# Multiterminal Source Coding Under Logarithmic Loss

Thomas A. Courtade, *Member, IEEE*, and Tsachy Weissman, *Fellow, IEEE*

***Abstract*— We consider the classical two-encoder multiterminal source coding problem where distortion is measured under logarithmic loss. We provide a single-letter description of the achievable rate distortion region for all discrete memoryless sources with finite alphabets. By doing so, we also give the rate distortion region for the $m$-encoder CEO problem (also under logarithmic loss). Several applications and examples are given.**

***Index Terms*— Logarithmic loss, lossy compression, multiterminal source coding, rate-distortion, rate region.**

## I. Introduction

A COMPLETE characterization of the achievable rate distortion region for the two-encoder source coding problem depicted in Figure 1 has remained an open problem for over three decades. Following tradition, we refer to this two-encoder source coding network as the *multiterminal source coding problem* throughout this paper. Several special cases have been solved for general source alphabets and distortion measures:

- The lossless case where $D_1 = 0, D_2 = 0$. Slepian and Wolf solved this case in their seminal work [3].
- The case where one source is recovered losslessly: i.e., $D_1 = 0$, $D_2 = D_{\max}$. This case corresponds to the source coding with side information problem of Ahlswede-Körner-Wyner [4], [5].
- The Wyner-Ziv case [6] where $Y_2$ is available to the decoder as side information and $Y_1$ should be recovered with distortion at most $D_1$.
- The Berger-Yeung case (which subsumes the previous three cases) [7] where $D_1$ is arbitrary and $D_2 = 0$.

Despite the apparent progress, other seemingly fundamental cases, such as when $D_1$ is arbitrary and $D_2 = D_{\max}$, remain unsolved except perhaps in very special cases.

Recently, the achievable rate distortion region for the quadratic Gaussian multiterminal source coding problem was completely characterized by Wagner, Tavildar, and Viswanath [8], who built upon significant prior work (cf. [9]–[13]). Until now, this was the only case for which the entire achievable rate distortion region was known. While this is a very important result, it is again a special case from a theoretical point of view: a specific choice of source distribution, and a specific choice of distortion measure.

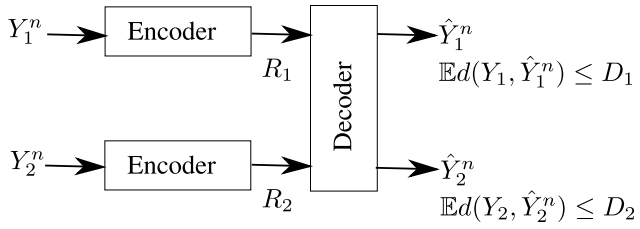


Fig. 1. The multiterminal source coding network.

In the present paper, we determine the achievable rate distortion region of the multiterminal source coding problem for all discrete memoryless sources with finite alphabets. However, as in [8], we restrict our attention to a specific distortion measure.

At a high level, the roadmap for our argument is similar to that of [8]. In particular, both arguments couple the multiterminal source coding problem to a parametrized family of CEO problems. Then, the parameter in the CEO problem is "tuned" to yield the converse result. Despite this apparent similarity, the proofs in [8] rely heavily on the previously known Gaussian CEO results [12], the Gaussian one-helper results [10], and the calculus performed on the closed-form entropy expressions which arise from the Gaussian source assumption. In our case we do not have this luxury, and our CEO tuning argument essentially relies on an existence lemma to yield the converse result. The success of our approach is largely due to the fact that the distortion measure we consider admits a lower bound in the form of a conditional entropy, much like the quadratic distortion measure for Gaussian sources.

### A. Our Contributions

In this paper, we give a single-letter characterization of the achievable rate distortion region for the multiterminal source coding problem under logarithmic loss. In the process of accomplishing this, we derive the achievable rate distortion region for the $m$-encoder CEO problem, also under logarithmic loss. In both settings, we obtain a stronger converse than is standard for rate distortion problems in the sense that augmenting the reproduction alphabet does not enlarge the rate distortion region. Notably, we make no assumptions on the source distributions, other than that the sources have finite alphabets. In both cases, the Berger-Tung inner bound on the rate distortion region is tight. To our knowledge,

Manuscript received October 13, 2011; revised November 19, 2012; accepted December 24, 2012. Date of publication November 1, 2013; date of current version December 20, 2013. This paper was presented at the 2012 IEEE International Symposium on Information Theory [1] and also forms part of T. A. Courtade's Ph.D. thesis [2]. This work was supported by the NSF Center for Science of Information under Grant Agreement CCF-0939370.

The authors are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: courtade@stanford.edu; tsachy@stanford.edu).

Communicated by Y. Oohama, Associate Editor for Source Coding.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2288257

this constitutes the first time that the entire achievable rate distortion region has been described for general finite-alphabet sources under nontrivial distortion constraints.

### B. Organization

This paper is organized as follows. In Section II we formally define the logarithmic loss function and the multiterminal source coding problem we consider. In Section III we define the CEO problem and give the rate distortion region under logarithmic loss. In Section IV we return to the multiterminal source coding problem and derive the rate distortion region for the two-encoder setting. Also in Sections III and IV, applications to estimation, horse racing, and list decoding are given. In Section V, we discuss connections between our results and the multiterminal source coding problem with arbitrary distortion measures. Section VI delivers our concluding remarks and discusses directions for future work.

## II. Problem Definition

Throughout this paper, we adopt notational conventions that are standard in the literature. Specifically, random variables are denoted by capital letters (e.g., $X$) and their corresponding alphabets are denoted by corresponding calligraphic letters (e.g., $\mathcal{X}$). We abbreviate a sequence $(X_1, X_2, \ldots, X_n)$ of $n$ random variables by $X^n$, and we denote the interval $(X_k, X_{k+1}, \ldots, X_j)$ by $X_k^j$. If the lower index is equal to 1, it will be omitted when there is no ambiguity (e.g., $X^j \triangleq X_1^j$). Frequently, random variables will appear with two subscripts (e.g., $Y_{i,j}$). In this case, we are referring to the $j^{th}$ instance of random variable $Y_i$. We overload our notation here slightly in that $Y_{i,1}^j$ is often abbreviated as $Y_i^j$. However, our meaning will always be clear from context. Throughout, we let $[x]^+ = \max\{x, 0\}$ for real-valued $x$.

Let $\{(Y_{1,j}, Y_{2,j})\}_{j=1}^n = (Y_1^n, Y_2^n)$ be a sequence of $n$ independent, identically distributed pairs of random variables with finite alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively, and joint pmf $p(y_1, y_2)$. That is, $(Y_1^n, Y_2^n) \sim \prod_{i=1}^n p(y_{1,j}, y_{2,j})$.

In this paper, we take the reproduction alphabet $\hat{\mathcal{Y}}_i$ to be equal to the set of probability distributions over the source alphabet $\mathcal{Y}_i$ for $i = 1, 2$. Thus, for a vector $\hat{Y}_i^n \in \hat{\mathcal{Y}}_i^n$, we will use the notation $\hat{Y}_{i,j}(y_i)$ to mean the $j^{th}$ coordinate $(1 \leq j \leq n)$ of $\hat{Y}_i^n$ (which is a probability distribution on $\mathcal{Y}_i$) evaluated for the outcome $y_i \in \mathcal{Y}_i$. In other words, the decoder generates 'soft' estimates of the source sequences.

We consider the logarithmic loss distortion measure defined as follows:

$$d(y_i, \hat{y}_i) = \log\left(\frac{1}{\hat{y}_i(y_i)}\right) \qquad \text{for } i = 1, 2.$$

Equivalently, $d(y_i, \hat{y}_i)$ is the relative entropy (i.e., Kullback-Leibler divergence) between the empirical distribution of the event $\{Y_i = y_i\}$ and the estimate $\hat{y}_i$. Using this definition for symbol-wise distortion, it is standard to define the distortion between sequences as

$$d(y_i^n, \hat{y}_i^n) = \frac{1}{n}\sum_{j=1}^n d(y_{i,j}, \hat{y}_{i,j}) \qquad \text{for } i = 1, 2.$$

We point out that the logarithmic loss function is a widely used penalty function in the theory of learning and prediction (cf. [14, Chapter 9]). Further, it is a particularly natural loss criterion in settings where the reconstructions are allowed to be 'soft', rather than deterministic values. Surprisingly, since distributed learning and estimation problems are some of the most oft-cited applications of lossy multiterminal source coding, it does not appear to have been studied in this context until the recent work [15]. However, we note that this connection has been established previously for the single-encoder case in the study of the information bottleneck method [16], [17]; we comment further on this connection in Section III. Beyond learning and prediction, a similar distortion measure has appeared before in the image processing literature [18]. As we demonstrate through several examples, the logarithmic loss distortion measure has a variety of useful applications in the context of multiterminal source coding.

A rate distortion code (of blocklength $n$) consists of encoding functions:

$$g_i^{(n)} : \mathcal{Y}_i^n \to \left\{1, \ldots, M_i^{(n)}\right\} \qquad \text{for } i = 1, 2$$

and decoding functions

$$\psi_i^{(n)} : \left\{1, \ldots, M_1^{(n)}\right\} \times \left\{1, \ldots, M_2^{(n)}\right\} \to \hat{\mathcal{Y}}_i^n \qquad \text{for } i = 1, 2.$$

A rate distortion vector $(R_1, R_2, D_1, D_2)$ is strict-sense achievable if there exists a blocklength $n$, encoding functions $g_1^{(n)}, g_2^{(n)}$ and a decoder $(\psi_1^{(n)}, \psi_2^{(n)})$ such that

$$R_i \geq \frac{1}{n}\log M_i^{(n)} \qquad \text{for } i = 1, 2$$
$$D_i \geq \mathbb{E}d(Y_i^n, \hat{Y}_i^n) \qquad \text{for } i = 1, 2.$$

Where

$$\hat{Y}_i^n = \psi_i^{(n)}(g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)) \qquad \text{for } i = 1, 2.$$

*Definition 1: Let $\mathcal{RD}^\star$ denote the set of strict-sense achievable rate distortion vectors and define the set of achievable rate distortion vectors to be its closure, $\overline{\mathcal{RD}^\star}$.*

Our ultimate goal in the present paper is to give a single-letter characterization of the region $\overline{\mathcal{RD}^\star}$. However, in order to do this, we first consider an associated CEO problem. In this sense, the roadmap for our argument is similar to that of [8]. Specifically, both arguments couple the multiterminal source coding problem to a parametrized family of CEO problems. Then, the parameter in the CEO problem is "tuned" to yield the converse result. Despite this apparent similarity, the proofs are quite different since the results in [8] depend heavily on the peculiarities of the Gaussian distribution.

## III. The CEO Problem

In order to attack the general multiterminal problem, we begin by studying the CEO problem (See [9] for an introduction.). To this end, let $\{(X_j, Y_{1,j}, Y_{2,j})\}_{j=1}^n = (X^n, Y_1^n, Y_2^n)$ be a sequence of $n$ independent, identically distributed random variables distributed according to the joint pmf $p(x, y_1, y_2) = p(x)p(y_1|x)p(y_2|x)$. That is, $Y_1 \leftrightarrow X \leftrightarrow Y_2$ form a Markov chain, in that order.

In this section, we consider the reproduction alphabet $\hat{\mathcal{X}}$ to be equal to the set of probability distributions over the source alphabet $\mathcal{X}$. As before, for a vector $\hat{X}^n \in \hat{\mathcal{X}}^n$, we will use the notation $\hat{X}_j(x)$ to mean the $j^{th}$ coordinate of $\hat{X}^n$ (which is a probability distribution on $\mathcal{X}$) evaluated for the outcome $x \in \mathcal{X}$. As in the rest of this paper, $d(\cdot,\cdot)$ is the logarithmic loss distortion measure.

A rate distortion CEO code (of blocklength $n$) consists of encoding functions:

$$g_i^{(n)} : \mathcal{Y}_i^n \to \left\{1, \ldots, M_i^{(n)}\right\} \qquad \text{for } i = 1, 2$$

and a decoding function

$$\psi^{(n)} : \left\{1, \ldots, M_1^{(n)}\right\} \times \left\{1, \ldots, M_2^{(n)}\right\} \to \hat{\mathcal{X}}^n.$$

A rate distortion vector $(R_1, R_2, D)$ is strict-sense achievable for the CEO problem if there exists a blocklength $n$, encoding functions $g_1^{(n)}, g_2^{(n)}$ and a decoder $\psi^{(n)}$ such that

$$\begin{aligned} R_i &\geq \frac{1}{n} \log M_i^{(n)} \qquad \text{for } i = 1, 2 \\ D &\geq \mathbb{E} d(X^n, \hat{X}^n). \end{aligned}$$

Where

$$\hat{X}^n = \psi^{(n)}(g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)).$$

*Definition 2: Let $\mathcal{RD}_{CEO}^{\star}$ denote the set of strict-sense achievable rate distortion vectors and define the set of achievable rate distortion vectors to be its closure, $\overline{\mathcal{RD}}_{CEO}^{\star}$.*

### A. Inner Bound

*Definition 3: Let $(R_1, R_2, D) \in \mathcal{RD}_{CEO}^{i}$ if and only if there exists a joint distribution of the form*

$$p(x, y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)p(q)$$

*where $|\mathcal{U}_1| \leq |\mathcal{Y}_1|$, $|\mathcal{U}_2| \leq |\mathcal{Y}_2|$, and $|\mathcal{Q}| \leq 4$, which satisfies*

$$\begin{aligned} R_1 &\geq I(Y_1; U_1|U_2, Q) \\ R_2 &\geq I(Y_2; U_2|U_1, Q) \\ R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Q) \\ D &\geq H(X|U_1, U_2, Q). \end{aligned}$$

*Theorem 1: $\mathcal{RD}_{CEO}^{i} \subseteq \overline{\mathcal{RD}}_{CEO}^{\star}$. That is, all rate distortion vectors $(R_1, R_2, D) \in \mathcal{RD}_{CEO}^{i}$ are achievable.*

Before proceeding with the proof, we cite the following variant of a well-known inner bound:

*Proposition 1 (Berger-Tung Inner Bound [19], [20]): The rate distortion vector $(R_1, R_2, D)$ is achievable if*

$$\begin{aligned} R_1 &\geq I(U_1; Y_1|U_2, Q) \\ R_2 &\geq I(U_2; Y_2|U_1, Q) \\ R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Q) \\ D &\geq \mathbb{E}\left[d(X, f(U_1, U_2, Q))\right] \end{aligned}$$

*for a joint distribution*

$$p(x)p(y_1|x)p(y_2|x)p(u_1|y_1, q)p(u_2|y_2, q)p(q)$$

*and reproduction function*

$$f : \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Q} \to \hat{\mathcal{X}}.$$

The proof of this proposition is a standard exercise in information theory, and is therefore omitted. The interested reader is directed to the text [21] for a modern, detailed treatment. The proposition follows from what is commonly called the Berger-Tung achievability scheme. In this encoding scheme, each encoder quantizes its observation $Y_i^n$ to a codeword $U_i^n$, such that the empirical distribution of the entries in $(Y_i^n, U_i^n)$ is very close to the true distribution $p(y_i, u_i)$. In order to communicate their respective quantizations to the decoder, the encoders essentially perform Slepian-Wolf coding. For this reason, the Berger-Tung achievability scheme is also referred to as a "quantize-and-bin" coding scheme.

*Proof of Theorem 1:* Given Proposition 1, the proof of Theorem 1 is immediate. Indeed, if we apply Proposition 1 with the reproduction function $f(U_1, U_2, Q) \triangleq \{\Pr[X = x|U_1, U_2, Q]\}_{x \in \mathcal{X}}$, we note that

$$\mathbb{E}\left[d(X, f(U_1, U_2, Q))\right] = H(X|U_1, U_2, Q),$$

which yields the desired result. ∎

Thus, from the proof of Theorem 1, we see that our inner bound $\mathcal{RD}_{CEO}^{i}$ simply corresponds to a specialization of the general Berger-Tung inner bound to the case of logarithmic loss.

### B. A Matching Outer Bound

A particularly useful property of the logarithmic loss distortion measure is that the expected distortion is lower-bounded by a conditional entropy. A similar property is enjoyed by Gaussian random variables under quadratic distortion. In particular, if $G$ is Gaussian, and $\hat{G}$ is such that $\mathbb{E}(\hat{G} - G)^2 \leq D$, then $\frac{1}{2}\log(2\pi e)D \geq h(G|\hat{G})$. The case for logarithmic loss is similar, and we state it formally in the following lemma which is crucial in the proof of the converse.

*Lemma 1: Let $Z = (g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n))$ be the argument of the reproduction function $\psi^{(n)}$. Then $n\mathbb{E}d(X^n, \hat{X}^n) \geq H(X^n|Z)$.*

*Proof:* By definition of the reproduction alphabet, we can consider the reproduction $\hat{X}^n$ to be a probability distribution on $\mathcal{X}^n$ conditioned on the argument $Z$. In particular, if $\hat{x}^n = \psi^{(n)}(z)$, define $s(x^n|z) \triangleq \prod_{j=1}^n \hat{x}_j(x_j)$. It is readily verified that $s$ is a probability measure on $\mathcal{X}^n$. Then, we obtain the following lower bound on the expected distortion conditioned on $Z = z$:

$$\begin{aligned} &\mathbb{E}\left[d(X^n, \hat{X}^n)|Z = z\right] \\ &\quad = \frac{1}{n}\sum_{j=1}^{n}\sum_{x^n \in \mathcal{X}^n} p(x^n|z)\log\left(\frac{1}{\hat{x}_j(x_j)}\right) \\ &\quad = \frac{1}{n}\sum_{x^n \in \mathcal{X}^n} p(x^n|z)\sum_{j=1}^{n}\log\left(\frac{1}{\hat{x}_j(x_j)}\right) \\ &\quad = \frac{1}{n}\sum_{x^n \in \mathcal{X}^n} p(x^n|z)\log\left(\frac{1}{s(x^n|z)}\right) \\ &\quad = \frac{1}{n}\sum_{x^n \in \mathcal{X}^n} p(x^n|z)\log\left(\frac{p(x^n|z)}{s(x^n|z)}\right) + \frac{1}{n}H(X^n|Z = z) \\ &\quad = \frac{1}{n}D\left(p(x^n|z)\|s(x^n|z)\right) + \frac{1}{n}H(X^n|Z = z) \\ &\quad \geq \frac{1}{n}H(X^n|Z = z), \end{aligned}$$

where $p(x^n|z) = \Pr(X^n = x^n|Z = z)$ is the true conditional distribution. Averaging both sides over all values of $Z$, we obtain the desired result. ∎

*Definition 4: Let* $(R_1, R_2, D) \in \mathcal{RD}^o_{CEO}$ *if and only if there exists a joint distribution of the form*

$$p(x)p(y_1|x)p(y_2|x)p(u_1|y_1, q)p(u_2|y_2, q)p(q),$$

*which satisfies*

$$\left.\begin{aligned} R_1 &\geq [I(Y_1; U_1|X, Q) + H(X|U_2, Q) - D]^+ \\ R_2 &\geq [I(Y_2; U_2|X, Q) + H(X|U_1, Q) - D]^+ \\ R_1 + R_2 &\geq [I(U_1; Y_1|X, Q) \\ &\quad + I(U_2; Y_2|X, Q) + H(X) - D]^+ \\ D &\geq H(X|U_1, U_2, Q). \end{aligned}\right\} \quad (1)$$

*Theorem 2: If* $(R_1, R_2, D)$ *is strict-sense achievable for the CEO problem, then* $(R_1, R_2, D) \in \mathcal{RD}^o_{CEO}$.

