger). \quad (35)$$

Consider the tuple $(A^\dagger, B^\dagger, C^\dagger, X^\dagger, Y_1^\dagger, Y_2^\dagger)$. 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 $(A^\dagger, B^\dagger, C^\dagger, X^\dagger, Y_1^\dagger, Y_2^\dagger)$ with another that shares the same Markov chain (31) and marginals $(A^\dagger, C^\dagger, X^\dagger)$, $(B^\dagger, C^\dagger, X^\dagger)$ and $(X^\dagger, Y_1^\dagger, Y_2^\dagger)$, but imposes the new chain

$$A^\dagger \leftrightarrow (C^\dagger, X^\dagger) \leftrightarrow B^\dagger. \quad (36)$$

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 $\mathcal{A}$ and $\mathcal{B}$ to be bounded independently. With a slight abuse of notation, we retain the same notation $(A^\dagger, B^\dagger, C^\dagger, X^\dagger, Y_1^\dagger, Y_2^\dagger)$ 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 $(C^\dagger, X^\dagger)$ given $A^\dagger = a$. Let $\mathcal{P}_2$ denote the set of all joint distributions on $\mathcal{C}^\dagger \times \mathcal{X}$. For each and every (c, x) in $\mathcal{C}^\dagger \times \mathcal{X}$ but one, define $g_{c,x} : \mathcal{P}_2 \longrightarrow [0, 1]$ by setting

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

Here

$$\mathbb{E}_{A^\dagger}\big\{g_{c,x}(q_{A^\dagger})\big\} = \mathbb{P}[(C^\dagger, X^\dagger) = (c, x)]$$

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

$$g_5(q) := H(\mathsf{X}|\mathsf{C}, \mathsf{Y}_1)$$

and

$$g_6(q) := \sum_{c\in\mathcal{C}^\dagger} \sum_{y_1\in\mathcal{Y}_1} \min_{\hat{x}\in\hat{\mathcal{X}}_1} \sum_{x\in\mathcal{X}} \sum_{y_2\in\mathcal{Y}_2} q(c, x)p(y_1, y_2|x)\delta_1(\hat{x}, x),$$

where the joint distribution of $(\mathsf{C}, \mathsf{X}, \mathsf{Y}_1, \mathsf{Y}_2)$ is understood as follows: (C, X) is distributed according to q, and $(\mathsf{Y}_1, \mathsf{Y}_2)$ conditionally depends on X via the true side information channel. We have

$$\mathbb{E}_{A^\dagger}\big\{g_5(q_{A^\dagger})\big\} = H(X^\dagger|A^\dagger, C^\dagger, Y_1^\dagger).$$

and

$$\mathbb{E}_{A^\dagger}\big\{g_6(q_{A^\dagger})\big\} = \min_{\phi_1^\dagger:\mathcal{A}\times\mathcal{C}^\dagger\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\delta_1\big(X, \phi_1^\dagger(A^\dagger, C^\dagger, Y_1^\dagger)\big).$$

The Support Lemma asserts that there exists a random variable $A^\ddagger$ defined on an alphabet $\mathcal{A}^\ddagger$ with cardinality

$$|\mathcal{A}^\ddagger| \leq |\mathcal{C}^\dagger||\mathcal{X}| + 1$$

together with a collection of $|\mathcal{A}^\ddagger|$ distributions $\{q_a^\ddagger\}$ from $\mathcal{P}_2$ (indexed by the elements a of $\mathcal{A}^\ddagger$) such that

$$\mathbb{E}_{A^\ddagger}\big\{g_{c,x}(q_{A^\ddagger})\big\} = \mathbb{E}_{A^\dagger}\big\{g_{c,x}(q_{A^\dagger})\big\} \quad (37)$$

and

$$\mathbb{E}_{A^\ddagger}\big\{g_j(q_{A^\ddagger})\big\} = \mathbb{E}_{A^\dagger}\big\{g_j(q_{A^\dagger})\big\}, \quad j = 5, 6. \quad (38)$$

The new variable $A^\ddagger$, the distributions $\{q_a^\ddagger\}$, the true side information channel, the conditional distribution $P(B^\dagger|X^\dagger, C^\dagger)$, and the Markov chains (31) and (36) come together to specify a tuple $(A^\ddagger, B^\ddagger, C^\ddagger, X^\ddagger, Y_1^\ddagger, Y_2^\ddagger)$ on $\mathcal{A}^\ddagger\times\mathcal{B}\times\mathcal{C}^\dagger\times\mathcal{X}\times\mathcal{Y}_1\times\mathcal{Y}_2$.

The equalities in (37) ensure that $(C^\ddagger, X^\ddagger)$ and $(C^\dagger, X^\dagger)$ have the same distribution. By construction, we also have that $(B^\ddagger, C^\ddagger, X^\ddagger, Y_1^\ddagger, Y_2^\ddagger)$ and $(B^\dagger, C^\dagger, X^\dagger, Y_1^\dagger, Y_2^\dagger)$ have the same distribution, and therefore

$$\begin{aligned}
&\max\Big\{I(X^\ddagger;C^\ddagger|Y_1^\ddagger), I(X^\ddagger;C^\ddagger|Y_2^\ddagger)\Big\} + H(X^\ddagger|C^\ddagger, Y_1^\ddagger) \\
&\qquad\qquad + I(X^\ddagger; B^\ddagger|C^\ddagger, Y_2^\ddagger) \\
&= \max\Big\{I(X^\dagger;C^\dagger|Y_1^\dagger), I(X^\dagger;C^\dagger|Y_2^\dagger)\Big\} + H(X^\dagger|C^\dagger, Y_1^\dagger) \\
&\qquad\qquad + I(X^\dagger; B^\dagger|C^\dagger, Y_2^\dagger). \quad (39)
\end{aligned}$$

