Successive Refinement with Conditionally Less Noisy Side Information

Roy Timo
Institute for Telecommunications Research
University of South Australia
roy.timo@unisa.edu.au

Tobias J. Oechtering
ACCESS Linnaeus Center
KTH Royal Institute of Technology
oech@kth.se

Michèle Wigger
Comm. and Electr. Department
Telecom ParisTech
michele.wigger@telecom-paristech.fr

***Abstract*—We consider the successive refinement of information problem with decoder side information. The rate-distortion region is unknown in general; Steinberg & Merhav and Tian & Diggavi solved it in the special case of *degraded* side information. We extend this special case to a new setup, *conditionally less noisy* side information, and we give a single-letter solution when one distortion function is deterministic.**

I. Background & Motivation

The *successive refinement of information* problem models scenarios in which data is compressed and decompressed in stages. For example, suppose that a video file is broadcast to many users with different fidelity requirements: Some users have large screens and require high fidelity, while others have small screens and require low fidelity. It is advantageous in such situations to first broadcast low fidelity video to all users, and later send refinements to those requiring high fidelity.

Let us briefly recall the basic problem setup for a discrete memoryless source. The source emits a string of n independent and identically distributed (i.i.d.) random variables $\boldsymbol{X} = X_1, X_2, \ldots, X_n$. An encoder maps $\boldsymbol{X}$ to an initial description (bit stream) at a relatively low rate R_1, from which a decoder reconstructs a coarse approximation $\hat{\boldsymbol{X}}_{\mathbf{1}}$ of $\boldsymbol{X}$. The encoder then generates a second description at rate R_2, and another decoder reconstructs a fine approximation $\hat{\boldsymbol{X}}_{\mathbf{2}}$ using both descriptions — the average distortion of $\hat{\boldsymbol{X}}_{\mathbf{2}}$ being less than that of $\hat{\boldsymbol{X}}_{\mathbf{1}}$.

An interesting information-theoretic problem is to determine those rates at which it is theoretically possible to compress $\boldsymbol{X}$. More formally, a rate-distortion (RD) tuple (R_1, R_2, D_1, D_2) is said to be *achievable* if there exists a compression scheme with refinement rates R_1 and R_2 such that the reconstructions $\hat{\boldsymbol{X}}_1$ and $\hat{\boldsymbol{X}}_2$ have average distortions D_1 and D_2 respectively. The *RD region* is the set of all achievable RD tuples, and the problem is to find a computable expression [1, p. 262] thereof.

Another interesting problem is to determine when successive compression incurs no rate penalty, relative to an optimal source code for each user operating in isolation. Specifically, a source is said to be *successively refinable* if for every pair of distortions (D_1, D_2) it is possible to reliably operate at rates corresponding to Shannon's RD function $R_X(\cdot)$ [2]; that is, $R_1 = R_X(D_1)$ and $R_1 + R_2 = R_X(D_2)$.

A plethora of work has been devoted to the aforementioned problems. For example, studies of the RD region date back to Gray and Wyner's seminal paper [3]. Koshelev [4], Equitz and Cover [2] established necessary and sufficient conditions for a source to be successively refinable. Lastras and Berger [5] showed that all memoryless sources with squared error distortion are nearly successively refinable. Successive refinement is a special case of multiple-descriptions coding [6], see El Gamal and Kim [7, Sec. 13.5] for a textbook treatment.

In this paper, we wish to determine the RD region when the decoders have side information (prior knowledge about the source). For example, in video compression $\boldsymbol{X}$ can be the current video frame and the side information can be the previous frames reconstructed by the decoders. Formally, we assume a Wyner and Ziv [8] setup for the side information: A discrete memoryless source emits an i.i.d. string $(\boldsymbol{X}, \boldsymbol{Y_1}, \boldsymbol{Y_2}) = (X_1, Y_{1,1}, Y_{2,1}), (X_2, Y_{1,2}, Y_{2,2}), \ldots, (X_n, Y_{1,n}, Y_{2,n})$. Here $\boldsymbol{X}$ is the source to be compressed, and $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ serve as side information about $\boldsymbol{X}$ at the first and second decoders respectively. Neither $\boldsymbol{Y_1}$ nor $\boldsymbol{Y_2}$ is known to the encoder.

