# Distributed and Cascade Lossy Source Coding With a Side Information "Vending Machine"

Behzad Ahmadi, *Student Member, IEEE*, and Osvaldo Simeone, *Member, IEEE*

***Abstract*—Source coding with a side information "vending machine" is a recently proposed framework in which the statistical relationship between the side information and the source, instead of being given and fixed as in the classical Wyner–Ziv problem, can be controlled by the decoder. This control action is selected by the decoder based on the message encoded by the source node. Unlike conventional settings, the message can thus carry not only information about the source to be reproduced at the decoder, but also control information aimed at improving the quality of the side information. In this paper, the analysis of the tradeoffs between rate, distortion, and cost associated with the control actions is extended from the previously studied point-to-point setup to two basic multiterminal models. First, a distributed source coding model is studied, in which two encoders communicate over rate-limited links to a decoder, whose side information can be controlled. The control actions are selected by the decoder based on the messages encoded by both source nodes. For this setup, inner bounds are derived on the rate-distortion-cost region for both cases in which the side information is available causally and noncausally at the decoder. These bounds are shown to be tight under specific assumptions, including the scenario in which the sequence observed by one of the nodes is a function of the source observed by the other and the side information is available causally at the decoder. Then, a cascade scenario in which three nodes are connected in a cascade and the last node has controllable side information is also investigated. For this model, the rate-distortion-cost region is derived for general distortion requirements and under the assumption of causal availability of side information at the last node.**

***Index Terms*—Cascade source coding, distributed source coding, observation costs, rate-distortion theory, side information, side information vending machine.**

## I. Introduction

Permuter and Weissman [1] introduced the notion of a side information "vending machine." To illustrate the idea, consider the setting in Fig. 1, as studied in [1]. Here, unlike the conventional Wyner–Ziv setup (see, e.g., [2, Ch. 12]), the joint distribution of the side information $Y^n$ available at the decoder (Node 2) and of the source $X^n$ observed at the encoder (Node 1) is not given. Instead, it can be controlled through the selection of an "action" $A^n$, so that, for a given action $A$ and source symbol $X^n$, the side information $Y^n$ is distributed according to a given conditional distribution $p(y|a,x)$. Action $A^n$ is selected by the decoder based on the message $M$, of $R$ bits per source symbol, received from the encoder, and is subject to a cost constraint. The latter limits the "quality" of the side information that can be collected by the decoder.

The source coding problem with a vending machine provides a useful model for scenarios in which acquiring data as side information is costly and thus should be done effectively. Examples include computer networks, in which data must be obtained from remote data bases, and sensor networks, where data are acquired via measurements. The key aspect of this model is that the message $M$ produced by the encoder plays a double role. In fact, on the one hand, it needs to carry the description of the source $X^n$ itself, as in, e.g., the standard Wyner–Ziv model. On the other hand, it can also carry *control* information aimed at enabling the decoder to make an appropriate selection of action $A^n$. The goal of such a selection is to obtain a side information $Y^n$ that is better suited to provide partial information about the source $X^n$ to the decoder. This in turn can potentially reduce the rate $R$ necessary for the decoder to reconstruct source $X^n$ at a given distortion level (or, vice versa, to reduce the distortion level for a given rate $R$).

The performance of the system in Fig. 1 is expressed in terms of the interplay among three metrics, namely the rate $R$, the cost budget $\Gamma$ on the action $A$, and the distortion $D$ of the reconstruction $\hat{X}^n$ at the decoder. This tradeoff is summarized by the *rate-distortion-cost* function $R(D,\Gamma)$. This function characterizes the infimum of all rates $R$ for which a distortion level $D$ can be achieved under an action cost budget $\Gamma$, by allowing encoding of an arbitrary number $n$ of source symbols $X^n = (X_1,\ldots,X_n)$. This function is derived in [1] for both cases in which the side information $Y^n$ is available "noncausally" to the decoder, as in the standard Wyner–Ziv model, or "causally," as introduced in [3]. In the former case [see Fig. 1(a)], the estimated sequence $\hat{X}^n = (\hat{X}_1,\ldots,\hat{X}_n)$ is a function of message $M$ and of the entire side information sequence $Y^n = (Y_1,\ldots,Y_n)$, while, in the latter [see Fig. 1(b)], each estimated sample $\hat{X}_i$ is a function of message $M$ and the side information as received up to time $i$, i.e., $Y^i = (Y_1,\ldots,Y_i)$ for $i = 1,\ldots,n$. We note that the model with causal side information is appropriate, for instance, when there are delay constraints on the reproduction at the decoder or when the decoder operates by filtering the side information sequence. We refer to [3, Sec. I] for an extensive discussion on these points.

Following [1], recent works [4] and [5] generalized the characterization of the rate-distortion-cost function for the models

Manuscript received September 29, 2011; revised February 06, 2013; accepted April 25, 2013. Date of publication June 17, 2013; date of current version September 11, 2013. This work was supported by the U.S. National Science Foundation under Grant CCF-0914899. This paper was presented in part at the 2012 IEEE International Symposium on Information Theory.

The authors are with the Department of Electrical and Computer Engineering and CWCSPR, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: behzad.ahmadi@njit.edu; osvaldo.simeone@njit.edu).

Communicated by Y. Oohama, Associate Editor for Source Coding.
Digital Object Identifier 10.1109/TIT.2013.2269134

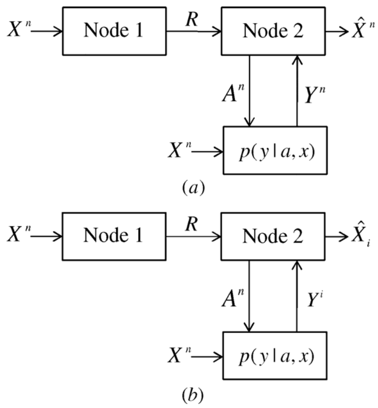


Fig. 1. Source coding with a vending machine at the decoder [1] with (a) "noncausal" side information and (b) "causal" side information.

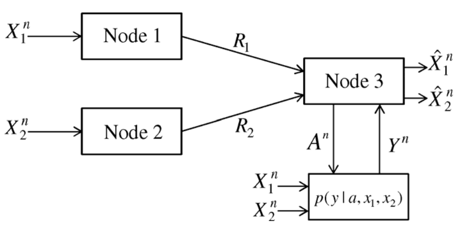


Fig. 2. Distributed source coding with a side information vending machine at the decoder.

in Fig. 1 to a setup analogous to the so-called Kaspi–Heegard–Berger problem [6], [7], in which the side information vending machine may or may not be available at the decoder. This entails the presence of *two decoders*, rather than only one as in Fig. 1, one with access to the vending machine and one without any side information. In [4] and [5], the authors also solved the more general case in which both decoders have access to the same vending machine, and either the side informations produced by the vending machine at the two decoders satisfy a degradedness condition, or lossless source reconstructions are required at the decoders. The papers [8], [9] studied the setting of Fig. 1 but under the additional constraints of common reconstruction, in the sense of [10], in [8], and of secrecy with respect to an "eavesdropping" node in [9], providing characterizations of the corresponding achievable performance. The impact of actions that adapt to the previously measured samples of the side information is studied in [11]. Finally, real-time constraints are investigated in [12].

### *A. Contributions and Overview*

In this paper, we study two multiterminal extensions of the setup in Fig. 1, namely the distributed source coding setting of Fig. 2, and the *cascade* model of Fig. 3. The analysis of these scenarios is motivated by the observation that they constitute key components of computer and sensor networks. In fact, as discussed above, an important aspect of these networks is the need to effectively acquire side information data, which can be modeled by including a side information vending machine. We overview the two extensions and the corresponding main results below.

*1) Distributed Source Coding With a Side Information Vending Machine (see Section II):* In the distributed source coding setting of Fig. 2, *two encoders* (Node 1 and Node 2), which measure correlated sources $X_1^n$ and $X_2^n$, respectively, communicate over rate-limited links, of rates $R_1$ and $R_2$, respectively, to a single decoder (Node 3). The decoder has side information $Y^n$ on sources $X_1^n$ and $X_2^n$, which can be controlled through an action $A^n$. The action sequence is selected by the decoder based on the messages $M_1$ and $M_2$ received from Nodes 1 and 2, respectively, and needs to satisfy a cost constraint of $\Gamma$. Inner bounds are derived to the rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ under noncausal and causal side information by combining the strategies proposed in [1] with the Berger–Tung strategy [13] and its extension to the Wyner–Ziv setup [14]. These bounds are shown to be tight under specific assumptions, including the scenario where the sequence observed by one of the nodes is a function of the source observed by the other and the side information is available causally at the decoder.

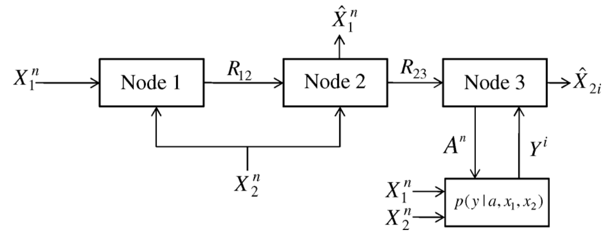


Fig. 3. Cascade source coding with a side information vending machine. Side information is assumed to be available "causally" to the decoder.

*2) Cascade Source Coding With a Side Information Vending Machine (see Section III):* In the cascade model of Fig. 3, Node 1 is connected via a rate-limited link, of rate $R_{12}$, to Node 2, which is in turn communicates with Node 3 with rate $R_{23}$. Source $X_1^n$ is measured by Node 1 and the correlated source $X_2^n$ by both Nodes 1 and 2. Similarly to the distributed coding setting described above, Node 3 has side information $Y^n$ on sources $X_1^n$ and $X_2^n$, which can be controlled via an action $A^n$. Action $A^n$ is selected by Node 3 based on the message received from Node 2 and needs to satisfy a cost constraint of $\Gamma$. We derive the set $\mathcal{R}(D_1, D_2, \Gamma)$ of all achievable rates $(R_{12}, R_{23})$ for given distortion constraints $(D_1, D_2)$ on the reconstructions $\hat{X}_1^n$ and $\hat{X}_2^n$ at Nodes 2 and 3, respectively, and for cost constraint $\Gamma$. This characterization is obtained under the assumption that the side information $Y^n$ is available causally at Node 3. It is mentioned that, following the submission of this paper, the analysis of the case with noncausal side information at Node 3 was carried out in [15].