*Proof:* Suppose the point $(R_1, R_2, D)$ is strict-sense achievable. Let $A$ be a nonempty subset of $\{1, 2\}$, and let $F_i = g_i^{(n)}(Y_i^n)$ be the message sent by encoder $i \in \{1, 2\}$. Define $U_{i,j} \triangleq (F_i, Y_i^{j-1})$ and $Q_j \triangleq (X^{j-1}, X_{j+1}^n) = X^n \backslash X_j$. To simplify notation, let $Y_A = \cup_{i\in A} Y_i$ (similarly for $U_A$ and $F_A$).

With this notation established, we have the following string of inequalities:

$$\begin{aligned} n\sum_{i\in A} R_i &\geq \sum_{i\in A} H(F_i) \geq H(F_A) \\ &\geq I(Y_A^n; F_A|F_{A^c}) \\ &= I(X^n, Y_A^n; F_A|F_{A^c}) && (2) \\ &= I(X^n; F_A|F_{A^c}) + \sum_{i\in A} I(F_i; Y_i^n|X^n) && (3) \\ &= H(X^n|F_{A^c}) - H(X^n|F_1, F_2) && (4) \\ &\quad + \sum_{i\in A}\sum_{j=1}^n I(Y_{i,j}; F_i|X^n, Y_i^{j-1}) \\ &\geq H(X^n|F_{A^c}) - nD && (5) \\ &\quad + \sum_{i\in A}\sum_{j=1}^n I(Y_{i,j}; F_i|X^n, Y_i^{j-1}) && (6) \\ &= \sum_{j=1}^n H(X_j|F_{A^c}, X^{j-1}) - nD && (7) \\ &\quad + \sum_{i\in A}\sum_{j=1}^n I(Y_{i,j}; F_i|X^n, Y_i^{j-1}) \\ &= \sum_{j=1}^n H(X_j|F_{A^c}, X^{j-1}) - nD && (8) \\ &\quad + \sum_{i\in A}\sum_{j=1}^n I(Y_{i,j}; U_{i,j}|X_j, Q_j) && (9) \\ &\geq \sum_{j=1}^n H(X_j|U_{A^c,j}, Q_j) - nD && (10) \\ &\quad + \sum_{i\in A}\sum_{j=1}^n I(Y_{i,j}; U_{i,j}|X_j, Q_j). && (11) \end{aligned}$$

The nontrivial steps above can be justified as follows:

- (2) follows since $F_A$ is a function of $Y_A^n$.
- (3) follows since $F_i$ is a function of $Y_i^n$ and $F_1 \leftrightarrow X^n \leftrightarrow F_2$ form a Markov chain (since $Y_1^n \leftrightarrow X^n \leftrightarrow Y_2^n$ form a Markov chain).
- (6) follows since $nD \geq H(X^n|F_1, F_2)$ by Lemma 1.
- (9) follows from the Markov chain $Y_{i,j} \leftrightarrow X^n \leftrightarrow Y_i^{j-1}$, which follows from the i.i.d. nature of the source sequences.
- (11) simply follows from the fact that conditioning reduces entropy.

Therefore, dividing both sides by $n$, we have:

$$\begin{aligned} \sum_{i\in A} R_i \geq \frac{1}{n}\sum_{j=1}^n H(X_j|U_{A^c,j}, Q_j) \\ + \sum_{i\in A}\frac{1}{n}\sum_{j=1}^n I(Y_{i,j}; U_{i,j}|X_j, Q_j) - D. \end{aligned}$$

Also, using Lemma 1 and the fact that conditioning reduces entropy:

$$D \geq \frac{1}{n}H(X^n|F_1, F_2) \geq \frac{1}{n}\sum_{j=1}^n H(X_j|U_{1,j}, U_{2,j}, Q_j).$$

Observe that $Q_j$ is independent of $(X_j, Y_{1,j}, Y_{2,j})$ and, conditioned on $Q_j$, we have the long Markov chain $U_{1,j} \leftrightarrow Y_{1,j} \leftrightarrow X_j \leftrightarrow Y_{2,j} \leftrightarrow U_{2,j}$. Finally, by a standard time-sharing argument, we conclude by observing that if $(R_1, R_2, D)$ is strict-sense achievable for the CEO problem, then

$$\begin{aligned} R_1 &\geq I(Y_1; U_1|X, Q) + H(X|U_2, Q) - D \\ R_2 &\geq I(Y_2; U_2|X, Q) + H(X|U_1, Q) - D \\ R_1 + R_2 &\geq I(U_1; Y_1|X, Q) + I(U_2; Y_2|X, Q) + H(X) - D \\ D &\geq H(X|U_1, U_2, Q). \end{aligned}$$

for some joint distribution of the form

$$p(q)p(x, y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q). \quad (12)$$

Since $R_1, R_2 \geq 0$, the theorem follows. ∎

*Theorem 3:* $\mathcal{RD}^o_{CEO} = \mathcal{RD}^i_{CEO} = \overline{\mathcal{RD}}^\star_{CEO}$.

*Proof:* We first remark that the cardinality bounds on the alphabets in the definition of $\mathcal{RD}^i_{CEO}$ can be imposed without any loss of generality. This is a consequence of [22, Lemma 2.2] and is discussed in detail in Appendix A.

Therefore, it will suffice to show $\mathcal{RD}^o_{CEO} \subseteq \mathcal{RD}^i_{CEO}$ without considering the cardinality bounds. To this end, fix $p(q)$, $p(u_1|y_1, q)$, and $p(u_2|y_2, q)$ and consider the extreme

points[1] of polytope defined by the inequalities (1):

$$P_1 = \Big(0, 0, I(Y_1; U_1|X, Q) + I(Y_2; U_2|X, Q) + H(X)\Big)$$
$$P_2 = \Big(I(Y_1; U_1|Q), 0, I(U_2; Y_2|X, Q) + H(X|U_1, Q)\Big)$$
$$P_3 = \Big(0, I(Y_2; U_2|Q), I(U_1; Y_1|X, Q) + H(X|U_2, Q)\Big)$$
$$P_4 = \Big(I(Y_1; U_1|Q), I(Y_2; U_2|U_1, Q), H(X|U_1, U_2, Q)\Big)$$
$$P_5 = \Big(I(Y_1; U_1|U_2, Q), I(Y_2; U_2|Q), H(X|U_1, U_2, Q)\Big),$$

where the point $P_j$ is a triple $(R_1^{(j)}, R_2^{(j)}, D^{(j)})$. We say a point $(R_1^{(j)}, R_2^{(j)}, D^{(j)})$ is *dominated by* a point in $\mathcal{RD}_{CEO}^{i}$ if there exists some $(R_1, R_2, D) \in \mathcal{RD}_{CEO}^{i}$ for which $R_1 \le R_1^{(j)}$, $R_2 \le R_2^{(j)}$, and $D \le D^{(j)}$. Observe that each of the extreme points $P_1, \ldots, P_5$ is dominated by a point in $\mathcal{RD}_{CEO}^{i}$:

- First, observe that $P_4$ and $P_5$ are both in $\mathcal{RD}_{CEO}^{i}$, so these points are not problematic.
- Next, observe that the point $(0, 0, H(X))$ is in $\mathcal{RD}_{CEO}^{i}$, which can be seen by setting all auxiliary random variables to be constant. This point dominates $P_1$.
- By using auxiliary random variables $(\hat{U}_1, \hat{U}_2, Q) = (U_1, \emptyset, Q)$, the point $(I(Y_1; U_1|Q), 0, H(X|U_1, Q))$ is in $\mathcal{RD}_{CEO}^{i}$, and dominates the point $P_2$. By a symmetric argument, the point $P_3$ is also dominated by a point in $\mathcal{RD}_{CEO}^{i}$.

Since $\mathcal{RD}_{CEO}^{o}$ is the convex hull of all such extreme points (i.e., the convex hull of the union of extreme points over all appropriate joint distributions), the theorem is proved. ∎

*Remark 1: Theorem 3 can be extended to the general case of m-encoders. Details are provided in Appendix B.*

### C. A Stronger Converse Result for the CEO Problem

As defined, our reproduction sequence $\hat{X}^n$ is an $n$-tuple of distributions on $\mathcal{X}$, which we identify with a product distribution on $\mathcal{X}^n$ in the natural way. However, for a blocklength $n$ code, we can allow $\hat{X}^n$ to be *any* probability distribution on $\mathcal{X}^n$ and the converse result still holds. In this case, we define the sequence distortion as follows:

$$d(x^n, \hat{x}^n) = \frac{1}{n} \log\left(\frac{1}{\hat{x}^n(x^n)}\right),$$

which is compatible with the original definition when $\hat{X}^n$ is a product distribution. The reader can verify that the result of Lemma 1 is still true for this more general distortion alphabet by setting $s(x^n|z) = \hat{x}^n(x^n)$ in the corresponding proof. Since Lemma 1 is the key tool in the CEO converse result, this implies that the converse holds even if $\hat{X}^n$ is allowed to be any probability distribution on $\mathcal{X}^n$ (rather than being restricted to the set of product distributions).

[1]For two encoders, it is easy enough to enumerate the extreme points by inspection. However, this can be formalized by a submodularity argument, which is given in Appendix B.

When this stronger converse result is taken together with the achievability result, we observe that restricting $\hat{X}^n$ to be a product distribution is in fact optimal and can achieve all points in $\overline{\mathcal{RD}}_{CEO}^{\star}$.

### D. An Example: Distributed Compression of a Posterior Distribution

Suppose two sensors observe sequences $Y_1^n$ and $Y_2^n$ respectively, which are conditionally independent given a hidden sequence $X^n$. The sensors communicate with a fusion center through rate-limited links of capacity $R_1$ and $R_2$ respectively. Given sequences $(Y_1^n, Y_2^n)$ are observed, the sequence $X^n$ cannot be determined in general, so the fusion center would like to estimate the posterior distribution $p(x^n|Y_1^n, Y_2^n)$. Since the communication links are rate-limited, the fusion center cannot necessarily compute $p(x^n|Y_1^n, Y_2^n)$ exactly. In this case, the fusion center would like to generate an estimate $\hat{p}(x^n|g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n))$ that should approximate $p(x^n|Y_1^n, Y_2^n)$ in the sense that, on average:

$$D\Big(p(x^n|y_1^n, y_2^n)\Big\|\hat{p}(x^n|g_1^{(n)}(y_1^n), g_2^{(n)}(y_2^n))\Big) \le n\varepsilon,$$

where, consistent with standard notation (e.g. [23]), we write $D(p(x^n|y_1^n, y_2^n)\|\hat{p}(x^n|g_1^{(n)}(y_1^n), g_2^{(n)}(y_2^n)))$ as shorthand for

$$\sum_{x^n, y_1^n, y_2^n} p(x^n, y_1^n, y_2^n) \log \frac{p(x^n|y_1^n, y_2^n)}{\hat{p}(x^n|g_1^{(n)}(y_1^n), g_2^{(n)}(y_2^n))}.$$

The relevant question here is the following. What is the minimum distortion $\varepsilon$ that is attainable given $R_1$ and $R_2$?

Considering the CEO problem for this setup, we have:

$$\begin{aligned}\mathbb{E}d(\hat{X}^n, X^n) &= \frac{1}{n}\sum_{(x^n, y_1^n, y_2^n)} p(x^n, y_1^n, y_2^n) \log\left(\frac{1}{\hat{x}^n(x^n)}\right) \\ &= \frac{1}{n} D\Big(p(x^n|y_1^n, y_2^n)\Big\|\hat{x}^n(x^n)\Big) + \frac{1}{n} H(X^n|Y_1^n, Y_2^n).\end{aligned}$$

Identifying $\hat{p}(x^n|g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)) \leftarrow \hat{X}^n(x^n)$, we have:

$$\begin{aligned}&D\Big(p(x^n|y_1^n, y_2^n)\Big\|\hat{p}(x^n|g_1^{(n)}(y_1^n), g_2^{(n)}(y_2^n))\Big) \\ &\quad = n\mathbb{E}d(\hat{X}^n, X^n) - nH(X|Y_1, Y_2).\end{aligned}$$

Thus, finding the minimum possible distortion reduces to an optimization problem over $\overline{\mathcal{RD}}_{CEO}^{\star}$. In particular, the minimum attainable distortion $\varepsilon^*$ is given by

$$\varepsilon^* = \inf\Big\{D : (R_1, R_2, D) \in \overline{\mathcal{RD}}_{CEO}^{\star}\Big\} - H(X|Y_1, Y_2). \quad (13)$$

Moreover, the minimum distortion is obtained by estimating each $x_j$ separately. In other words, there exists an optimal (essentially, for large $n$) estimate $\hat{p}^*(x^n|\cdot, \cdot)$ (which is itself a function of optimal encoding functions $g_1^{*(n)}(\cdot)$ and $g_2^{*(n)}(\cdot)$) that can be expressed as a product distribution

$$\hat{p}^*(x^n|\cdot, \cdot) = \prod_{j=1}^{n} \hat{p}_j^*\left(x_j|g_1^{*(n)}(\cdot), g_2^{*(n)}(\cdot)\right).$$

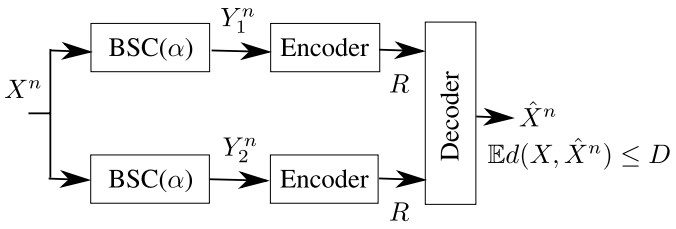


Fig. 2. An example CEO problem where $X \sim \text{Bernoulli}(\frac{1}{2})$, $\Pr(Y_i = X) = (1-\alpha)$, and both encoders are subject to the same rate constraint.

For this choice of $\hat{p}^*(x^n|\cdot,\cdot)$, we have the following relationship:

$$\frac{1}{n}\sum_{j=1}^{n} D\Big(p(x_j|y_{1,j}, y_{2,j})\Big\|\hat{p}_j^*\left(x_j|g_1^{*(n)}(y_1^n), g_2^{*(n)}(y_2^n)\right)\Big) = \varepsilon^*.$$

In light of this fact, we can apply Markov's inequality to obtain an estimate on peak component-wise distortion. Namely the number of coordinates $j$ for which

$$D\Big(p(x_j|y_{1,j}, y_{2,j})\Big\|\hat{p}_j^*\left(x_j|g_1^{*(n)}(y_1^n), g_2^{*(n)}(y_2^n)\right)\Big) \geq \zeta$$

is at most $n\varepsilon^*/\zeta$.

To make this example more concrete, consider the scenario depicted in Figure 2, where $X \sim \text{Bernoulli}(\frac{1}{2})$ and $Y_i$ is the result of passing $X$ through a binary symmetric channel with crossover probability $\alpha$ for $i = 1, 2$. To simplify things, we constrain the rates of each encoder to be at most $R$ bits per channel use.

By performing a brute-force search over a fine mesh of conditional distributions $\{p(u_i|y_i)\}_{i=1}^2$, we numerically approximate the set of $(R, D)$ pairs such that $(R, R, D)$ is in the achievable region $\overline{\mathcal{RD}}_{CEO}^{\star}$ corresponding to the network in Figure 2. The lower convex envelope of these $(R, D)$ pairs is plotted in Figure 3 for $\alpha \in \{0.01, 0.1, 0.25\}$. Continuing our example above for this concrete choice of source parameters, we compute the minimum achievable Kullback-Leibler distance $\varepsilon^*$ according to (13). The result is given in Figure 4.

These numerical results are intuitively satisfying in the sense that, if $Y_1, Y_2$ are high-quality estimates of $X$ (e.g., $\alpha = 0.01$), then a small increase in the allowable rate $R$ results in a large relative improvement of $\hat{p}(x|\cdot,\cdot)$, the decoder's estimate of $p(x|Y_1, Y_2)$. On the other hand, if $Y_1, Y_2$ are poor-quality estimates of $X$ (e.g., $\alpha = 0.25$), then we require a large increase in the allowable rate $R$ in order to obtain an appreciable improvement of $\hat{p}(x|\cdot,\cdot)$.

One field where this example is directly applicable is machine learning. In this case, $X_j$ could represent the class of object $j$, and $Y_{1,j}, Y_{2,j}$ are observable attributes. In machine learning, one typically estimates the probability that an object belongs to a particular class given a set of observable attributes. For this type of estimation problem, relative entropy is a natural penalty criterion.

Another application is to horse-racing with conditionally independent, rate-limited side information sequences. In this case, the doubling rate of the gambler's wealth can be expressed in terms of the logarithmic loss distortion measure. This example is consistent with the original interpretation of the CEO problem, where the CEO makes consecutive business decisions (investments) pertaining to outcomes $X^n$, with the objective of maximizing the wealth of the company. We omit the details.

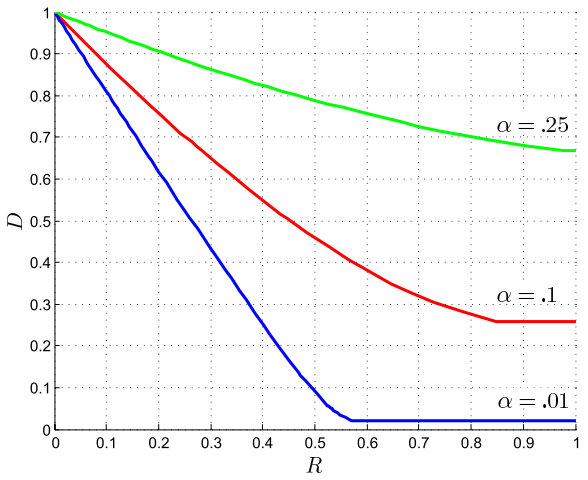


Fig. 3. The distortion-rate function of the network in Figure 2 computed for $\alpha \in \{0.01, 0.1, 0.25\}$.

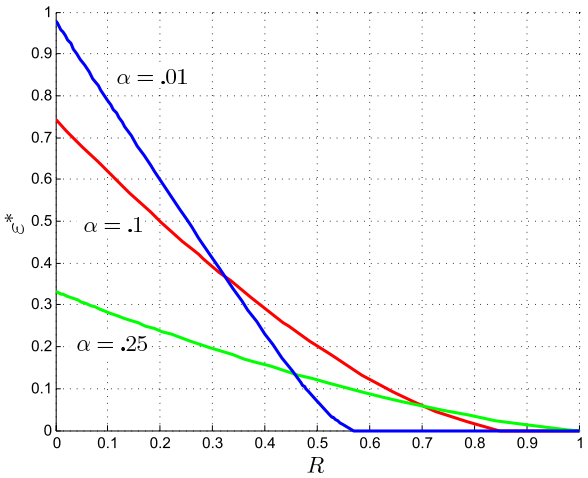


Fig. 4. The minimum achievable Kullback-Leibler distance computed according to (13), i.e., the curves here are those of Figure 3, lowered by the constant $H(X|Y_1, Y_2)$.