The first such study was undertaken by Steinberg and Merhav [9] for *degraded* side information; that is, assuming $X -\!\!\circ\!\!- Y_2 -\!\!\circ\!\!- Y_1$ forms a Markov chain. Tian and Diggavi [10] extended the setup, approach and results of [9] from two stages to multiple stages. We extend [9] and [10] to the "conditionally less nosy" setup defined next. The definition is motivated by similar literature on degraded and less noisy broadcast channels [11].

Let L be an auxiliary random variable jointly distributed with (X, Y_1, Y_2). We say that Y_2 is *conditionally less noisy* than Y_1 given L, abbreviated as $(Y_2 \succeq Y_1 \mid L)$, if

$$I(W; Y_2|L) \geq I(W; Y_1|L) \tag{1}$$

holds for every auxiliary random variable W such that $W -\!\!\circ\!\!- (X, L) -\!\!\circ\!\!- (Y_1, Y_2)$ forms a Markov chain.

The above definition generalises that used in our preliminary work [13] on a lossless special case of successive refinement. The next proposition shows that degraded is a special case of conditionally less noisy. It can also be shown that the reverse implication is not true: conditionally less noisy does not imply degraded. Proofs of both assertions are given in [12].

R. Timo is supported by the Australian Research Council Discovery Grant DP120102123. M. Wigger is partly supported by the city of Paris under the programme "Emergences."

Proposition 1: If $L -\!\!\circ\!\!- X -\!\!\circ\!\!- (Y_1, Y_2)$ forms a Markov chain and the side information is degraded, then $(Y_2 \succeq Y_1 \mid L)$.

The main result of the paper (Theorem 2 in Section II) gives a single-letter characterisation of the RD region when one distortion function is deterministic and the side information is conditionally less noisy.

An important special case of successive refinement with side information is the *Kaspi / Heegard-Berger problem* [13]–[16]. Here, the encoder only sends one description about $\boldsymbol{X}$ to both decoders, and the problem is to determine the smallest rate at which given distortions can be achieved. Our results for successive-refinement automatically carry over to this setup.

Notation: All random variables are discrete and finite and written in uppercase, e.g., X. The alphabet of a random variable is written in matching calligraphic font, e.g. $\mathcal{X}$ is the alphabet of 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 $\mathcal{X}$.

II. Formal Problem Statement and Main Result

Let (X, Y_1, Y_2) have a joint distribution $P_{XY_1Y_2}$ on $\mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2$, and let $(\boldsymbol{X}, \boldsymbol{Y_1}, \boldsymbol{Y_2})$ be a string of n i.i.d. copies of (X, Y_1, Y_2).

A successive-refinement n-block code consists of three (possibly stochastic) maps, f, g_1, and g_2. The first map f defines an *encoder* and takes the form

$$f\colon \boldsymbol{\mathcal{X}} \longrightarrow \mathcal{M}_1 \times \mathcal{M}_2,$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are finite *message index* sets. The other two maps, g_1 and g_2, define *decoders* and take the form

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

where $\boldsymbol{\hat{\mathcal{X}}_1}$ and $\boldsymbol{\hat{\mathcal{X}}_2}$ are n-fold Cartesian products of reconstruction alphabets $\hat{\mathcal{X}}_1$ and $\hat{\mathcal{X}}_2$. The encoder computes $(M_1, M_2) := f(\boldsymbol{X})$. The first index M_1 is sent to both decoders, and the second index M_2 is sent only to Decoder 2. Decoders 1 and 2 reconstruct $\boldsymbol{\hat{X}_1} := g_1(M_1, \boldsymbol{Y_1})$ and $\boldsymbol{\hat{X}_2} := g_2(M_1, M_2, \boldsymbol{Y_2})$ respectively.

The two rates of a code (f, g_1, g_2) are defined by

$$\kappa_j := \frac{1}{n} \log_2 |\mathcal{M}_j|, \quad j = 1, 2,$$

and the two average distortions by

$$\Delta_j := \mathbb{E}\, \frac{1}{n} \sum_{i=1}^{n} \delta_j(X, \hat{X}_{j,i}), \quad j = 1, 2,$$

where $\delta_j : \mathcal{X} \times \hat{\mathcal{X}}_j \longrightarrow [0, \infty)$ is a bounded distortion function.

An RD tuple (R_1, R_2, D_1, D_2) is said to be *achievable* if there exists an n-block code (f, g_1, g_2) — for some sufficiently large blocklength n — satisfying $R_1 \geq \kappa_1$, $R_2 \geq \kappa_2$, $D_1 \geq \Delta_1$ and $D_2 \geq \Delta_2$. The closure of the set of all achievable RD tuples (R_1, R_2, D_1, D_2) is called the *RD region* $\mathcal{R}$.

