Distributed Information-Theoretic Biclustering of Two Memoryless Sources

Georg Pichler, Pablo Piantanida, and Gerald Matz

***Abstract*—A novel multi-terminal source coding problem motivated by biclustering applications is investigated. In this setting, two separate encoders observe two dependent memoryless processes X^n and Z^n, respectively. The encoders' goal is to find rate-limited functions $f(\mathsf{X}^n)$ and $g(\mathsf{Z}^n)$ that maximize asymptotically the mutual information $\mathrm{I}(f(\mathsf{X}^n); g(\mathsf{Z}^n)) \geq n\mu$. We derive non-trivial inner and outer bounds on the optimal characterization of the achievable rates for this problem. Applications also arise in the context of distributed hypothesis testing against independence under communication constraints.**

I. Introduction

The recent decades witnessed a rapid proliferation of data available in digital form in a myriad of repositories such as internet fora, blogs, web applications, news, emails, and the significant social media bandwagon. A significant part of this information is unstructured and it is thus hard to find the relevant information. This results in a growing need for a fundamental understanding and efficient methods for analyzing data and discovering relevant knowledge from it in the form of structured information while reducing the entropy of data.

When specifying certain hidden (unobserved) features of interest, the problem then consists of extracting those relevant characteristics from data sets, while ignoring other, irrelevant features. Formulating this idea in terms of lossy source compression [1], we can assess the value of a coding scheme based on its rate and the information it provides about specific unobserved features.

A. Information Bottleneck

The *Information Bottleneck (IB) method* [2] has been successfully applied to machine learning and communications problems (e.g., [3], [4]). It can be understood as an alternative to lossy source coding that quantifies fidelity in terms of mutual information with a relevance variable [5] instead of a distortion measure. The idea is to identify relevant information from a vector X^n of observed samples as being the information that these samples provides about another hidden signal Z^n, e.g., the information that a posting in an Internet forum provides about the posters opinion regarding a particular topic might be considered relevant.

The IB method tries to find a (lossy) description $f(\mathsf{X}^n)$ subject to complexity requirements while preserving the maximum possible information (measure of relevance) about the unobserved quantity Z^n. More precisely, the statistician designs $\mathsf{U} = f(\mathsf{X}^n)$ to simultaneously satisfy the constraints:

$$\mathrm{I}(\mathsf{U};\mathsf{Z}^n) \geq n\mu \;\text{ subject to }\; \mathrm{I}(\mathsf{U};\mathsf{X}^n) \leq nR \qquad (1)$$

which provide asymptotic tradeoffs between statistical rates of *relevance* μ and *complexity* R.

Although the IB method was originally introduced from purely conceptual statistical considerations without a proof of optimality, it also poses an information-theoretic problem as was shown in [6]. An equivalent definition of this problem, as will be shown in Section IV, is based on noisy lossy source coding [7] via the logarithmic loss distortion [8].

B. Distributed Biclustering

Applying a function f to the process X^n, as done in the IB method, can be interpreted as a clustering of the possible outcomes of the experiment X^n. The goal then is to make the clustering as coarse as possible (minimizing complexity), while on the other hand maximizing the information that the cluster index provides regarding Z^n (maximizing relevance). The two variables X^n and Z^n play inherently different roles in this problem. However, a more general and symmetric formulation can be derived by simultaneously clustering Z^n. This principle is called *biclustering* (or *co-clustering*). It seems [9, Section 3.2.4] that the concept was first explicitly considered in [10], although not by that name. Given an $N \times M$ data matrix $A = (a_{nm})$, the goal of a biclustering algorithm [11] is to find partitions $B_k \subseteq [1:N]$ and $C_l \subseteq [1:M]$, $k = [1:K]$, $l = [1:L]$ such that all the "biclusters" $(a_{nm})_{n\in B_k, m\in C_l}$ are homogeneous in some specific sense. The criteria of "homogeneity" depends on the application. This method received renewed attention when Cheng and Church [12] applied biclustering to gene expression data. Several biclustering algorithms have since been developed in this field (e.g., see [13] and references therein).

G. Pichler and G. Matz are with the Institute of Telecommunications, TU Wien, Vienna, Austria.

P. Piantanida is with Laboratoire de Signaux et Systèmes (L2S, UMR8506), CentraleSupélec-CNRS-Université Paris-Sud, Gif-sur-Yvette, France.

Part of this work was supported by the FP7 Network of Excellence in Wireless COMmunications NEWCOM# and by the WWTF under grant ICT12-054 (TINCOIN).