### E. An Example: Joint Estimation of the Encoder Observations

Suppose one wishes to estimate the encoder observations $(Y_1, Y_2)$. In this case, the rate region simplifies considerably. In particular, if we tolerate a distortion $D$ in our estimate of the pair $(Y_1, Y_2)$, then the achievable rate region is the same as the Slepian-Wolf rate region with each rate constraint relaxed by $D$ bits. Formally:

*Theorem 4: If $X = (Y_1, Y_2)$, then $\overline{\mathcal{RD}}_{CEO}^{\star}$ consists of all vectors $(R_1, R_2, D)$ satisfying*

$$\begin{aligned} R_1 &\geq [H(Y_1|Y_2) - D]^+ \\ R_2 &\geq [H(Y_2|Y_1) - D]^+ \\ R_1 + R_2 &\geq [H(Y_1, Y_2) - D]^+ \\ D &\geq 0. \end{aligned}$$

*Proof:* First, note that Theorem 3 implies that $\overline{\mathcal{RD}}^{\star}_{CEO}$ is equivalent to the the union of $(R_1, R_2, D)$ triples satisfying (1) taken over all joint distributions $p(q)p(x, y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)$. Now, since $X = (Y_1, Y_2)$, each of the inequalities (1) can be lower bounded as follows:

$$\begin{aligned}
R_1 &\geq I(Y_1; U_1|Y_1, Y_2, Q) + H(Y_1, Y_2|U_2, Q) - D \\
&= H(Y_2|U_2, Q) + H(Y_1|Y_2) - D \\
&\geq H(Y_1|Y_2) - D \\
R_2 &\geq I(Y_2; U_2|Y_1, Y_2, Q) + H(Y_1, Y_2|U_1, Q) - D \\
&= H(Y_1|U_1, Q) + H(Y_2|Y_1) - D \\
&\geq H(Y_2|Y_1) - D \\
R_1 + R_2 &\geq I(U_1; Y_1|Y_1, Y_2, Q) + I(U_2; Y_2|Y_1, Y_2, Q) \\
&\quad + H(Y_1, Y_2) - D \\
&= H(Y_1, Y_2) - D \\
D &\geq H(Y_1, Y_2|U_1, U_2, Q) \geq 0.
\end{aligned}$$

Finally, observe that by setting $U_i = Y_i$ for $i = 1, 2$, we can achieve any point in this relaxed region (again, a consequence of Theorem 3). ∎

We remark that this result was first proved in [15] by Courtade and Wesel using a different method.

### F. An Example: The Information Bottleneck Method

If we consider the CEO problem with a single observed source (i.e., $Y_2 = \emptyset$), then the achievable rate distortion region given by Theorem 3 is characterized by all $(R_1, D)$ pairs satisfying

$$\begin{aligned}
R_1 &\geq I(Y_1; U_1) \\
D &\geq H(X|U_1)
\end{aligned}$$

for some $U_1$ satisfying the Markov chain $X \leftrightarrow Y_1 \leftrightarrow U_1$. Alternatively, by making the substitution $\tau = H(X) - D$, this tradeoff can be characterized as follows:

$$R_1(\tau) = \min_{p(u_1|y_1): I(U_1;X) \geq \tau} I(Y_1; U_1). \tag{14}$$

Expression (14) is known as the *Information Bottleneck Function* (cf. [24]). Intuitively, $U_1$ is a description of $X$ which is generated (stochastically) from the observation $Y_1$. The function $R_1(\tau)$ describes the tradeoff between the complexity and the accuracy of the description $U_1$. Ideally, $U_1$ should capture the relevant information about $X$ present in the observation $Y_1$.

The concept of the Information Bottleneck was first introduced by Tishby et al. in [16], and the first formal rate distortion theorem on the topic was later proved by Gilad-Bachrach et al. in [24]. We remark that algorithms motivated by the Information Bottleneck Method have been successfully applied to a wide variety of problems. Examples include word clustering for text classification [25], galaxy spectra classification [26], neural code analysis [27], and speech recognition [28]. Since Theorem 3 (and the $m$-encoder extension given in Appendix B) generalize the tradeoff (14) to a distributed setting, our results could be applied to similar problems. Particularly those for which processing and computation occurs in a distributed or parallel manner.

## IV. Multiterminal Source Coding

With Theorem 3 in hand, we are now in a position to characterize the achievable rate distortion region $\overline{\mathcal{RD}}^{\star}$ for the multiterminal source coding problem under logarithmic loss. As before, we prove an inner bound first.

### A. Inner Bound

*Definition 5: Let $(R_1, R_2, D_1, D_2) \in \mathcal{RD}^i$ if and only if there exists a joint distribution of the form*

$$p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)p(q)$$

*where $|\mathcal{U}_1| \leq |\mathcal{Y}_1|$, $|\mathcal{U}_2| \leq |\mathcal{Y}_2|$, and $|\mathcal{Q}| \leq 5$, which satisfies*

$$\begin{aligned}
R_1 &\geq I(Y_1; U_1|U_2, Q) \\
R_2 &\geq I(Y_2; U_2|U_1, Q) \\
R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Q) \\
D_1 &\geq H(Y_1|U_1, U_2, Q) \\
D_2 &\geq H(Y_2|U_1, U_2, Q).
\end{aligned}$$

*Theorem 5: $\mathcal{RD}^i \subseteq \overline{\mathcal{RD}}^{\star}$. That is, all rate distortion vectors in $\mathcal{RD}^i$ are achievable.*

Again, we require an appropriate version of the Berger-Tung inner bound:

*Proposition 2 (Berger-Tung Inner Bound [19], [20]): The rate distortion vector $(R_1, R_2, D_1, D_2)$ is achievable if*

$$\begin{aligned}
R_1 &\geq I(U_1; Y_1|U_2, Q) \\
R_2 &\geq I(U_2; Y_2|U_1, Q) \\
R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Q) \\
D_1 &\geq \mathbb{E}\left[d(Y_1, f_1(U_1, U_2, Q))\right] \\
D_2 &\geq \mathbb{E}\left[d(Y_2, f_2(U_1, U_2, Q))\right].
\end{aligned}$$

*for a joint distribution*

$$p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)p(q)$$

*and reproduction functions*

$$f_i : \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Q} \to \hat{\mathcal{Y}}_i, \qquad \text{for } i = 1, 2.$$

*Proof of Theorem 5:* To prove the theorem, we simply apply Proposition 2 with the reproduction functions $f_i(U_1, U_2, Q) := \Pr[Y_i = y_i|U_1, U_2, Q]$. ∎

Hence, we again see that our inner bound $\mathcal{RD}^i \subseteq \overline{\mathcal{RD}}^{\star}$ is nothing more than the Berger-Tung inner bound specialized to the setting when distortion is measured under logarithmic loss.

### B. A Matching Outer Bound

The main result of this paper is the following theorem.

*Theorem 6: $\mathcal{RD}^i = \overline{\mathcal{RD}}^{\star}$.*

*Proof:* As before, we note that the cardinality bounds on the alphabets in the definition of $\mathcal{RD}^i$ can be imposed without any loss of generality. This is discussed in detail in Appendix A.

Assume $(R_1, R_2, D_1, D_2)$ is strict-sense achievable. Observe that proving that $(R_1, R_2, D_1, D_2) \in \mathcal{RD}^i$ will prove the theorem, since $\mathcal{RD}^i \subseteq \overline{\mathcal{RD}}^{\star}$ and $\overline{\mathcal{RD}}^{\star}$ is closed by definition.

Define $\mathcal{P}(R_1, R_2)$ to be the set of joint distributions of the form

$$p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)p(q)$$

with $|\mathcal{U}_1| \leq |\mathcal{Y}_1|$, $|\mathcal{U}_2| \leq |\mathcal{Y}_2|$, and $|\mathcal{Q}| \leq 4$ satisfying

$$\begin{aligned} R_1 &\geq I(U_1; Y_1|U_2, Q) \\ R_2 &\geq I(U_2; Y_2|U_1, Q) \\ R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Q). \end{aligned}$$

We remark that $\mathcal{P}(R_1, R_2)$ is compact. We also note that it will suffice to show the existence of a joint distribution in $\mathcal{P}(R_1, R_2)$ satisfying $H(Y_1|U_1, U_2, Q) \leq D_1$ and $H(Y_2|U_1, U_2, Q) \leq D_2$ to prove that $(R_1, R_2, D_1, D_2) \in \mathcal{RD}^i$.

With foresight, consider random variable $X$ defined as follows

$$X = \begin{cases} (Y_1, 1) & \text{with probability } t \\ (Y_2, 2) & \text{with probability } 1 - t. \end{cases} \tag{15}$$

In other words, $X = (Y_B, B)$, where $B$ is a Bernoulli random variable independent of $Y_1, Y_2$. Observe that $Y_1 \leftrightarrow X \leftrightarrow Y_2$ form a Markov chain, and thus, we are able to apply Theorem 3.

Since $(R_1, R_2, D_1, D_2)$ is strict-sense achievable, the decoder can construct reproductions $\hat{Y}_1^n, \hat{Y}_2^n$ satisfying

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}d(Y_{i,j}, \hat{Y}_{i,j}) \leq D_i \qquad \text{for } i = 1, 2.$$

Fix the encoding operations and set $\hat{X}_j((y_1, 1)) = t\hat{Y}_{1,j}(y_1)$ and $\hat{X}_j((y_2, 2)) = (1-t)\hat{Y}_{2,j}(y_2)$. Then for the CEO problem defined by $(X, Y_1, Y_2)$:

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}d(X_j, \hat{X}_j) \\ &\quad = \frac{t}{n}\sum_{j=1}^{n} \mathbb{E}\log\left(\frac{1}{t\hat{Y}_{1,j}(Y_{1,j})}\right) \qquad (16) \\ &\qquad + \frac{1-t}{n}\sum_{j=1}^{n} \mathbb{E}\log\left(\frac{1}{(1-t)\hat{Y}_{2,j}(Y_{2,j})}\right) \\ &\quad = h_2(t) + \frac{t}{n}\sum_{j=1}^{n} \mathbb{E}d(Y_{1,j}, \hat{Y}_{1,j}) \qquad (17) \\ &\qquad + \frac{1-t}{n}\sum_{j=1}^{n} \mathbb{E}d(Y_{2,j}, \hat{Y}_{2,j}) \\ &\quad \leq h_2(t) + tD_1 + (1-t)D_2 \end{aligned}$$

where $h_2(t)$ is the binary entropy function. Hence, for this CEO problem, distortion $h_2(t)+tD_1+(1-t)D_2$ is achievable and Theorem 3 implies existence of a joint distribution[2] $P_t \in \mathcal{P}(R_1, R_2)$ satisfying

$$\begin{aligned} h_2(t) + tD_1 + (1-t)D_2 &\geq H(X|U_1^{(t)}, U_2^{(t)}, Q^{(t)}) \\ &= h_2(t) + tH(Y_1|U_1^{(t)}, U_2^{(t)}, Q^{(t)}) \\ &\quad + (1-t)H(Y_2|U_1^{(t)}, U_2^{(t)}, Q^{(t)}), \end{aligned}$$

where the second equality follows by definition of $X$ in (15).

Now, we "tune" the parameter $t$ to yield the desired result. Defining $H_1(P_t) \triangleq H(Y_1|U_1^{(t)}, U_2^{(t)}, Q^{(t)})$ and $H_2(P_t) \triangleq H(Y_2|U_1^{(t)}, U_2^{(t)}, Q^{(t)})$, we note the following two facts:

1) By continuity of entropy, the functions $H_1(\cdot)$ and $H_2(\cdot)$ are continuous on the compact domain $\mathcal{P}(R_1, R_2)$.
2) The above argument proves the existence of a function $\varphi : [0, 1] \to \mathcal{P}(R_1, R_2)$ which satisfies

$$tH_1(\varphi(t)) + (1-t)H_2(\varphi(t)) \leq tD_1 + (1-t)D_2$$

for all $t \in [0, 1]$.

These two facts satisfy the requirements of Lemma 7 (see Appendix D), and hence there exists $P_{t_1} \in \mathcal{P}(R_1, R_2)$, $P_{t_2} \in \mathcal{P}(R_1, R_2)$, and $\theta \in [0, 1]$ for which

$$\begin{aligned} \theta H_1(P_{t_1}) + (1-\theta)H_1(P_{t_2}) &\leq D_1 \\ \theta H_2(P_{t_1}) + (1-\theta)H_2(P_{t_2}) &\leq D_2. \end{aligned}$$

Timesharing[3] between distributions $P_{t_1}$ and $P_{t_2}$ with probabilities $\theta$ and $(1-\theta)$, respectively, yields a distribution $P^* \in \mathcal{P}(R_1, R_2)$ which satisfies $H_1(P^*) \leq D_1$ and $H_2(P^*) \leq D_2$. This proves the theorem. ■

### C. A Stronger Converse

For the CEO problem, we are able to obtain a stronger converse result as discussed in Section III-C. We can obtain a similar result for the multiterminal source coding problem. Indeed, the converse result we just proved continues to hold even when $\hat{y}_i^n$ is allowed to be any probability measure on $\mathcal{Y}_i^n$, rather than a product distribution. The proof of this fact is somewhat involved and can be found in Appendix E.

We note that the proof of this strengthened converse result (i.e., Theorem 12 in Appendix E) offers a direct proof of the converse of Theorem 6, and as such we do not require a CEO result (Theorem 3) or the "tuning argument" given by Lemma 7. At the heart of this alternative proof lies the Csiszár sum identity (and a careful choice of auxiliary random variables) which provides a coupling between the attainable distortions for each source. In the original proof of Theorem 6, this coupling is accomplished by the tuning argument through Lemma 7.

Interestingly, the two proofs are similar in spirit, with the key differences being the use of the Csiszár sum identity versus the tuning argument. Intuitively, the original tuning argument given in the above proof of Theorem 6 allows a simpler choice of auxiliary random variables which leads to a more elegant and transparent proof, but appears incapable of establishing the strengthened converse. On the other hand, applying the Csiszár sum identity requires a very careful choice of auxiliary random

[2]Henceforth, we use the superscript $(t)$ to explicitly denote the dependence of the auxiliary random variables on the distribution parametrized by $t$.

[3]The timesharing scheme can be embedded in the timesharing variable $Q$, increasing the cardinality of $\mathcal{Q}$ by a factor of two.

variables which, in turn, affords a finer degree of control over various quantities.

### D. An Example: The Daily Double

The *Daily Double* is a single bet that links together wagers on the winners of two consecutive horse races. Winning the Daily Double is dependent on both wagers winning together. In general, the outcomes of two consecutive races can be correlated (e.g. due to track conditions), so a gambler can potentially use this information to maximize his expected winnings. Let $\mathcal{Y}_1$ and $\mathcal{Y}_2$ be the set of horses running in the first and second races respectively. If horses $y_1$ and $y_2$ win their respective races, then the payoff is $o(y_1, y_2)$ dollars for each dollar invested in outcome $(Y_1, Y_2) = (y_1, y_2)$. The quantity $o(y_1, y_2)$ is called the *odds* function.

There are two betting strategies one can follow:

1) The gambler can wager a fraction $b_1(y_1)$ of his wealth on horse $y_1$ winning the first race and parlay his winnings by betting a fraction $b_2(y_2)$ of his wealth on horse $y_2$ winning the second race. In this case, the gambler's wealth relative is $b_1(Y_1)b_2(Y_2)o(Y_1, Y_2)$ upon learning the outcome of the Daily Double. We refer to this betting strategy as the *product-wager*.
2) The gambler can wager a fraction $b(y_1, y_2)$ of his wealth on horses $(y_1, y_2)$ winning the first and second races, respectively. In this case, the gambler's wealth relative is $b(Y_1, Y_2)o(Y_1, Y_2)$ upon learning the outcome of the Daily Double. We refer to this betting strategy as the *joint-wager*.

Clearly the joint-wager includes the product-wager as a special case. However, the product-wager requires less effort to place, so the question is: how do the two betting strategies compare?

To make things interesting, suppose the gamblers have access to rate-limited information about the first and second race outcomes at rates $R_1, R_2$ respectively.[4] Further, assume that $R_1 \le H(Y_1)$, $R_2 \le H(Y_2)$, and $R_1 + R_2 \le H(Y_1, Y_2)$. For $(R_1, R_2)$ and $p(y_1, y_2)$ given, let $\mathcal{P}(R_1, R_2)$ denote the set of joint pmf's of the form

$$p(q, y_1, y_2, u_1, u_2) = p(q)p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)$$

which satisfy

$$\begin{aligned} R_1 &\ge I(Y_1; U_1|U_2, Q) \\ R_2 &\ge I(Y_2; U_2|U_1, Q) \\ R_1 + R_2 &\ge I(Y_1, Y_2; U_1, U_2|Q) \end{aligned}$$

for alphabets $\mathcal{U}_1, \mathcal{U}_2, \mathcal{Q}$ satisfying $|\mathcal{U}_i| \le |\mathcal{Y}_i|$ and $|Q| \le 5$.

Typically, the quality of a bet is measured by the associated doubling rate (cf. [23]). Theorem 6 implies that the optimal doubling rate for the product-wager is given by:

$$\begin{aligned} W^*_{\text{p-w}}(p(y_1, y_2)) &= \sum_{y_1, y_2} p(y_1, y_2) \log b^*_1(y_1) b^*_2(y_2) o(y_1, y_2) \\ &= \mathbb{E} \log o(Y_1, Y_2) \\ &\quad - \inf_{p \in \mathcal{P}(R_1, R_2)} \Big\{ H(Y_1|U_1, U_2, Q) + H(Y_2|U_1, U_2, Q) \Big\}. \end{aligned}$$

[4]For example, the separately encoded side information could come from two different experts, each of which are knowledgeable about only one race.

Likewise, Theorem 4 implies that the optimal doubling rate for the joint-wager is given by:

$$\begin{aligned} &W^*_{\text{j-w}}(p(y_1, y_2)) \\ &= \sum_{y_1, y_2} p(y_1, y_2) \log b^*(y_1, y_2) o(y_1, y_2) \\ &= \mathbb{E} \log o(Y_1, Y_2) + \min \Big\{ R_1 - H(Y_1|Y_2), R_2 - H(Y_2|Y_1), \\ &\qquad\qquad R_1 + R_2 - H(Y_1, Y_2) \Big\}. \end{aligned}$$

It is important to note that we do not require the side informations to be the same for each type of wager, rather, the side informations are only provided at the same rates. Thus, the gambler placing the joint-wager receives side information at rates $(R_1, R_2)$ that maximizes his doubling rate, while the gambler placing the product-wager receives (potentially different) side information at rates $(R_1, R_2)$ that maximizes his doubling rate. However, as we will see shortly, for any rates $(R_1, R_2)$, there always exists rate-limited side information which simultaneously allows each type of gambler to attain their maximum doubling rate.