*Notation:* For $a, b$ integer with $a \leq b$, we define $[a, b]$ as the interval $[a, a+1, \ldots, b]$ and $x_a^b = (x_a, \ldots, x_b)$; if instead $a > b$, we set $[a, b] = \emptyset$ and $x_a^b = \emptyset$. We will also write $x_1^b$ for $x^b$ for simplicity of notation. Random variables are denoted with capital letters and corresponding values with lowercase letters. Given random variables, or more generally vectors, $X$ and $Y$, we will use the notation $p_X(x)$ or $p(x)$ for $\Pr[X = x]$, and $p_{X|Y}(x|y)$ or $p(x|y)$ for $\Pr[X = x|Y = y]$, where the latter notations are used when the meaning is clear from the context. Given set $\mathcal{X}$, we define as $\mathcal{X}^n$ the $n$-fold Cartesian product of $\mathcal{X}$. Function $\delta(x)$ represents the Kronecker delta function, i.e., $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise.

## II. Distributed Source Coding With a Side Information Vending Machine

In this section, we first detail the system model for the problem of distributed source coding with a side information vending machine in Section II-A. Then, we propose an achievable strategy in Section II-B for both the cases with noncausal and causal side information at the decoder. In Sections II-C and II-D, scenarios are discussed in which the achievable strategies match given outer bounds. A numerical example is then developed in Section II-E.

### A. System Model

The problem of distributed lossy source coding with a vending machine and noncausal side information is illustrated in Fig. 2. It is defined by the probability mass functions (pmfs) $p_{X_1X_2}(x_1,x_2)$ and $p_{Y|AX_1X_2}(y|a,x_1,x_2)$ and discrete alphabets $\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y},\mathcal{A},\hat{\mathcal{X}}_1,\hat{\mathcal{X}}_2$ as follows. The source sequences $X_1^n$ and $X_2^n$ with $X_1^n\in\mathcal{X}_1^n$ and $X_2^n\in\mathcal{X}_2^n$, respectively, are such that the tuples $(X_{1i},X_{2i})$ for $i\in[1,n]$ are independent identically distributed (i.i.d.) with joint pmf $p_{X_1X_2}(x_1,x_2)$. Node 1 measures sequences $X_1^n$ and encodes it into message $M_1$ of $nR_1$ bits, while Node 2 measures sequences $X_2^n$ and encodes it into message $M_2$ of $nR_2$ bits. Node 3 wishes to reconstruct the two sources within given distortion requirements, to be discussed below, as $\hat{X}_1^n\in\hat{\mathcal{X}}_1^n$ and $\hat{X}_2^n\in\hat{\mathcal{X}}_2^n$.

To this end, Node 3 selects an action sequence $A^n$, where $A^n\in\mathcal{A}^n$, based on the messages $M_1$ and $M_2$ received from Nodes 1 and 2, respectively. The side information sequence $Y^n$ is then realized as the output of a memoryless channel with inputs $(A^n,X_1^n,X_2^n)$. Specifically, given $A^n$, $X_1^n$, and $X_2^n$, the sequence $Y^n$ is distributed as

$$p(y^n|a^n,x_1^n,x_2^n)=\prod_{i=1}^{n}p_{Y|AX_1X_2}(y_i|a_i,x_{1i},x_{2i}).\quad(1)$$

The overall cost of an action sequence $a^n$ is defined by a per-symbol cost function $\Lambda$: $\mathcal{A}\rightarrow[0,\Lambda_{\max}]$ with $0\leq\Lambda_{\max}<\infty$, as

$$\Lambda^n(a^n)=\frac{1}{n}\sum_{i=1}^{n}\Lambda(a_i).\quad(2)$$

The estimated sequences $\hat{X}_1^n$ and $\hat{X}_2^n$ are obtained as a function of both messages $M_1$ and $M_2$ and of the side information $Y^n$. The estimates $\hat{X}_1^n$ and $\hat{X}_2^n$ are constrained to satisfy distortion constraints defined by two per-symbol distortion measures, namely $d_j(x_1,x_2,y,\hat{x}_j)$: $\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{Y}\times\hat{\mathcal{X}}_j\rightarrow[0,D_{\max}]$ for $j=1,2$ with $0\leq D_{\max}<\infty$. Based on such scalar measures, the overall distortion for the estimated sequences $\hat{x}_1^n$ and $\hat{x}_2^n$ is defined as

$$d_j^n(x_1^n,x_2^n,y^n,\hat{x}_j^n)=\frac{1}{n}\sum_{i=1}^{n}d_j(x_{1i},x_{2i},y_i,\hat{x}_{ji}),\quad\text{for } j=1,2.\quad(3)$$

Note that, based on (3), the estimate $\hat{X}_j^n$ for $j=1,2$ can be required to be a lossy version of an arbitrary (per-letter) function of both sources $X_1^n$ and $X_2^n$ and of the side information sequence $Y^n$. A formal description of the operations at encoders and decoder, and of cost and distortion constraints, is presented below for both the cases in which the side information is available causally or noncausally at the decoder.

*Definition 1:* An $(n,R_1,R_2,D_1,D_2,\Gamma)$ code for the case of *noncasual* side information at Node 3 consists of two source encoders

$$\mathrm{g}_1\colon\mathcal{X}_1^n\rightarrow[1,2^{nR_1}],$$
$$\text{and } \mathrm{g}_2\colon\mathcal{X}_2^n\rightarrow[1,2^{nR_2}],\quad(4)$$

which map the sequences $X_1^n$ and $X_2^n$ into messages $M_1$ and $M_2$ at Node 1 and Node 2, respectively; an "action" function

$$\ell\colon[1,2^{nR_1}]\times[1,2^{nR_2}]\rightarrow\mathcal{A}^n,\quad(5)$$

which maps the message $(M_1,M_2)$ into an action sequence $A^n$ at Node 3; and two decoding functions

$$\mathrm{h}_1\colon[1,2^{nR_1}]\times[1,2^{nR_2}]\times\mathcal{Y}^n\rightarrow\hat{\mathcal{X}}_1^n,\quad(6)$$
$$\text{and } \mathrm{h}_2\colon[1,2^{nR_1}]\times[1,2^{nR_2}]\times\mathcal{Y}^n\rightarrow\hat{\mathcal{X}}_2^n,\quad(7)$$

which map the messages $M_1$ and $M_2$, and the side information sequence $Y^n$ into the estimated sequences $\hat{X}_1^n$ and $\hat{X}_2^n$ at Node 3, such that the action cost constraint $\Gamma$ is satisfied as

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\Lambda(A_i)\right]\leq\Gamma,\quad(8)$$

and the distortion constraints $D_1$ and $D_2$ hold, namely

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_j(X_{1i},X_{2i},Y_i,\hat{X}_{ji})\right]\leq D_j,\ \text{for } j=1,2.\quad(9)$$

*Definition 2:* An $(n,R_1,R_2,D_1,D_2,\Gamma)$ code for the case of *causal* side information at Node 3 is as in Definition 1 with the only difference that, in lieu of (6) and (7), we have the sequence of decoding functions

$$\mathrm{h}_{1i}\colon[1,2^{nR_1}]\times[1,2^{nR_2}]\times\mathcal{Y}^i\rightarrow\hat{\mathcal{X}}_{1i},\quad(10)$$
$$\text{and } \mathrm{h}_{2i}\colon[1,2^{nR_1}]\times[1,2^{nR_2}]\times\mathcal{Y}^i\rightarrow\hat{\mathcal{X}}_{2i},\quad(11)$$

for $i\in[1,n]$, which map the message $(M_1,M_2)$ and the measured sequence $Y^i$ into the $i$th estimated symbol $\hat{X}_{ji}=\mathrm{h}_{ji}(M_1,M_2,Y^i)$ for $j=1,2$ at Node 3.

*Definition 3:* Given a distortion-cost tuple $(D_1,D_2,\Gamma)$, a rate pair $(R_1,R_2)$ is said to be achievable for the case with noncausal or causal side information if, for any $\epsilon>0$ and sufficiently large $n$, there exists a corresponding $(n,R_1,R_2,D_1+\epsilon,D_2+\epsilon,\Gamma+\epsilon)$ code.

*Definition 4:* The *rate-distortion-cost region* $\mathcal{R}_{NC}(D_1,D_2,\Gamma)$ is defined as the closure of all rate pairs $(R_1,R_2)$ that are achievable with noncausal side information given the distortion-cost tuple $(D_1,D_2,\Gamma)$. The rate-distortion-cost region $\mathcal{R}_C(D_1,D_2,\Gamma)$ is similarly defined for the case of casual side information.

### B. Achievable Strategies

In this section, we obtain inner bounds to the rate-distortion-cost regions for the cases with noncausal and causal side information.

*Proposition 1:* The rate-distortion-cost region with noncausal side information at Node 3 satisfies the inclusion

$\mathcal{R}_{NC}(D_1, D_2, \Gamma) \supseteq \mathcal{R}^a_{NC}(D_1, D_2, \Gamma)$, where the region $\mathcal{R}^a_{NC}(D_1, D_2, \Gamma)$ is given by the union of the set of all of rate tuples $(R_1, R_2)$ that satisfy the inequalities

$$R_1 \geq I(X_1; V_1|V_2, Q) + I(X_1; U_1|V_1, V_2, U_2, Y, Q) \tag{12a}$$

$$R_2 \geq I(X_2; V_2|V_1, Q) + I(X_2; U_2|V_1, V_2, U_1, Y, Q) \tag{12b}$$

$$\text{and } R_1 + R_2 \geq I(X_1, X_2; V_1, V_2|Q) + I(X_1, X_2; U_1, U_2|V_1, V_2, Y, Q), \tag{12c}$$

for some joint pmfs that factorizes as

$$\begin{aligned} &p(q, x_1, x_2, y, v_1, v_2, u_1, u_2, a, \hat{x}_1, \hat{x}_2) \\ &= p(q)p(x_1, x_2)p(v_1, u_1|x_1, q)p(v_2, u_2|x_2, q)\delta(a - \mathrm{a}(v_1, v_2, q)) \\ &\quad p(y|a, x_1, x_2)\delta(\hat{x}_1 - \hat{\mathrm{x}}_1(u_1, u_2, y, q)) \\ &\quad \delta(\hat{x}_2 - \hat{\mathrm{x}}_2(u_1, u_2, y, q)), \end{aligned} \tag{13}$$

with pmfs $p(q)$ and $p(v_1, u_1|x_1, q)$ and $p(v_2, u_2|x_2, q)$ and deterministic functions $\mathrm{a}\colon \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{Q} \to \mathcal{A}$, $\hat{\mathrm{x}}_j\colon \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y} \times Q \to \hat{\mathcal{X}}_j$ for $j = 1, 2$, such that the action and the distortion constraints

$$\mathrm{E}\left[\Lambda(A)\right] \leq \Gamma \tag{14a}$$

$$\text{and} \mathrm{E}\left[d_j(X_1, X_2, Y, \hat{X}_j)\right] \leq D_j, \text{ for } j = 1, 2, \tag{14b}$$

hold. Finally, any extreme point of the region $\mathcal{R}^a_{NC}(D_1, D_2, \Gamma)$ can be obtained by limiting the cardinalities of the random variables $(V_1, V_2, U_1, U_2)$ as $|\mathcal{V}_j| \leq |\mathcal{X}_j| + 6$ and $|\mathcal{U}_j| \leq |\mathcal{X}_j|\,|\mathcal{V}_j| + 5$, for $j = 1, 2$.