If in the above problem description we remove the message index M_2, we obtain the Kaspi/Heegard-Berger problem [14, Sec. IIV], [15]. It can be easily shown that

$$R_{\text{HB}}(D_1, D_2) = \min\{R_1 \geq 0 : (R_1, 0, D_1, D_2) \in \mathcal{R}\} \quad (2)$$

is the RD function for the Kaspi/Heegard-Berger problem.

Our new results for $\mathcal{R}$ and $R(D_1, D_2)$ hold when the first reconstruction is 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.

The distortion function δ_1 is said to be a *deterministic* [6], [17] 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

$$\delta_1(x, \hat{x}) = \begin{cases} 0 & \text{if } \hat{x} = \psi(x) \\ 1 & \text{otherwise.} \end{cases}$$

We henceforth assume δ_1 is deterministic, we let

$$\tilde{X} := \psi(X), \quad (3)$$

and we assume δ_2 is arbitrary. Also, we denote the restriction of $\mathcal{R}$ to $D_1 = 0$ by

$$\mathcal{R}\big|_{D_1=0} := \big\{(R_1, R_2, 0, D_2) \in \mathcal{R}\big\}.$$

and the restriction of $R_{\text{HB}}(D_1, D_2)$ to $D_1 = 0$ by

$$R_{\text{HB}}(0, D_2) := \min\{R_1 \geq 0 : (R_1, 0, 0, D_2) \in \mathcal{R}\}. \quad (4)$$

Finally, let us define

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

where the minimisation is over all auxiliary A such that

- $A -\!\!\circ\!\!- X -\!\!\circ\!\!- Y_2$ forms a Markov chain;
- the cardinality of the alphabet of A satisfies $|\mathcal{A}| \leq |\mathcal{X}| + 1$;
- there exists a map $\phi_2 : \mathcal{A} \times \tilde{\mathcal{X}} \times \mathcal{Y}_2 \longrightarrow \hat{\mathcal{X}}_2$ such that $D_2 \geq \mathbb{E}\, \delta_2(X, \phi_2(A, \tilde{X}, Y_2))$.

The function $S(D_2)$ is non-increasing, convex and continuous in D_2, see [8, Thm. A2].

Theorem 2: If δ_1 is deterministic, $H(\tilde{X}|Y_1) \geq H(\tilde{X}|Y_2)$ and $(Y_2 \succeq Y_1 \mid \tilde{X})$, then $\mathcal{R}|_{D_1=0}$ is equal to the set of RD tuples $(R_1, R_2, 0, D_2)$ satisfying

$$R_1 \geq H(\tilde{X}|Y_1) \quad (5a)$$

$$R_1 + R_2 \geq H(\tilde{X}|Y_1) + S(D_2). \quad (5b)$$

Corollary 2.1: Under the conditions of Theorem 2, we have

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

Theorem 2 is proved in Section III-D ahead. In [12] we extend Theorem 2 to three receivers and additionally present results on Tian and Diggavi's "side-information scalable source coding" setup [17].

III. Proof of Theorem 2 and Comparison to Degraded Side Information

A. An Achievable RD Region

A single-letter achievable inner bound to the RD region $\mathcal{R}$ was derived in [16, Thm. 1]. From this result, we now distil a simpler achievable RD region for deterministic δ_1 and $D_1 = 0$.

Let $\mathcal{R}_{\text{in}}$ denote the set of all $(R_1, R_2, 0, D_2)$ satisfying

$$R_1 \geq H(\tilde{X}|Y_1) \quad (6a)$$

$$R_1 + R_2 \geq \max\big\{H(\tilde{X}|Y_1), H(\tilde{X}|Y_2)\big\} + S(D_2). \quad (6b)$$

Lemma 3 (Achievable): If δ_1 is deterministic, then

$$\mathcal{R}\big|_{D_1=0} \supseteq \mathcal{R}_{\text{in}}.$$

Lemma 3 holds for arbitrary side information. Its proof is based on the following scheme. The encoder first sends $\tilde{\boldsymbol{X}}$ to both decoders with rate close to $\max\{H(\tilde{X}|Y_1), H(\tilde{X}|Y_2)\}$; this is possible by, e.g., Sgarro [18]. The encoder then sends a lossy copy of $\boldsymbol{X}$ to decoder 2 at rate $S(D_2)$; this is possible by, e.g., Wyner and Ziv [8].