By combining the expressions for $W^*_{\text{p-w}}(p(y_1, y_2))$ and $W^*_{\text{j-w}}(p(y_1, y_2))$, we find that the difference in doubling rates is given by:

$$\begin{aligned} &\Delta(R_1, R_2) = W^*_{\text{j-w}}(p(y_1, y_2)) - W^*_{\text{p-w}}(p(y_1, y_2)) \\ &= \min \Big\{ R_1 - H(Y_1|Y_2), R_2 - H(Y_2|Y_1), \\ &\qquad R_1 + R_2 - H(Y_1, Y_2) \Big\} \\ &\quad + \inf_{p \in \mathcal{P}(R_1, R_2)} \{ H(Y_1|U_1, U_2, Q) + H(Y_2|U_1, U_2, Q) \} \end{aligned} \tag{18}$$

$$\begin{aligned} &= \inf_{p \in \mathcal{P}(R_1, R_2)} \min \Big\{ R_1 - I(Y_1; U_1|U_2, Q) + I(Y_1; Y_2) \\ &\qquad\qquad - I(Y_1; U_2, Q) + H(Y_2|U_1, U_2, Q), \\ &\qquad\qquad R_2 - I(Y_2; U_2|U_1, Q) + I(Y_2; Y_1) \\ &\qquad\qquad - I(Y_2; U_1, Q) + H(Y_1|U_1, U_2, Q), \\ &\qquad\qquad R_1 + R_2 - I(Y_1, Y_2; U_1, U_2|Q) \\ &\qquad\qquad + I(Y_1; Y_2|U_1, U_2, Q) \Big\} \\ &= \inf_{p \in \mathcal{P}(R_1, R_2)} I(Y_1; Y_2|U_1, U_2, Q). \end{aligned} \tag{19}$$

The final equality (19) follows since

- $R_1 \ge I(Y_1; U_1|U_2, Q)$ and $R_2 \ge I(Y_2; U_2|U_1, Q)$ for any $p \in \mathcal{P}(R_1, R_2)$.
- $I(Y_2; Y_1) \ge I(Y_2; U_1, Q)$ and $I(Y_1; Y_2) \ge I(Y_1; U_2, Q)$ for any $p \in \mathcal{P}(R_1, R_2)$ by the data processing inequality.
- The infimum in (18) is attained by a $p \in \mathcal{P}(R_1, R_2)$ satisfying $R_1 + R_2 = I(Y_1, Y_2; U_1, U_2|Q)$. See Lemma 10 in Appendix F for details.
- By definition of conditional mutual information,

$$H(Y_i|U_1, U_2, Q) \ge I(Y_1; Y_2|U_1, U_2, Q)$$

for $i = 1, 2$.

Let $p^* \in \mathcal{P}(R_1, R_2)$ be the distribution that attains the infimum in (18) (such a $p^*$ always exists), then (19) yields

$$\begin{aligned} &W^*_{\text{j-w}}(p(y_1, y_2)) - W^*_{\text{p-w}}(p(y_1, y_2)) \\ &= \mathbb{E}_{p^*} \log \left[o(Y_1, Y_2) p^*(Y_1, Y_2|U_1, U_2, Q)\right] \\ &- \mathbb{E}_{p^*} \log \left[o(Y_1, Y_2) p^*(Y_1|U_1, U_2, Q) p^*(Y_2|U_1, U_2, Q)\right]. \end{aligned}$$

Hence, we can interpret the auxiliary random variables corresponding to $p^*$ as optimal rate-limited side informations for *both* betting strategies. Moreover, optimal bets for each strategy are given by

1) $b^*(y_1, y_2) = p^*(y_1, y_2|u_1, u_2, q)$ for the joint-wager, and
2) $b_1^*(y_1) = p^*(y_1|u_1, u_2, q)$, $b_2^*(y_2) = p^*(y_2|u_1, u_2, q)$ for the product-wager.

Since $\mathcal{P}(R_1, R_2) \subseteq \mathcal{P}(R_1', R_2')$ for $R_1 \leq R_1'$ and $R_2 \leq R_2'$, the function $\Delta(R_1, R_2)$ is nonincreasing in $R_1$ and $R_2$. Thus, the benefits of using the joint-wager over the product-wager diminish in the amount of side-information available. It is also not difficult to show that $\Delta(R_1, R_2)$ is jointly convex in $(R_1, R_2)$.

Furthermore, for rate-pairs $(R_1, R_2)$ and $(R_1', R_2')$ satisfying $R_1 < R_1'$ and $R_2 < R_2'$, there exist corresponding optimal joint- and product-wagers $b^*(y_1, y_2)$ and $b_1^*(y_1)b_2^*(y_2)$, and $b^{*'}(y_1, y_2)$ and $b_1^{*'}(y_1)b_2^{*'}(y_2)$, respectively, satisfying

$$\begin{aligned} &D\Big(b^{*'}(y_1, y_2)\Big|\Big|b_1^{*'}(y_1)b_2^{*'}(y_2)\Big) \\ &\quad < D\Big(b^*(y_1, y_2)\Big|\Big|b_1^*(y_1)b_2^*(y_2)\Big). \end{aligned} \tag{20}$$

So, roughly speaking, the joint-wager and product-wager look "more alike" as the amount of side information is increased. The proof of the strict inequality in (20) can be inferred from the proof of Lemma 10 in Appendix F.

### E. An Application: List Decoding

In the previous example, we did not take advantage of the stronger converse result which we proved in Appendix E (see the discussion in Section IV-C). In this section, we give an application that requires this strengthened result.

Formally, a 2-list code (of blocklength $n$ consists) of encoding functions:

$$g_i^{(n)} : \mathcal{Y}_i^n \to \left\{1, \ldots, M_i^{(n)}\right\} \qquad \text{for } i = 1, 2$$

and list decoding functions

$$\begin{aligned} L_1^{(n)} &: \left\{1, \ldots, M_1^{(n)}\right\} \times \left\{1, \ldots, M_2^{(n)}\right\} \to 2^{\mathcal{Y}_1^n} \\ L_2^{(n)} &: \left\{1, \ldots, M_1^{(n)}\right\} \times \left\{1, \ldots, M_2^{(n)}\right\} \to 2^{\mathcal{Y}_2^n}. \end{aligned}$$

A list decoding tuple $(R_1, R_2, \Delta_1, \Delta_2)$ is achievable if, for any $\epsilon > 0$, there exists a 2-list code of blocklength $n$ satisfying the rate constraints

$$\begin{aligned} \frac{1}{n} \log M_1^{(n)} &\leq R_1 + \epsilon \\ \frac{1}{n} \log M_2^{(n)} &\leq R_2 + \epsilon, \end{aligned}$$

and the probability of list-decoding error constraints

$$\begin{aligned} \Pr\left[Y_1^n \notin L_1^{(n)}\left(g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)\right)\right] &\leq \epsilon, \\ \Pr\left[Y_2^n \notin L_2^{(n)}\left(g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)\right)\right] &\leq \epsilon. \end{aligned}$$

with list sizes

$$\begin{aligned} \frac{1}{n} \log |L_1^{(n)}| &\leq \Delta_1 + \epsilon \\ \frac{1}{n} \log |L_2^{(n)}| &\leq \Delta_2 + \epsilon. \end{aligned}$$

With a 2-list code so defined, the following theorem shows that the 2-list decoding problem and multiterminal source coding problem under logarithmic loss are equivalent (inasmuch as the achievable regions are identical):

*Theorem 7: The list decoding tuple $(R_1, R_2, \Delta_1, \Delta_2)$ is achievable if and only if*

$$\begin{aligned} R_1 &\geq I(U_1; Y_1|U_2, Q) \\ R_2 &\geq I(U_2; Y_2|U_1, Q) \\ R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Q) \\ \Delta_1 &\geq H(Y_1|U_1, U_2, Q) \\ \Delta_2 &\geq H(Y_2|U_1, U_2, Q) \end{aligned}$$

*for some joint distribution*

$$p(y_1, y_2, u_1, u_2, q) = p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)p(q),$$

*where $|\mathcal{U}_1| \leq |\mathcal{Y}_1|$, $|\mathcal{U}_2| \leq |\mathcal{Y}_2|$, and $|\mathcal{Q}| \leq 5$.*

*Remark 2: We note that a similar connection to list decoding can be made for other multiterminal scenarios, in particular the CEO problem.*

To prove the theorem, we require a slightly modified version of [29, Lemma 1]:

*Lemma 2: If the list decoding tuple $(R_1, R_2, \Delta_1, \Delta_2)$ is achieved by a sequence of 2-list codes $\{g_1^{(n)}, g_2^{(n)}, L_1^{(n)}, L_2^{(n)}\}_{n\to\infty}$, then*

$$\begin{aligned} H(Y_1^n|g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)) &\leq |L_1^{(n)}| + n\epsilon_n \\ H(Y_2^n|g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)) &\leq |L_2^{(n)}| + n\epsilon_n, \end{aligned}$$

*where $\epsilon_n \to 0$ as $n \to \infty$.*

*Proof:* The proof is virtually identical to that of [29, Lemma 1], and is therefore omitted. ■

*Proof of Theorem 7:* The direct part is straightforward. Indeed, for a joint distribution $p(y_1, y_2, u_1, u_2, q) = p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)p(q)$, apply the Berger-Tung achievability scheme and take $L_i^{(n)}$ to be the set of $y_i^n$ sequences which are jointly typical with the decoded quantizations $(U_1^n, U_2^n)$. This set has cardinality no larger than $2^{n(H(Y_i|U_1,U_2,Q)+\epsilon)}$, which proves achievability.

To see the converse, note that setting

$$\hat{Y}_i^n = \Pr\left[Y_i^n|g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n)\right]$$

achieves a logarithmic loss of $\frac{1}{n}H(Y_i^n|g_1^{(n)}(Y_1^n), g_2^{(n)}(Y_2^n))$ for source $i$ in the setting where reproductions are not restricted to product distributions. Applying the strengthened converse of Theorem 6 together with Lemma 2 yields the desired result. ■

## V. Relationship to the General Multiterminal Source Coding Problem

In this section, we relate our results for logarithmic loss to multiterminal source coding problems with arbitrary distortion measures and reproduction alphabets.

As before, we let $\{Y_{1,j}, Y_{2,j}\}_{j=1}^{n}$ be a sequence of $n$ independent, identically distributed random variables with finite alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively, and joint pmf $p(y_1, y_2)$.

In this section, the reproduction alphabets $\breve{\mathcal{Y}}_i$, $i = 1, 2$, are arbitrary. We also consider generic distortion measures:

$$\breve{d}_i : \mathcal{Y}_i \times \breve{\mathcal{Y}}_i \to \mathbb{R}^+ \qquad \text{for } i = 1, 2,$$

where $\mathbb{R}^+$ denotes the set of nonnegative real numbers. The sequence distortion is then defined as follows:

$$\breve{d}_i(y_i^n, \breve{y}_i^n) = \frac{1}{n}\sum_{j=1}^{n} \breve{d}_i(y_{i,j}, \breve{y}_{i,j}).$$

We will continue to let $d(\cdot,\cdot)$ and $\hat{\mathcal{Y}}_1, \hat{\mathcal{Y}}_2$ denote the logarithmic loss distortion measure and the associated reproduction alphabets, respectively.

A rate distortion code (of blocklength $n$) consists of encoding functions:

$$\breve{g}_i^{(n)} : \mathcal{Y}_i^n \to \left\{1, \ldots, M_i^{(n)}\right\} \qquad \text{for } i = 1, 2$$

and decoding functions

$$\breve{\psi}_i^{(n)} : \left\{1, \ldots, M_1^{(n)}\right\} \times \left\{1, \ldots, M_2^{(n)}\right\} \to \breve{\mathcal{Y}}_i^n \qquad \text{for } i = 1, 2.$$

A rate distortion vector $(R_1, R_2, D_1, D_2)$ is strict-sense achievable if there exists a blocklength $n$, encoding functions $\breve{g}_1^{(n)}, \breve{g}_2^{(n)}$ and a decoder $(\breve{\psi}_1^{(n)}, \breve{\psi}_2^{(n)})$ such that

$$R_i \geq \frac{1}{n}\log M_i^{(n)} \qquad \text{for } i = 1, 2 \tag{21}$$

$$D_i \geq \mathbb{E}\breve{d}_i(Y_i^n, \breve{Y}_i^n) \qquad \text{for } i = 1, 2. \tag{22}$$

Where

$$\breve{Y}_i^n = \breve{\psi}_i^{(n)}(\breve{g}_1^{(n)}(Y_1^n), \breve{g}_2^{(n)}(Y_2^n)) \qquad \text{for } i = 1, 2.$$

For these functions, we define the quantity

$$\beta_i\left(\breve{g}_1^{(n)}, \breve{g}_2^{(n)}, \breve{\psi}_1^{(n)}, \breve{\psi}_2^{(n)}\right) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\log\left(\sum_{y_i \in \mathcal{Y}_i} 2^{-\breve{d}_i(y_i, \breve{Y}_{i,j})}\right) \qquad \text{for } i = 1, 2. \tag{23}$$

Now, let $\beta_i(R_1, R_2, D_1, D_2)$ be the infimum of the $\beta_i\left(\breve{g}_1^{(n)}, \breve{g}_2^{(n)}, \breve{\psi}_1^{(n)}, \breve{\psi}_2^{(n)}\right)$'s, where the infimum is taken over all codes that achieve the rate distortion vector $(R_1, R_2, D_1, D_2)$.

At this point it is instructive to pause and consider some examples.

*Example 1 (Binary Sources and Hamming Distortion): For $i = 1, 2$, let $\breve{\mathcal{Y}}_i = \mathcal{Y}_i = \{0, 1\}$ and let $\breve{d}_i$ be the $\alpha$-scaled Hamming distortion measure*:

$$\breve{d}_i(y_i, \breve{y}_i) = \begin{cases} 0 & \text{if } \breve{y}_i = y_i, \\ \alpha & \text{if } \breve{y}_i \neq y_i. \end{cases}$$

*In this case,*

$$\sum_{y_i \in \mathcal{Y}_i} 2^{-\breve{d}_i(y_i, \breve{Y}_{i,j})} = 2^0 + 2^{-\alpha}, \tag{24}$$

*so $\beta_i(R_1, R_2, D_1, D_2) = \log(1 + 2^{-\alpha})$ for any $(R_1, R_2, D_1, D_2)$. This notion that $\beta_i(R_1, R_2, D_1, D_2)$ is a constant extends to all distortion measures for which the columns of the $|\mathcal{Y}_i| \times |\breve{\mathcal{Y}}_i|$ distortion matrix are permutations of one another.*

*Example 2 (Binary Sources and Erasure Distortion): For $i = 1, 2$, let $\mathcal{Y}_i = \{0, 1\}$, $\breve{\mathcal{Y}}_i = \{0, 1, e\}$ and let $\breve{d}_i$ be the standard erasure distortion measure*:

$$\breve{d}_i(y_i, \breve{y}_i) = \begin{cases} 0 & \text{if } \breve{y}_i = y_i \\ 1 & \text{if } \breve{y}_i = e \\ \infty & \text{if } \breve{y}_i \in \{0, 1\} \text{ and } \breve{y}_i \neq y_i. \end{cases}$$

*In this case,*

$$\sum_{y_i \in \mathcal{Y}_i} 2^{-\breve{d}_i(y_i, \breve{Y}_{i,j})} = \begin{cases} 2^{-\infty} + 2^0 = 1 & \text{if } \breve{Y}_{i,j} \in \{0, 1\} \\ 2^{-1} + 2^{-1} = 1 & \text{if } \breve{Y}_{i,j} = e. \end{cases} \tag{25}$$

*so $\beta_i(R_1, R_2, D_1, D_2) = 0$ for any $(R_1, R_2, D_1, D_2)$. This result can easily be extended to erasure distortion on larger alphabets by setting the penalty to $\log|\mathcal{Y}_i|$ when $\breve{Y}_i = e$.*

*Theorem 8: Suppose $(R_1, R_2, D_1, D_2)$ is strict-sense achievable for the general multiterminal source coding problem. Then*

$$\left.\begin{aligned} R_1 &\geq I(U_1; Y_1|U_2, Q) \\ R_2 &\geq I(U_2; Y_2|U_1, Q) \\ R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Q) \\ D_1 &\geq H(Y_1|U_1, U_2, Q) \\ &\quad - \beta_1(R_1, R_2, D_1, D_2) \\ D_2 &\geq H(Y_2|U_1, U_2, Q) \\ &\quad - \beta_2(R_1, R_2, D_1, D_2) \end{aligned}\right\} \tag{26}$$

*for some joint distribution $p(y_1, y_2)p(q)p(u_1|y_1, q)$ $p(u_2|y_2, q)$ with $|\mathcal{U}_i| \leq |\mathcal{Y}_i|$ and $|\mathcal{Q}| \leq 5$.*

*Proof:* Since $(R_1, R_2, D_1, D_2)$ is strict-sense achievable, there exists a blocklength $n$, encoding functions $\breve{g}_1^{(n)}, \breve{g}_2^{(n)}$ and a decoder $(\breve{\psi}_1^{(n)}, \breve{\psi}_2^{(n)})$ satisfying (21)-(22). Given these functions, the decoder can generate reproductions $\breve{Y}_1^n, \breve{Y}_2^n$ satisfying the average distortion constraints (22). From the reproduction $\breve{Y}_i^n$, we construct the reproduction $\hat{Y}_i^n$ as follows:

$$\hat{Y}_j(y_i) = \frac{2^{-\breve{d}_i(y_i, \breve{Y}_{i,j})}}{\sum_{y_i' \in \mathcal{Y}_i} 2^{-\breve{d}_i(y_i', \breve{Y}_{i,j})}}.$$

Now, using the logarithmic loss distortion measure, observe that $\hat{Y}_i^n$ satisfies

$$\begin{aligned} \mathbb{E}d(Y_i^n, \hat{Y}_i^n) &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\log\left(2^{\breve{d}_i(Y_{i,j}, \breve{Y}_{i,j})}\right) \\ &\quad + \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\log\left(\sum_{y_i' \in \mathcal{Y}_i} 2^{-\breve{d}_i(y_i', \breve{Y}_{i,j})}\right) \\ &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\breve{d}_i(Y_{i,j}, \breve{Y}_{i,j}) \\ &\quad + \beta_i\left(\breve{g}_1^{(n)}, \breve{g}_2^{(n)}, \breve{\psi}_1^{(n)}, \breve{\psi}_2^{(n)}\right) \\ &\leq D_i + \beta_i\left(\breve{g}_1^{(n)}, \breve{g}_2^{(n)}, \breve{\psi}_1^{(n)}, \breve{\psi}_2^{(n)}\right) \\ &:= \tilde{D}_i. \end{aligned}$$

Thus, $(R_1, R_2, \tilde{D}_1, \tilde{D}_2)$ is achievable for the multiterminal source coding problem with the logarithmic loss distortion measure. Applying Theorem 6 and taking the infimum over all coding schemes that achieve $(R_1, R_2, D_1, D_2)$ proves the theorem. ∎

This outer bound is interesting because the region is defined over the same set of probability distributions that define the Berger-Tung inner bound. While the $\beta_i$'s can be difficult to compute in general, we have shown that they can be readily determined for many popular distortion measures. As an application, we now give a quantitative approximation of the rate distortion region for binary sources subject to Hamming distortion constraints. Before proceeding, we prove the following lemma.