In addition, (38) ensures that

$$H(X^\ddagger|A^\ddagger, C^\ddagger, Y_1^\ddagger) = H(X^\dagger|A^\dagger, C^\dagger, Y_1^\dagger) \quad (40)$$

and

$$\begin{aligned}
&\min_{\phi_1^\ddagger:\mathcal{A}^\ddagger\times\mathcal{C}^\dagger\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\,\delta_1\big(X^\ddagger, \phi_1^\ddagger(A^\ddagger, C^\ddagger, Y_1^\ddagger)\big) \\
&\quad = \min_{\phi_1^\dagger:\mathcal{A}\times\mathcal{C}^\dagger\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\,\delta_1\big(X^\dagger, \phi_1(A^\dagger, C^\dagger, Y_1^\dagger)\big). \quad (41)
\end{aligned}$$

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

$$\begin{aligned}
&\max\Big\{I(X^\ddagger;C^\ddagger|Y_1^\ddagger), I(X^\ddagger;C^\ddagger|Y_2^\ddagger)\Big\} + I(X^\ddagger; A^\ddagger|C^\ddagger, Y_1^\ddagger) \\
&\qquad\qquad + I(X^\ddagger; B^\ddagger|C^\ddagger, Y_2^\ddagger) \\
&= \max\Big\{I(X;C|Y_1), I(X;C|Y_2)\Big\} + I(X;A|C,Y_1) \\
&\qquad\qquad + I(X;B|C,Y_2). \quad (42)
\end{aligned}$$

as well as

$$\begin{aligned}
&\min_{\phi_1^\ddagger:\mathcal{A}^\ddagger\times\mathcal{C}^\dagger\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\,\delta_1\big(X^\ddagger, \phi_1^\ddagger(A^\ddagger, C^\ddagger, Y_1^\ddagger)\big) \\
&\quad = \min_{\phi_1:\mathcal{A}\times\mathcal{C}\times\mathcal{Y}_1\to\hat{\mathcal{X}}_1} \mathbb{E}\,\delta_1\big(X, \phi_1(A, C, Y_1)\big) \quad (43a)
\end{aligned}$$

and

$$\begin{aligned}
&\min_{\phi_2^\ddagger:\mathcal{B}\times\mathcal{C}^\dagger\times\mathcal{Y}_2\to\hat{\mathcal{X}}_2} \mathbb{E}\,\delta_2\big(X^\ddagger, \phi_2^\ddagger(B^\ddagger, C^\ddagger, Y_2^\ddagger)\big) \\
&\quad = \min_{\phi_2:\mathcal{B}\times\mathcal{C}\times\mathcal{Y}_2\to\hat{\mathcal{X}}_2} \mathbb{E}\,\delta_2\big(X, \phi_2(B, C, Y_2)\big), \quad (43b)
\end{aligned}$$

as desired.

Using analogous arguments as above, we can find a random vector $(A', B', C', X', Y_1', Y_2')$ over $\mathcal{A}^\ddagger\times\mathcal{B}'\times\mathcal{C}^\dagger\times\mathcal{X}\times\mathcal{Y}_1\times\mathcal{Y}_2$, where the cardinality of the alphabet $\mathcal{B}'$ satisfies

$$|\mathcal{B}'| \le |\mathcal{C}^\dagger||\mathcal{X}| + 1,$$

and such that (42) and (43) are satisfied when the tuple $(A^\ddagger, B^\ddagger, C^\ddagger, X^\ddagger, Y_1^\ddagger, Y_2^\ddagger)$ is replaced by the new tuple $(A', B', C', X', Y_1', Y_2')$. This concludes the proof of the cardinality bounds. ■

Appendix C
Proof of Lemma 4

A. Assertion (i)

Consider any auxiliary random variable W for which

$$W \leftrightarrow (X, L) \leftrightarrow (Y_1, Y_2) \quad (44)$$

is a Markov chain. We have

$$\begin{aligned}
I(W;Y_2|L) &= H(W|L) - H(W|L,Y_2) \\
&\overset{\text{(a)}}{=} H(W|L) - H(W|L,Y_2,Y_1) \\
&\ge H(W|L) - H(W|L,Y_1) \\
&= I(W;Y_1|L),
\end{aligned}$$

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. ■

B. Assertion (ii)

Take any auxiliary random variable W for which

$$W \leftrightarrow (X_1, X_2) \leftrightarrow (Y_1, Y_2).$$

Consider Definition 6 with $L = X_1$. We have

$$\begin{aligned}
0 &\le I(W;Y_1|X_1) \\
&= H(Y_1|X_1) - H(Y_1|W,X_1) \\
&\overset{\text{(a)}}{=} H(Y_1|X_1,X_2) - H(Y_1|W,X_1) \\
&\overset{\text{(b)}}{=} H(Y_1|X_1,X_2) - H(Y_1|W,X_1,X_2) \\
&= I(W;Y_1|X_1,X_2) \\
&\overset{\text{(c)}}{=} 0,
\end{aligned}$$

where the indicated steps apply the following Markov chains:

$$\begin{aligned}
&\text{(a)} \quad X_2 \leftrightarrow X_1 \leftrightarrow Y_1 \\
&\text{(b)} \quad X_2 \leftrightarrow (W, X_1) \leftrightarrow Y_1 \\
&\text{(c)} \quad W \leftrightarrow (X_1, X_2) \leftrightarrow (Y_1, Y_2).
\end{aligned}$$