B. The RD Region for Degraded Side Information

The side information is said to be *degraded* if

$$X -\!\!\circ\!\!- Y_2 -\!\!\circ\!\!- Y_1 \tag{7}$$

forms a Markov chain. Steinberg and Merhav [9] and Tian and Diggavi [10] exploited (7) to derive single-letter expressions for the RD region. The result of [10] is summarised next for deterministic δ_1 and $D_1 = 0$.

Theorem 4 (Thm. 1, [10]): If the side information is degraded and δ_1 deterministic, then $\mathcal{R}|_{D_1=0} = \mathcal{R}_{\text{in}}$ and (6) simplifies to

$$\begin{aligned} R_1 &\geq H(\tilde{X}|Y_1) \\ R_1 + R_2 &\geq H(\tilde{X}|Y_1) + S(D_2). \end{aligned}$$

The chain (7) only plays a minor role in Tian and Diggavi's achievability proof for Theorem 4. Indeed, essentially the same random-coding argument can be used to prove Lemma 3. On the other hand, Tian and Diggavi's converse crucially relies on (7), see, for example, the equalities in [10, Eqn. (56)].

C. Beyond Degraded Side Information

We now broaden the scope of Theorem 4 to that of Theorem 2. The main effort required will be to derive a new converse without (7). The next lemma does precisely this, and its proof is the topic of Section IV.

Let $\mathcal{R}_{\text{out}}$ denote the set of all $(R_1, R_2, 0, D_2)$ satisfying

$$R_1 \geq H(\tilde{X}|Y_1) \tag{8a}$$

$$\begin{aligned} R_1 + R_2 \geq{}& H(\tilde{X}|Y_1) + S(D_2) \\ &+ \min_{W} \big\{ I(W;Y_2|\tilde{X}) - I(W;Y_1|\tilde{X}) \big\}, \end{aligned} \tag{8b}$$

where the minimisation is taken over an auxiliary random variable W such that $|\mathcal{W}| \leq |\mathcal{X}|$ and $W -\!\!\circ\!\!- X -\!\!\circ\!\!- (Y_1, Y_2)$.

Lemma 5 (Converse): If δ_1 is deterministic, then

$$\mathcal{R}\big|_{D_1=0} \subseteq \mathcal{R}_{\text{out}}.$$

The proof of Lemma 5 is given later in Section IV.

D. Proof of Theorem 2

By the assumption in the theorem, $H(\tilde{X}|Y_1) \geq H(\tilde{X}|Y_2)$, and thus by Lemma 3 the region in Theorem 2 is achievable. It remains to prove the converse. The outer bound in Lemma 5 and the region in Theorem 2 differ only in the sum rate constraints (5b) and (8b). But when $(Y_2 \succeq Y_1 \mid \tilde{X})$ then by (1)

$$\min_{W} \big\{ I(W;Y_2|\tilde{X}) - I(W;Y_1|\tilde{X}) \big\} = 0, \tag{9}$$

and (8b) coincides with (5b). This establishes the converse. ■

Theorem 4 is a special case of Theorem 2. Specifically, note that $X -\!\!\circ\!\!- Y_2 -\!\!\circ\!\!- Y_1$ implies $(Y_2 \succeq Y_1 \mid \tilde{X})$ and $H(\tilde{X}|Y_1) \geq H(\tilde{X}|Y_2)$ by Proposition 1 and the data processing lemma.

As a final remark, we note that our proof of 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 $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_2})$ as i.i.d. decoder side information.

IV. Proof of Lemma 5

We first state the solution to an entropy-characterisation problem: express the difference of two n-letter conditional mutual informations in a single-letter form. This problem and solution will play an important role in the proof of Lemma 5.

We will need the following notation: For a string of n random variables $\boldsymbol{A} = A_1, A_2, \ldots, A_n$, let $A_j^k := A_j, A_{j+1}, \ldots, A_k$ for $1 \leq j \leq k \leq n$.