*Lemma 3: Suppose $(R_1, R_2, \tilde{D}_1, \tilde{D}_2)$ is strict-sense achievable for the multiterminal source coding problem with binary sources and $\breve{d}_i$ equal to the $\alpha_i$-scaled Hamming distortion measure, for $i = 1, 2$. Then the Berger-Tung achievability scheme can achieve a point $(R_1, R_2, D_1, D_2)$ satisfying*

$$D_i - \tilde{D}_i \leq \left(\frac{\alpha_i}{2} - 1\right) H_i + \log(1 + 2^{-\alpha_i})$$

*for some $H_i \in [0, 1]$, $i = 1, 2$.*

*Proof:* By Theorem 8, $(R_1, R_2, \tilde{D}_1, \tilde{D}_2)$ satisfy (26) for some joint distribution $p(y_1, y_2)p(q)p(u_1|y_1, q)p(u_2|y_2, q)$. For this distribution, define the reproduction functions

$$\breve{Y}_i(U_1, U_2, Q) = \arg\max_{y_i} p(y_i|U_1, U_2, Q) \qquad \text{for } i = 1, 2 \quad (27)$$

Then, observe that for $i = 1, 2$:

$$\begin{aligned}
&\mathbb{E}\breve{d}_i(Y_i, \breve{Y}_i) \qquad (28)\\
&= \sum_{u_1,u_2,q} p(u_1, u_2, q)\Big[\alpha_i \cdot \min_{y_i} p(y_i|u_1, u_2, q)\\
&\qquad\qquad + 0 \cdot \max_{y_i} p(y_i|u_1, u_2, q)\Big]\\
&= \alpha_i \sum_{u_1,u_2,q} p(u_1, u_2, q) \cdot \min_{y_i} p(y_i|u_1, u_2, q)\\
&\leq \frac{\alpha_i}{2} \sum_{u_1,u_2,q} p(u_1, u_2, q) H(Y_i|U_1, U_2, Q = u_1, u_2, q) \quad (29)\\
&= \frac{\alpha_i}{2} H(Y_i|U_1, U_2, Q).
\end{aligned}$$

Where (29) follows from the fact that $2p \leq h_2(p)$ for $0 \leq p \leq 0.5$. Thus, $D_i = \frac{\alpha_i}{2} H(Y_i|U_1, U_2, Q)$ is achievable for rates $(R_1, R_2)$ using the Berger-Tung achievability scheme. Combining this with the fact that $\tilde{D}_i \geq H(Y_i|U_1, U_2, Q) - \log(1 + 2^{-\alpha_i})$, we see that

$$\begin{aligned}
D_i - \tilde{D}_i \leq\ & \frac{\alpha_i}{2} H(Y_i|U_1, U_2, Q)\\
& - H(Y_i|U_1, U_2, Q) + \log(1 + 2^{-\alpha_i}).
\end{aligned}$$

∎

Lemma 3 allows us to give a quantitative outer bound on the achievable rate distortion region in terms of the Berger-Tung inner bound.

*Corollary 1: Suppose $(R_1, R_2, \tilde{D}_1^{(1)}, \tilde{D}_2^{(1)})$ is strict-sense achievable for the multiterminal source coding problem with binary sources and $\breve{d}_i$ equal to the standard 1-scaled Hamming distortion measure, for $i = 1, 2$. Then the Berger-Tung achievability scheme can achieve a point $(R_1, R_2, D_1^{(1)}, D_2^{(1)})$, where*

$$D_i^{(1)} - \tilde{D}_i^{(1)} \leq \frac{1}{2} \log\left(\frac{5}{4}\right) < 0.161 \qquad \text{for } i = 1, 2.$$

*Proof:* For rates $(R_1, R_2)$, note that distortions $(\tilde{D}_1, \tilde{D}_2)$ are strict-sense achievable for the $\alpha_i$-scaled Hamming distortion measures if and only if distortions $(\tilde{D}_1^{(1)}, \tilde{D}_2^{(1)}) = (\frac{1}{\alpha_1}\tilde{D}_1, \frac{1}{\alpha_2}\tilde{D}_2)$ are strict-sense achievable for the 1-scaled Hamming distortion measure. Likewise, the point $(R_1, R_2, D_1, D_2)$ is achieved by the Berger-Tung coding scheme for the $\alpha_i$-scaled Hamming distortion measures if and only if $(R_1, R_2, \frac{1}{\alpha_1}D_1, \frac{1}{\alpha_2}D_2)$ is achieved by the Berger-Tung coding scheme for the 1-scaled Hamming distortion measure.

Thus, applying Lemma 3, we can use the Berger-Tung achievability scheme to achieve a point $(R_1, R_2, D_1^{(1)}, D_2^{(1)})$ satisfying

$$\begin{aligned}
D_i^{(1)} - \tilde{D}_i^{(1)} &= \frac{1}{\alpha_i}\left(D_i - \tilde{D}_i\right)\\
&\leq \frac{1}{\alpha_i}\left(\frac{\alpha_i}{2} - 1\right) H_i + \frac{1}{\alpha_i}\log(1 + 2^{-\alpha_i})\\
&= \left(\frac{1}{2} - \frac{1}{\alpha_i}\right) H_i + \frac{1}{\alpha_i}\log(1 + 2^{-\alpha_i}) \quad (30)
\end{aligned}$$

for some $H_i \in [0, 1]$. We can optimize (30) over $\alpha_i$ to find the minimum gap for a given $H_i$. Maximizing over $H_i \in [0, 1]$ then gives the worst-case gap. Straightforward calculus yields the saddle-point:

$$\begin{aligned}
&\max_{H_i \in [0,1]} \inf_{\alpha_i > 0} \left\{\left(\frac{1}{2} - \frac{1}{\alpha_i}\right) H_i + \frac{1}{\alpha_i}\log(1 + 2^{-\alpha_i})\right\}\\
&\quad = \inf_{\alpha_i > 0} \max_{H_i \in [0,1]} \left\{\left(\frac{1}{2} - \frac{1}{\alpha_i}\right) H_i + \frac{1}{\alpha_i}\log(1 + 2^{-\alpha_i})\right\}\\
&\quad = \frac{1}{2}\log\left(\frac{5}{4}\right) < 0.161,
\end{aligned}$$

which is achieved for $\alpha_i = 2$ and any $H \in [0, 1]$. ∎

*Remark 3: We note briefly that this estimate can potentially be improved if one knows more about the source distribution.*

## VI. Concluding Remarks

For the CEO problem, our results can be extended to an arbitrary number of encoders. This extension is proved in Appendix B. Hence, one immediate direction for further work would be to extend our other results to more than two encoders.

We remark that generalizing the results for the two-encoder source coding problem with distortion constraints on $Y_1$ and $Y_2$ poses a significant challenge. The obvious point of difficulty in the proof is extending the tuning argument to higher dimensions so that it yields a distribution with the desired properties. In fact, a "quick-fix" to the tuning argument alone would not be sufficient since this would imply that the Berger-Tung inner bound is tight for more than two encoders. This is known to be false (even for logarithmic loss) since the Berger-Tung achievability scheme is not optimal for the lossless modulo-sum problem studied by Körner and Marton in [30].

## ACKNOWLEDGMENT

The authors would like to thank Professors Suhas Diggavi, Aaron Wagner and Rick Wesel for helpful discussions.

## APPENDIX A
## CARDINALITY BOUNDS ON AUXILIARY RANDOM VARIABLES

In order to obtain tight cardinality bounds on the auxiliary random variables used throughout this paper, we refer to a recent result by Jana. In [22], [31], the author carefully applies the Caratheodory-Fenchel-Eggleston theorem in order to obtain tight cardinality bounds on the auxiliary random variables in the Berger-Tung inner bound. This result extends the results and techniques employed by Gu and Effros for the Wyner-Ahlswede-Körner problem [32], and by Gu, Jana, and Effros for the Wyner-Ziv problem [33]. We now state Jana's result, appropriately modified for our purposes:

Consider an arbitrary joint distribution $p(v, y_1, \ldots, y_m)$ with random variables $V, Y_1, \ldots, Y_m$ coming from alphabets $\mathcal{V}, \mathcal{Y}_1, \ldots, \mathcal{Y}_m$ respectively.

Let $d_l : \mathcal{V} \times \hat{\mathcal{V}}_l \to \mathbb{R}$, $1 \leq l \leq L$ be arbitrary distortion measures defined for possibly different reproduction alphabets $\hat{\mathcal{V}}_l$.

*Definition 6: Define $\mathcal{A}^\star$ to be the set of $(m + L)$-vectors $(R_1, \ldots, R_m, D_1, \ldots, D_L)$ satisfying the following conditions:*

1) *auxiliary random variables $U_1, \ldots, U_m$ exist such that*

$$\sum_{i \in \mathcal{I}} R_i \geq I(Y_{\mathcal{I}}; U_{\mathcal{I}} | U_{\mathcal{I}^c}), \quad \text{for all } \mathcal{I} \subseteq \{1, \ldots, m\}, \text{ and}$$

2) *mappings $\psi_l : \mathcal{U}_1 \times \cdots \times \mathcal{U}_m \to \hat{\mathcal{V}}_l$, $1 \leq l \leq L$ exist such that*

$$\mathbb{E} d_l(V, \psi_l(U_1, \ldots, U_m)) \leq D_l$$

*for some joint distribution*

$$p(v, y_1, \ldots, y_m) \prod_{j=1}^{m} p(u_j | y_j).$$

*Lemma 4 (Lemma 2.2 from [22]): Every extreme point of $\mathcal{A}^\star$ corresponds to some choice of auxiliary variables $U_1, \ldots, U_m$ with alphabet sizes $|\mathcal{U}_j| \leq |\mathcal{Y}_j|$, $1 \leq j \leq m$.*

In order to obtain the cardinality bounds for the CEO problem, we simply let $L = 1$, $V = X$, and $\hat{\mathcal{V}}_1 = \hat{\mathcal{X}}$. Defining

$$d_1(x, \hat{x}) = \log\left(\frac{1}{\hat{x}(x)}\right),$$

we see that $\overline{\mathcal{RD}}^\star_{CEO} = \text{conv}\,(\mathcal{A}^\star)$, where $\text{conv}\,(\mathcal{A}^\star)$ denotes the convex hull of $\mathcal{A}^\star$. Therefore, Lemma 4 implies that all extreme points of $\overline{\mathcal{RD}}^\star_{CEO}$ are achieved with a choice of auxiliary random variables $U_1, \ldots, U_m$ with alphabet sizes $|\mathcal{U}_j| \leq |\mathcal{Y}_j|$, $1 \leq j \leq m$. By timesharing between extreme points, any point in $\overline{\mathcal{RD}}^\star_{CEO}$ can be achieved for these alphabet sizes.

Obtaining the cardinality bounds for the multiterminal source coding problem proceeds in a similar fashion. In particular, let $L = m = 2$, $V = (Y_1, Y_2)$, and $\hat{\mathcal{V}}_j = \hat{\mathcal{Y}}_j$, $j = 1, 2$. Defining

$$d_j((y_1, y_2), \hat{y}_j) = \log\left(\frac{1}{\hat{y}_j(y_j)}\right) \quad \text{for } j = 1, 2,$$

we see that $\overline{\mathcal{RD}}^\star = \text{conv}\,(\mathcal{A}^\star)$. In this case, Lemma 4 implies that all extreme points of $\overline{\mathcal{RD}}^\star$ are achieved with a choice of auxiliary random variables $U_1, U_2$ with alphabet sizes $|\mathcal{U}_j| \leq |\mathcal{Y}_j|$, $1 \leq j \leq 2$. By timesharing between extreme points, any point in $\overline{\mathcal{RD}}^\star$ can be achieved for these alphabet sizes.

In order to obtain cardinality bounds on the timesharing variable $Q$, we can apply Caratheodory's theorem (cf. [34]). In particular, if $C \subset \mathbb{R}^n$ is compact, then any point in $\text{conv}(C)$ is a convex combination of at most $n+1$ points of $C$. Taking $C$ to be the closure of the set of extreme points of $\mathcal{A}^\star$ is sufficient for our purposes (boundedness of $C$ can be dealt with by a standard truncation argument).

*Remark 4: The well-known support lemma (cf. [21], [35]) provides an alternative, albeit suboptimal, method for bounding the cardinalities of the auxiliary random variables. Indeed, a standard application of the support lemma implies all points in $\overline{\mathcal{RD}}^\star$ are achieved by auxiliaries satisfying $|\mathcal{U}_j| \leq |\mathcal{Y}_j| + 3$ for $1 \leq j \leq 2$. Furthermore, all points in $\overline{\mathcal{RD}}^\star_{CEO}$ are achieved by auxiliaries satisfying $|\mathcal{U}_j| \leq |\mathcal{Y}_j| + 2$ for $1 \leq j \leq 2$. In both cases, the respective bounds on $\mathcal{Q}$ remain unchanged.*

## APPENDIX B
## EXTENSION OF CEO RESULTS TO $m$ ENCODERS

In this appendix, we prove the generalization of Theorem 3 to $m$ encoders, which essentially amounts to extending the argument in the proof of Theorem 3 to the general case. We begin by stating the $m$-encoder generalizations of Theorems 1 and 2, the proofs of which are trivial extensions of the proofs given for the two-encoder case and are therefore omitted.

*Definition 7: Let $\mathcal{R}^i_{CEO,m}$ be the set of all $(R_1, \ldots, R_m, D)$ satisfying*

$$\sum_{i \in \mathcal{I}} R_i \geq I(Y_{\mathcal{I}}; U_{\mathcal{I}} | U_{\mathcal{I}^c}, Q) \text{ for all } \mathcal{I} \subseteq \{1, \ldots, m\}$$
$$D \geq H(X | U_1, \ldots, U_m, Q).$$

*for some joint distribution $p(q)p(x)\prod_{i=1}^m p(y_i|x)p(u_i|y_i, q)$.*

*Theorem 9: All rate distortion vectors $(R_1, \ldots, R_m, D) \in \mathcal{R}^i_{CEO,m}$ are achievable.*

*Definition 8: Let $\mathcal{R}^o_{CEO,m}$ be the set of $(R_1, \ldots, R_m, D)$ satisfying*

$$\sum_{i \in \mathcal{I}} R_i \geq \left[\sum_{i \in \mathcal{I}} I(U_i; Y_i | X, Q) + H(X | U_{\mathcal{I}^c}, Q) - D\right]^+ \quad (31)$$

*for all $\mathcal{I} \subseteq \{1, \ldots, m\}$, and*

$$D \geq H(X | U_1, \ldots, U_m, Q). \quad (32)$$

*for some joint distribution $p(q)p(x)\prod_{i=1}^m p(y_i|x)p(u_i|y_i, q)$.*

*Theorem 10: If $(R_1, \ldots, R_m, D)$ is strict-sense achievable, then $(R_1, \ldots, R_m, D) \in \mathcal{R}^o_{CEO,m}$.*

Given the definitions of $\mathcal{R}^i_{CEO,m}$ and $\mathcal{R}^o_{CEO,m}$, the generalization of Theorem 3 to $m$ encoders is an immediate consequence of the following lemma:

*Lemma 5:* $\mathcal{R}^o_{CEO,m} \subseteq \mathcal{R}^i_{CEO,m}$.

*Proof:* Suppose $(R_1, \ldots, R_m, D) \in \mathcal{R}^o_{CEO,m}$, then by definition there exists $p(q)$ and conditional distributions $\{p(u_i|y_i, q)\}_{i=1}^m$ so that (31) and (32) are satisfied. For the joint distribution corresponding to $p(q)$ and conditional distributions $p\{(u_i|y_i, q)\}_{i=1}^m$, define $\mathcal{P}_D \subset \mathbb{R}^m$ to be the polytope defined by the inequalities (31). Now, to show $(R_1, \ldots, R_m, D) \in \mathcal{R}^i_{CEO,m}$, it suffices to show that each extreme point of $\mathcal{P}_D$ is dominated by a point in $\mathcal{R}^i_{CEO,m}$ that achieves distortion at most $D$.

To this end, define the set function $f : 2^{[m]} \to \mathbb{R}$ as follows:

$$\begin{aligned} f(\mathcal{I}) &:= I(Y_{\mathcal{I}}; U_{\mathcal{I}}|U_{\mathcal{I}^c}, Q) - (D - H(X|U_1, \ldots, U_m, Q)) \\ &= \sum_{i\in\mathcal{I}} I(U_i; Y_i|X, Q) + H(X|U_{\mathcal{I}^c}, Q) - D, \end{aligned}$$

where the equality follows since $U_{\mathcal{I}} \leftrightarrow Y_{\mathcal{I}} \leftrightarrow X \leftrightarrow Y_{\mathcal{I}^c} \leftrightarrow U_{\mathcal{I}^c}$ form a Markov chain for each $\mathcal{I} \subseteq \{1, \ldots, m\}$ conditioned on $Q$.

It can be verified that the function $f$ and the function $f^+(\mathcal{I}) = \max\{f(\mathcal{I}), 0\}$ are supermodular functions (see Appendix C). By construction, $\mathcal{P}_D$ is equal to the set of $(R_1, \ldots, R_m)$ which satisfy:

$$\sum_{i\in\mathcal{I}} R_i \geq f^+(\mathcal{I}).$$

It follows by basic results in submodular optimization (see Appendix C) that, for a linear ordering $i_1 \prec i_2 \prec \cdots \prec i_m$ of $\{1, \ldots, m\}$, an extreme point of $\mathcal{P}_D$ can be greedily computed as follows for $j = 1, \ldots, m$:

$$\tilde{R}_{i_j} = f^+(\{i_1, \ldots, i_j\}) - f^+(\{i_1, \ldots, i_{j-1}\}).$$

Furthermore, all extreme points of $\mathcal{P}_D$ can be enumerated by looking over all linear orderings $i_1 \prec i_2 \prec \cdots \prec i_m$ of $\{1, \ldots, m\}$. Each ordering of $\{1, \ldots, m\}$ is analyzed in the same manner, hence we assume (for notational simplicity) that the ordering we consider is the natural ordering $i_j = j$.