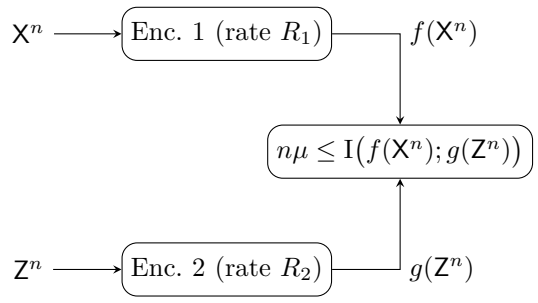

Fig. 1: Biclustering of two memoryless sources

Dhillon *et al.* [14] adopted an information-theoretic approach that is similar to the IB framework. They used mutual information to characterize the quality of a biclustering. For the special case, where A represents the joint probability distribution of two discrete random variables X and Z, i.e., $a_{nm} = \mathrm{P}\{\mathsf{X} = n, \mathsf{Z} = m\}$, the goal is to find functions $f\colon [1 : N] \to [1 : K]$ and $g\colon [1 : M] \to [1 : L]$ that maximize $\mathrm{I}\big(f(\mathsf{X}); g(\mathsf{Z})\big)$ for given K and L. In much the same way as the setup defined by the multi-letter constrains in (1), we show that biclustering fits into an information-theoretic framework.

C. Related Work and Contribution

In [15], Witsenhausen and Wyner investigated lower bounds for conditional entropy when simultaneously requiring another conditional entropy to fall below a threshold. Their work was a generalization of an earlier result [16] and also related to [17]–[20]. As a matter of fact, the conditional entropy bound [15] turns out to be an equivalent formulation of the multi-letter constrains in (1) and hence, the results in [15] provide a single-letter characterization of the IB method [2], [6]. A similar observation also holds from the characterization of the image of sets via noisy channels [21].

In this paper, we introduce and study the distributed biclustering problem from a formal information-theoretic perspective. Given distributed, dependent but memoryless samples X^n and Z^n observed at different encoders, the aim is to extract a description from each sample, such that the descriptions are maximally informative about each other. Each encoder tries to find a (lossy) description of its observation subject to complexity requirements (coding rate), such that the mutual information between the two descriptions is maximized. The goal is to characterize the optimal tradeoff between the rates of *relevance* and of *complexity*. As one can see in the schematic in Figure 1, no decoding takes place, but the rate-limited descriptions $f(\mathsf{X}^n)$ and $g(\mathsf{Z}^n)$ are compared directly.

This problem seems to have a formidable mathematical complexity. It appears to be closely related to hypothesis testing against independence with multiterminal data compression [22], which is not yet solved in general [23]. The solved case of full side-information [24] corresponds to the special case of the IB method.

We first provide an outer and an inner bound on the achievable region. In general there is a gap between the two bounds. The outer bound follows from standard information-theoretic manipulations, while the inner bound uses methods from [22].

Notation and Conventions

We denote random quantities and realizations by capital, sans-serif and lowercase letters, respectively. Random variables are assumed to be supported on finite sets. Bold type is used for vectors of length n and calligraphic type for sets. We use the same letter for the random variable and for its support set, e.g., the random variable X takes values in $\mathcal{X}$. For convenience we will use $[1 : N]$ to denote the set $\{1, 2, \ldots, N\}$. We use $\overline{\mathcal{A}}$ to denote the topological closure of a set $\mathcal{A}$ and $\mathcal{A}^{\mathrm{c}}$ for the complement of a set (or event). The convex hull of a set is written as $\mathrm{conv}(\mathcal{A})$. Given a random variable X, we write p_{X} for its pmf. We use the notation of [25, Chapter 2] for information-theoretic quantities. However, all logarithms in this paper are base e and therefore all information theoretic quantities are measured in nats. The symbol $\mathrm{h_b}(p) \triangleq -p \log p - (1-p) \log (1-p)$ is used for the binary entropy function and $a * b \triangleq a(1-b)+(1-a)b$ is the binary convolution operation. The notation $\mathsf{X} \multimap \mathsf{Y} \multimap \mathsf{Z}$ indicates that X, Y, and Z form a Markov chain in this order. When generating codebooks we will assume that the codebook size is an integer to keep the notation simple.