Thus, we have that

$$I(W;Y_1|X_1) = 0$$

and therefore $I(W;Y_1|X_1)$ is no larger than $I(W;Y_2|X_1)$. ■

Appendix D
Proof of Lemma 5

Fix a distortion $D_2 \ge 0$ 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

$$R + \epsilon \ge \frac{1}{n}\log|\mathcal{M}|, \quad (45)$$

$$\epsilon \ge \mathbb{E}\,\frac{1}{n}\sum_{i=1}^n \delta_1(X_i, \hat{X}_{1,i}), \quad (46)$$

and

$$D_2 + \epsilon \ge \mathbb{E}\,\frac{1}{n}\sum_{i=1}^n \delta_2(X_i, \hat{X}_{2,i}). \quad (47)$$

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

$$P_{\mathrm{e},i} := \mathbb{P}\big[\hat{X}_{1,i} \ne \tilde{X}_i\big].$$

The probability $P_{\mathrm{e},i}$ can be expressed as $P_{\mathrm{e},i} = \mathbb{E}\delta_1(X_i, \hat{X}_{1,i})$, so by (46)

$$\frac{1}{n}\sum_{i=1}^{n} P_{\mathrm{e},i} \le \epsilon. \tag{48}$$

Consider next the conditional entropy $H(\tilde{\boldsymbol{X}}|M, \boldsymbol{Y_1})$. Starting from the fact that $\hat{\boldsymbol{X}}_{\boldsymbol{1}}$ is determined by $(M, \boldsymbol{Y_1})$, we have

$$\begin{aligned}
H(\tilde{\boldsymbol{X}}|M, \boldsymbol{Y_1}) &\overset{\text{(a)}}{=} H(\tilde{\boldsymbol{X}}|M, \boldsymbol{Y_1}, \hat{\boldsymbol{X}}_{\boldsymbol{1}}) \le H(\tilde{\boldsymbol{X}}|\hat{\boldsymbol{X}}_{\boldsymbol{1}}) \\
&\overset{\text{(b)}}{\le} \sum_{i=1}^{n} H(\tilde{X}_i|\hat{X}_{1,i}) \\
&\overset{\text{(c)}}{\le} \sum_{i=1}^{n} \big(h(P_{\mathrm{e},i}) + P_{\mathrm{e},i}\log|\tilde{\mathcal{X}}|\big) \\
&\overset{\text{(d)}}{\le} n\ h\left(\frac{1}{n}\sum_{i=1}^{n} P_{\mathrm{e},i}\right) + \left(\sum_{i=1}^{n} P_{\mathrm{e},i}\right)\log|\tilde{\mathcal{X}}| \\
&\overset{\text{(e)}}{\le} nh(\epsilon) + n\epsilon\log|\tilde{\mathcal{X}}| \\
&\overset{\text{(f)}}{=} n\varepsilon(\epsilon),
\end{aligned} \tag{49}$$

where (a) applies the Markov chain

$$\tilde{\boldsymbol{X}} \leftrightarrow (M, \boldsymbol{Y_1}) \leftrightarrow \hat{\boldsymbol{X}}_{\boldsymbol{1}};$$

(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

$$\varepsilon(\epsilon) := h(\epsilon) + \epsilon\log|\tilde{\mathcal{X}}|.$$

Finally, we notice that $\varepsilon(\epsilon) \to 0$ as $\epsilon \to 0$.

Now consider the rate condition (45). We have

$$\begin{aligned}
R + \epsilon &\ge \frac{1}{n}\log_2|\mathcal{M}| \\
&\ge \frac{1}{n}H(M) \\
&\ge \frac{1}{n}H(M|\boldsymbol{Y_1}) \\
&\ge \frac{1}{n}I(\boldsymbol{X}, \tilde{\boldsymbol{X}}; M|\boldsymbol{Y_1}) \\
&= \frac{1}{n}\Big(I(\tilde{\boldsymbol{X}}; M|\boldsymbol{Y_1}) + I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})\Big) \\
&\overset{\text{(a)}}{\ge} \frac{1}{n}\Big(H(\tilde{\boldsymbol{X}}|\boldsymbol{Y_1}) - n\varepsilon(\epsilon) + I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})\Big) \\
&\overset{\text{(b)}}{=} H(\tilde{X}|Y_1) - \varepsilon(\epsilon) + \frac{1}{n}I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1}),
\end{aligned} \tag{50}$$

where (a) substitutes (49) and (b) invokes the fact that $(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{Y_1})$ 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 $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_2})$ instead of $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})$, we obtain

$$\begin{aligned}
&I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1}) \\
&\overset{\text{(a)}}{=} I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) - H(M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) + H(M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1}) \\
&= I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) + I(M; \boldsymbol{Y_2}|\tilde{\boldsymbol{X}}) - I(M; \boldsymbol{Y_1}|\tilde{\boldsymbol{X}})
\end{aligned} \tag{51}$$

where (a) invokes that M is a function of $\boldsymbol{X}$ or, in the more general case of stochastic encoders, that

$$M \leftrightarrow \boldsymbol{X} \leftrightarrow (\tilde{\boldsymbol{X}}, \boldsymbol{Y_1}, \boldsymbol{Y_2}).$$