Let $j$ be the first index for which $\tilde{R}_j > 0$. Then, by construction,

$$\tilde{R}_k = I(U_k; Y_k|U_{k+1}, \ldots, U_m, Q) \text{ for all } k > j.$$

Furthermore, we must have $f(\{1, \ldots, j'\}) \leq 0$ for all $j' < j$. Thus, $\tilde{R}_j$ can be expressed as

$$\begin{aligned} \tilde{R}_j &= \sum_{i=1}^{j} I(Y_i; U_i|X, Q) + H(X|U_{j+1}, \ldots, U_m, Q) - D \\ &= I(Y_j; U_j|U_{j+1}, \ldots, U_m, Q) + f(\{1, \ldots, j-1\}) \\ &= (1-\theta) I(Y_j; U_j|U_{j+1}, \ldots, U_m, Q), \end{aligned}$$

where $\theta \in [0, 1)$ is defined as:

$$\begin{aligned} \theta &= \frac{-f(\{1, \ldots, j-1\})}{I(Y_j; U_j|U_{j+1}, \ldots, U_m, Q)} \\ &= \frac{1}{I(Y_j; U_j|U_{j+1}, \ldots, U_m, Q)} \Big[ D - H(X|U_1, \ldots, U_m, Q) \\ &\quad - I(U_1, \ldots, U_{j-1}; Y_1, \ldots, Y_{j-1}|U_j, \ldots, U_m, Q) \Big]. \end{aligned}$$

By the results of Theorem 9, the rates $(\tilde{R}_1, \ldots, \tilde{R}_m)$ permit the following coding scheme: For a fraction $(1-\theta)$ of the time, a codebook can be used that allows the decoder to recover $U_j^n, \ldots, U_m^n$ with high probability. The other fraction $\theta$ of the time, a codebook can be used that allows the decoder to recover $U_{j+1}^n, \ldots, U_m^n$ with high probability. As $n \to \infty$, this coding scheme can achieve distortion

$$\begin{aligned} \tilde{D} &= (1-\theta) H(X|U_j, \ldots, U_m, Q) \\ &\quad + \theta H(X|U_{j+1}, \ldots, U_m, Q) \\ &= H(X|U_j, \ldots, U_m, Q) + \theta I(X; U_j|U_{j+1}, \ldots, U_m, Q) \\ &= H(X|U_j, \ldots, U_m, Q) + \frac{I(X; U_j|U_{j+1}, \ldots, U_m, Q)}{I(Y_j; U_j|U_{j+1}, \ldots, U_m, Q)} \\ &\quad \times \Big[ D - H(X|U_1, \ldots, U_m, Q) \\ &\quad - I(U_1, \ldots, U_{j-1}; Y_1, \ldots, Y_{j-1}|U_j, \ldots, U_m, Q) \Big] \\ &\leq H(X|U_j, \ldots, U_m, Q) + D - H(X|U_1, \ldots, U_m, Q) \\ &\quad - I(U_1, \ldots, U_{j-1}; Y_1, \ldots, Y_{j-1}|U_j, \ldots, U_m, Q) \quad (33) \\ &= D + I(X; U_1, \ldots U_{j-1}|U_j, \ldots, U_m, Q) \\ &\quad - I(U_1, \ldots, U_{j-1}; Y_1, \ldots, Y_{j-1}|U_j, \ldots, U_m, Q) \\ &= D - I(U_1, \ldots, U_{j-1}; Y_1, \ldots, Y_{j-1}|X, U_j, \ldots, U_m, Q) \\ &\leq D. \quad (34) \end{aligned}$$

In the preceding string of inequalities (33) follows since $U_j$ is conditionally independent of everything else given $(Y_j, Q)$, and (34) follows from the non-negativity of mutual information.

Therefore, for every extreme point $(\tilde{R}_1, \ldots, \tilde{R}_m)$ of $\mathcal{P}_D$, the point $(\tilde{R}_1, \ldots, \tilde{R}_m, D)$ lies in $\mathcal{R}^i_{CEO,m}$. This proves the lemma. ∎

Finally, we remark that the results of Appendix A imply that it suffices to consider auxiliary random variables $U_1, \ldots, U_m$ with alphabet sizes $|\mathcal{U}_j| \leq |\mathcal{Y}_j|$, $1 \leq j \leq m$ (or, $|\mathcal{U}_j| \leq |\mathcal{Y}_j| + 2^{m-1}$ if one applies the support lemma). The timesharing variable $Q$ requires an alphabet size bounded by $|\mathcal{Q}| \leq m+2$.

## Appendix C
## Supermodular Functions

In this appendix, we review some basic results in submodular optimization that were used in Appendix B to prove Lemma 5. We tailor our statements toward supermodularity, since this is the property we require in Appendix B.

We begin by defining a supermodular function.

*Definition 9: Let $E = \{1, \ldots, n\}$ be a finite set. A function $s : 2^E \to \mathbb{R}$ is* supermodular *if for all $S, T \subseteq E$*

$$s(S) + s(T) \leq s(S \cap T) + s(S \cup T). \quad (35)$$

One of the fundamental results in submodular optimization is that a greedy algorithm minimizes a linear function over a supermodular polyhedron. By varying the linear function to be minimized, all extreme points of the supermodular polyhedron can be enumerated. In particular, define the supermodular polyhedron $\mathcal{P}(s) \subset \mathbb{R}^n$ to be the set of $x \in \mathbb{R}^n$ satisfying

$$\sum_{i\in T} x_i \geq s(T) \text{ for all } T \subseteq E.$$

The following theorem provides an algorithm that enumerates the extreme points of $\mathcal{P}(s)$.

*Theorem 11 (See [36]–[38]): For a linear ordering $e_1 \prec e_2 \prec \cdots \prec e_n$ of the elements in $E$, Algorithm C.1 returns an extreme point $v$ of $\mathcal{P}(s)$. Moreover, all extreme points of $\mathcal{P}(s)$ can be enumerated by considering all linear orderings of the elements of $E$.*

**Algorithm C.1:** Greedy$(s, E, \prec)$

***comment:*** *Returns extreme point $v$ of $\mathcal{P}(s)$ corresponding to the ordering $\prec$.*

***for*** $i = 1, \ldots n$

Set $v_i = s(\{e_1, e_2, \ldots, e_i\}) - s(\{e_1, e_2, \ldots, e_{i-1}\})$

***return*** $(v)$

*Proof:* See [36]–[38]. ∎

Theorem 11 is the key tool we employ to establish Lemma 5. In order to apply it, we require the following lemma.

*Lemma 6: For any joint distribution of the form $p(q)p(x)\prod_{i=1}^{m} p(y_i|x)p(u_i|y_i, q)$ and fixed $D \in \mathbb{R}$, define the set function $f : 2^{[m]} \to \mathbb{R}$ as:*

$$\begin{aligned} f(\mathcal{I}) &:= I(Y_{\mathcal{I}}; U_{\mathcal{I}}|U_{\mathcal{I}^c}, Q) \\ &\quad -(D - H(X|U_1, \ldots, U_m, Q)) \qquad (36) \\ &= \sum_{i\in\mathcal{I}} I(U_i; Y_i|X, Q) + H(X|U_{\mathcal{I}^c}, Q) - D, \end{aligned}$$

*and the corresponding non-negative set function $f^+ : 2^{[m]} \to \mathbb{R}$ as $f^+ = \max\{f, 0\}$. The functions $f$ and $f^+$ are supermodular.*

*Proof:* In order to verify that $f$ is supermodular, it suffices to check that the function $f'(\mathcal{I}) = I(Y_{\mathcal{I}}; U_{\mathcal{I}}|U_{\mathcal{I}^c}, Q)$ is supermodular since the latter two terms in (36) are constant. To this end, consider sets $T, S \subseteq \{1, \ldots, m\}$ and observe that:

$$\begin{aligned} &f'(S) + f'(T) = I(Y_S; U_S|U_{S^c}, Q) + I(Y_T; U_T|U_{T^c}, Q) \\ &= H(U_S|U_{S^c}, Q) - H(U_S|Y_S, Q) + H(U_T|U_{T^c}, Q) \\ &\quad - H(U_T|Y_T, Q) \\ &= H(U_S|U_{S^c}, Q) + H(U_T|U_{T^c}, Q) \\ &\quad - H(U_{S\cup T}|Y_{S\cup T}, Q) - H(U_{S\cap T}|Y_{S\cap T}, Q) \qquad (37) \\ &= H(U_{S\setminus T}|U_{S^c}, Q) + H(U_{S\cap T}|U_{(S\cap T)^c}, Q) \\ &\quad + H(U_T|U_{T^c}, Q) - H(U_{S\cup T}|Y_{S\cup T}, Q) \\ &\quad - H(U_{S\cap T}|Y_{S\cap T}, Q) \qquad (38) \\ &= H(U_{S\setminus T}|U_{S^c}, Q) + H(U_T|U_{T^c}, Q) - H(U_{S\cup T}|Y_{S\cup T}, Q) \\ &\quad + I(U_{S\cap T}; Y_{S\cap T}|U_{(S\cap T)^c}, Q) \\ &\leq H(U_{S\setminus T}|U_{(S\cup T)^c}, Q) + H(U_T|U_{T^c}, Q) \\ &\quad - H(U_{S\cup T}|Y_{S\cup T}, Q) + I(U_{S\cap T}; Y_{S\cap T}|U_{(S\cap T)^c}, Q) \qquad (39) \\ &= I(U_{S\cup T}; Y_{S\cup T}|U_{(S\cup T)^c}, Q) \\ &\quad + I(U_{S\cap T}; Y_{S\cap T}|U_{(S\cap T)^c}, Q) \\ &= f'(S \cap T) + f'(S \cup T). \end{aligned}$$

The labeled steps above can be justified as follows:

- (37) follows since $U_i$ is conditionally independent of everything else given $(Y_i, Q)$.
- (38) is simply the chain rule.
- (39) follows since conditioning reduces entropy.

Next, we show that $f^+ = \max\{f, 0\}$ is supermodular. Observe first that $f$ is monotone increasing, i.e., if $S \subset T$, then $f(S) \leq f(T)$. Thus, fixing $S, T \subseteq \{1, \ldots, m\}$, we can assume without loss of generality that

$$f(S \cap T) \leq f(S) \leq f(T) \leq f(S \cup T).$$

If $f(S \cap T) \geq 0$, then (35) is satisfied for $s = f^+$ by the supermodularity of $f$. On the other hand, if $f(S \cup T) \leq 0$, then (35) is a tautology for $s = f^+$. Therefore, it suffices to check the following three cases:

- Case 1: $f(S \cap T) \leq 0 \leq f(S) \leq f(T) \leq f(S \cup T)$. In this case, the supermodularity of $f$ and the fact that $f^+ \geq f$ imply:
$$\begin{aligned} f^+(S \cup T) + f^+(S \cap T) &\geq f(S \cup T) + f(S \cap T) \\ &\geq f(S) + f(T) \\ &= f^+(S) + f^+(T). \end{aligned}$$
- Case 2: $f(S \cap T) \leq f(S) \leq 0 \leq f(T) \leq f(S \cup T)$. Since $f$ is monotone increasing, we have:
$$\begin{aligned} f^+(S \cup T) + f^+(S \cap T) &= f(S \cup T) + 0 \\ &\geq f(T) + 0 = f^+(S) + f^+(T). \end{aligned}$$
- Case 3: $f(S \cap T) \leq f(S) \leq f(T) \leq 0 \leq f(S \cup T)$. By definition of $f^+$:
$$\begin{aligned} f^+(S \cup T) + f^+(S \cap T) &= f(S \cup T) + 0 \\ &\geq 0 + 0 = f^+(S) + f^+(T). \end{aligned}$$

Hence, $f^+ = \max\{f, 0\}$ is supermodular. ∎

## Appendix D
## Amplifying a Pointwise Convexity Constraint

*Lemma 7: Let $r_1, r_2 \in \mathbb{R}$ be given, and suppose $f_1 : K \to \mathbb{R}$ and $f_2 : K \to \mathbb{R}$ are continuous functions defined on a compact domain $K \subset \mathbb{R}^n$. If there exists a function $h : [0, 1] \to K$ satisfying*

$$t\,(f_1 \circ h)\,(t) + (1 - t)\,(f_2 \circ h)\,(t) \leq tr_1 + (1 - t)r_2 \quad (40)$$

*for all $t \in [0, 1]$, then there exists $x_1^*, x_2^* \in K$ and $t^* \in [0, 1]$ for which*

$$\begin{aligned} t^* f_1(x_1^*) + (1 - t^*) f_1(x_2^*) &\leq r_1 \\ t^* f_2(x_1^*) + (1 - t^*) f_2(x_2^*) &\leq r_2. \end{aligned}$$

We remark that, in our application of Lemma 7, we will take $K$ is taken to be a closed subset of a finite-dimensional probability simplex and $f_1, f_2$ to be conditional entropies evaluated for probability distributions in $K$.

*Proof of Lemma 7:* Since $f_1, f_2$ are continuous and $K$ is compact, there exists $M < \infty$ such that $f_1$ and $f_2$ are bounded from above and below by $M$ and $-M$, respectively. Fix $\epsilon > 0$, and partition the interval $[0, 1]$ as $0 = t_1 < t_2 < \cdots < t_m = 1$, such that $|t_{j+1} - t_j| < \frac{\epsilon}{M}$. For convenience define $x_{t_j} := h(t_j)$ when $t_j$ is in the partition.

Now, for $i = 1, 2$ define piecewise-linear functions $g_1(t), g_2(t)$ on [0,1] by:

$$g_i(t) = \left\{ \begin{array}{ll} f_i(x_{t_j}) & \text{if } \exists j \in \{1, \ldots, m\} \\ & \text{such that } t = t_j, \\ \theta f_i(x_{t_j}) & \text{if } \exists j \in \{1, \ldots, m\} \\ +(1-\theta) f_i(x_{t_{j+1}}) & \text{s.t. } t \in (t_j, t_{j+1}), \end{array} \right\} \quad (41)$$

where $\theta \in (0, 1)$ is chosen so that $t = \theta t_j + (1-\theta)t_{j+1}$ when $t$ is in the interval $(t_j, t_{j+1})$.

With $g_1(t)$ and $g_2(t)$ defined in this manner, suppose $t = \theta t_j + (1-\theta)t_{j+1}$ for some $j$ and $\theta$. Then straightforward algebra yields:

$$\begin{aligned} & tg_1(t) + (1-t)g_2(t) \\ &= (\theta t_j + (1-\theta)t_{j+1}) \left(\theta f_1(x_{t_j}) + (1-\theta) f_1(x_{t_{j+1}})\right) \\ &\quad + (1 - \theta t_j - (1-\theta)t_{j+1}) \left(\theta f_2(x_{t_j}) + (1-\theta) f_2(x_{t_{j+1}})\right) \\ &= \theta^2 \left[t_j f_1(x_{t_j}) + (1-t_j) f_2(x_{t_j})\right] \\ &\quad + (1-\theta)^2 \left[t_{j+1} f_1(x_{t_{j+1}}) + (1-t_{j+1}) f_2(x_{t_{j+1}})\right] \\ &\quad + \theta(1-\theta) \left[(1-t_j) f_2(x_{t_{j+1}}) + (1-t_{j+1}) f_2(x_{t_j}) \right. \\ &\qquad \left. + t_{j+1} f_1(x_{t_j}) + t_j f_1(x_{t_{j+1}})\right] \\ &\leq \theta^2 \left[t_j f_1(x_{t_j}) + (1-t_j) f_2(x_{t_j})\right] \\ &\quad + (1-\theta)^2 \left[t_{j+1} f_1(x_{t_{j+1}}) + (1-t_{j+1}) f_2(x_{t_{j+1}})\right] \\ &\quad + \theta(1-\theta) \left[(1-t_{j+1}) f_2(x_{t_{j+1}}) + (1-t_j) f_2(x_{t_j}) \right. \\ &\qquad \left. + t_j f_1(x_{t_j}) + t_{j+1} f_1(x_{t_{j+1}})\right] + \epsilon \\ &\leq \theta^2 \left[t_j r_1 + (1-t_j) r_2\right] \\ &\quad + (1-\theta)^2 \left[t_{j+1} r_1 + (1-t_{j+1}) r_2\right] \\ &\quad + \theta(1-\theta) \left[(1-t_{j+1}) r_2 + (1-t_j) r_2 \right. \\ &\qquad \left. + t_j r_1 + t_{j+1} r_1\right] + \epsilon \\ &= (\theta t_j + (1-\theta)t_{j+1}) r_1 + (1 - \theta t_j - (1-\theta)t_{j+1}) r_2 + \epsilon \\ &= t r_1 + (1-t) r_2 + \epsilon, \end{aligned} \quad (42)$$

where the first inequality follows since $|t_{j+1} - t_j|$ is small, and the second inequality follows from the the fact that (40) holds for each $t_j$ in the partition. Notably, this implies that it is impossible to have

$$g_1(t) > r_1 + \epsilon \quad \text{and} \quad g_2(t) > r_2 + \epsilon$$

hold simultaneously for any $t \in [0, 1]$, else we would obtain a contradiction to (42). Also, since we included the endpoints $t_1 = 0$ and $t_m = 1$ in the partition, we have the following two inequalities:

$$g_1(1) \leq r_1, \text{ and } g_2(0) \leq r_2.$$

Combining these observations with the fact that $g_1(t)$ and $g_2(t)$ are continuous, there must exist some $t^* \in [0, 1]$ for which

$$g_1(t^*) \leq r_1 + \epsilon, \quad \text{and} \quad g_2(t^*) \leq r_2 + \epsilon$$

simultaneously. An illustration of this is given in Figure 5, which is a mere variation on the classical intermediate value theorem.

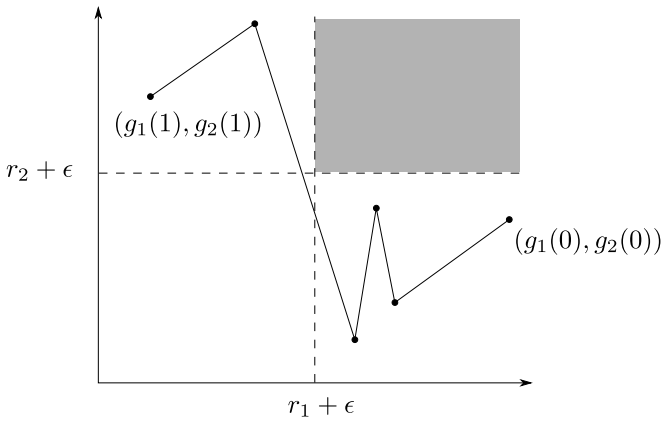


Fig. 5. A parametric plot of the function $\varphi : t \mapsto (g_1(t), g_2(t))$. Since $\varphi(t)$ is continuous, starts with $g_2(0) \leq r_2 + \epsilon$, ends with $g_1(1) \leq r_1 + \epsilon$, and doesn't intersect the shaded area, $\varphi(t)$ must pass through the lower-left region.