We will use the notion of robust typicality [26], also used in [22], which our results are heavily based upon. For a random variable X and $\delta \geq 0$, define the typical set as the set of random variables $\mathcal{T}_{[\mathsf{X}]\delta} \triangleq \{\widetilde{\mathsf{X}} | \forall x \in \mathcal{X} : |\mathrm{p}_{\mathsf{X}}(x) - \mathrm{p}_{\widetilde{\mathsf{X}}}(x)| \leq \delta \mathrm{p}_{\mathsf{X}}(x)\}$. For $n \in \mathbb{N}$ let the set of typical sequences be $\mathcal{T}^n_{[\mathsf{X}]\delta} \triangleq \{\hat{\boldsymbol{x}} \in \mathcal{X}^n | \hat{\mathsf{X}} \in \mathcal{T}_{[\mathsf{X}]\delta}\}$ where $\hat{\mathsf{X}}$ is the type variable corresponding to $\hat{\boldsymbol{x}}$ [27, Definition 2.1].

II. Problem Statement and Results

In this section, we present the biclustering problem as a distributed source coding problem and provide bounds for its achievable region.

Let X and Z be two random variables with joint distribution $\mathrm{p}_{\mathsf{X},\mathsf{Z}}(x, z)$, taking values in the finite sets $\mathcal{X}$ and $\mathcal{Z}$, respectively. The random n-vectors $(\mathbf{X}, \mathbf{Z})$ consist of i.i.d. copies of (X, Z). An (n, R_1, R_2) code (f, g) consists of two functions $f\colon \mathcal{X}^n \to \mathcal{M}_1$, and $g\colon \mathcal{Z}^n \to \mathcal{M}_2$ where $\mathcal{M}_k$ is an arbitrary finite set with $\log |\mathcal{M}_k| \leq nR_k$ for each $k \in \{1, 2\}$.

Definition 1 (Relevance)**.** *For an (n, R_1, R_2)-code (f, g), we define the* co-information of f and g *as*

$$\text{co-in}(f; g) \triangleq \frac{1}{n} \mathrm{I}(f(\mathbf{X}); g(\mathbf{Z})) \ .$$

Definition 2 (Rate region)**.** *A point (μ, R_1, R_2) of real values is said to be* achievable *if there exists an (n, R_1, R_2) code (f, g) for some $n \in \mathbb{N}$ such that*

$$\text{co-in}(f; g) \geq \mu \ .$$

Furthermore, let $\mathcal{R}$ be the set of all achievable points.

Our goal is to bound the achievable region $\overline{\mathcal{R}}$.

Definition 3. *Define $\mathcal{R}^o$ as the set of all points (μ, R_1, R_2) such that*

$$\begin{aligned} R_1 &\geq \mathrm{I}(\mathsf{U};\mathsf{X}) \\ R_2 &\geq \mathrm{I}(\mathsf{V};\mathsf{Z}) \\ \mu &\leq \mathrm{I}(\mathsf{U};\mathsf{X}) + \mathrm{I}(\mathsf{V};\mathsf{Z}) - \mathrm{I}(\mathsf{UV};\mathsf{XZ}) \end{aligned} \tag{2}$$

for some random variables U and V such that $\mathsf{U}\multimap\mathsf{X}\multimap\mathsf{Z}$ and $\mathsf{X}\multimap\mathsf{Z}\multimap\mathsf{V}$.

Theorem 4 (Outer bound)**.** $\mathcal{R} \subseteq \mathcal{R}^o$.

Proof: See Appendix A. ∎

In order to show an achievability for the biclustering problem, we employ the techniques derived in [22] by using a duality between this problem and the characterization of optimal error exponents for hypothesis testing problems. Essentially, we will need the proof of [22, Theorem 6], which itself is based on [22, Lemma 8]. These two results are extensions of [22, Theorem 2] and [22, Lemma 4] and the proofs follow very closely. However, since these results were stated without proof and they are employed in a different context here, we will provide the proofs for completeness sake. The more technical parts are deferred to the appendix.

Definition 5. *The region $\mathcal{R}^i$ is defined as the set of points (μ, R_1, R_2) such that*

$$R_1 \geq \mathrm{I}(\mathsf{U};\mathsf{X}) \tag{3}$$

$$R_2 \geq \mathrm{I}(\mathsf{V};\mathsf{Z}) \tag{4}$$

$$\mu \leq \mathrm{I}(\mathsf{U};\mathsf{V}) \tag{5}$$

for some random variables U,V, such that $\mathsf{U}\multimap\mathsf{X}\multimap\mathsf{Z}\multimap\mathsf{V}$.