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:

$$\begin{aligned}
&I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) \\
&= \sum_{i=1}^{n} I(X_i; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}, X_1^{i-1}) \\
&\overset{\text{(a)}}{=} \sum_{i=1}^{n} I(X_i; M, \tilde{X}_1^{i-1}, \tilde{X}_{i+1}^{n}, Y_{2,1}^{i-1}, Y_{2,i+1}^{n}, X_1^{i-1}|\tilde{X}_i, Y_{2,i}) \\
&\ge \sum_{i=1}^{n} I(X_i; M, Y_{2,1}^{i-1}, Y_{2,i+1}^{n}|\tilde{X}_i, Y_{2,i}) \\
&\overset{\text{(b)}}{=} \sum_{i=1}^{n} I(X_i; B_i|\tilde{X}_i, Y_{2,i}),
\end{aligned} \tag{52}$$

where (a) follows because $(\boldsymbol{X}, \boldsymbol{Y_2}, \tilde{\boldsymbol{X}})$ i.i.d. and therefore

$$H(X_i|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}, X_1^{i-1}) = H(X_i|\tilde{X}_i, Y_{2,i}),$$

and in (b) we define

$$B_i := (M, Y_{2,1}^{i-1}, Y_{2,i+1}^{n}).$$

Continuing on from (52), we have

$$\begin{aligned}
\frac{1}{n}I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) &\ge \frac{1}{n}\sum_{i=1}^{n} I(X_i; B_i|\tilde{X}_i, Y_{2,i}) \\
&\overset{\text{(a)}}{\ge} \frac{1}{n}\sum_{i=1}^{n} S\big(\mathbb{E}\delta_2(X_i, \hat{X}_{2,i})\big) \\
&\overset{\text{(b)}}{\ge} S\left(\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\delta_2(X_i, \hat{X}_{2,i})\right) \\
&\overset{\text{(c)}}{\ge} S(D_2 + \epsilon),
\end{aligned} \tag{53}$$

where

(a) follows from the definition of $S(D_2)$ upon noticing that the i-th reconstructed symbol, $\hat{X}_{2,i}$, 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

$$D_2 + \epsilon \ge \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\delta_2(X_i, \hat{X}_{2,i}).$$

Consider (50), (51) and (53). We have

$$\begin{aligned}
R + \epsilon \ge\ & H(\tilde{X}|Y_1) - \varepsilon(\epsilon) + S(D_2 + \epsilon) \\
&+ \frac{1}{n}\Big(I(M; \boldsymbol{Y_2}|\tilde{\boldsymbol{X}}) - I(M; \boldsymbol{Y_1}|\tilde{\boldsymbol{X}})\Big).
\end{aligned}$$

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

$$W \leftrightarrow X \leftrightarrow (Y_1, Y_2),$$

$|\mathcal{W}| \leq |\mathcal{X}|$, and

$$R + \epsilon \geq H(\tilde{X}|Y_1) - \varepsilon(\epsilon) + S(D_2 + \epsilon) + I(W; Y_2|\tilde{X}) - I(W; Y_1|\tilde{X}).$$

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 3 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

$$\begin{aligned} H(M) &\geq I(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{Y_1}, \boldsymbol{Y_2}; M) \\ &\geq I(\tilde{\boldsymbol{X}}; M|\boldsymbol{Y_1}) + I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1}, \boldsymbol{Y_2}) \\ &\overset{(a)}{=} H(\tilde{\boldsymbol{X}}|\boldsymbol{Y_1}) - H(\tilde{\boldsymbol{X}}|M, \boldsymbol{Y_1}) + I(\boldsymbol{X}; M|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) \\ &\overset{(b)}{\geq} n\Big(H(\tilde{X}|Y_1) - \varepsilon(\epsilon) + S(D_2 + \epsilon)\Big), \end{aligned} \tag{54}$$

where (a) applies $M \leftrightarrow (\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) \leftrightarrow \boldsymbol{Y_1}$ and (b) repeats the steps in (49), (53), where $\varepsilon(\epsilon)$ can be chosen so that $\varepsilon(\epsilon) \to 0$ 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

$$(Y_1, Y_2, Y_3) \leftrightarrow X \leftrightarrow (U_{123}, U_{12}, U_{13}, U_{23}, U_1, U_2, U_3); \tag{55}$$

forms a Markov chain, and there exist three deterministic mappings

$$\phi_j : \mathcal{U}_j \times \mathcal{Y}_j \longrightarrow \hat{\mathcal{X}}_j, \quad j = 1, 2, 3,$$

with

$$D_j \geq \mathbb{E}\ \delta_j\big(X, \phi_j(U_j, Y_j)\big).$$

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

$$\begin{aligned} U_{123} &= U_1 = A_1 \\ U_{23} &= U_2 = A_2 \\ U_3 &= A_3. \end{aligned}$$

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 11. ■

Appendix G
Proof of Lemma 9

We have

$$\begin{aligned} R_1 + \epsilon &\geq \frac{1}{n} H(M_1) \\ &\geq \frac{1}{n} I(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}; M_1|\boldsymbol{Y_1}) \\ &\overset{(a)}{\geq} \frac{1}{n}\big(H(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}|\boldsymbol{Y_1}) - n\varepsilon_1(\epsilon)\big) \\ &\overset{(b)}{=} H(\tilde{X}_1|Y_1) - \varepsilon_1(\epsilon), \end{aligned} \tag{58}$$