A. An Entropy-Characterisation Problem

Consider a tuple of random variables (R, S_1, S_2, T, L) with an arbitrary joint distribution. Let $(\boldsymbol{R}, \boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}, \boldsymbol{L})$ denote a string of n i.i.d. copies of (R, S_1, S_2, T, L). Further, suppose that J is jointly distributed with the n-string $(\boldsymbol{R}, \boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}, \boldsymbol{L})$ and $J -\!\!\circ\!\!- (\boldsymbol{R}, \boldsymbol{L}) -\!\!\circ\!\!- (\boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T})$ forms a Markov chain. Consider the difference $I(J;\boldsymbol{S_2}|\boldsymbol{L}) - I(J;\boldsymbol{S_1}|\boldsymbol{L})$ of n-letter conditional mutual informations. We wish to know whether this difference can be expressed in a *single-letter* form. The next lemma answers this question in the affirmative; its proof is given in Appendix A.

Lemma 6: 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 $\mathcal{W}$ such that $|\mathcal{W}| \leq |\mathcal{R}||\mathcal{L}|$,

$$\begin{aligned} I(J;\boldsymbol{S_2}|\boldsymbol{L}) &- I(J;\boldsymbol{S_1}|\boldsymbol{L}) \\ &= n\big(I(W;S_2|L) - I(W;S_1|L)\big) \end{aligned} \tag{10}$$

and $W -\!\!\circ\!\!- (R, L) -\!\!\circ\!\!- (S_1, S_2, T)$ forms a Markov chain.

Remark 1: The proof of the lemma may remind the reader of steps applied in the converse to the capacity of less noisy broadcast channels. Notice however, that our Lemma 6 does not apply to channel coding problems where the input and output sequences can have memory and thus are no candidates for our n-tuples $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. In fact, when the n-tuples $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are not i.i.d., then in our lemma, the right-hand side of (10) has the extra term $\sum_{i=1}^{n} \big(I(S_{1i}, L_i; S_{1,1}^{i-1}, L_1^{i-1}) - I(S_{2i}L_i; S_{2,i+1}^{n}, L_{i+1}^{n}) \big)$, which can be positive or negative. So, also this extended form of our lemma does not directly lead to a converse for capacity problems.

B. Proof of Lemma 5

Let $(R_1, R_2, 0, D_2)$ lie in the RD region $\mathcal{R}|_{D_1=0}$. Then, for every $\epsilon > 0$ the tuple $(R_1 + \epsilon, R_2 + \epsilon, \epsilon, D_2 + \epsilon)$ is achievable i.e., for sufficiently large n we can find an n-block code (f, g_1, g_2) with $\kappa_1 \leq R_1 + \epsilon$, $\kappa_2 \leq R_2 + \epsilon$, $\Delta_1 \leq \epsilon$ and $\Delta_2 \leq D_2 + \epsilon$. For this n-block code, we have

$$\begin{aligned} R_1 + \epsilon &\geq \frac{1}{n} H(M_1) \\ &\geq \frac{1}{n} I(\tilde{\boldsymbol{X}}; M_1 | \boldsymbol{Y_1}) \\ &\overset{\text{(a)}}{\geq} \frac{1}{n}\big(H(\tilde{\boldsymbol{X}}|\boldsymbol{Y_1}) - n\varepsilon_1(n,\epsilon)\big) && (11) \\ &\overset{\text{(b)}}{=} H(\tilde{X}|Y_1) - \varepsilon_1(n,\epsilon), && (12) \end{aligned}$$

where (a) applies Fano's inequality and $\varepsilon_1(n, \epsilon)$ can be chosen so that $\varepsilon_1(n, \epsilon) \to 0$ as $\epsilon \to 0$ (for details see [12]) ; and (b) follows because the pair $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})$ is i.i.d. Moreover,

$$\begin{aligned} & R_1 + R_2 + \epsilon \\ &\quad \geq \frac{1}{n} H(M_1, M_2) \\ &\quad \geq \frac{1}{n} I(\tilde{\boldsymbol{X}}, \boldsymbol{X}; M_1, M_2 | \boldsymbol{Y_1}) \\ &\quad = \frac{1}{n}\Big(I(\tilde{\boldsymbol{X}}; M_1, M_2|\boldsymbol{Y_1}) + I(\boldsymbol{X}; M_1, M_2|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})\Big) \\ &\quad \overset{\text{(a)}}{\geq} \frac{1}{n}\Big(H(\tilde{\boldsymbol{X}}|\boldsymbol{Y_1}) - H(\tilde{\boldsymbol{X}}|\boldsymbol{Y_1}, M_1) \\ &\qquad\qquad + I(\boldsymbol{X}; M_1, M_2|\tilde{\boldsymbol{X}}, \boldsymbol{Y_1})\Big) \\ &\quad \overset{\text{(b)}}{=} \frac{1}{n}\Big(nH(\tilde{X}|Y_1) - n\varepsilon_1(n,\epsilon) + I(\boldsymbol{X}; M_1, M_2|\tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) \\ &\qquad\qquad + I(\boldsymbol{Y_2}; M_1, M_2|\tilde{\boldsymbol{X}}) - I(\boldsymbol{Y_1}; M_1, M_2|\tilde{\boldsymbol{X}})\Big) && (13) \end{aligned}$$