*Remark 1:* If we set $p(y|a, x_1, x_2) = p(y|x_1, x_2)$, so that the side information is action-independent, Proposition 1 reduces to the extension of the Berger–Tung scheme [13] to the Wyner–Ziv setup studied in [14, Th. 2]. Moreover, in the special case in which there is only one encoder, the achievable rate coincides with that derived in [1, Th. 1].

The proof of Proposition 1 follows easily from standard arguments, and thus, it is only briefly discussed here. The proposed scheme combines the Berger–Tung distributed source coding strategy [13] and the distributed Wyner–Ziv approach proposed in [14, Th. II] with the layered two-stage coding scheme that is proved to be optimal in [1] for the special case of a single encoder. Throughout the discussion, we neglect the time-sharing variable $Q$ for simplicity. This can be handled in the standard way (see, e.g., [2, Sec. 4.5.3]). The encoding scheme at Nodes 1 and 2 multiplexes two descriptions, which are obtained in two encoding stages. In the first encoding stage, the distributed source coding strategy of [13], conventionally referred to as the Berger–Tung scheme, is adopted by Nodes 1 and 2 to convey descriptions $V_1^n$ and $V_2^n$, respectively, to Node 3. In order for the decoder to be able to recover these descriptions, the rates $R'_1$ and $R'_2$ allocated by Nodes 1 and 2 have to satisfy the conditions [2], [13, Ch. 13]

$$R'_1 \geq I(X_1; V_1|V_2) \tag{15a}$$

$$R'_2 \geq I(X_2; V_2|V_1) \tag{15b}$$

$$\text{and } R'_1 + R'_2 \geq I(X_1, X_2; V_1, V_2). \tag{15c}$$

Having decoded the descriptions $(V_1^n, V_2^n)$, Node 3 selects the action sequence $A^n$ as the per-symbol function $A_i = \mathrm{a}(V_{1i}, V_{2i})$ for $i \in [1, n]$. Node 3 thus measures the side information sequence $Y^n$. The sequences $(Y^n, V_1^n, V_2^n)$ can then be regarded as side information available at the decoder. Therefore, in the second encoding stage, the distributed Wyner–Ziv scheme proposed in [14, Th. 2] is used to convey the descriptions $U_1^n$ and $U_2^n$ by Nodes 1 and 2, respectively, to Node 3. Note that the fact that sequences $(Y^n, V_1^n, V_2^n)$ are not i.i.d. does not affect the achievability of the rate region derived in [14]. This is because, as shown in [2, Lemma 3.1], the packing lemma leveraged to ensure the correctness of the decoding process applies for an arbitrary distribution of the sequences $(Y^n, V_1^n, V_2^n)$. In order for the decoder to correctly retrieve the descriptions $U_1^n$ and $U_2^n$, the rates $R''_1$ and $R''_2$ allocated by Nodes 1 and 2 must satisfy the inequalities [14]

$$R''_1 \geq I(X_1; U_1|V_1, V_2, U_2, Y) \tag{16a}$$

$$R''_2 \geq I(X_2; U_2|V_1, V_2, U_1, Y) \tag{16b}$$

$$\text{and } R''_1 + R''_2 \geq I(X_1, X_2; U_1, U_2|V_1, V_2, Y). \tag{16c}$$

Nodes 1 and 2 multiplex the source indices obtained in the two phases, and hence, the overall rates are $R_1 = R'_1 + R''_1$ and $R_2 = R'_2 + R''_2$. Using these equalities, along with (15) and (16), leads to (12). Finally, the decoder $j$ estimates $\hat{X}_j^n$ with $j = 1, 2$ sample by sample as a function of $U_{1i}, U_{2i}$ and $Y_i$. The proof of the cardinality bounds follows from standard arguments and is sketched in Appendix A.[1] We now turn to a similar achievable strategy for the case with causal side information.

*Proposition 2:* The rate-distortion-cost region with causal side information at Node 3 satisfies the inclusion $\mathcal{R}_C(D_1, D_2, \Gamma) \supseteq \mathcal{R}^a_C(D_1, D_2, \Gamma)$, where the region $\mathcal{R}^a_C(D_1, D_2, \Gamma)$ is given by the union of the set of all of rate tuples $(R_1, R_2)$ that satisfy the inequalities

$$R_1 \geq I(X_1; U_1|U_2, Q) \tag{17a}$$

$$R_2 \geq I(X_2; U_2|U_1, Q) \tag{17b}$$

$$\text{and } R_1 + R_2 \geq I(X_1, X_2; U_1, U_2|Q), \tag{17c}$$

for some joint pmfs that factorize as

$$\begin{aligned} &p(q, x_1, x_2, y, u_1, u_2, a, \hat{x}_1, \hat{x}_2) \\ &= p(q)p(x_1, x_2)p(u_1|x_1, q)p(u_2|x_2, q)\delta(a - \mathrm{a}(u_1, u_2, q)) \\ &\quad p(y|a, x_1, x_2)\delta(\hat{x}_1 - \hat{\mathrm{x}}_1(u_1, u_2, y, q)) \\ &\quad \delta(\hat{x}_2 - \hat{\mathrm{x}}_2(u_1, u_2, y, q)), \end{aligned} \tag{18}$$

with pmfs $p(q)$, $p(u_1|x_1, q)$ and $p(u_2|x_2, q)$ and deterministic functions $\mathrm{a}\colon \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Q} \to \mathcal{A}$ and $\hat{\mathrm{x}}_j\colon \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y} \times Q \to \hat{\mathcal{X}}_j$ for $j = 1, 2$, such that the action and the distortion constraints (14a) and (14b) hold, respectively. Finally, any extreme point in the region $\mathcal{R}^a_C(D_1, D_2, \Gamma)$ can be obtained by constraining the cardinalities of random variables $(U_1, U_2)$ as $|\mathcal{U}_1| \leq |\mathcal{X}_1| + 5$ and $|\mathcal{U}_2| \leq |\mathcal{X}_2| + 5$.

[1] It is noted that, using the approach of [16], it may be possible to improve the cardinality bounds. This aspect is not further explored here.

The proof follows by similar arguments as the ones in the proof of Proposition 1 with the only difference that only one stage of encoding is sufficient. Specifically, as in Proposition 1, Berger–Tung coding is adopted to convey the descriptions $U_1^n$ and $U_2^n$ to Node 3. Note that, with causal side information, there is no advantage in having a second encoding stage, since the side information sequence cannot be leveraged for binning in contrast to the case with noncausal side information [2], [3, Ch. 12]. The cardinality bounds follow from arguments similar to Appendix A.

### C. Degraded Source Sets and Causal Side Information

In this section, we consider the special case in which the sequence observed by Node 2 is a symbol-by-symbol function of the source observed at Node 1 [17, Sec. V.] (see also [18]). In other words, we can write $X_{1i} = (X'_{1i}, X_{2i})$ for $i \in [1, n]$, where $X_1'^n$ is an i.i.d. sequence independent of $X_2^n$. We refer to this setup as having degraded source sets. Moreover, we assume that the side information $Y$ is available causally at Node 3. The next proposition proves that the achievable strategy of Proposition 2 is optimal in this case.

*Proposition 3:* The rate-distortion-cost region $\mathcal{R}_C(D_1, D_2, \Gamma)$ for the setup with degraded source sets and with causal side information at Node 3 satisfies $\mathcal{R}_C(D_1, D_2, \Gamma) = \mathcal{R}_C^a(D_1, D_2, \Gamma)$.

For the proof of converse, we refer the reader to Appendix B.

*Remark 2:* Proposition 3 generalizes to the case with action-dependent side information the result in [17, Sec. V] for the case with no side information.

### D. One-Distortion Criterion and Noncausal Side Information

In this section, we consider a variation on the setup of source coding with action-dependent noncausal side information described in Definition 1. Specifically, Node 3 selects the action sequence $A^n$ based only on the message $M_1$ received from Node 1. In other words, the action function (5) is modified to

$$\ell \colon [1, 2^{nR_1}] \to \mathcal{A}^n, \tag{19}$$

which maps the message $M_1$ into an action sequence $A^n$ at Node 3. This may be the case in scenarios in which there is a hierarchy between Nodes 1 and 2, e.g., in a sensor network, and the functionality of remote control of the side information is assigned solely to Node 1. The next proposition characterizes the rate-distortion-cost function $\mathcal{R}_{NC}(D_1, 0, \Gamma)$ under the mentioned assumption when Hamming distortion is selected for $\hat{X}_2$. That is, we choose the distortion measure $d_2(x_2, \hat{x}_2)$ as $d_H(x_2, \hat{x}_2) = 0$ if $x_2 = \hat{x}_2$ and $d_H(x_2, \hat{x}_2) = 1$ otherwise. This implies that we impose the constraint of vanishingly small per-symbol Hamming distortion between source $X_2^n$ and estimate $\hat{X}_2^n$, or equivalently the constraint $\frac{1}{n}\sum_{i=1}^{n}\Pr[\hat{X}_{2i} \neq X_{2i}] \to 0$ for $n \to \infty$. We will refer to this assumption by saying that source sequence $X_2^n$ must be recovered losslessly at the decoder.