where (a) applies Fano's inequality in the same way as (49), where $\varepsilon_1(\epsilon)$ can be chosen so that $\varepsilon_1(\epsilon) \to 0$ as $\epsilon \to 0$; and (b) follows because the pair $(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \boldsymbol{Y}_1)$ 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 $(M_1, M_2) \leftrightarrow (\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \boldsymbol{X}) \leftrightarrow (\boldsymbol{Y_1}, \boldsymbol{Y_2})$;
(b) $\tilde{\boldsymbol{X}}_{\boldsymbol{2}}$ is determined by $\boldsymbol{X}$;
(c) exploits the fact that $(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \tilde{\boldsymbol{X}}_{\boldsymbol{2}}, \boldsymbol{Y_1}, \boldsymbol{Y_2})$ is i.i.d. and applies Fano's inequality twice, in a manner similar to (49), where $\varepsilon_1(\epsilon)$ and $\varepsilon_2(\epsilon)$ 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

$$(M_1, M_2, M_3) \leftrightarrow (\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \tilde{\boldsymbol{X}}_{\boldsymbol{2}}, \boldsymbol{X}) \leftrightarrow (\boldsymbol{Y_2}, \boldsymbol{Y_3}). \tag{61}$$

Consider the first conditional mutual information on the right hand side of (62). We have

$$\begin{aligned} &\frac{1}{n} I(\boldsymbol{X}; M_1, M_2, M_3|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \tilde{\boldsymbol{X}}_{\boldsymbol{2}}, \boldsymbol{Y_3}) \\ &\overset{(a)}{\geq} \frac{1}{n} \sum_{i=1}^{n} I(X_i; M_1, M_2, M_3, Y_{3,1}^{i-1}, Y_{3,i+1}^{n}|\tilde{X}_{1,i}, \tilde{X}_{2,i}, Y_{3,i}) \\ &\overset{(b)}{=} \frac{1}{n} \sum_{i=1}^{n} I(X_i; C_i|\tilde{X}_{1,i}, \tilde{X}_{2,i}, Y_{3,i}) \\ &\overset{(c)}{\geq} \sum_{i=1}^{n} S'\big(\mathbb{E}\delta_3(X_i, \hat{X}_{3,i})\big) \\ &\overset{(d)}{\geq} S'\left(\mathbb{E}\frac{1}{n}\sum_{i=1}^{n} \delta_3(X_i, \tilde{X}_{3,i})\right) \\ &\overset{(e)}{\geq} S'(D_3 + \epsilon), \end{aligned} \tag{63}$$

where (a) follows from the same reasoning as step (a) of (52); in (b), we define

$$C_i := \big(M_1, M_2, M_3, Y_{3,1}^{i-1}, Y_{3,i+1}^{n}\big);$$

$$\begin{aligned} R_1 \geq\ & I(X;U_{123}) - I(U_{123};Y_1) \\ & + I(X;U_{12}|U_{123}) - I(U_{12};Y_1|U_{123}) \\ & + I(X,U_{12};U_{13}|U_{123}) - I(U_{13};U_{12}Y_1|U_{123}) \\ & + I(X;U_1|U_{123},U_{12},U_{13}) - I(U_1;Y_1|U_{123},U_{12},U_{13}) \end{aligned} \tag{57a}$$

$$\begin{aligned} R_1 + R_2 \geq\ & I(X;U_{123}) - \min\big\{I(U_{123};Y_1), I(U_{123};Y_2)\big\} \\ & + I(X;U_{12}|U_{123}) - \min\big\{I(U_{12};Y_1|U_{123}), I(U_{12};Y_2|U_{123})\big\} \\ & + I(X,U_{12};U_{13}|U_{123}) - I(U_{13};U_{12},Y_1|U_{123}) \\ & + I(X,U_{12},U_{13};U_{23}|U_{123}) - I(U_{23};U_{12},Y_2|U_{123}) \\ & + I(X;U_1|U_{123},U_{12},U_{13}) - I(U_1;Y_1|U_{123},U_{12},U_{13}) \\ & + I(X;U_2|U_{123},U_{12},U_{23}) - I(U_2;Y_2|U_{123},U_{12},U_{23}) \end{aligned} \tag{57b}$$

$$\begin{aligned} R_1 + R_2 + R_3 \geq\ & I(X;U_{123}) - \min\big\{I(U_{123};Y_1), I(U_{123};Y_2), I(U_{123};Y_3)\big\} \\ & + I(X;U_{12}|U_{123}) - \min\big\{I(U_{12};Y_1|U_{123}), I(U_{12};Y_2|U_{123})\big\} \\ & + I(X,U_{12};U_{13}|U_{123}) - \min\big\{I(U_{13};U_{12},Y_1|U_{123}), I(U_{13};Y_3|U_{123})\big\} \\ & + I(X,U_{12},U_{13};U_{23}|U_{123}) - \min\big\{I(U_{23};U_{12},Y_2|U_{123}), I(U_{23};U_{13},Y_3|U_{123})\big\} \\ & + I(X;U_1|U_{123},U_{12},U_{13}) - I(U_1;Y_1|U_{123},U_{12},U_{13}) \\ & + I(X;U_2|U_{123},U_{12},U_{23}) - I(U_2;Y_2|U_{123},U_{12},U_{23}) \\ & + I(X;U_3|U_{123},U_{13},U_{23}) - I(U_3;Y_3|U_{123},U_{13},U_{23}). \end{aligned} \tag{57c}$$