where (a) holds because conditioning does not increase entropy; (b) applies Fano's inequality as in (11), and it uses that $(\tilde{\boldsymbol{X}}, \boldsymbol{Y_1}, \boldsymbol{Y_2})$ is i.i.d. and $(M_1, M_2) -\!\!\circ\!\!- (\tilde{\boldsymbol{X}}, \boldsymbol{X}) -\!\!\circ\!\!- (\boldsymbol{Y_1}, \boldsymbol{Y_2})$.

Consider the first conditional mutual information on the right hand side of (13). We have

$$\begin{aligned} & \frac{1}{n} I(\boldsymbol{X}; M_1, M_2 | \tilde{\boldsymbol{X}}, \boldsymbol{Y_2}) \\ &\quad \overset{\text{(a)}}{\geq} \frac{1}{n}\sum_{i=1}^{n} I(X_i; M_1, M_2, Y_{2,1}^{i-1}, Y_{2,i+1}^{n} | \tilde{X}_i, Y_{2,i}) && (14) \\ &\quad \overset{\text{(b)}}{=} \frac{1}{n}\sum_{i=1}^{n} I(X_i; C_i | \tilde{X}_i, Y_{2,i}) && (15) \\ &\quad \overset{\text{(c)}}{\geq} \sum_{i=1}^{n} S\big(\mathbb{E}\delta_2(X_i, \hat{X}_{2,i})\big) && (16) \\ &\quad \overset{\text{(d)}}{\geq} S\left(\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\delta_2(X_i, \hat{X}_{2,i})\right) && (17) \\ &\quad \overset{\text{(e)}}{\geq} S(D_2 + \epsilon). && (18) \end{aligned}$$

The reasoning behind each step is as follows: (a) $(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{Y_2})$ is i.i.d.; (b) define $C_i := \big(M_1, M_2, Y_{2,1}^{i-1}, Y_{2,i+1}^{n}\big)$; (c) from the definition of $S(\cdot)$ and since $\hat{X}_{2,i}$ can be written as a function of $(C_i, Y_{2,i})$; and (d) and (e) by convexity and monotonicity of $S(\cdot)$. From (13) and (18), we obtain

$$\begin{aligned} & R_1 + R_2 + \epsilon \\ &\quad \geq H(\tilde{X}|Y_1) + S(D_2 + \epsilon) - \varepsilon_1(n, \epsilon) \\ &\qquad + \frac{1}{n}\Big(I(M_1, M_2; \boldsymbol{Y_2}|\tilde{\boldsymbol{X}}) - I(M_1, M_2; \boldsymbol{Y_1}|\tilde{\boldsymbol{X}})\Big). && (19) \end{aligned}$$

Apply Lemma 6 to (19) with $J = (M_1, M_2)$, $R = X$, $S_1 = Y_1$, $S_2 = Y_2$, and $L = \tilde{X}$: There exists an auxiliary W satisfying $|\mathcal{W}| \leq |\mathcal{X}|$ and $W -\!\!\circ\!\!- X -\!\!\circ\!\!- (Y_1, Y_2)$ such that

$$\begin{aligned} R_1 + R_2 + \epsilon \geq H(\tilde{X}|Y_1) &+ S(D_2 + \epsilon) - \varepsilon_1(n, \epsilon) \\ &+ I(W; Y_2|\tilde{X}) - I(W; Y_1|\tilde{X}) && (20) \end{aligned}$$

The proof follows now by (12) and (20), by taking the limit $\epsilon \to 0$ together with the continuity of $S(\cdot)$. ■

Appendix A
Proof of Lemma 6

The proof will make use of the following telescoping identity. For any string of arbitrarily distributed random variables, (A_1, B_1), (A_2, B_2), …, (A_n, B_n), we have [19, Sec. G]

$$\sum_{i=1}^{n} I(A_1^{i}; B_{i+1}^{n}) = \sum_{i=1}^{n} I(A_1^{i-1}; B_i^{n}), \tag{21}$$

with the notational convention $I(A_1^n; B_{n+1}^n) \triangleq 0$ and $I(A_1^{-1}; B_0^n) \triangleq 0$.