Note that in the definition of $\mathcal{R}^o$, we can rewrite (2) as $\mu \leq \mathrm{I}(\mathsf{U};\mathsf{V}) - \mathrm{I}(\mathsf{U};\mathsf{V}|\mathsf{XZ})$. Thus, (2) would reduce to (5) provided that the minimizing distribution satisfies $\mathsf{U}\multimap\mathsf{X}\multimap\mathsf{Z}\multimap\mathsf{V}$.

Theorem 6 (Inner bound)**.** $\mathcal{R}^i \subseteq \overline{\mathcal{R}}$.

Remark 1. It seems impossible to bound the mutual information between two encodings using the standard tools of typicality coding and the Markov lemma [28]. This is due to the fact that mutual information relates only to the distribution of the encodings and not their values. Typicality coding provides statements about the values resulting from the decoding (i.e., typicality with high probability), but doesn't provide information about the distribution of the encoding.

Before proving Theorem 6, we need the following lemma, which is a slightly adapted version of [22, Lemma 8].

Lemma 7 ([22, Lemma 8])**.** *Let $\varepsilon > 0$ and $\mathsf{U}\multimap\mathsf{X}\multimap\mathsf{Z}\multimap\mathsf{V}$. Then, we can choose some suitably small $\delta > 0$ and suitably large $n, M_1, M_2 \in \mathbb{N}$ with*

$$\exp\big(n\mathrm{I}(\mathsf{U};\mathsf{X})\big) < M_1 \leq \exp\big(n(\mathrm{I}(\mathsf{U};\mathsf{X}) + \varepsilon)\big) \tag{6}$$

$$\exp\big(n\mathrm{I}(\mathsf{V};\mathsf{Z})\big) < M_2 \leq \exp\big(n(\mathrm{I}(\mathsf{V};\mathsf{Z}) + \varepsilon)\big) \tag{7}$$

such that the following properties hold: We can find $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{M_1} \in \mathcal{T}^n_{[\mathsf{U}]\delta}$, $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{M_2} \in \mathcal{T}^n_{[\mathsf{V}]\delta}$, pairwise disjoint sets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{M_1}$ and pairwise disjoint sets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{M_2}$, such that $\mathcal{C}_i \subseteq \mathcal{T}^n_{[\mathsf{X}|\mathsf{U}]\delta}(\boldsymbol{u}_i)$ and $\mathcal{D}_j \subseteq \mathcal{T}^n_{[\mathsf{Z}|\mathsf{V}]\delta}(\boldsymbol{v}_j)$ with the properties that

$$\sum_{i,j=1}^{M_1,M_2} \mathbb{1}_{\mathcal{T}^n_{[\mathsf{UV}]\delta}}(\boldsymbol{u}_i, \boldsymbol{v}_j) \cdot \mathrm{P}\{(\mathbf{X},\mathbf{Z}) \in \mathcal{C}_i \times \mathcal{D}_j\} \geq 1 - \varepsilon \tag{8}$$

and

$$\sum_{i,j=1}^{M_1,M_2} \mathbb{1}_{\mathcal{T}^n_{[\mathsf{UV}]\delta}}(\boldsymbol{u}_i, \boldsymbol{v}_j) \leq \exp\big(n(\mathrm{I}(\mathsf{UV};\mathsf{XZ}) + \varepsilon)\big) \ . \tag{9}$$

Proof: See Appendix B. ∎

In the course of the proof we will furthermore need the following set of random variables.

Definition 8. *For two random variables U and V, such that $\mathsf{U}\multimap\mathsf{X}\multimap\mathsf{Z}\multimap\mathsf{V}$, define for $\delta \geq 0$ the set of random variables*

$$\begin{aligned} \mathcal{L}_\delta(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V}) \triangleq \{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}} : (\widetilde{\mathsf{U}},\widetilde{\mathsf{X}}) \in \mathcal{T}_{[\mathsf{UX}]\delta}, \\ (\widetilde{\mathsf{V}},\widetilde{\mathsf{Z}}) \in \mathcal{T}_{[\mathsf{VZ}]\delta}, (\widetilde{\mathsf{U}},\widetilde{\mathsf{V}}) \in \mathcal{T}_{[\mathsf{UV}]\delta}\} \end{aligned}$$

and set $\mathcal{L}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V}) \triangleq \mathcal{L}_0(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})$.

Note that $\mathcal{L}_\delta(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V}) \subseteq \mathcal{L}_{\delta'}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})$ for $\delta \leq \delta'$.