$$\begin{aligned} R_1 + R_2 + \epsilon &\geq \frac{1}{n} H(M_1,M_2) \geq \frac{1}{n} I(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \boldsymbol{X}; M_1, M_2|\boldsymbol{Y_1}) \\ &= \frac{1}{n}\Big(I(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}; M_1, M_2|\boldsymbol{Y_1}) + I(\boldsymbol{X}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \boldsymbol{Y_1})\Big) \\ &\overset{\text{(a)}}{=} \frac{1}{n}\Big(I(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}; M_1, M_2|\boldsymbol{Y_1}) + I(\boldsymbol{X}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \boldsymbol{Y_2}) + I(\boldsymbol{Y_2}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}) - I(\boldsymbol{Y_1}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})\Big) \\ &\overset{\text{(b)}}{=} \frac{1}{n}\Big(I(\tilde{\boldsymbol{X}}_{\boldsymbol{1}}; M_1, M_2|\boldsymbol{Y}_1) + I(\tilde{\boldsymbol{X}}_{\boldsymbol{2}}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \boldsymbol{Y_2}) + I(\boldsymbol{X}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \tilde{\boldsymbol{X}}_{\boldsymbol{2}}, \boldsymbol{Y_2}) \\ &\qquad + I(\boldsymbol{Y_2}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}) - I(\boldsymbol{Y_1}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})\Big) \\ &\overset{\text{(c)}}{\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}\Big(I(\boldsymbol{X}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \tilde{\boldsymbol{X}}_{\boldsymbol{2}}, \boldsymbol{Y_2}) \\ &\qquad + I(\boldsymbol{Y_2}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}) - I(\boldsymbol{Y_1}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})\Big) \qquad (59) \\ &\overset{\text{(d)}}{\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}\Big(I(\boldsymbol{Y_2}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}) - I(\boldsymbol{Y_1}; M_1, M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})\Big). \qquad (60) \end{aligned}$$

---

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

$$R = X,\ S_1 = Y_1,\ S_2 = Y_2,\ T = Y_3 \text{ and } L = \tilde{X}_1,$$

and once for

$$R = X,\ S_1 = Y_2,\ S_2 = Y_3,\ T = Y_1 \text{ and } L = (\tilde{X}_1, \tilde{X}_2).$$

We conclude that there exist auxiliary random variables W_1, W_2 and W_3 with

$$|\mathcal{W}_1|, |\mathcal{W}_2|, |\mathcal{W}_3| \leq |\mathcal{X}|,$$

and

$$W_j \leftrightarrow X \leftrightarrow (Y_1, Y_2, Y_3), \qquad j = 1, 2, 3,$$

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

$$\begin{aligned} R_1 + R_2 + \epsilon \geq\ & H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) + I(W_1; Y_2|\tilde{X}_1) \\ & - I(W_1; Y_1|\tilde{X}_1) - \varepsilon_1(\epsilon) - \varepsilon_2(\epsilon) \end{aligned} \tag{65}$$

and

$$\begin{aligned} & R_1 + R_2 + R_3 + \epsilon \\ & \geq H(\tilde{X}_1|Y_1) + H(\tilde{X}_2|\tilde{X}_1, Y_2) + S'(D_3 + \epsilon) - \varepsilon_2(\epsilon) - \varepsilon_1(\epsilon) \\ & \quad + I(W_3; Y_3|\tilde{X}_1, \tilde{X}_2) - I(W_3; Y_2|\tilde{X}_1, \tilde{X}_2) + I(W_2; Y_2|\tilde{X}_1) \\ & \quad - I(W_2; Y_1|\tilde{X}_1). \end{aligned} \tag{66}$$

$$R_1+R_2+R_3+\epsilon \geq H(\tilde{X}_1|Y_1)+H(\tilde{X}_2|\tilde{X}_1,Y_2)-\varepsilon_1(\epsilon)-\varepsilon_2(\epsilon)+\frac{1}{n}\Big(I(\boldsymbol{X};M_1,M_2,M_3|\tilde{\boldsymbol{X}}_{\boldsymbol{1}},\tilde{\boldsymbol{X}}_{\boldsymbol{2}},\boldsymbol{Y_2}) + I(\boldsymbol{Y_2};M_1,M_2,M_3|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})-I(\boldsymbol{Y_1};M_1,M_2,M_3|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})\Big)$$
$$\overset{\text{(a)}}{=} H(\tilde{X}_1|Y_1)+H(\tilde{X}_2|\tilde{X}_1,Y_2)-\varepsilon_1(\epsilon)-\varepsilon_2(\epsilon)+\frac{1}{n}\Big(I(\boldsymbol{X};M_1,M_2,M_3|\tilde{\boldsymbol{X}}_{\boldsymbol{1}},\tilde{\boldsymbol{X}}_{\boldsymbol{2}},\boldsymbol{Y_3}) + I(M_1,M_2,M_3;\boldsymbol{Y_3}|\tilde{\boldsymbol{X}}_{\boldsymbol{1}},\tilde{\boldsymbol{X}}_{\boldsymbol{2}})-I(M_1,M_2,M_3;\boldsymbol{Y_2}|\tilde{\boldsymbol{X}}_{\boldsymbol{1}},\tilde{\boldsymbol{X}}_{\boldsymbol{2}}) + I(\boldsymbol{Y_2};M_1,M_2,M_3|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})-I(\boldsymbol{Y_1};M_1,M_2,M_3|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})\Big) \quad (62)$$

$$R_1+R_2+R_3+\epsilon \geq H(\tilde{X}_1|Y_1)+H(\tilde{X}_2|\tilde{X}_1,Y_2)+S'(D_3+\epsilon)+\frac{1}{n}\Big(I(M_1,M_2,M_3;\boldsymbol{Y_3}|\tilde{\boldsymbol{X}}_{\boldsymbol{1}},\tilde{\boldsymbol{X}}_{\boldsymbol{2}}) - I(M_1,M_2,M_3;\boldsymbol{Y_2}|\tilde{\boldsymbol{X}}_{\boldsymbol{1}},\tilde{\boldsymbol{X}}_{\boldsymbol{2}})\Big)+\frac{1}{n}\Big(I(M_1,M_2,M_3;\boldsymbol{Y_2}|\tilde{\boldsymbol{X}}_{\boldsymbol{1}}) - I(M_1,M_2,M_3;\boldsymbol{Y_1}|\tilde{\boldsymbol{X}}_{\boldsymbol{1}})\Big)-\varepsilon_1(\epsilon)-\varepsilon_2(\epsilon). \quad (64)$$