*Proposition 4:* If the action function is given by (19) and $X_2^n$ must be recovered losslessly at Node 3, the rate-distortion-cost region $\mathcal{R}_{NC}(D_1, 0, \Gamma)$ is given by union of the set of all of rate tuples $(R_1, R_2)$ that satisfy the inequalities

$$R_1 \geq I(X_1; A|Q) + I(X_1; U_1|A, X_2, Y, Q) \tag{20a}$$

$$R_2 \geq H(X_2|A, Y, U_1, Q) \tag{20b}$$

$$\text{and } R_1 + R_2 \geq I(X_1; A|Q) + H(X_2|A, Y, Q) + I(X_1; U_1|A, X_2, Y, Q), \tag{20c}$$

for some joint pmfs that factorizes as

$$p(q, x_1, x_2, y, u_1, a, \hat{x}_1) = p(q)p(x_1,x_2)p(a,u_1|x_1,q)p(y|a,x_1,x_2)\delta(\hat{x}_1 - \hat{\mathrm{x}}_1(u_1,x_2,y,q)), \tag{21}$$

with pmfs $p(q)$ and $p(a, u_1|x_1, q)$ and deterministic function $\hat{\mathrm{x}}_1(u_1, x_2, y, q)$, such that the action and the distortion constraints

$$\mathrm{E}\left[\Lambda(A)\right] \leq \Gamma \tag{22a}$$

$$\text{and } \mathrm{E}\left[d_1(X_1, X_2, Y, \hat{X}_1)\right] \leq D_1 \tag{22b}$$

hold. Finally, $Q$ and $U_1$ are auxiliary random variables whose alphabet cardinality can be constrained as $|Q| \leq 6$ and $|\mathcal{U}_1| \leq 6\,|\mathcal{X}_1|\,|\mathcal{A}| + 3$ without loss of optimality.

*Remark 3:* In the case in which there is no side information, Proposition 4 reduces to [19, Th. 1].

For the proof of converse, we refer the reader to Appendix C. The achievability follows from Proposition 1 by setting $V_2 = \emptyset$, $V_1 = A$, and $U_2 = X_2$.

*Remark 4:* Extension of the result in proposition to an arbitrary number $K$ of encoders can be found in [20].

### E. Binary Example

We now focus on a specific numerical example in order to illustrate the result derived in Propositions 1 and 4 and the advantage of selecting actions at Node 3 based on the message received from one of the nodes. Specifically, we assume that all alphabets are binary and that $(X_1, X_2)$ is a doubly symmetric binary source (DSBS) characterized by probability $p$, with $0 \leq p \leq 1/2$, so that $p(x_1) = p(x_2) = 1/2$ for $x_1, x_2 \in \{0, 1\}$ and $\Pr[X_1 \neq X_2] = p$. Moreover, we adopt Hamming distortion for both sources to reconstruct both $X_1$ and $X_2$ losslessly in the sense discussed above. Note that this implies that we set $d_1(x_1, x_2, y, \hat{x}_1) = d_H(x_1, \hat{x}_1)$ and $D_1 = 0$. The side information $Y_i$ is such that

$$Y_i = \begin{cases} \mathrm{f}(X_{1i}, X_{2i}), & \text{if } A_i = 1 \\ 1, & \text{if } A_i = 0, \end{cases} \tag{23}$$

where $\mathrm{f}(x_1, x_2)$ is a deterministic function to be specified. Therefore, when action $A_i = 1$ is selected, then $Y_i = \mathrm{f}(X_{1i}, X_{2i})$ is measured at the receiver, while with $A_i = 0$ no useful information is collected by the decoder. The action sequence $A^n$ must satisfy the cost constraint (8), where the cost function is defined as $\Lambda(A_i) = 1$ if $A_i = 1$ and $\Lambda(A_i) = 0$ if $A_i = 0$. It follows that, given (23), a cost $\Gamma$ implies that the decoder can observe $\mathrm{f}(X_{1i}, X_{2i})$ only for at

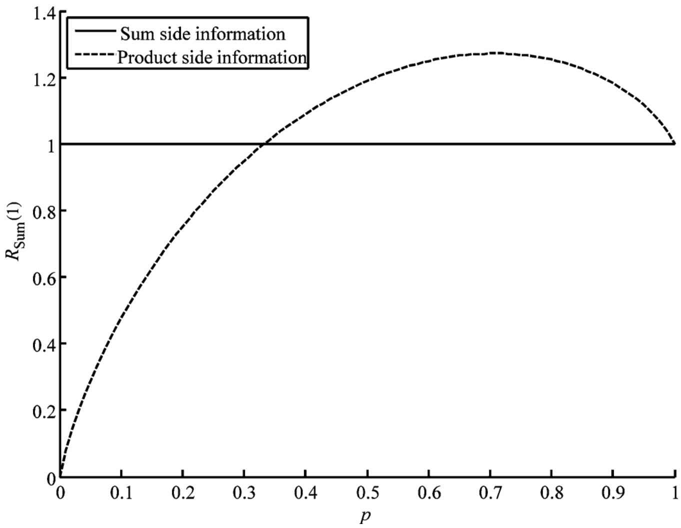


Fig. 4. Sum-rates versus $p$ for sum and product side informations ($\Gamma = 1$).

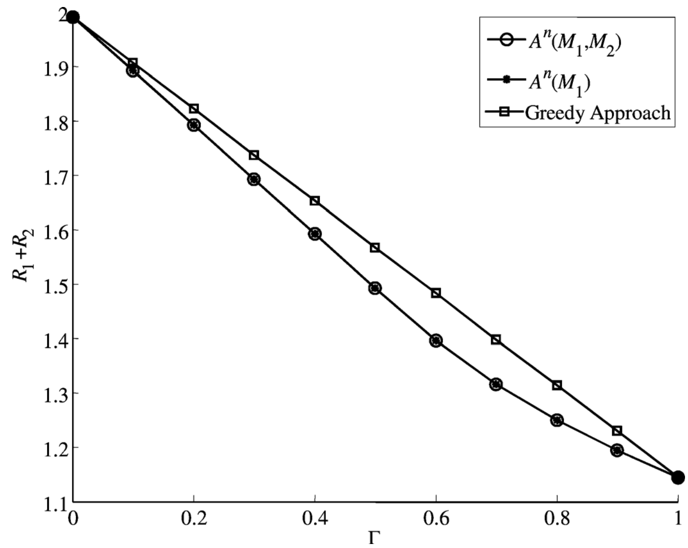


Fig. 5. Sum-rates versus the action cost $\Gamma$ for product side information ($p = 0.45$).

most $n\Gamma$ symbols. As for the function $\mathrm{f}(x_1, x_2)$, we consider two cases, namely $\mathrm{f}(x_1, x_2) = x_1 \oplus x_2$, where $\oplus$ is the binary sum and $\mathrm{f}(x_1, x_2) = x_1 \odot x_2$, where $\odot$ is the binary product. We assume that the side information is available noncausally at the decoder.

To start with, observe that the sum-rate is a nonincreasing function of the action cost $\Gamma$, and hence, the minimum sum-rate is obtained when $\Gamma = 1$. With $\Gamma = 1$, it is clearly optimal to set $A = 1$, irrespective of the value of $X_1$. In this case, from the Slepian–Wolf theorem, the sum rate equals $R_{\mathrm{sum}}(1) = H(X_1, X_2|Y)$. Specifically, with sum side information, we get

$$R^{\oplus}_{\mathrm{sum}}(1) = 1, \tag{24}$$

since we have $R^{\oplus}_{\mathrm{sum}}(1) = H(X_1, X_2|X_1 \oplus X_2) = H(X_1|X_1 \oplus X_2) = H(X_1)$, where the second equality follows from the chain rule and the third from the crypto-lemma [21, Lemma 2]. Instead, with product side information, we obtain

$$R^{\odot}_{\mathrm{sum}}(1) = H\left(\frac{1-p}{1+p}, \frac{p}{1+p}, \frac{p}{1+p}\right)\left(\frac{1+p}{2}\right), \tag{25}$$

where we have used the definition $H(p_1, p_2, \ldots, p_k) = -\sum_{i=1}^{k} p_k \log_2 p_k$. Equation (25) follows since

$$\begin{aligned} R^{\odot}_{\mathrm{sum}}(1) &= H(X_1, X_2|X_1 \odot X_2) \\ &= H(X_1, X_2|X_1 \odot X_2 = 0)\Pr[X_1 \odot X_2 = 0], \end{aligned} \tag{26}$$

where the second equality is a consequence of the fact that $X_1 \odot X_2 = 1$ implies that $X_1 = 1$ and $X_2 = 1$. Sum-rate (25) is then obtained by evaluating (26) for the DSBS at hand. Fig. 4 shows the sum-rates (24) and (25), demonstrating that, if $p$ is sufficiently small, namely if $p \lesssim 0.33$, we have $R^{\odot}_{\mathrm{sum}}(1) < R^{\oplus}_{\mathrm{sum}}(1)$, and thus, product side information is more informative than the sum, while for $p \gtrsim 0.33$, the opposite is true (and for $p = 1$, they are equally informative).

Considering a general cost budget $0 \le \Gamma \le 1$, in order to emphasize the role of both data and control information for the system performance, we now evaluate the sum-rate attainable by imposing that the action $A$ be selected by Node 3 *a priori*, that is, without any control from Node 1. This can be easily seen to be given by [1]

$$\begin{aligned} R_{\mathrm{sum,\ greedy}}(\Gamma) &= \Gamma H(X_1, X_2|Y) + (1-\Gamma)H(X_1, X_2) \\ &= \Gamma H(X_1, X_2|Y) + (1-\Gamma)(1 + H(p)). \end{aligned} \tag{27}$$

This sum-rate will be compared below with the performance of the scheme in Proposition 1, in which the actions are selected based on both messages $(M_1, M_2)$, and that of Proposition 4, in which the actions are selected based only on message $M_1$.

Fig. 5 depicts the mentioned sum-rates[2] versus the action cost $\Gamma$ for $p = 0.45$ and product side information. It can be seen that the greedy approach suffers from a significant performance loss with respect to the approaches in which actions are selected based on the messages received from one encoder or both encoders. It can also be observed that no gains are obtained by selecting the actions based on both messages. The fact that choosing the action based on the message received from Node 1 provides performance benefits can be explained as follows. If $X_1 = 0$, the value of the side information is always $Y = X_1 \odot X_2 = 0$ irrespective of the value of $X_2$. Therefore, if $X_1 = 0$, the side information is less informative than if $X_1 = 1$, and hence, it may be advantageous to save on the action cost by setting $A = 0$. Consequently, choosing actions based on the message received from Node 1 can result in a lower sum-rate.

The scenario with sum side information is considered in Fig. 6 for $p = 0.1$. A first observation is that, as proved in Appendix D, choosing the action based only on $M_1$ cannot improve the sum-rate with respect to the greedy case. This contrasts with the product side information case, and is due to the fact that $X_1$ is independent of the side information $Y$. Instead, choosing the

[2]The sum-rate from Proposition 1 is calculated by assuming binary auxiliary variables $V_1$ and $V_2$ and performing global optimization.