Proof of Theorem 6: Select U and V such that $\mathsf{U}\multimap\mathsf{X}\multimap\mathsf{Z}\multimap\mathsf{V}$ and (3) to (5) hold. Fix $\varepsilon > 0$, let M_1 and M_2 satisfy (6) and (7) and let $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{M_1} \in \mathcal{T}^n_{[\mathsf{U}]\delta}$, $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{M_1}$, $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{M_2} \in \mathcal{T}^n_{[\mathsf{V}]\delta}$, and $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{M_2}$ be those given in Lemma 7 for a suitably small $\delta > 0$ and n large enough. Define the sets $\mathcal{C} \triangleq \bigcup_{i=1}^{M_1} \mathcal{C}_i$ and $\mathcal{D} \triangleq \bigcup_{j=1}^{M_2} \mathcal{D}_j$. Then define the code (f, g) as

$$f(\boldsymbol{x}) \triangleq \begin{cases} i & \boldsymbol{x} \in \mathcal{C}_i \\ 0 & \boldsymbol{x} \notin \mathcal{C} \end{cases}, \qquad g(\boldsymbol{z}) \triangleq \begin{cases} j & \boldsymbol{z} \in \mathcal{D}_j \\ 0 & \boldsymbol{z} \notin \mathcal{D} \end{cases}.$$

This is an $(n, \mathrm{I}(\mathsf{U};\mathsf{X}) + c\varepsilon, \mathrm{I}(\mathsf{V};\mathsf{Z}) + c\varepsilon)$ code for any constant $c > 1$ if n is large enough. We now need to analyze co-in$(f; g)$. Let $\overline{\mathsf{X}}$ and $\overline{\mathsf{Z}}$ be random variables distributed according to $\mathrm{p}_{\overline{\mathsf{X}},\overline{\mathsf{Z}}}(x, z) = \mathrm{p}_{\mathsf{X}}(x)\mathrm{p}_{\mathsf{Z}}(z)$, i.e., with the same marginals as X and Z, but independent. Defining $\mathsf{W}_1 \triangleq f(\mathbf{X})$, $\mathsf{W}_2 \triangleq g(\mathbf{Z})$, $\overline{\mathsf{W}}_1 \triangleq f(\overline{\mathbf{X}})$, $\overline{\mathsf{W}}_2 \triangleq g(\overline{\mathbf{Z}})$ and $\mathcal{F} \triangleq \{(i, j) : (\boldsymbol{u}_i, \boldsymbol{v}_j) \in \mathcal{T}^n_{[\mathsf{UV}]\delta}\}$, we have

$$\begin{aligned} n \cdot \text{co-in}(f; g) &= \mathrm{I}(f(\mathbf{X}); g(\mathbf{Z})) \\ &= \sum_{i,j} \mathrm{p}_{\mathsf{W}_1,\mathsf{W}_2}(i, j) \log \frac{\mathrm{p}_{\mathsf{W}_1,\mathsf{W}_2}(i, j)}{\mathrm{p}_{\mathsf{W}_1}(i)\mathrm{p}_{\mathsf{W}_2}(j)} \\ &= \sum_{i,j\in\mathcal{F}} \mathrm{p}_{\mathsf{W}_1,\mathsf{W}_2}(i, j) \log \frac{\mathrm{p}_{\mathsf{W}_1,\mathsf{W}_2}(i, j)}{\mathrm{p}_{\mathsf{W}_1}(i)\mathrm{p}_{\mathsf{W}_2}(j)} \\ &\quad + \sum_{i,j\in\mathcal{F}^{\mathrm{c}}} \mathrm{p}_{\mathsf{W}_1,\mathsf{W}_2}(i, j) \log \frac{\mathrm{p}_{\mathsf{W}_1,\mathsf{W}_2}(i, j)}{\mathrm{p}_{\mathsf{W}_1}(i)\mathrm{p}_{\mathsf{W}_2}(j)} \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \mathrm{P}\{(\mathsf{W}_1,\mathsf{W}_2)\in\mathcal{F}\}\log\frac{\mathrm{P}\{(\mathsf{W}_1,\mathsf{W}_2)\in\mathcal{F}\}}{\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}}\\
&+\mathrm{P}\{(\mathsf{W}_1,\mathsf{W}_2)\in\mathcal{F}^{\mathrm{c}}\}\log\frac{\mathrm{P}\{(\mathsf{W}_1,\mathsf{W}_2)\in\mathcal{F}^{\mathrm{c}}\}}{\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}^{\mathrm{c}}\big\}}\\
&\geq -\mathrm{P}\{(\mathsf{W}_1,\mathsf{W}_2)\in\mathcal{F}\}\log\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}\\
&-\mathrm{h_b}(\mathrm{P}\{(\mathsf{W}_1,\mathsf{W}_2)\in\mathcal{F}\})\\
&\overset{(8)}{\geq} -(1-\varepsilon)\log\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}\\
&-\mathrm{h_b}(\mathrm{P}\{(\mathsf{W}_1,\mathsf{W}_2)\in\mathcal{F}\})\\
&\geq -\log(2)-(1-\varepsilon)\log\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}\ , \qquad (10)
\end{aligned}
$$