The converse proof follows by (58), (65), and (66), by letting $\epsilon \to 0$, and by the continuity of $S'(D_3)$ in D_3. ■

APPENDIX H
PROOFS OF THEOREM 12

A. Assertion (i)

Achievability: The rate constraints follow from (6) upon setting $A_1 = \tilde{X}_1$ and $A_{12} = A_2 = \tilde{X}_2$ and invoking the assumptions $\tilde{X}_2 = \psi'(\tilde{X}_1)$ and $H(\tilde{X}_2|Y_1) \leq H(\tilde{X}_2|Y_2)$.

Converse: The lower bound on R_1 is trivial. The lower bound on the sum rate R_1+R_2 follows by, now familiar, arguments:

$$\begin{aligned}
&R_1+R_2+\epsilon \geq \frac{1}{n}H(M_1,M_2) \geq \frac{1}{n}I(\boldsymbol{X},\tilde{\boldsymbol{X}}_{\boldsymbol{2}};M_1,M_2|\boldsymbol{Y_2})\\
&=\frac{1}{n}\Big(I(\tilde{\boldsymbol{X}}_{\boldsymbol{2}};M_1,M_2|\boldsymbol{Y_2})+I(\boldsymbol{X};M_1,M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{2}},\boldsymbol{Y_2})\Big)\\
&=\frac{1}{n}\Big(I(\tilde{\boldsymbol{X}}_{\boldsymbol{2}};M_1,M_2|\boldsymbol{Y_2})+I(\boldsymbol{X};M_1,M_2|\tilde{\boldsymbol{X}}_{\boldsymbol{2}},\boldsymbol{Y_1})\\
&\qquad +I(M_1,M_2;\boldsymbol{Y_1}|\tilde{\boldsymbol{X}}_{\boldsymbol{2}})-I(M_1,M_2;\boldsymbol{Y_2}|\tilde{\boldsymbol{X}}_{\boldsymbol{2}})\Big)\\
&\overset{\text{(a)}}{\geq} H(\tilde{X}_2|Y_2)+H(\tilde{X}_1|\tilde{X}_2,Y_1)-\varepsilon(\epsilon)\\
&\qquad +\frac{1}{n}\Big(I(M_1,M_2;\boldsymbol{Y_1}|\tilde{\boldsymbol{X}}_{\boldsymbol{2}})-I(M_1,M_2;\boldsymbol{Y_2}|\tilde{\boldsymbol{X}}_{\boldsymbol{2}})\Big)\\
&\overset{\text{(b)}}{=} H(\tilde{X}_2|Y_2)+H(\tilde{X}_1|\tilde{X}_2,Y_1)-\varepsilon(\epsilon)+I(W;Y_1|\tilde{X}_2)\\
&\qquad -I(W;Y_2|\tilde{X}_2)\\
&\overset{\text{(c)}}{\geq} H(\tilde{X}_2|Y_2)+H(\tilde{X}_1|\tilde{X}_2,Y_1)-\varepsilon(\epsilon),
\end{aligned}$$

where (a) applies Fano's inequality and that $\tilde{X}_1$ can be computed as a function of X and $\varepsilon(\epsilon) \to 0$ as $\epsilon \to 0$; (b) uses Lemma 1; and (c) invokes the assumption $(Y_1 \succeq Y_2 \mid \tilde{X}_2)$. ■

B. Assertion (ii)

Achievability: The rate constraints follow from (6) upon setting $A_{12} = \tilde{X}_1$, $A_2 = \tilde{X}_2$ and $A_1 =$ constant and invoking the assumptions $\tilde{X}_1 = \psi'(\tilde{X}_2)$ and $H(\tilde{X}_1|Y_1) \leq H(\tilde{X}_1|Y_2)$.

Converse: The converse holds because for $j = 1, 2$, we have $R_j \geq H(\tilde{X}_j|Y_j) \geq 0$. ■

Roy Timo (S'06-M'09) is an Alexander von Humboldt research fellow with the Institute for Communications Engineering at the Technische Universität München (TUM). Prior to joining TUM, Dr. Timo was a research fellow with the Institute for Telecommunications Research at the University of South Australia and postdoctoral researcher with the Communications and Electronics Department at Telecom ParisTech.

Dr. Timo received the Bachelor of Engineering (Hons.) degree from The Australian National University (ANU) in July 2005, and the Ph.D. degree in Engineering from The ANU in December 2009. He was a NICTA-enhanced ANU Ph.D. candidate at NICTA's Canberra Research Laboratory. He received the best student paper awards at the 2007 Australian Communications Theory Workshop and the 2007 IEEE Australasian Telecommunications Networking and Applications Conference. He is a member of the IEEE Information Theory Society.