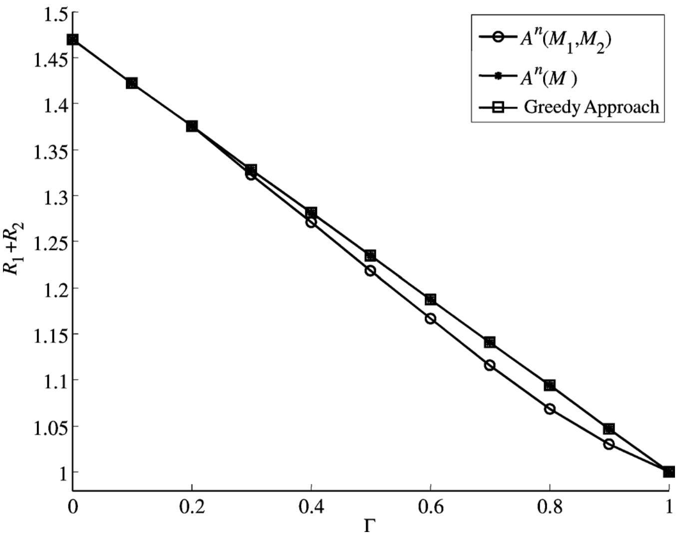


Fig. 6. Sum-rates versus the action cost $\Gamma$ for sum side information ($p = 0.1$).

actions based on both messages allows us to save on the necessary communication sum-rate.

## III. Cascade Source Coding With a Side Information Vending Machine

In this section, we first describe the system model for the setting of Fig. 3 of cascade source coding with a side information vending machine. We recall that side information $Y$ is here assumed to be available causally at the decoder (Node 3). The corresponding model with noncausal side information is studied in [15]. We then present the characterization of the corresponding rate-distortion-cost performance in Section III-B.

### A. System Model

The problem of cascade lossy computing with causal observation costs at second user, illustrated in Fig. 3, is defined by the pmfs $p_{X_1X_2}(x_1,x_2)$ and $p_{Y|AX_1X_2}(y|a,x_1,x_2)$ and discrete alphabets $\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y},\mathcal{A},\hat{\mathcal{X}}_1,\hat{\mathcal{X}}_2$, as follows. The source sequences $X_1^n$ and $X_2^n$ with $X_1^n \in \mathcal{X}_1^n$ and $X_2^n \in \mathcal{X}_2^n$, respectively, are such that the pairs $(X_{1i},X_{2i})$ for $i \in [1,n]$ are i.i.d. with joint pmf $p_{X_1X_2}(x_1,x_2)$. Node 1 measures sequences $X_1^n$ and $X_2^n$ and encodes them in a message $M_{12}$ of $nR_{12}$ bits, which is delivered to Node 2. Node 2 estimates a sequence $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ within given distortion requirements to be discussed below. Moreover, Node 2 encodes the message $M_{12}$, received from Node 1, and the locally available sequence $X_2^n$ in a message $M_{23}$ of $nR_{23}$ bits, which is delivered to node 3. Node 3 wishes to estimate a sequence $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ within given distortion requirements to be discussed. To this end, Node 3 receives message $M_{23}$ and, based on this, selects an action sequence $A^n$, where $A^n \in \mathcal{A}^n$. The action sequence affects the quality of the measurement $Y^n$ of sequence $X_1^n$ and $X_2^n$ obtained at the Node 3. Specifically, given $A^n$, $X_1^n$, and $X_2^n$, the sequence $Y^n$ is distributed as in (1). The cost of the action sequence is defined by a cost function $\Lambda\colon \mathcal{A} \rightarrow [0,\Lambda_{\max}]$ with $0 \leq \Lambda_{\max} < \infty$, as in (2). The estimated sequence $\hat{X}_2^n$ with $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ is then obtained as a function of $M_{23}$ and $Y^n$.

Estimated sequences $\hat{X}_j^n$ for $j=1,2$ must satisfy distortion constraints defined by functions $d_j(x_1,x_2,y,\hat{x}_j)\colon \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y} \times \hat{\mathcal{X}}_j \rightarrow [0,D_{\max}]$ with $0 \leq D_{\max} < \infty$ for $j=1,2$, respectively. A formal description of the operations at encoder and decoder follows.

*Definition 5:* An $(n,R_{12},R_{23},D_1,D_2,\Gamma)$ code for the setup of Fig. 3 consists of two source encoders, namely

$$\mathrm{g}_1\colon \mathcal{X}_1^n \times \mathcal{X}_2^n \rightarrow [1,2^{nR_{12}}], \tag{28}$$

which maps the sequences $X_1^n$ and $X_2^n$ into a message $M_{12}$;

$$\mathrm{g}_2\colon \mathcal{X}_2^n \times [1,2^{nR_{12}}] \rightarrow [1,2^{nR_{23}}] \tag{29}$$

which maps the sequence $X_2^n$ and message $M_{12}$ into a message $M_{23}$; an "action" function

$$\ell\colon [1,2^{nR_{23}}] \rightarrow \mathcal{A}^n, \tag{30}$$

which maps the message $M_{23}$ into an action sequence $A^n$; a decoding function

$$\mathrm{h}_1\colon [1,2^{nR_{12}}] \times \mathcal{X}_2^n \rightarrow \hat{\mathcal{X}}_1^n, \tag{31}$$

which maps the message $M_{12}$ and the measured sequence $X_2^n$ into the estimated sequence $\hat{X}_1^n$; and a sequence of decoding functions

$$\mathrm{h}_{2i}\colon [1,2^{nR_{23}}] \times \mathcal{Y}^i \rightarrow \hat{\mathcal{X}}_2, \tag{32}$$

for $i \in [1,n]$ which maps the message $M_{23}$ and the measured sequence $Y^i$ into the $i$th estimated symbol $\hat{X}_{2i} = \mathrm{h}_{2i}(M_{23},Y^i)$, such that the action cost constraint $\Gamma$ and distortion constraints $D_j$ for $j=1,2$ are satisfied, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] \leq \Gamma \tag{33}$$

$$\text{and } \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_j(X_{1i},X_{2i},Y_i,\hat{X}_{ji})\right] \leq D_j \text{ for } j=1,2, \tag{34}$$

respectively.

*Definition 6:* Given a distortion-cost tuple $(D_1,D_2,\Gamma)$, a rate tuple $(R_{12},R_{23})$ is said to be achievable if, for any $\epsilon > 0$, and sufficiently large $n$, there exists a $(n,R_{12},R_{23},D_1+\epsilon,D_2+\epsilon,\Gamma+\epsilon)$ code.

*Definition 7:* The *rate-distortion-cost region* $\mathcal{R}(D_1,D_2,\Gamma)$ is defined as the closure of all rate tuples $(R_{12},R_{23})$ that are achievable given the distortion-cost tuple $(D_1,D_2,\Gamma)$.

*Remark 5:* For side information $Y$ independent of the action $A$ given $X_1$ and $X_2$, i.e., for $p(y|a,x_1,x_2) = p(y|x_1,x_2)$, the rate-distortion region $\mathcal{R}(D_1,D_2,\Gamma)$ has been derived in [22].

### B. Rate-Distortion-Cost Region

We have the following characterization of the rate-distortion-cost region.

*Proposition 5:* The rate-distortion-cost region $\mathcal{R}(D_1,D_2,\Gamma)$ for the setup of Fig. 3 is given by the union of all rate pairs $(R_{12},R_{23})$ satisfying the inequalities

$$R_{12} \geq I(X_1;U,A,\hat{X}_1|X_2) \tag{35a}$$

$$\text{and } R_{23} \geq I(X_1,X_2;U,A), \tag{35b}$$

for some joint pmfs that factorize as

$$p(x_1, x_2, y, a, u, \hat{x}_1, \hat{x}_2) = p(x_1, x_2)p(a, u, \hat{x}_1|x_1, x_2)p(y|a, x_1, x_2) \cdot \delta(\hat{x}_2 - \hat{\mathrm{x}}_2(u, y)), \tag{36}$$

with pmf $p(a, u, \hat{x}_1|x_1, x_2)$ and deterministic function $\hat{\mathrm{x}}_2(u, y)$, such that the action and the distortion constraints

$$\mathrm{E}[\Lambda(A)] \leq \Gamma \tag{37}$$
$$\text{and } \mathrm{E}[d_j(X_1, X_2, Y, \hat{X}_j)] \leq D_j, \text{ for } j = 1, 2, \tag{38}$$

respectively, hold. Finally, $U$ is an auxiliary random variable whose alphabet cardinality can be constrained as $|\mathcal{U}| \leq |\mathcal{X}_1|\,|\mathcal{X}_2| + 4$, without loss of optimality.

*Remark 6:* If $p(y|a, x_1, x_2) = p(y|x_1, x_2)$, Proposition 5 reduces to [22, Th. 1].

The proof of converse is provided in Appendix E. The coding strategy that proves achievability is a combination of the techniques proposed in [1] and [22, Th. 1]. Here, we briefly outline the main ideas, since the technical details follow from standard arguments. In the scheme at hand, Node 1 first maps sequences $X_1^n$ and $X_2^n$ into the action sequence $A^n$ and an auxiliary codeword $U^n$ using the standard joint typicality criterion. This mapping operation requires a codebook of rate $I(X_1, X_2; U, A)$ (see, e.g., [2, Ch. 3]). Then, given the so-obtained sequences $A^n$ and $U^n$, source sequences $X_1^n$ and $X_2^n$ are further mapped into the estimate $\hat{X}_1^n$ for Node 2 so that the sequences $(X_1^n, X_2^n, A^n, U^n, \hat{X}_1^n)$ are jointly typical. This requires rate $I(X_1, X_2; \hat{X}_1|U, A)$ [2, Ch. 3]. Leveraging the side information $X_2^n$ available at Node 2, conveying the codewords $A^n$, $\hat{X}_1^n$ and $U^n$ to Node 2 requires rate $I(X_1, X_2; U, A) + I(X_1, X_2; \hat{X}_1|U, A) - I(U, A, \hat{X}_1; X_2)$ [2, Ch. 12], which equals the right-hand side of (35a). Node 2 conveys $U^n$ and $A^n$ to Node 3 by simply forwarding the index received from Node 1 (of rate $I(X_1, X_2; U, A)$). Finally, Node 3 estimates $\hat{X}_2^n$ through a symbol-by-symbol function as $\hat{X}_{2i} = \hat{\mathrm{x}}_2(U_i, Y_i)$ for $i \in [1, n]$.

## IV. Concluding Remarks

In the setting of source coding with a side information vending machine introduced in [1], the decoder can control the quality of the side information through a control, or action, sequence that is selected based on the message encoded by the source node. Since this message must also carry information directly related to the source to be reproduced at the decoder, a key aspect of the model is the interplay between encoding data and control information.