where (a) follows from the log sum inequality [25, Theorem 2.7.1]. For each $i\in[1:M_1]$ and $j\in[1:M_2]$ define

$$\mathcal{S}(i,j)\triangleq\{\boldsymbol{u}_i\}\times\mathcal{C}_i\times\mathcal{D}_j\times\{\boldsymbol{v}_j\}$$

and

$$\mathfrak{S}\triangleq\bigcup_{i,j\in\mathcal{F}}\mathcal{S}(i,j)\ .$$

Pick any $(\hat{\boldsymbol{u}},\hat{\boldsymbol{x}},\hat{\boldsymbol{z}},\hat{\boldsymbol{v}})\in\mathfrak{S}$ and let $\hat{\mathsf{U}}$, $\hat{\mathsf{X}}$, $\hat{\mathsf{Z}}$, and $\hat{\mathsf{V}}$ be the type variables corresponding to $\hat{\boldsymbol{u}}$, $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{z}}$, and $\hat{\boldsymbol{v}}$ [27, Definition 2.1]. We then have

$$
\begin{aligned}
\mathrm{p}_{\overline{\mathsf{X}},\overline{\mathsf{Z}}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{z}})=\exp\Big(&-n\big(\mathrm{H}(\hat{\mathsf{X}}\hat{\mathsf{Z}})\\
&+\mathrm{D_{KL}}(\hat{\mathsf{X}}\hat{\mathsf{Z}}\|\overline{\mathsf{X}}\overline{\mathsf{Z}})\big)\Big) \qquad (11)
\end{aligned}
$$

from the properties of types [27, Lemma 2.6]. Let $K(i,j)$ be the number of elements in $\mathcal{S}(i,j)$ with type $(\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}})$, then [27, Lemma 2.5]

$$K(i,j)\leq\exp\Big(n\mathrm{H}(\hat{\mathsf{X}}\hat{\mathsf{Z}}|\hat{\mathsf{U}}\hat{\mathsf{V}})\Big).$$

Let K be the number of elements of $\mathfrak{S}$ with type $(\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}})$. Then

$$
\begin{aligned}
K&=\sum_{(i,j)\in\mathcal{F}}K(i,j)\\
&\leq\sum_{(i,j)\in\mathcal{F}}\exp\Big(n\mathrm{H}(\hat{\mathsf{X}}\hat{\mathsf{Z}}|\hat{\mathsf{U}}\hat{\mathsf{V}})\Big)\\
&\overset{(9)}{\leq}\exp\Big(n\big(\mathrm{I}(\mathsf{UV};\mathsf{XZ})+\mathrm{H}(\hat{\mathsf{X}}\hat{\mathsf{Z}}|\hat{\mathsf{U}}\hat{\mathsf{V}})+\varepsilon\big)\Big) \qquad (12)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}\\
&\overset{(11)}{=}\sum_{\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}}}K\cdot\exp\Big(-n\big(\mathrm{H}(\hat{\mathsf{X}}\hat{\mathsf{Z}})+\mathrm{D_{KL}}(\hat{\mathsf{X}}\hat{\mathsf{Z}}\|\overline{\mathsf{X}}\overline{\mathsf{Z}})\big)\Big)\\
&\overset{(12)}{\leq}\sum_{\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}}}\exp\Big(-n\big(k(\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}})-\varepsilon\big)\Big)
\end{aligned}
$$

where the sum is over all types that occur in $\mathfrak{S}$ and

$$
\begin{aligned}
k(\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}})\triangleq\ &\mathrm{I}(\hat{\mathsf{U}}\hat{\mathsf{V}};\hat{\mathsf{X}}\hat{\mathsf{Z}})\\
&-\mathrm{I}(\mathsf{UV};\mathsf{XZ})+\mathrm{D_{KL}}(\hat{\mathsf{X}}\hat{\mathsf{Z}}\|\overline{\mathsf{X}}\overline{\mathsf{Z}})\ .
\end{aligned}
$$