We first prove (10). Notice that

$$I(J; \boldsymbol{S}_2|\boldsymbol{L}) - I(J; \boldsymbol{S}_1|\boldsymbol{L}) = I(J; \boldsymbol{S}_2, \boldsymbol{L}) - I(J; \boldsymbol{S}_1, \boldsymbol{L}), \tag{22}$$

by the chain rule for mutual information. Expand the first mutual information term $I(J; \boldsymbol{S_2}, \boldsymbol{L})$ on the right hand side of (22) as follows:

$$\begin{aligned} I(J; \boldsymbol{S_2}, \boldsymbol{L}) &\overset{\text{(a)}}{=} \sum_{i=1}^{n} I(J; S_{2,i}, L_i | S_{2,1}^{i-1}, L_1^{i-1}) \\ &\overset{\text{(b)}}{=} \sum_{i=1}^{n} I(J, S_{2,1}^{i-1}, L_1^{i-1}; S_{2,i}, L_i) \\ &\overset{\text{(c)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{1,i+1}^{n}, S_{2,1}^{i-1}, L_1^{i-1}, L_{i+1}^{n}; S_{2,i}, L_i) \\ &\qquad - I(S_{1,i+1}^{n}, L_{i+1}^{n}; S_{2,i}, L_i | J, S_{2,1}^{i-1}, L_1^{i-1})\Big) \\ &\overset{\text{(d)}}{=} \sum_{i=1}^{n} \Big(I(W_i; S_{2,i}, L_i) \\ &\qquad - I(S_{1,i+1}^{n}, L_{i+1}^{n}; S_{2,i}, L_i | J, S_{2,1}^{i-1}, L_1^{i-1})\Big) \end{aligned} \tag{23}$$

where (a) and (c) follow from the chain rule for mutual information; (b) exploits the fact that the source is i.i.d. and therefore $H(S_{2,i}, L_i | S_{2,1}^{i-1}, L_1^{i-1}) = H(S_{2,i}, L_i)$; and, finally, in (d) we define and substitute the random variable

$$W_i \triangleq (J, S_{1,i+1}^{n}, S_{2,1}^{i-1}, L_1^{i-1}, L_{i+1}^{n}). \tag{24}$$

Expand the second mutual information term $I(J; \boldsymbol{S_1}, \boldsymbol{L})$ on the right hand side of (22) using the telescoping identity (21):

$$I(J; \boldsymbol{S_1}, \boldsymbol{L}) \overset{\text{(a)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{2,1}^{i-1}, L_1^{i-1}; S_{1,i}^{n}, L_i^{n})$$

$$
\begin{aligned}
&\qquad - I(J, S_{2,1}^{i}, L_1^{i}; S_{1,i+1}^{n}, L_{i+1}^{n})\Big)\\
&\overset{\text{(b)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{2,1}^{i-1}, L_1^{i-1}; S_{1,i}, L_i | S_{1,i+1}^{n}, L_{i+1}^{n})\\
&\qquad - I(S_{2,i}, L_i; S_{1,i+1}^{n}, L_{i+1}^{n} | J, S_{2,1}^{i-1}, L_1^{i-1})\Big)\\
&\overset{\text{(c)}}{=} \sum_{i=1}^{n} \Big(I(J, S_{1,i+1}^{n}, S_{2,1}^{i-1}, L_1^{i-1}, L_{i+1}^{n}; S_{1,i}, L_i)\\
&\qquad - I(S_{2,i}, L_i; S_{1,i+1}^{n}, L_{i+1}^{n} | J, S_{2,1}^{i-1}, L_1^{i-1})\Big)\\
&\overset{\text{(d)}}{=} \sum_{i=1}^{n} \Big(I(W_i; S_{1,i}, L_i)\\
&\qquad - I(S_{2,i}, L_i; S_{1,i+1}^{n}, L_{i+1}^{n} | J, S_{2,1}^{i-1}, L_1^{i-1})\Big),
\end{aligned}
\tag{25}
$$

where (a) invokes the telescoping identity (21) and the chain rule for mutual information; (b) again uses the chain rule; (c) exploits the i.i.d.-ness of the source; and in (d) we substitute for W_i. Subtract (25) from (23) to obtain

$$
\begin{aligned}
&I(J; \boldsymbol{S_2}, \boldsymbol{L}) - I(J; \boldsymbol{S_1}, \boldsymbol{L})\\
&\qquad = \sum_{i=1}^{n} I(W_i; S_{2,i}, L_i) - I(W_i; S_{1,i}, L_i).
\end{aligned}
\tag{26}
$$