In this study, we have generalized the original work [1] to two standard multiterminal scenarios, namely distributed source coding and cascade source coding. For the former, we obtained inner bounds to the rate-distortion-cost regions for the cases with noncausal and causal side information at the decoder. These bounds have been found to be tight in two special cases. We have also provided some numerical example to shed some light on the advantages of an optimized tradeoff between data and control transmission. As for the cascade source coding problem, a single-letter characterizations of achievable rate-distortion-cost tradeoffs has been derived under the assumption of causal side information at the decoder.

A number of open problems have been left unsolved by this study, including the identification of more general conditions under which the inner bounds of Propositions 1 and 2 are tight. The technical challenges that we have faced in this task are related to the well-known issues that arise when identifying auxiliary random variables that satisfy the desired Markov chain conditions in distributed source coding problems (see, e.g., [2, Ch. 13]).

## Appendix A

Using standard inequalities, it can be seen that the rate region (12) evaluated with a constant $Q$ is a contra-polymatroid, as the Berger–Tung region (17) (see e.g., [23]). Moreover, the role of the variable $Q$ is that of performing the convexification of the union of all regions of tuples $(R_1, R_2, D_1, D_2, \Gamma)$ that satisfy (12) and (14) for some fixed $Q$. It follows from [23] that every extreme point of region of achievable tuples $(R_1, R_2, D_1, D_2, \Gamma)$ satisfies the equations

$$R_1 = I(X_1; V_1|V_2) + I(X_1; U_1|U_2, V_1, V_2, Y) \tag{39a}$$
$$R_2 = I(X_2; V_2) + I(X_2; U_2|V_1, V_2, Y) \tag{39b}$$

along with (14), where both relationships are satisfied with equality, or

$$R_1 = I(X_1; V_1) + I(X_1; U_1|V_1, V_2, Y) \tag{40a}$$
$$R_2 = I(X_2; V_2|V_1) + I(X_2; U_2|U_1, V_1, V_2, Y) \tag{40b}$$

along with (14) satisfied with equality. Applying the Fenchel–Eggleston–Caratheodory theorem to the right-hand side of the equations above and to (14) concludes the proof (See [2, Appendix C] and [13]).

## Appendix B
## Proof of the Converse for Proposition 3

In this section, the proof of converse for Proposition 3 is given. For any $(n, R_1, R_2, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code, we have the following inequalities:

$$\begin{aligned}
nR_1 &\geq H(M_1) \geq H(M_1|M_2) \\
&\overset{(a)}{=} I(M_1; X_1^n, X_2^n|M_2) \\
&= \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_2) \\
&\quad - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2) \\
&\overset{(b)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_2) \\
&\quad - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2, Y^{i-1}) \\
&\overset{(c)}{\geq} \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_2, Y^{i-1}) \\
&\quad - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2, Y^{i-1}) \\
&\overset{(d)}{=} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; U_{1i}|U_{2i}),
\end{aligned}$$

where $(a)$ follows because $M_1$ is a function of $(X_1^n, X_2^n)$ given that $X_2^n$ is a function of $X_1^n$ by assumption; $(b)$ follows since $(X_{1i}, X_{2i})$—$(X_1^{i-1}, X_2^{i-1}, M_1, M_2)$—$Y^{i-1}$ forms a Markov chain; $(c)$ follows by the fact that conditioning decreases entropy; and $(d)$ follows by defining $U_{ji} = (X_1^{i-1}, X_2^{i-1}, Y^{i-1}, M_j)$ for $j = 1, 2$. We also have a similar chain of inequalities for $R_2$. As for the sum-rate $R_1 + R_2$, we have

$$\begin{aligned} n(R_1+R_2) &\geq H(M_1, M_2) \\ &\overset{(a)}{=} I(M_1, M_2; X_1^n, X_2^n) \\ &= \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}) \\ &- H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2) \\ &\overset{(b)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}) \\ &- H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2, Y^{i-1}) \\ &\overset{(c)}{\geq} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; U_{1i}, U_{2i}), \end{aligned}$$

where $(a)$ follows because $(M_1, M_2)$ are functions of $(X_1^n, X_2^n)$; $(b)$ follows since $(X_{1i}, X_{2i})$—$(X_1^{i-1}, X_2^{i-1}, M_1, M_2)$—$Y^{i-1}$ forms a Markov chain; and $(c)$ follows using the definition of $U_{ji}$ for $j = 1, 2$. Next, let $Q$ be a uniform random variable over the interval $[1, n]$ and independent of $(X_1^n, X_2^n, U_1^n, U_2^n, Y^n)$ and define $U_j \triangleq (Q, U_{jQ})$, for $j = 1, 2$, $X_1 \triangleq X_{1Q}$, $X_2 \triangleq X_{2Q}$, $Y \triangleq Y_Q$. Note that $\hat{X}_j$ is a function of $U_1, U_2$ and $Y$ for $j = 1, 2$. Moreover, from (8) and (9), we have

$$\Gamma + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}[\Lambda(A)] \tag{41}$$

$$\begin{aligned} \text{and } D_j + \epsilon &\geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_j(X_{1i}, X_{2i}, Y_i.\hat{X}_{ji},)\right] \\ &= \mathrm{E}[d_1(X_1, X_2, Y, \hat{X}_j)], \text{ for } j = 1, 2. \end{aligned} \tag{42}$$

## Appendix C
## Proof of the Converse for Proposition 4

In this section, the proof of converse for Proposition 4 is given. Fix a code $(n, R_1, R_2, D_1 + \epsilon, \epsilon, \Gamma)$ for an $\epsilon > 0$, whose existence for all sufficiently large $n$ is required by the definition of achievability.

From the distortion constraint for $\hat{X}_2$, we have the inequality

$$\epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[d_H(X_{2i}, \hat{X}_{2i})] \overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n} p_{e,2i}, \tag{43}$$

where we have defined $p_{e,2i} = \Pr[X_{2i} \neq \hat{X}_{2i}]$, and $(a)$ follows from the definition of the metric $d_H(x, \hat{x})$ as the Hamming distortion. Moreover, we also have the following chain of inequalities:

$$\begin{aligned} H(X_2^n|\hat{X}_2^n) &\overset{(a)}{\leq} \sum_{i=1}^{n} H(X_{2i}|\hat{X}_{2i}) \overset{(b)}{\leq} \sum_{i=1}^{n} H(p_{e,i}) + p_{e,i}\log\left|\hat{\mathcal{X}}_{2i}\right| \\ &\overset{(c)}{\leq} nH\left(\frac{1}{n}\sum_{i=1}^{n} p_{e,i}\right) + n\left(\frac{1}{n}\sum_{i=1}^{n} p_{e,i}\right)\log\left|\hat{\mathcal{X}}_{2i}\right| \\ &\overset{(d)}{\leq} nH(\epsilon) + n\epsilon\log\left|\hat{\mathcal{X}}_{ji}\right| \\ &\triangleq n\delta(\epsilon), \end{aligned} \tag{44}$$

where $(a)$ follows by conditioning reduces entropy; $(b)$ follows by Fano's inequality; $(c)$ follows by Jensen's inequality; and $(d)$ follows by (43), where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Note that, in the following, we use the convention in [2, Ch. 3] of defining as $\delta(\epsilon)$ any function such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$.

For rate $R_1$, we then have the following series of inequalities:

$$\begin{aligned} nR_1 &\geq H(M_1) \overset{(a)}{=} H(M_1, A^n) \\ &= H(A^n) + H(M_1|A^n) \\ &\overset{(b)}{\geq} H(A^n) - H(A^n|X_1^n, X_2^n) + H(M_1|A^n, Y^n, X_2^n) \\ &- H(M_1|A^n, Y^n, X_1^n, X_2^n) \\ &= I(A^n; X_1^n, X_2^n) + I(M_1; X_1^n|A^n, Y^n, X_2^n) \\ &= I(A^n; X_1^n, X_2^n) + H(X_1^n|A^n, Y^n, X_2^n) \\ &- H(X_1^n|A^n, Y^n, X_2^n, M_1) \\ &= H(X_1^n, X_2^n) - H(X_1^n, X_2^n|A^n) \\ &+ H(X_1^n, X_2^n, Y^n|A^n) - H(Y^n, X_2^n|A^n) \\ &- H(X_1^n|A^n, Y^n, X_2^n, M_1) \\ &= H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n, X_2^n|A^n) \\ &- H(X_1^n|A^n, Y^n, X_2^n, M_1), \end{aligned} \tag{45}$$

where $(a)$ follows because $A^n$ is a function of $M_1$ and $(b)$ follows because entropy is nonnegative and conditioning decreases entropy. For the first three terms in (45), we have

$$\begin{aligned} &H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n, X_2^n|A^n) \\ &= H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n|A^n) \\ &- H(X_2^n|A^n, Y^n) \\ &\overset{(a)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) + H(Y_i|Y^{i-1}, A^n, X_1^n, X_2^n) \\ &- H(Y_i|Y^{i-1}, A^n) - H(X_{2i}|X_2^{i-1}, A^n, Y^n) \\ &\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) + H(Y_i|A_i, X_{1i}, X_{2i}) - H(Y_i|A_i) \\ &- H(X_{2i}|A_i, Y_i) \\ &= \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - I(Y_i; X_{1i}, X_{2i}|A_i) - H(X_{2i}|A_i, Y_i) \\ &= \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|A_i) + H(X_{1i}, X_{2i}|A_i, Y_i) \\ &- H(X_{2i}|A_i, Y_i) \\ &= \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i) + H(X_{1i}|A_i, Y_i, X_{2i}), \end{aligned} \tag{46}$$

where $(a)$ follows by the chain rule for entropy and the fact that $X_1^n, X_2^n$ are i.i.d. and $(b)$ follows since $Y_i$—$(A_i, X_{1i}, X_{2i})$—$(Y^{i-1}, A^{n\backslash i}, X_1^{n\backslash i}, X_2^{n\backslash i})$ forms a Markov chain, by the definition of problem, and since conditioning reduces entropy.

Combining (45) and (46), and defining $U_{1i} = (A^{n\backslash i}, Y^{n\backslash i}, X_2^{n\backslash i}, M_1)$, we obtain

$$\begin{aligned}
nR_1 &\overset{(a)}{\geq} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i) + H(X_{1i}|A_i, Y_i, X_{2i}) \\
&\quad - H(X_{1i}|X_1^{i-1}, A^n, Y^n, X_2^n, M_1) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} I(X_{1i}; A_i) + H(X_{1i}|A_i, Y_i, X_{2i}) \\
&\quad - H(X_{1i}|A^n, Y^n, X_2^n, M_1) \\
&\overset{(c)}{=} \sum_{i=1}^{n} (X_{1i}; A_i) + I(X_{1i}; U_{1i}|A_i, Y_i, X_{2i}), \qquad (47)
\end{aligned}$$

where $(a)$ follows by the chain rule for entropy; $(b)$ follows because mutual information is nonnegative and due to the fact that conditioning decreases entropy; and $(c)$ follows by the definition of mutual information and definition of $U_{1i}$.