We can further bound

$$
\begin{aligned}
\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}&\leq(n+1)^{|\mathcal{U}||\mathcal{X}||\mathcal{Z}||\mathcal{V}|}\\
&\cdot\max_{\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}}}\exp\Big(-n\big(k(\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}})-\varepsilon\big)\Big) \qquad (13)
\end{aligned}
$$

where the maximum is over all types occurring in $\mathfrak{S}$. For any type $(\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}})$ in $\mathfrak{S}$, we have by definition $(\hat{\mathsf{U}},\hat{\mathsf{X}},\hat{\mathsf{Z}},\hat{\mathsf{V}})\in\mathcal{L}_\delta(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})$ and thus, by (13),

$$
\begin{aligned}
&\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}\leq(n+1)^{|\mathcal{U}||\mathcal{X}||\mathcal{Z}||\mathcal{V}|}\\
&\cdot\max_{(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}_\delta(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})}\exp\Big(-n\big(k(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})-\varepsilon\big)\Big)\ . \qquad (14)
\end{aligned}
$$

Combining (10) and (14) we showed for n large enough

$$
\begin{aligned}
\text{co-in}(f;g)&\geq-\frac{\log(2)}{n}-\frac{1-\varepsilon}{n}\log\mathrm{P}\big\{(\overline{\mathsf{W}}_1,\overline{\mathsf{W}}_2)\in\mathcal{F}\big\}\\
&\geq-\varepsilon+(1-\varepsilon)\\
&\quad\cdot\min_{(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}_\delta(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})}k(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})-\varepsilon\\
&\geq\min_{(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}_\delta(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})}k(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})-C\varepsilon \qquad (15)
\end{aligned}
$$

for some constant C. As $k(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})$ is continuous as a function of $\mathrm{p}_{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}}}$ and (15) holds for arbitrarily small δ if n is large enough, we obtain

$$\text{co-in}(f;g)\geq\min_{(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})}k(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})-C'\varepsilon$$

for some constant C'. Observe that for $(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})$, basic manipulations yield

$$k(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})=\mathrm{D_{KL}}(\widetilde{\mathsf{U}}\widetilde{\mathsf{X}}\widetilde{\mathsf{Z}}\widetilde{\mathsf{V}}\|\overline{\mathsf{U}\mathsf{X}\mathsf{Z}\mathsf{V}})$$

where $\overline{\mathsf{U}}$ and $\overline{\mathsf{V}}$ are random variables uniquely determined by the requirements $\overline{\mathsf{U}}\text{-}\circ\text{-}\overline{\mathsf{X}}\text{-}\circ\text{-}\overline{\mathsf{Z}}\text{-}\circ\text{-}\overline{\mathsf{V}}$, $\mathrm{p}_{\overline{\mathsf{X}},\overline{\mathsf{U}}}=\mathrm{p}_{\mathsf{X},\mathsf{U}}$, and $\mathrm{p}_{\overline{\mathsf{Z}},\overline{\mathsf{V}}}=\mathrm{p}_{\mathsf{Z},\mathsf{V}}$. Observing that $\mathrm{D_{KL}}(\widetilde{\mathsf{U}}\widetilde{\mathsf{X}}\widetilde{\mathsf{Z}}\widetilde{\mathsf{V}}\|\overline{\mathsf{U}\mathsf{X}\mathsf{Z}\mathsf{V}})=\mathrm{I}(\widetilde{\mathsf{X}}\widetilde{\mathsf{U}};\widetilde{\mathsf{Z}}\widetilde{\mathsf{V}})$, we showed, that $(\mu,R_1,R_2)\in\mathcal{R}$ with $\mu=\min_{(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})}\mathrm{I}(\widetilde{\mathsf{X}}\widetilde{\mathsf{U}};\widetilde{\mathsf{Z}}\widetilde{\mathsf{V}})-C'\varepsilon$, $R_1=\mathrm{I}(\mathsf{X};\mathsf{U})+c\varepsilon$, and $R_2=\mathrm{I}(\mathsf{Z};\mathsf{V})+c\varepsilon$. To complete the proof we only need to show the equality

$$\mathrm{I}(\mathsf{U};\mathsf{V})=\min_{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}}\in\mathcal{L}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})}\mathrm{I}(\widetilde{\mathsf{X}}\widetilde{\mathsf{U}};\widetilde{\mathsf{Z}}\widetilde{\mathsf{V}})\ .$$