Applying this result, we can find a sequence $\{x_1^{(n)}, x_2^{(n)}, t^{(n)}\}_{n=1}^{\infty}$ satisfying

$$\begin{aligned} t^{(n)} f_1(x_1^{(n)}) + (1-t^{(n)}) f_1(x_2^{(n)}) &\leq r_1 + \frac{1}{n} \\ t^{(n)} f_2(x_1^{(n)}) + (1-t^{(n)}) f_2(x_2^{(n)}) &\leq r_2 + \frac{1}{n} \end{aligned}$$

for each $n \geq 1$. Since $K \times K \times [0, 1]$ is sequentially compact, there exists a convergent subsequence $\{n_j\}_{j=1}^{\infty}$ such that $(x_1^{(n_j)}, x_2^{(n_j)}, t^{(n_j)}) \to (x_1^*, x_2^*, t^*) \in K \times K \times [0, 1]$. The continuity of $f_1$ and $f_2$ then apply to yield the desired result.■

## APPENDIX E
## STRENGTHENING THE CONVERSE OF THEOREM 6

In this appendix, we prove a stronger version of the converse of Theorem 6. To be precise, let $\hat{\mathcal{Y}}_1^{*n}$ and $\hat{\mathcal{Y}}_2^{*n}$ denote the set of probability measures on $\mathcal{Y}_1^n$ and $\mathcal{Y}_2^n$, respectively. Let $d_1^*, d_2^*$ be the (extended)-log loss distortion measures defined as follows:

$$\begin{aligned} d_1^*(y_1^n, \hat{y}_1^n) &= \frac{1}{n} \log \left( \frac{1}{\hat{y}_1^n(y_1^n)} \right) \\ d_2^*(y_2^n, \hat{y}_2^n) &= \frac{1}{n} \log \left( \frac{1}{\hat{y}_2^n(y_2^n)} \right), \end{aligned}$$

where $\hat{y}_1^n(y_1^n)$ is the probability assigned to outcome $y_1^n \in \mathcal{Y}_1^n$ by the probability measure $\hat{y}_1^n \in \hat{\mathcal{Y}}_1^{*n}$. Similarly for $\hat{y}_2^n(y_2^n)$. Note that this extends the standard definition of logarithmic loss to sequence reproductions.

*Definition 10: We say that a tuple $(R_1, R_2, D_1, D_2)$ is sequence-achievable if, for any $\epsilon > 0$, there exist encoding functions*

$$\begin{aligned} f_1 &: \mathcal{Y}_1^n \to \{1, \ldots, 2^{nR_1}\} \\ f_2 &: \mathcal{Y}_2^n \to \{1, \ldots, 2^{nR_2}\}, \end{aligned}$$

*and decoding functions*

$$\begin{aligned} \phi_1 &: \{1, \ldots, 2^{nR_1}\} \times \{1, \ldots, 2^{nR_2}\} \to \hat{\mathcal{Y}}_1^{*n} \\ \phi_2 &: \{1, \ldots, 2^{nR_1}\} \times \{1, \ldots, 2^{nR_2}\} \to \hat{\mathcal{Y}}_2^{*n}, \end{aligned}$$

*which satisfy*

$$\mathbb{E}\, d_1^*(Y_1^n, \hat{Y}_1^n) \le D_1 + \epsilon$$
$$\mathbb{E}\, d_2^*(Y_2^n, \hat{Y}_2^n) \le D_2 + \epsilon,$$

*where*

$$\hat{Y}_1^n = \phi_1(f_1(Y_1^n), f_2(Y_2^n))$$
$$\hat{Y}_2^n = \phi_2(f_1(Y_1^n), f_2(Y_2^n)).$$

*Theorem 12: If $(R_1, R_2, D_1, D_2)$ is sequence-achievable, then $(R_1, R_2, D_1, D_2) \in \mathcal{RD}^i = \overline{\mathcal{RD}^\star}$.*

*Proof:* The theorem is an immediate consequence of Theorem 6 and Lemmas 8 and 9, which are given below. ■

*Remark 5: We refer to Theorem 12 as the "strengthened converse" of Theorem 6. Indeed, it states that enlarging the set of possible reproduction sequences to include non-product distributions cannot attain better performance than when the decoder is restricted to choosing a reproduction sequence from the set of product distributions.*

*Lemma 8: If $(R_1, R_2, \tilde{D}_1, D_2)$ is sequence-achievable, then there exists a joint distribution*

$$p(y_1, y_2, u_1, u_2, q) = p(q)p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)$$

*and a $D_1 \le \tilde{D}_1$ which satisfies*

$$D_1 \ge H(Y_1|U_1, U_2, Q)$$
$$D_2 \ge D_1 + H(Y_2|U_1, U_2, Q) - H(Y_1|U_1, U_2, Q),$$

*and*

$$R_1 \ge H(Y_1|U_2, Q) - D_1$$
$$R_2 \ge I(Y_2; U_2|Y_1, Q) + H(Y_1|U_1, Q) - D_1$$
$$R_1 + R_2 \ge I(Y_2; U_2|Y_1, Q) + H(Y_1) - D_1.$$

*Proof:* For convenience, let $F_1 = f_1(Y_1^n)$ and $F_2 = f_2(Y_2^n)$, where $f_1, f_2$ are the encoding functions corresponding to a scheme which achieves $(R_1, R_2, \tilde{D}_1, D_2)$ (in the sequence-reproduction sense). Define $D_1 = \frac{1}{n}H(Y_1^n|F_1, F_2)$, so that:

$$nD_1 = H(Y_1^n|F_1, F_2). \tag{43}$$

Since $n\tilde{D}_1 \ge H(Y_1^n|F_1, F_2)$ by the strengthened version[5] of Lemma 1, we have $D_1 \le \tilde{D}_1$ as desired. By definition of $D_1$, we immediately obtain the following inequality:

$$nD_1 = \sum_{i=1}^n H(Y_{1,i}|F_1, F_2, Y_{1,i+1}^n)$$
$$\ge \sum_{i=1}^n H(Y_{1,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n). \tag{44}$$

Next, recall the Csiszár sum identity:

$$\sum_{i=1}^n I(Y_{1,i+1}^n; Y_{2,i}|Y_2^{i-1}, F_1, F_2)$$
$$= \sum_{i=1}^n I(Y_2^{i-1}; Y_{1,i}|Y_{1,i+1}^n, F_1, F_2).$$

[5]See the comment in Section III-C.

This, together with (43), implies the following inequality:

$$nD_2 \ge nD_1 + \sum_{i=1}^n \Big[H(Y_{2,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n) - H(Y_{1,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n)\Big], \tag{45}$$

which we can verifiy as follows:

$$nD_2 \ge H(Y_2^n|F_1, F_2) = \sum_{i=1}^n H(Y_{2,i}|F_1, F_2, Y_2^{i-1})$$
$$= \sum_{i=1}^n \Big[H(Y_{2,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n) + I(Y_{1,i+1}^n; Y_{2,i}|F_1, F_2, Y_2^{i-1})\Big]$$
$$= \sum_{i=1}^n \Big[H(Y_{2,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n) + I(Y_2^{i-1}; Y_{1,i}|Y_{1,i+1}^n, F_1, F_2)\Big]$$
$$= H(Y_1^n|F_1, F_2) + \sum_{i=1}^n \Big[H(Y_{2,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n) - H(Y_{1,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n)\Big]$$
$$= nD_1 + \sum_{i=1}^n \Big[H(Y_{2,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n) - H(Y_{1,i}|F_1, F_2, Y_2^{i-1}, Y_{1,i+1}^n)\Big].$$

Next, observe that we can lower bound $R_1$ as follows:

$$nR_1 \ge H(F_1) \ge I(Y_1^n; F_1|F_2)$$
$$= \sum_{i=1}^n H(Y_{1,i}|F_2, Y_1^{i-1}) - H(Y_1^n|F_1, F_2)$$
$$\ge \sum_{i=1}^n H(Y_{1,i}|F_2, Y_1^{i-1}, Y_2^{i-1}) - nD_1 \tag{46}$$
$$= \sum_{i=1}^n H(Y_{1,i}|F_2, Y_2^{i-1}) - nD_1 \tag{47}$$
$$\ge \sum_{i=1}^n H(Y_{1,i}|F_2, Y_2^{i-1}, Y_{1,i+1}^n) - nD_1. \tag{48}$$

In the above string of inequalities, (46) follows from (43) and the fact that conditioning reduces entropy. Equality (47) follows since $Y_{1,i} \leftrightarrow F_2, Y_2^{i-1} \leftrightarrow Y_1^{i-1}$ form a Markov chain (in that order).

Next, we can obtain a lower bound on $R_2$:

$$\begin{aligned}
nR_2 &\geq H(F_2) \geq H(F_2|F_1) = H(F_2|F_1, Y_1^n) + I(Y_1^n; F_2|F_1) \\
&\geq I(Y_2^n; F_2|F_1, Y_1^n) + I(Y_1^n; F_2|F_1) \\
&= I(Y_2^n; F_2|Y_1^n) + I(Y_1^n; F_2|F_1) \qquad (49) \\
&= \sum_{i=1}^{n} I(Y_{2,i}; F_2|Y_1^n, Y_2^{i-1}) \\
&\quad + \sum_{i=1}^{n} H(Y_{1,i}|F_1, Y_{1,i+1}^n) - nD_1 \qquad (50) \\
&\geq \sum_{i=1}^{n} I(Y_{2,i}; F_2|Y_1^n, Y_2^{i-1}) \\
&\quad + \sum_{i=1}^{n} H(Y_{1,i}|F_1, Y_2^{i-1}, Y_{1,i+1}^n) - nD_1 \\
&= \sum_{i=1}^{n} I(Y_{2,i}; F_2, Y_1^{i-1}, Y_2^{i-1}|Y_{1,i}, Y_2^{i-1}, Y_{1,i+1}^n) \\
&\quad + \sum_{i=1}^{n} H(Y_{1,i}|F_1, Y_2^{i-1}, Y_{1,i+1}^n) - nD_1 \qquad (51) \\
&\geq \sum_{i=1}^{n} I(Y_{2,i}; F_2, Y_2^{i-1}|Y_{1,i}, Y_2^{i-1}, Y_{1,i+1}^n) \\
&\quad + \sum_{i=1}^{n} H(Y_{1,i}|F_1, Y_2^{i-1}, Y_{1,i+1}^n) - nD_1. \qquad (52)
\end{aligned}$$

In the above string of inequalities, (50) follows from (43) and the chain rule. (51) follows from the i.i.d. property of the sources, and (52) follows by monotonicity of mutual information.

A lower bound on the sum-rate $R_1 + R_2$ can be obtained as follows:

$$\begin{aligned}
n(R_1 + R_2) &\geq H(F_1) + H(F_2) \geq H(F_2) + H(F_1|F_2) \\
&\geq I(F_2; Y_1^n, Y_2^n) + I(F_1; Y_1^n|F_2) \\
&= I(F_2; Y_1^n) + I(F_2; Y_2^n|Y_1^n) + I(F_1; Y_1^n|F_2) \\
&= I(F_2; Y_2^n|Y_1^n) + I(F_1, F_2; Y_1^n) \\
&\geq \sum_{i=1}^{n} I(Y_{2,i}; F_2, Y_2^{i-1}|Y_{1,i}, Y_2^{i-1}, Y_{1,i+1}^n) \\
&\quad + \sum_{i=1}^{n} H(Y_{1,i}) - nD_1. \qquad (53)
\end{aligned}$$

Where (53) follows in a manner similar to (49)-(52) in the lower bound on $R_2$.

Now, define $U_{1,i} \triangleq F_1$, $U_{2,i} \triangleq (F_2, Y_2^{i-1})$, and $Q_i \triangleq (Y_2^{i-1}, Y_{1,i+1}^n)$. Then we can summarize our results so far as follows. Inequalities (44) and (45) become

$$\begin{aligned}
D_1 &\geq \frac{1}{n}\sum_{i=1}^{n} H(Y_{1,i}|U_{1,i}, U_{2,i}, Q_i) \\
D_2 &\geq D_1 + \frac{1}{n}\sum_{i=1}^{n} H(Y_{2,i}|U_{1,i}, U_{2,i}, Q_i) \\
&\quad - H(Y_{1,i}|U_{1,i}, U_{2,i}, Q_i),
\end{aligned}$$

and inequalities (48), (52), and (53) can be written as:

$$\begin{aligned}
R_1 &\geq \frac{1}{n}\sum_{i=1}^{n} H(Y_{1,i}|U_{2,i}, Q_i) - D_1 \\
R_2 &\geq \frac{1}{n}\sum_{i=1}^{n} I(Y_{2,i}; U_{2,i}|Y_{1,i}, Q_i) \\
&\quad + H(Y_{1,i}|U_{1,i}, Q_i) - D_1 \\
R_1 + R_2 &\geq \frac{1}{n}\sum_{i=1}^{n} I(Y_{2,i}; U_{2,i}|Y_{1,i}, Q_i) + H(Y_{1,i}) - D_1.
\end{aligned}$$

Next, we note that $U_{1,i} \leftrightarrow Y_{1,i} \leftrightarrow Y_{2,i} \leftrightarrow U_{2,i}$ form a Markov chain (in that order) conditioned on $Q_i$. Moreover, $Q_i$ is independent of $Y_{1,i}, Y_{2,i}$. Hence, a standard timesharing argument proves the lemma. ∎

*Lemma 9: Fix* $(R_1, R_2, D_1, D_2)$. *If there exists a joint distribution of the form*

$$p(y_1, y_2, u_1, u_2, q) = p(q)p(y_1, y_2)p(u_1|y_1, q)p(u_2|y_2, q)$$

*which satisfies*

$$D_1 \geq H(Y_1|U_1, U_2, Q) \qquad (54)$$

$$D_2 \geq D_1 + H(Y_2|U_1, U_2, Q) - H(Y_1|U_1, U_2, Q), \qquad (55)$$

*and*

$$R_1 \geq H(Y_1|U_2, Q) - D_1 \qquad (56)$$

$$R_2 \geq I(Y_2; U_2|Y_1, Q) + H(Y_1|U_1, Q) - D_1 \qquad (57)$$

$$R_1 + R_2 \geq I(Y_2; U_2|Y_1, Q) + H(Y_1) - D_1, \qquad (58)$$

*then* $(R_1, R_2, D_1, D_2) \in \mathcal{RD}^i$.

*Proof:* Let $\mathcal{P}$ denote the polytope of rate pairs which satisfy the inequalities (56)-(58). It suffices to show that if $(r_1, r_2)$ is a vertex of $\mathcal{P}$, then $(r_1, r_2, D_1, D_2) \in \mathcal{RD}^i$. For convenience, let $[x]^+ = \max\{x, 0\}$. There are only two extreme points of $\mathcal{P}$:

$$\begin{aligned}
r_1^{(1)} &= \Big[H(Y_1|U_2, Q) - D_1\Big]^+ \\
r_2^{(1)} &= I(Y_2; U_2|Y_1, Q) + H(Y_1) - D_1 - r_1^{(1)},
\end{aligned}$$

and

$$\begin{aligned}
r_1^{(2)} &= I(Y_2; U_2|Y_1, Q) + H(Y_1) - D_1 - r_2^{(2)}, \\
r_2^{(2)} &= \Big[I(Y_2; U_2|Y_1, Q) + H(Y_1|U_1, Q) - D_1\Big]^+.
\end{aligned}$$

We first analyze the extreme point $(r_1^{(1)}, r_2^{(1)})$:

- Case 1.1: $r_1^{(1)} = 0$. In this case, we have $r_2^{(1)} = I(Y_2; U_2|Y_1, Q) + H(Y_1) - D_1$. This can be expressed as:

$$r_2^{(1)} = (1 - \theta)I(Y_2; U_2|Q),$$

where

$$\theta = \frac{D_1 - I(Y_2; U_2|Y_1, Q) - H(Y_1) + I(Y_2; U_2|Q)}{I(Y_2; U_2|Q)}.$$

Since $r_1^{(1)} = 0$, we must have $D_1 \geq H(Y_1|U_2, Q)$. This implies that

$$\theta \geq \frac{H(Y_1|U_2, Q)I(Y_2;U_2|Y_1, Q) - H(Y_1) + I(Y_2; U_2|Q)}{I(Y_2; U_2|Q)} = 0.$$

Also, we can assume without loss of generality that $D_1 \leq H(Y_1)$, hence $\theta \in [0, 1]$. Applying the Berger-Tung achievability scheme, we can achieve the following distortions:

$$\begin{aligned} D_1^\theta &= \theta H(Y_1) + (1-\theta)H(Y_1|U_2, Q) \\ &= H(Y_1|U_2, Q) + \theta I(Y_1; U_2|Q) \\ &\leq H(Y_1|U_2, Q) + D_1 - I(Y_2; U_2|Y_1, Q) \\ &\quad - H(Y_1) + I(Y_2; U_2|Q) \qquad (59) \\ &= D_1 - I(Y_2; U_2|Y_1, Q) - I(Y_1; U_2|Q) \\ &\quad + I(Y_2; U_2|Q) \\ &= D_1, \end{aligned}$$

where (59) follows since $I(Y_1; U_2|Q) \leq I(Y_2; U_2|Q)$ by the data processing inequality.

$$\begin{aligned} D_2^\theta &= \theta H(Y_2) + (1-\theta)H(Y_2|U_2, Q) \\ &= H(Y_2|U_2, Q) + \theta I(Y_2; U_2|Q) \\ &= H(Y_2|U_2, Q) + D_1 - I(Y_2; U_2|Y_1, Q) \\ &\quad - H(Y_1) + I(Y_2; U_2|Q) \\ &= H(Y_2) + D_1 - I(Y_2; U_2|Y_1, Q) - H(Y_1) \\ &= H(Y_2|Y_1, U_2, Q) + D_1 - H(Y_1|Y_2) \\ &= H(Y_2|Y_1, U_1, U_2, Q) + D_1 - H(Y_1|Y_2) \qquad (60) \\ &\leq H(Y_2|Y_1, U_1, U_2, Q) + D_1 - H(Y_1|Y_2, U_1, U_2, Q) \\ &= H(Y_2|U_1, U_2, Q) + D_1 - H(Y_1|U_1, U_2, Q) \\ &\leq D_2, \qquad (61) \end{aligned}$$

where (60) follows since $U_1 \leftrightarrow (Y_1, U_2, Q) \leftrightarrow Y_2$, and (61) follows from (55).

- Case 1.2: $r_1^{(1)} \geq 0$. In this case, we have $r_2^{(1)} = I(Y_2; U_2|Y_1, Q) + I(Y_1; U_2|Q) = I(Y_2; U_2|Q)$. Also, we can write $r_1^{(1)}$ as:

$$r_1^{(1)} = (1-\theta)I(Y_1; U_1|U_2, Q),$$

where

$$\theta = \frac{D_1 - H(Y_1|U_2, Q) + I(Y_1; U_1|U_2, Q)}{I(Y_1; U_1|U_2, Q)}.$$

Since $r_1^{(1)} \geq 0$, we must have $D_1 \leq H(Y_1|U_2, Q)$. This implies that

$$\theta \leq \frac{H(Y_1|U_2, Q) - H(Y_1|U_2, Q) + I(Y_1; U_1|U_2, Q)}{I(Y_1; U_1|U_2, Q)} = 1.$$

Also, (54) implies that $D_1 \geq H(Y_1|U_1, U_2, Q)$, hence $\theta \in [0, 1]$. Applying the Berger-Tung achievability scheme, we can achieve the following distortions:

$$\begin{aligned} D_1^\theta &= \theta H(Y_1|U_2, Q) + (1-\theta)H(Y_1|U_1, U_2, Q) \\ &= H(Y_1|U_1, U_2, Q) + \theta I(Y_1; U_1|U_2, Q) \\ &= H(Y_1|U_1, U_2, Q) + D_1 - H(Y_1|U_2, Q) \\ &\quad + I(Y_1; U_1|U_2, Q) \\ &= D_1, \end{aligned}$$

and

$$\begin{aligned} D_2^\theta &= \theta H(Y_2|U_2, Q) + (1-\theta)H(Y_2|U_1, U_2, Q) \\ &= H(Y_2|U_1, U_2, Q) + \theta I(Y_2; U_1|U_2, Q) \\ &\leq H(Y_2|U_1, U_2, Q) + D_1 \\ &\quad - H(Y_1|U_2, Q) + I(Y_1; U_1|U_2, Q) \qquad (62) \\ &= H(Y_2|U_1, U_2, Q) + D_1 - H(Y_1|U_1, U_2, Q) \\ &\leq D_2, \qquad (63) \end{aligned}$$

where (62) follows since $I(Y_2; U_1|U_2, Q) \leq I(Y_1; U_1|U_2, Q)$ by the data processing inequality, and (63) follows from (55).