Next, we consider the rate $R_2$. We have

$$\begin{aligned}
nR_2 &\geq H(M_2) \geq H(M_2|A^n, Y^n, M_1) \\
&\quad - H(M_2|A^n, Y^n, M_1, X_2^n) \\
&= I(M_2; X_2^n|A^n, Y^n, M_1) \\
&= H(X_2^n|A^n, Y^n, M_1) - H(X_2^n|A^n, Y^n, M_1, M_2) \\
&\overset{(a)}{\geq} H(X_2^n|A^n, Y^n, M_1) - n\delta(\epsilon) \\
&= \sum_{i=1}^{n} H(X_{2i}|X_2^{i-1}, A^n, Y^n, M_1) - n\delta(\epsilon) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_{2i}|A_i, Y_i, U_{1i}) - n\delta(\epsilon), \qquad (48)
\end{aligned}$$

where $(a)$ follows because from (44), $H(X_2^n|A^n, Y^n, M_1, M_2) \leq H(X_2^n|\hat{X}_2^n) \leq n\delta(\epsilon)$, given that $\hat{X}_2^n$ is a function of $M_1,\ M_2$ and $Y^n$ and $(b)$ follows using the definition of $U_{1i}$ and due to the fact that conditioning decreases entropy. For the sum-rate $R_1 + R_2$, we also have the following series of inequalities:

$$\begin{aligned}
n(R_1+R_2) &\geq H(M_1, M_2) \overset{(a)}{=} H(M_1, M_2, A^n) \\
&= H(A^n) + H(M_1, M_2|A^n) \\
&\geq H(A^n) - H(A^n|X_1^n, X_2^n) + H(M_1, M_2|A^n, Y^n) \\
&\quad - H(M_1, M_2|A^n, Y^n, X_1^n, X_2^n) \\
&= I(A^n; X_1^n, X_2^n) + I(M_1, M_2; X_1^n, X_2^n|A^n, Y^n) \\
&= I(A^n; X_1^n, X_2^n) + H(X_1^n, X_2^n|A^n, Y^n) \\
&\quad - H(X_1^n, X_2^n|A^n, Y^n, M_1, M_2) \\
&= H(X_1^n, X_2^n) - H(X_1^n, X_2^n|A^n) \\
&\quad + H(X_1^n, X_2^n, Y^n|A^n) - H(Y^n|A^n) \\
&\quad - H(X_2^n|A^n, Y^n, M_1, M_2) \\
&\quad - H(X_1^n|A^n, Y^n, X_2^n, M_1, M_2) \\
&\overset{(b)}{\geq} H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n|A^n) \\
&\quad - H(X_1^n|A^n, Y^n, X_2^n, M_1, M_2) - n\delta(\epsilon), \qquad (49)
\end{aligned}$$

where $(a)$ follows because $A^n$ is a function of $M_1$; and $(b)$ follows as in $(a)$ of (48). For the first three terms in (49), we have

$$\begin{aligned}
&H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n|A^n) \\
&\overset{(a)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) + H(Y_i|Y^{i-1}, A^n, X_1^n, X_2^n) \\
&\quad - H(Y_i|Y^{i-1}, A^n) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) + H(Y_i|A_i, X_{1i}, X_{2i}) - H(Y_i|A_i) \\
&= \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - I(Y_i; X_{1i}, X_{2i}|A_i) \\
&= \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|A_i) + H(X_{1i}, X_{2i}|A_i, Y_i) \\
&= \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i) + H(X_{2i}|A_i, Y_i) + H(X_{1i}|A_i, Y_i, X_{2i}),
\end{aligned} \qquad (50)$$

where $(a)$ follows from the chain rule for entropy and by the chain rule for entropy and the fact that $(X_1^n, X_2^n)$ are i.i.d.; and $(b)$ follows since $Y_i$—$(A_i, X_{\{1,2\}i})$—$(Y^{i-1}, A^{n\backslash i}, X_1^{n\backslash i}, X_2^{n\backslash i})$ forms a Markov chain, by the definition of problem, and since conditioning reduces entropy. Combining (49) and (50), and using the definition of $U_{1i}$, we obtain

$$\begin{aligned}
n(R_1 + R_2) &\overset{(a)}{\geq} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i) + H(X_{2i}|A_i, Y_i) \\
&\quad + H(X_{1i}|A_i, Y_i, X_{2i}) \\
&\quad - H(X_{1i}|X_1^{i-1}, A^n, Y^n, X_2^n, M_1, M_2) - n\delta(\epsilon) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} I(X_{1i}; A_i) + H(X_{2i}|A_i, Y_i) + H(X_{1i}|A_i, Y_i, X_{2i}) \\
&\quad - H(X_{1i}|A^n, Y^n, X_2^n, M_1) - n\delta(\epsilon) \\
&\overset{(c)}{\geq} \sum_{i=1}^{n} (X_{1i}; A_i) + H(X_{2i}|A_i, Y_i) + I(X_{1i}; U_{1i}|A_i, Y_i, X_{2i}) \\
&\quad - n\delta(\epsilon), \qquad (51)
\end{aligned}$$

where $(a)$ follows by the chain rule for entropy; $(b)$ follows because mutual information is nonnegative and due to the fact that conditioning decreases entropy; and $(c)$ follows by the definition of mutual information and definition of $U_{1i}$ and the fact that conditioning decreases entropy.

Moreover, $(X_{2i}, Y_i) - (X_{1i}, A_i) - U_{1i}$ forms a Markov chain. This can be seen by using the principle of $d$-separation [24, Sec. A.9] from Fig. 7, which represents the joint distribution of all the variables at hand.

Let $Q$ be a uniform random variable over the interval $[1, n]$ and independent of $(X_1^n, X_2^n, A^n,\ U_1^n, Y^n, \hat{X}_1^n)$ and define $U_1 \triangleq (Q, U_{1Q})$, $X_1 \triangleq X_{1Q}$, $X_2 \triangleq X_{2Q}$, $Y \triangleq Y_Q$, $A \triangleq A_Q$,

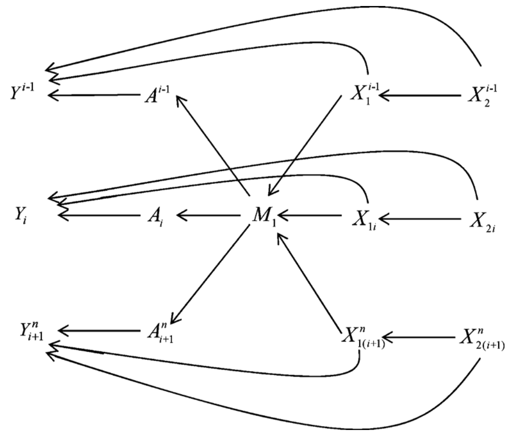


Fig. 7. Bayesian network representing the joint pmf of variables $(M_1, X_1^n, X_2^n, A^n, Y^n)$ for the model in Fig. 2.

and $\hat{X}_1 \triangleq \hat{X}_{1Q}$. Note that $\hat{X}_1$ is a function of $U_1$, $X_2$, and $Y$. Moreover, from (8) and (9), we have

$$\begin{aligned}\Gamma + \epsilon &\geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}[\Lambda(A)] \\ \text{and } D_1 + \epsilon &\geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_1(X_{1i}, X_{2i}, Y_i.\hat{X}_{1i})\right] \\ &= \mathrm{E}[d_1(X_1, X_2, Y, \hat{X}_1)]. \end{aligned} \tag{52}$$

Finally, since (47), (48), and (51) are convex with respect to $p(a, u_1|x_1, q)$ for fixed $p(q)$, $p(x_1, x_2)$, and $p(y|a, x_1, x_2)$, we have that inequalities (20) hold, which completes the proof of (20a) and (22b). The cardinality bounds are proved by using the Fenchel–Eggleston–Caratheodory theorem in the standard way.

## Appendix D
## Greedy Actions Are Optimal With Sum Side Information

Here, we prove equality

$$R^{\oplus}_{\text{sum, greedy}}(\Gamma) = R^{\oplus}_{\text{sum}}(\Gamma) \tag{53}$$

which shows that no gain is accrued by choosing the actions based only on message $M_1$ with the sum side information. Fix the pmf $p(a|x_1)$ that achieves the minimum in the sum-rate obtained from (20c), namely

$$R^{\oplus}_{\text{sum}}(\Gamma) = \min\, I(X_1; A) + H(X_1, X_2|A, Y),$$

where the mutual information is calculated with respect to the distribution

$$p(x_1, x_2, y, a) = p(x_1, x_2)p(a|x_1)p(y|a, x_1, x_2), \tag{54}$$

and the minimum is taken over all distributions $p(a|x_1)$ such that $\mathrm{E}\left[\Lambda(A)\right] = \mathrm{E}\left[A\right] \leq \Gamma$. Note that for such a pmf $p(a|x_1)$, we have $\mathrm{E}[A] = p(a) = \Gamma$, as it can be easily seen. We then have the following series of equalities:

$$\begin{aligned} &R^{\oplus}_{\text{sum, greedy}}(\Gamma) - R^{\oplus}_{\text{sum}}(\Gamma) \\ &\overset{(a)}{=} \Gamma H(X_1, X_2|X_1 \oplus X_2) + (1-\Gamma)H(X_1, X_2) \\ &- H(X_1, X_2|A, X_1 \oplus X_2) - I(X_1; A) \\ &\overset{(b)}{=} \Gamma H(X_1|X_1 \oplus X_2) + (1-\Gamma)(1 + H(p)) \\ &- \Gamma H(X_1, X_2|A = 1, X_1 \oplus X_2) \\ &- (1-\Gamma)H(X_1, X_2|A = 0) - I(X_1; A) \\ &\overset{(c)}{=} \Gamma H(X_1) + (1-\Gamma)(1 + H(p)) - \Gamma H(X_1|A = 1) \\ &- (1-\Gamma)H(X_1|A = 0) \\ &- (1-\Gamma)H(X_2|X_1, A = 0) - I(X_1; A) \\ &\overset{(d)}{=} \Gamma + (1-\Gamma)(1 + H(p)) - H(X_1|A) \\ &- (1-\Gamma)H(X_2|X_1) - I(X_1; A) \\ &= \Gamma + (1-\Gamma)(1 + H(p)) - H(X_1|A) \\ &- (1-\Gamma)H(p) - 1 + H(X_1|A) = 0, \end{aligned}$$

where $(a)$ follows by the definition (27); $(b)$ follows using the chain rule for entropy and from the definition of conditional entropy; $(c)$ follows by the crypto-lemma [21, Lemma 2]; and $(d)$ follows from the fact that $X_2 - X_1 - A$ forms a Markov chain.