We now *single-letterize* the quantity on the right hand side of (26). To this end, we 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 (26) by n, we have

$$
\begin{aligned}
&\frac{1}{n}\left(\sum_{i=1}^{n} I(W_i; S_{2,i}, L_i) - I(W_i; S_{1,i}, L_i) \right)\\
&\overset{\text{(a)}}{=} \frac{1}{n}\sum_{i=1}^{n} \Big(I(W_i; S_{2,i}, L_i | Q = i) - I(W_i; S_{1,i}, L_i | Q = i) \Big)\\
&\overset{\text{(b)}}{=} I(W_Q; S_{2,Q}, L_Q | Q) - I(W_Q; S_{1,Q}, L_Q | Q)\\
&\overset{\text{(c)}}{=} I(W_Q, Q; S_{2,Q}, L_Q) - I(W_Q, Q; S_{1,Q}, L_Q)\\
&\overset{\text{(d)}}{=} I(W; S_2, L) - I(W; S_1, L),
\end{aligned}
\tag{27}
$$

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, finally, in (d) we define and substitute

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

From (26) and (27), we have

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

We also notice that

$$
W_i \mathrel{-\!\!\circ\!\!-} (R_i, L_i) \mathrel{-\!\!\circ\!\!-} (S_{1,i}, S_{2,i}, T_i),
\tag{28}
$$

forms a Markov chain for all $i = 1, 2, \ldots, n$. Each of the n Markov chains in (28) follows from (24), the n-letter chain

$$
J \mathrel{-\!\!\circ\!\!-} (\boldsymbol{R}, \boldsymbol{L}) \mathrel{-\!\!\circ\!\!-} (\boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{T}),
$$

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

$$
W \mathrel{-\!\!\circ\!\!-} (R, L) \mathrel{-\!\!\circ\!\!-} (S_1, S_2, T).
\tag{29}
$$

The proof of the cardinality bound is omitted. ■

Some final remarks. A consequence of the identity (21) is the classic *Csiszár sum* identity [7, Sec. 2.4],

$$
\sum_{i=1}^{n} I(A_i; B_{i+1}^{n} | A_1^{i-1}) = \sum_{i=1}^{n} I(B_i; A_1^{i-1} | B_{i+1}^{n}).
\tag{30}
$$

The proof of Lemma 6 can be manipulated so as to replace the *telescoping sum identity* step (25) with a *Csiszár sum identity* step. We feel, however, that the present proof is cleaner.

References

[1] I. Csiszár and J. Körner, *Information Theory: coding theorems for discrete memoryless systems*, 2nd ed. Cambridge University Press, 2011.

[2] W. H. R. Equitz and T. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 269–275, 1991.

[3] R. Gray and A. Wyner, "Source coding for a simple network," *Bell Sys. Tech. Journal*, vol. 53, no. 9, pp. 1681–1721, 1974.

[4] V. N. Koshelev, "Hierarchical coding of discrete sources," *Problemy Peredachi Informatsii*, vol. 16, no. 3, pp. 31–49, 1980.

[5] L. Lastras and T. Berger, "All sources are nearly successively refinable," *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 918–926, 2001.

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

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

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

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

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

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

[12] R. Timo, T. J. Oechtering, and M. Wigger, "Source Coding Problems with Conditionally Less Noisy Side Information," submitted to *IEEE Trans. Inform. Theory*, Dec. 2012. Available at http://arxiv.org/pdf/1212.2396.

[13] R. Timo, T. J. Oechtering, and M. Wigger, "Source coding with conditionally less noisy side information," in *IEEE Information Theory Workshop*, Lausanne, Switzerland, 2012.

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

[15] A. H. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 2031–2034, 1994.

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

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

[18] A. Sgarro, "Source coding with side information at several decoders," *IEEE Trans. Inform. Theory*, vol. 23, no. 2, pp. 179–182, 1977.

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

[20] ——, "Topics in multi-user information theory," *Found. Trends Commun. and Inform. Theory*, vol. 4, no. 45, pp. 265–444, 2008.