In a similar manner, we now analyze the second extreme point $(r_1^{(2)}, r_2^{(2)})$:

- Case 2.1: $r_2^{(2)} = 0$. In this case, we have $r_1^{(2)} = I(Y_2; U_2|Y_1, Q) + H(Y_1) - D_1$. This can be expressed as:

$$r_1^{(2)} = (1-\theta)I(Y_1; U_1|Q),$$

where

$$\theta = \frac{D_1 - I(Y_2; U_2|Y_1, Q) - H(Y_1) + I(Y_1; U_1|Q)}{I(Y_1; U_1|Q)}.$$

Since $r_2^{(2)} = 0$, we must have $D_1 \geq H(Y_1|U_1, Q) + I(Y_2; U_2|Y_1, Q)$. This implies that

$$\theta \geq \frac{1}{I(Y_1; U_1|Q)}\Big[H(Y_1|U_1, Q) + I(Y_2; U_2|Y_1, Q) - I(Y_2; U_2|Y_1, Q) - H(Y_1) + I(Y_1; U_1|Q)\Big] = 0.$$

Also, we can assume without loss of generality that $D_1 \leq H(Y_1)$, hence

$$\theta \leq \frac{H(Y_1) - I(Y_2; U_2|Y_1, Q) - H(Y_1) + I(Y_1; U_1|Q)}{I(Y_1; U_1|Q)} \leq 1,$$

and therefore $\theta \in [0, 1]$. Applying the Berger-Tung achievability scheme, we can achieve the following distortions:

$$\begin{aligned} D_1^\theta &= \theta H(Y_1) + (1-\theta)H(Y_1|U_1, Q) \\ &= H(Y_1|U_1, Q) + \theta I(Y_1; U_1|Q) \\ &= H(Y_1|U_1, Q) + D_1 - I(Y_2; U_2|Y_1, Q) \\ &\quad - H(Y_1) + I(Y_1; U_1|Q) \\ &= D_1 - I(Y_2; U_2|Y_1, Q) \\ &\leq D_1, \end{aligned}$$

and

$$\begin{aligned} D_2^\theta &= \theta H(Y_2) + (1-\theta)H(Y_2|U_1, Q) \\ &= H(Y_2|U_1, Q) + \theta I(Y_2; U_1|Q) \\ &\le H(Y_2|U_1, Q) + D_1 - I(Y_2; U_2|Y_1, Q) \\ &\quad - H(Y_1) + I(Y_1; U_1|Q) \qquad (64) \\ &= H(Y_2|Y_1, U_2, Q) + D_1 - H(Y_1|Y_2, U_1, Q) \\ &= H(Y_2|Y_1, U_1, U_2, Q) + D_1 \\ &\quad - H(Y_1|Y_2, U_1, U_2, Q) \qquad (65) \\ &= H(Y_2|U_1, U_2, Q) + D_1 - H(Y_1|U_1, U_2, Q) \\ &\le D_2, \qquad (66) \end{aligned}$$

where (64) follows since $I(Y_2; U_1|Q) \le I(Y_1; U_1|Q)$ by the data processing inequality, (65) follows since $U_1 \leftrightarrow (Y_1, U_2, Q) \leftrightarrow Y_2$ and $U_2 \leftrightarrow (Y_2, U_1, Q) \leftrightarrow Y_1$, and (66) follows from (55).

- Case 2.2: $r_2^{(2)} \ge 0$. In this case, we have $r_1^{(2)} = I(Y_1; U_1|Q)$. Also, we can write $r_2^{(2)}$ as:

$$r_2^{(2)} = (1-\theta) I(Y_2; U_2|U_1, Q),$$

where

$$\theta = \frac{1}{I(Y_2; U_2|U_1, Q)}\Big[D_1 - H(Y_1|U_1, Q) - I(Y_2; U_2|Y_1, Q) + I(Y_2; U_2|U_1, Q)\Big].$$

Since $r_2^{(2)} \ge 0$, we must have $D_1 \le H(Y_1|U_1, Q) + I(Y_2; U_2|Y_1, Q)$. This implies that $\theta \le 1$. Also, (54) implies that $D_1 \ge H(Y_1|U_1, U_2, Q)$, yielding

$$\begin{aligned} \theta &\ge \frac{1}{I(Y_2; U_2|U_1, Q)}\Big[H(Y_1|U_1, U_2, Q) - H(Y_1|U_1, Q) \\ &\qquad - I(Y_2; U_2|Y_1, Q) + I(Y_2; U_2|U_1, Q)\Big] \\ &= 0. \end{aligned}$$

Therefore, $\theta \in [0, 1]$. Applying the Berger-Tung achievability scheme, we can achieve the following distortions:

$$\begin{aligned} D_1^\theta &= \theta H(Y_1|U_1, Q) + (1-\theta)H(Y_1|U_1, U_2, Q) \\ &= H(Y_1|U_1, U_2, Q) + \theta I(Y_1; U_2|U_1, Q) \\ &\le H(Y_1|U_1, U_2, Q) + D_1 - H(Y_1|U_1, Q) \\ &\quad - I(Y_2; U_2|Y_1, Q) + I(Y_2; U_2|U_1, Q) \qquad (67) \\ &= D_1, \end{aligned}$$

where (67) follows since $I(Y_1; U_2|U_1, Q) \le I(Y_2; U_2|U_1, Q)$ by the data processing inequality.

$$\begin{aligned} D_2^\theta &= \theta H(Y_2|U_1, Q) + (1-\theta)H(Y_2|U_1, U_2, Q) \\ &= H(Y_2|U_1, U_2, Q) + \theta I(Y_2; U_2|U_1, Q) \\ &= H(Y_2|U_1, U_2, Q) + D_1 - H(Y_1|U_1, Q) \\ &\quad - I(Y_2; U_2|Y_1, Q) + I(Y_2; U_2|U_1, Q) \\ &= H(Y_2|U_1, U_2, Q) + D_1 - H(Y_1|U_1, U_2, Q) \\ &\le D_2, \qquad (68) \end{aligned}$$

where (68) follows from (55).

Thus, this proves that the Berger-Tung compression scheme can achieve any rate distortion tuple $(r_1, r_2, D_1, D_2)$ for $(r_1, r_2) \in \mathcal{P}$. Since $\mathcal{RD}^i$ is, by definition, the set of rate distortion tuples attainable by the Berger-Tung achievability scheme, we must have that $(R_1, R_2, D_1, D_2) \in \mathcal{RD}^i$. This proves the lemma. ∎

## APPENDIX F
## A LEMMA FOR THE DAILY DOUBLE

For a given joint distribution $p(y_1, y_2)$ on the finite alphabet $\mathcal{Y}_1 \times \mathcal{Y}_2$, let $\mathcal{P}(R_1, R_2)$ denote the set of joint pmf's of the form

$$p(q, y_1, y_2, u_1, u_2) = p(q)p(y_1, y_2)p(u_1|y_1, q)p(u_1|y_1, q)$$

which satisfy

$$\begin{aligned} R_1 &\ge I(Y_1; U_1|U_2, Q) \\ R_2 &\ge I(Y_2; U_2|U_1, Q) \\ R_1 + R_2 &\ge I(Y_1, Y_2; U_1, U_2|Q) \end{aligned}$$

for given finite alphabets $\mathcal{U}_1, \mathcal{U}_2, \mathcal{Q}$.

*Lemma 10: For $R_1, R_2$ satisfying $R_1 \le H(Y_1)$, $R_2 \le H(Y_2)$, and $R_1 + R_2 \le H(Y_1, Y_2)$, the infimum*

$$\inf_{p \in \mathcal{P}(R_1, R_2)} \{H(Y_1|U_1, U_2, Q) + H(Y_2|U_1, U_2, Q)\}$$

*is attained by some $p^* \in \mathcal{P}(R_1, R_2)$ which satisfies $R_1 + R_2 = I(Y_1, Y_2; U_1^*, U_2^*|Q^*)$, where $U_1^*, U_2^*, Q^*$ correspond to the auxiliary random variables defined by $p^*$.*

*Proof:* First, note that the infimum is always attained since $\mathcal{P}(R_1, R_2)$ is compact and the objective function is continuous on $\mathcal{P}(R_1, R_2)$. Therefore, let $U_1^*, U_2^*, Q^*$ correspond to the auxiliary random variables which attain the infimum.

If $H(Y_1|U_1^*, U_2^*, Q^*) + H(Y_2|U_1^*, U_2^*, Q^*) = 0$, then we must have $I(Y_1, Y_2; U_1^*, U_2^*|Q^*) = H(Y_1, Y_2)$. Thus, $R_1 + R_2 = I(Y_1, Y_2; U_1^*, U_2^*|Q^*)$.

Next, consider the case where $H(Y_1|U_1^*, U_2^*, Q^*) + H(Y_2|U_1^*, U_2^*, Q^*) > 0$. Assume for sake of contradiction that $R_1 + R_2 > I(Y_1, Y_2; U_1^*, U_2^*|Q^*)$. For any $p \in \mathcal{P}(R_1, R_2)$:

$$I(Y_1; U_1|U_2, Q) + I(Y_2; U_2|U_1, Q) \le I(Y_1, Y_2; U_1, U_2|Q).$$

Hence, at most one of the remaining rate constraints can be satisfied with equality. If none of the rate constraints are satisfied with equality, then define

$$(\tilde{U}_1, \tilde{U}_2) = \begin{cases} (U_1^*, U_2^*) & \text{with probability } 1-\epsilon \\ (Y_1, Y_2) & \text{with probability } \epsilon. \end{cases}$$

For $\epsilon > 0$ sufficiently small, the distribution $\tilde{p}$ corresponding to the auxiliary random variables $\tilde{U}_1, \tilde{U}_2, Q^*$ is still in $\mathcal{P}(R_1, R_2)$. However, $\tilde{p}$ satisfies

$$\begin{aligned} &H(Y_1|\tilde{U}_1, \tilde{U}_2, Q^*) + H(Y_2|\tilde{U}_1, \tilde{U}_2, Q^*) \\ &\qquad < H(Y_1|U_1^*, U_2^*, Q^*) + H(Y_2|U_1^*, U_2^*, Q^*), \end{aligned}$$

which contradicts the optimality of $p^*$.

Therefore, assume without loss of generality that

$$\begin{aligned} R_1 &= I(Y_1; U_1^*|U_2^*, Q^*) \\ R_1 + R_2 &> I(Y_1, Y_2; U_1^*, U_2^*|Q^*). \end{aligned}$$

This implies that $R_2 > I(Y_2; U_2^*|Q^*)$. Now, define

$$\tilde{U}_2 = \begin{cases} U_2^* & \text{with probability } 1-\epsilon \\ Y_2 & \text{with probability } \epsilon. \end{cases}$$

Note that for $\epsilon > 0$ sufficiently small:

$$I(Y_2; U_2^*|Q^*) < I(Y_2; \tilde{U}_2|Q^*) < R_2$$
$$I(Y_1, Y_2; U_1^*, U_2^*|Q^*) < I(Y_1, Y_2; U_1^*, \tilde{U}_2|Q^*) < R_1 + R_2,$$

and for any $\epsilon \in [0, 1]$:

$$R_1 = I(Y_1; U_1^*|U_2^*, Q^*) \geq I(Y_1; U_1^*|\tilde{U}_2, Q^*), \qquad (69)$$

and

$$H(Y_1|U_1^*, U_2^*, Q^*) + H(Y_2|U_1^*, U_2^*, Q^*) \geq H(Y_1|U_1^*, \tilde{U}_2, Q^*) + H(Y_2|U_1^*, \tilde{U}_2, Q^*). \qquad (70)$$

Since $R_2 \leq H(Y_2)$, as $\epsilon$ is increased from 0 to 1, at least one of the following must occur:

1) $I(Y_2; \tilde{U}_2|Q^*) = R_2$.
2) $I(Y_1, Y_2; U_1^*, \tilde{U}_2|Q^*) = R_1 + R_2$.
3) $I(Y_1; U_1|\tilde{U}_2, Q^*) < R_1$.

If either of events 1 or 2 occur first then the sum-rate constraint is met with equality (since they are equivalent in this case). If event 3 occurs first, then all rate constraints are satisfied with strict inequality and we can apply the above argument to contradict optimality of $p^*$. Since (70) shows that the objective is nonincreasing in $\epsilon$, there must exist a $\tilde{p} \in \mathcal{P}(R_1, R_2)$ which attains the infimum and satisfies the sum-rate constraint with equality. ■

## REFERENCES


[1] T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," in *Proc. IEEE ISIT*, Jul. 2012, pp. 761–765.

[2] T. A. Courtade, "Two problems in multiterminal information theory," Ph.D. dissertation, Dept. Electr. Eng., Univ. California, Los Angeles, CA, USA, 2012.

[3] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[4] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, Nov. 1975.

[5] A. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.

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

[7] 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.

[8] A. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic Gaussian two-encoder source-coding problem," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1938–1961, May 2008.

[9] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem [multiterminal source coding]," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996.

[10] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1912–1923, Nov. 1997.

[11] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.

[12] V. Prabhakaran, D. Tse, and K. Ramachandran, "Rate region of the quadratic Gaussian CEO problem," in *Proc. ISIT*, Jun./Jul. 2004, p. 117.

[13] Y. Oohama, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2577–2593, Jul. 2005.

[14] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge Univ. Press, 2006.

[15] T. Courtade and R. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *Proc. IEEE ISIT*, Aug. 2011, pp. 2040–2044.

[16] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999, pp. 368–377.

[17] P. Harremoes and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proc. IEEE ISIT*, Jun. 2007, pp. 566–570.

[18] T. Andre, M. Antonini, M. Barlaud, and R. Gray, "Entropy-based distortion measure for image coding," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 1157–1160.

[19] T. Berger, "The information theory approach to communications," in *Multiterminal Source Coding*, G. Longo, Ed. New York, NY, USA: Springer-Verlag, 1977.

[20] S.-Y. Tung, "Multiterminal source coding," Ph.D. dissertation, Dept. Electr. Eng., Cornell University, Ithaca, NY, USA, 1978.

[21] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[22] S. Jana, "Alphabet sizes of auxiliary random variables in canonical inner bounds," in *Proc. 43rd Annu. Conf. Information Sciences and Systems CISS*, Mar. 2009, pp. 67–71.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[24] R. Gilad-bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proc. COLT*, 2003, pp. 595–609.

[25] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *Proc. 23rd Eur. Colloq. Inf. Retr. Res.*, 2001, pp. 1–12.

[26] N. Slonim, R. Somerville, N. Tishby, and O. Lahav, "Objective classification of galaxy spectra using the information bottleneck method," *Monthly Notices R. Astronomical Soc.*, vol. 323, no. 2, pp. 270–284, 2001.

[27] A. Globerson, G. Chechik, N. Tishby, O. Steinberg, and E. Vaadia, "Distributional clustering of movements based on neural responses," 2001, unpublished.

[28] R. M. Hecht and N. Tishby, "Extraction of relevant speech features using the information bottleneck method," in *Proc. InterSpeech*, 2005, pp. 353–356.

[29] Y.-H. Kim, A. Sutivong, and T. Cover, "State amplification," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1850–1859, May 2008.

[30] J. Korner and K. Marton, "How to encode the modulo-two sum of binary sources (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 219–221, Mar. 1979.

[31] S. Jana, "Alphabet sizes of auxiliary random variables in canonical inner bounds," in *Proc. 43rd Annu. CISS* Mar. 2009, pp. 67–71, doi: 10.1109/CISS.2009.5054692, arXiv: 0810.1973 [cs.IT].

[32] W. Gu and M. Effros, "On approximating the rate region for source coding with coded side information," in *Proc. IEEE ITW*, Sep. 2007, pp. 432–435.

[33] W. Gu, S. Jana, and M. Effros, "On approximating the rate regions for lossy source coding with coded and uncoded side information," in *Proc. IEEE ISIT*, Jul. 2008, pp. 2162–2166.

[34] H. Witsenhausen, "Some aspects of convexity useful in information theory," *IEEE Trans. Inf. Theory*, vol. 26, no. 3, pp. 265–271, May 1980.

[35] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY, USA: Academic, 1981.

[36] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*. Berlin, Germany: Springer-Verlag, 2003.

[37] S. Fujishige, *Submodular Functions and Optimization*, 2nd ed. Berlin, Germany: Elsevier, 2010.

[38] S. McCormick, "Submodular function minimization," in *Handbooks in Operations Research and Management Science: Discrete Optimization*, vol. 12, K. Aardal, G. Nemhauser, and R. Weismantel, Eds. Amsterdam, The Netherlands: Elsevier, 2005.

**Thomas A. Courtade** (S'06–M'13) received the B.S. degree in Electrical Engineering from Michigan Technological University in 2007, and the M.S. and Ph.D. degrees in Electrical Engineering from UCLA in 2008 and 2012, respectively. In 2012, he was awarded the inaugural Postdoctoral Research Fellowship through the NSF Center for Science of Information. He currently holds this position, and resides at Stanford University. His honors include a Distinguished Ph.D. Dissertation award and an Excellence in Teaching award from the UCLA Department of Electrical Engineering and a Best Student Paper Award at the 2012 International Symposium on Information Theory.

**Tsachy Weissman** (S'99–M'02–SM'07–F'13) graduated summa cum laude with a B.Sc. in electrical engineering from the Technion in 1997, and earned his Ph.D. at the same place in 2001. He then worked at Hewlett-Packard Laboratories with the information theory group until 2003, when he joined Stanford University, where he is on the faculty of the Electrical Engineering department and incumbent of the STMicroelectronics chair in the School of Engineering.

He has spent leaves at the Technion, and at ETH Zurich. Tsachy's research is focused on information theory, statistical signal processing, the interplay between them, and their applications, with recent emphasis on applications in genomics. It has been recognized with various best paper awards, as well as awards for excellence in research and scholarship. He serves on the editorial boards of the IEEE TRANSACTIONS ON INFORMATION THEORY and Foundations and Trends in Communications and Information Theory.