## Appendix E
## Proof of the Converse for Proposition 5

In this section, we provide the proof of converse for Proposition 5. For any $(n, R_{12}, R_{23}, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code, we have the following inequalities:

$$\begin{aligned} nR_{12} &\geq H(M_{12}) \geq H(M_{12}|X_2^n) \overset{(a)}{=} H(M_{12}, M_{23}|X_2^n) \\ &\overset{(b)}{=} I(X_1^n; M_{12}, M_{23}|X_2^n) \\ &= \sum_{i=1}^{n} H(X_{1i}|X_1^{i-1}, X_2^n) \\ &- H(X_{1i}|X_1^{i-1}, X_2^n, M_{12}, M_{23}) \\ &\overset{(c)}{=} \sum_{i=1}^{n} H(X_{1i}|X_{2i}) \\ &- H(X_{1i}|X_1^{i-1}, X_2^n, A^n, M_{12}, M_{23}) \\ &\overset{(d)}{=} \sum_{i=1}^{n} H(X_{1i}|X_{2i}) \\ &- H(X_{1i}|X_1^{i-1}, X_2^n, Y^{i-1}, M_{12}, M_{23}, A^n, \hat{X}_1^n) \\ &\overset{(e)}{\geq} \sum_{i=1}^{n} H(X_{1i}|X_{2i}) - H(X_{1i}|X_{2i}, A_i, U_i, \hat{X}_{1i}) \\ &= \sum_{i=1}^{n} I(X_{1i}; A_i, U_i, \hat{X}_{1i}|X_{2i}), \end{aligned} \tag{55}$$

where $(a)$ follows because $M_{23}$ is a function of $(M_{12}, X_2^n)$; $(b)$ follows by definition of mutual information and since $M_{12}$ and $M_{23}$ are functions of $X_1^n$ and $X_2^n$; $(c)$ follows because $X_1^n$ and $X_2^n$ are i.i.d and since $A^n$ is a function of $M_{23}$; $(d)$ follows because $Y^{i-1} - (X_1^{i-1}, X_2^n, A^n, M_{12}, M_{23}) - X_{1i}$ forms a Markov chain and since $\hat{X}_1^n$ is a function of $M_{12}$ and $X_2^n$; and $(e)$ follows by defining $U_i = (X_1^{i-1}, X_2^{i-1}, Y^{i-1}, A^{n\backslash i}, M_{23})$ and since conditioning decreases entropy.

We also have the inequalities

$$\begin{aligned}
nR_{23} \geq H(M_{23}) &\overset{(a)}{=} I(X_1^n, X_2^n; M_{23}) \\
&\overset{(b)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) \\
&\quad - H(X_{1i}, X_{2i} | X_1^{i-1}, X_2^{i-1}, M_{23}) \\
&\overset{(c)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) \\
&\quad - H(X_{1i}, X_{2i} | X_1^{i-1}, X_2^{i-1}, A^n, M_{23}) \\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) \\
&\quad - H(X_{1i}, X_{2i} | X_1^{i-1}, X_2^{i-1}, Y^{i-1}, A^n, M_{23}) \\
&\overset{(e)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i} | A_i, U_i) \\
&= \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i, U_i),
\end{aligned} \tag{56}$$

where $(a)$ follows because $M_{23}$ is a function of $X_1^n$ and $X_2^n$; $(b)$ follows by the definition of mutual information and the chain rule for entropy and since $X_1^n$ and $X_2^n$ are i.i.d; $(c)$ follows because $A^n$ is a function of $M_{23}$; $(d)$ follows because $Y^{i-1} - (X_1^{i-1}, X_2^{i-1}, A^n, M_{23}) - (X_{1i}, X_{2i})$ forms a Markov chain; and $(e)$ follows by the definition of $U_i$.

Let $Q$ be a uniform random variable over $[1, n]$ and independent of $(X_1^n, X_2^n, Y^n, A^n, U^n, \hat{X}_1^n)$ and define $U \triangleq (Q, U_Q)$, $X_1 \triangleq X_{1Q}$, $X_2 \triangleq X_{2Q}$, $Y \triangleq Y_Q$, $A \triangleq A_Q$, $\hat{X}_1 \triangleq \hat{X}_{1Q}$, and $\hat{X}_2 \triangleq \hat{X}_{2Q}$. Note that $\hat{X}_2$ is a function of $U$ and $Y$. Moreover, from (33) and (34), we have

$$\Gamma + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}[\Lambda(A)] \tag{57}$$

$$\begin{aligned}
\text{and } D_j + \epsilon &\geq \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[d_j(X_{1i}, X_{2i}, Y_i, \hat{X}_{ji})\right] \\
&= \mathrm{E}[d_j(X_1, X_2, Y, \hat{X}_j)] \text{ for } j = 1, 2.
\end{aligned} \tag{58}$$

Finally, since (55) and (56) are convex with respect to $p(a, u, \hat{x}_1 | x_1, x_2)$ for fixed $p(x_1, x_2)$ and $p(y|a, x_1, x_2)$, we have from (55) and (56) that inequalities (35) hold. The cardinality bounds are proved by using the Fenchel–Eggleston–Caratheodory theorem in the standard way.

## References


[1] H. Permuter and T. Weissman, "Source coding with a side information "vending machine"," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4530–4544, Jul. 2011.

[2] A. El Gamal and Y. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, Dec. 2011.

[3] T. Weissman and A. El Gamal, "Source coding with limited-look-ahead side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5218–5239, Dec. 2006.

[4] B. Ahmadi and O. Simeone, "Robust coding for lossy computing with receiver-side observation costs," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, Jul.-Aug. 31–5, 2011, pp. 2939–2943.

[5] Y. Chia, H. Asnani, and T. Weissman, "Multi-terminal source coding with action dependent side information," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, Jul.-Aug. 31–5, 2011, pp. 2035–2039.

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

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

[8] K. Kittichokechai, T. J. Oechtering, and M. Skoglund, "Source coding with common reconstruction and action-dependent side information," in *Proc. IEEE Inf. Theory Workshop*, Dublin, Ireland, Aug. 2010, pp. 1–5.

[9] K. Kittichokechai, T. J. Oechtering, and M. Skoglund, "Secure source coding with action-dependent side information," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, Jul.-Aug. 31–5, 2011, pp. 1678–1682.

[10] Y. Steinberg, "Coding and common reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4995–5010, Nov. 2009.

[11] C. Choudhuri and U. Mitra, "How useful is adaptive action?," presented at the IEEE Globecom, Anaheim, CA, USA, Dec. 3–7, 2012.

[12] H. Asnani and T. Weissman, On real time coding with limited lookahead 2011 [Online]. Available: arXiv:1105.5755, submitted for publication

[13] S.-Y. Tung, "Multiterminal source coding," Ph.D. dissertation, Cornell Univ., Ithaca, NY, USA, 1978.

[14] M. Gastpar, "The Wyner–Ziv problem with multiple sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2762–2767, Nov. 2004.

[15] B. Ahmadi, C. Choudhuri, O. Simeone, and U. Mitra, Cascade source coding with a side information "Vending Machine" [Online]. Available: http://arxiv.org/abs/1207.2793

[16] S. Jana, "Alphabet sizes of auxiliary random variables in canonical inner bounds," in *Proc. Conf. Inf. Sci. Syst.*, Baltimore, MD, USA, Mar. 18–20, 2009, pp. 67–71.

[17] A. H. Kaspi and T. Berger, "Rate-distortion for correlated sources with partially separated encoders," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 6, pp. 828–840, Nov. 1982.

[18] A. B. Wagner, B. G. Kelly, and Y. Altug, "The lossy one-helper conjecture is false," in *Proc. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Sep.–Oct. 30–2, 2009, pp. 716–723.

[19] T. Berger and R. Yeung, "Multiterminal source encoding with one distortion criterion," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 228–236, Mar. 1989.

[20] B. Ahmadi and O. Simeone, Distributed and cascade lossy source coding with a side information "Vending Machine" 2011 [Online]. Available: http://arxiv.org/abs/1109.6665

[21] G. D. Forney Jr., "On the role of MMSE estimation in approaching the information- theoretic limits of linear Gaussian channels: Shannon meets Wiener," in *Proc. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Oct. 2003, pp. 430–439.

[22] Y.-K. Chia and T. Weissman, "Cascade and triangular source coding with causal side information," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, Jul.–Aug. 31–5, 2011, pp. 1683–1687.

[23] J. Chen, X. Zhang, T. Berger, and S. B. Wicker, "An upper bound on the sum rate distortion function and its corresponding rate allocation schemes for the CEO problem," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 977–987, Aug. 2004.

[24] G. Kramer, *Topics in Multi-User Information Theory*. Delft, The Netherlands: Now Publishers, 2008.

**Behzad Ahmadi** (S’10) received the B.Sc and M.Sc degrees in electrical and computer engineering from Isfahan University of Technology, Isfahan, Iran in August 2005 and February 2008, respectively. He is currently a Ph.D. candidate in electrical and computer engineering at the Center for Wireless Communications and Signal Processing Research (CWCSPR), New Jersey Institute of Technology (NJIT), Newark, NJ. His research interests include information theory and bio-medical signal processing. Behzad Ahmadi was a recipient of the Hashimoto Fellowship Scholarship 2011–12. He is also the recipient of Hashimoto prize for best Doctoral Dissertation at New Jersey Institute of Technology for the year 2012–13.

**Osvaldo Simeone** (M’01) received the M.Sc. degree (with honors) and the Ph.D. degree in information engineering from Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively. He is currently with the Center for Wireless Communications and Signal Processing Research (CWCSPR), New Jersey Institute of Technology (NJIT), Newark, where he is an Associate Professor. His current research interests concern the cross-layer analysis and design of wireless networks with emphasis on information-theoretic, signal processing, and queuing aspects. Specific topics of interest are: cognitive radio, cooperative communications, rate-distortion theory, ad hoc, sensor, mesh and hybrid networks, distributed estimation, and synchronization. Dr. Simeone is a co-recipient of Best Paper Awards of the IEEE SPAWC 2007 and IEEE WRECOM 2007. He currently serves as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.