Tobias J. Oechtering (S'01-M'08-SM'12) received his Dipl-Ing degree in Electrical Engineering and Information Technology in 2002 from RWTH Aachen University, Germany, his Dr-Ing degree in Electrical Engineering in 2007 from the Technische Universität Berlin, Germany, and his Docent degree in Communication Theory in 2012 from KTH Royal Institute of Technology. Between 2002 and 2008 he has been with Technische Universität Berlin and Fraunhofer Heinrich-Hertz Institute, Berlin, Germany. In 2008 he joined the Communication Theory department at KTH Royal Institute of Technology, Stockholm, Sweden and has been an Associate Professor since May 2013. Presently, he is serving as an editor for IEEE Communications Letters. Dr. Oechtering received the "Förderpreis 2009" from the Vodafone Foundation. His research interests include information and communication theory, physical layer security, statistical signal processing, as well as networked control.

Michèle Wigger (S'05-M'09) received the M.Sc. degree in electrical engineering (with distinction) and the Ph.D. degree in electrical engineering both from ETH Zurich in 2003 and 2008, respectively. In 2009 she was a postdoctoral researcher at the ITA center at the University of California, San Diego. Since December 2009 she has been an Assistant Professor at Telecom ParisTech, in Paris, France. Her research interests are in information and communications theory.

REFERENCES

[1] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[2] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Transactions on Information Theory*, vol. 31, no. 6, pp. 727–734, 1985.

[3] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Academic Press, 1981.

[4] T. Matsuta and T. Uyematsu, "A general formula of rate-distortion functions for source coding with side information at many decoders," in *proceedings IEEE International Symposium on Information Theory*, MIT, Cambridge, MA, 2012.

[5] S. Unal and A. B. Wagner, "General index coding with side information: three decoder case," in *IEEE International Symposium on Information Theory*, Istanbul, Turkey, 2013.

[6] T. Laich and M. Wigger, "Utility of encoder side information for the lossless Kaspi/Heegard-Berger problem," in *IEEE International Symposium on Information Theory*, Istanbul, Turkey, 2013.

[7] R. Timo, T. Chan, and A. Grant, "Rate distortion with side-information at many decoders," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5240–5257, 2011.

[8] A. Sgarro, "Source coding with side information at several decoders," *IEEE Transactions on Information Theory*, vol. 23, no. 2, pp. 179–182, 1977.

[9] R. Timo, A. Grant, and G. Kramer, "Lossy broadcasting with complementary side information," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 104 – 131, 2013.

[10] S. Watanabe, "The rate-distortion function for product of two sources with side-information at decoders," in *proceedings IEEE International Symposium on Information Theory*, St. Petersburg, Russia, 2011.

[11] ——, "The rate-distortion function for product of two sources with side-information at decoders," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5678 – 5691, 2013.

[12] T. Liaich, "The Kaspi / Heegard-Berger rroblem with an informed encoder," Thesis, Swiss Federal Institute of Technology, Zurich, 2012.

[13] J. Körner and K. Marton, "Comparison of two noisy channels," in *Topics in Information Theory*, Keszthely, Hungry, 1977.

[14] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[15] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner-Ziv problem," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1636–1654, 2004.

[16] C. Tian and S. Diggavi, "On multistage successive refinement for Wyner-Ziv source coding with degraded side informations," *IEEE Transactions on Information Theory*, vol. 53, no. 8, pp. 2946–2960, 2007.

[17] C. Tian and S. N. Diggavi, "Side-information scalable source coding," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5591–5608, 2008.

[18] R. Timo, A. Grant, T. Chan, and G. Kramer, "Source coding for a simple network with receiver side information," in *IEEE International Symposium on Information Theory*, Toronto, Canada, 2008.

[19] B. N. Vellambi and R. Timo, "Successive refinement with common receiver reconstructions," in *IEEE International Symposium on Information Theory*, Honolulu, USA, 2014.

[20] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1469–1482, 2006.

[21] J. Nayak, E. Tuncel, and D. Gunduz, "Wyner-Ziv coding over broadcast channels: digital schemes," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1782–1799, 2010.

[22] Y. Gao and E. Tuncel, "Wyner-Ziv coding over broadcast channels: hybrid digital / analog schemes," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 5660–5672, 2010.

[23] A. Maor and N. Merhav, "On successive refinement with causal side information at the decoders," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 332–343, 2008.

[24] R. Timo and B. N. Vellambi, "Two lossy source coding problems with causal side-information," in *proceedings IEEE International Symposium on Information Theory*, Seoul, Korea, 2009.

[25] B. Ahmadi, R. Tandon, O. Simeone, and H. V. Poor, "On the Heegard-Berger problem with common reconstruction constraints," in *proceedings IEEE International Symposium on Information Theory*, MIT, Cambridge, MA, 2012.

[26] B. N. Vellambi and R. Timo, "The Heegard-Berger problem with common receiver reconstructions," in *IEEE Information Theory Workshop*, Seville, Spain, 2013.

[27] R. W. Yeung, *Information theory and network coding*. Springer, 2008.

[28] A. El Gamal and T. Cover, "Achievable rates for multiple descriptions," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 851–857, 1982.

[29] C. Nair, "Capacity regions of two new classes of two-receiver broadcast channels," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4207–4214, 2010.

[30] J. Villard and P. Piantanida, "Secure multiterminal source coding with side information at the eavesdropper," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3668 – 3692, 2013.

[31] G. Kramer, "Teaching IT: an identity for the Gelfand-Pinsker converse," *IEEE Information Theory Society Newsletter*, vol. 61, no. 4, pp. 4–6, 2012.

[32] ——, "Topics in multi-user information theory," *Foundations and Trends in Communications and Information Theory*, vol. 4, no. 45, pp. 265–444, 2008.