The direction "$\leq$" is clear as $\mathrm{I}(\widetilde{\mathsf{X}}\widetilde{\mathsf{U}};\widetilde{\mathsf{Z}}\widetilde{\mathsf{V}})\geq\mathrm{I}(\widetilde{\mathsf{U}};\widetilde{\mathsf{V}})$ and for $(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})$ we have $\mathrm{p}_{\widetilde{\mathsf{U}},\widetilde{\mathsf{V}}}=\mathrm{p}_{\mathsf{U},\mathsf{V}}$. To see the other direction, define $(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})$ to be distributed according to $\mathrm{p}_{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}}}(u,x,z,v)=\mathrm{p}_{\mathsf{U},\mathsf{V}}(u,v)\mathrm{p}_{\mathsf{X}|\mathsf{U}}(x|u)\mathrm{p}_{\mathsf{Z}|\mathsf{V}}(z|v)$. Apparently $(\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}})\in\mathcal{L}(\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V})$ and $\mathrm{I}(\widetilde{\mathsf{X}}\widetilde{\mathsf{U}};\widetilde{\mathsf{Z}}\widetilde{\mathsf{V}})=\mathrm{I}(\mathsf{U};\mathsf{V})$. ∎

Remark 2. Perhaps unsurprisingly, common information [29] helps in maximizing the achievable region. As a simple consequence of Theorem 4, we have for any achievable point (μ,R_1,R_2) that $\mu\leq\min(R_1,R_2)$. Now let $\mathsf{Y}=\phi_1(\mathsf{X})=\phi_2(\mathsf{Z})$ (a.s.) be a common part of X and Z. Choosing $\mathsf{U}=\mathsf{V}=\mathsf{Y}$ in Definition 5, we see from Theorem 6 that $(\mathrm{H}(\mathsf{Y}),\mathrm{H}(\mathsf{Y}),\mathrm{H}(\mathsf{Y}))$ is achievable.

Using time-sharing with the trivially achievable point $(0,0,0)$ we see that the inner bound is tight when $\mu \le \mathrm{H}(\mathsf{Y})$.

The inner bound $\mathcal{R}^i$ can be further improved by convexification. Moreover, we introduce cardinality bounds to make it computable.

Proposition 9. *Let $\mathcal{R}'$ be the set of points (μ, R_1, R_2) such that*

$$\begin{aligned} R_1 &\ge \mathrm{I}(\mathsf{X};\mathsf{U}|\mathsf{Q}) \\ R_2 &\ge \mathrm{I}(\mathsf{Z};\mathsf{V}|\mathsf{Q}) \\ \mu &\le \mathrm{I}(\mathsf{U};\mathsf{V}|\mathsf{Q}) \end{aligned}$$

where U, V, and Q are random variables, such that $\mathrm{p}_{\mathsf{U},\mathsf{V},\mathsf{Q}|\mathsf{X},\mathsf{Z}} = \mathrm{p}_{\mathsf{Q}}\mathrm{p}_{\mathsf{U}|\mathsf{X},\mathsf{Q}}\mathrm{p}_{\mathsf{V}|\mathsf{Z},\mathsf{Q}}$ and $|\mathcal{U}| \le |\mathcal{X}|$, $|\mathcal{V}| \le |\mathcal{Z}|$, and $|\mathcal{Q}| \le 3$. We then have $\mathcal{R}' = \mathrm{conv}(\mathcal{R}^i) \subseteq \overline{\mathcal{R}}$.

Proof: See Appendix C. ∎

III. An Application Example: Doubly Symmetric Binary Source

In this section let (X,Z) be a doubly symmetric binary source [30, Example 10.1] with parameter p, i.e., $\mathsf{X} \sim \mathcal{B}(\frac{1}{2})$ is a Bernoulli random variable with parameter $\frac{1}{2}$, $\mathsf{N} \sim \mathcal{B}(p)$ and $\mathsf{Z} = \mathsf{X} \oplus \mathsf{N}$. $\mathcal{R}'$ as defined in Proposition 9 is the convex hull of all points (μ, R_1, R_2) such that there exist two stochastic binary (not necessarily symmetric) channels $\mathsf{X} \to \mathsf{U}$ and $\mathsf{Z} \to \mathsf{V}$ with

$$\begin{aligned} R_1 &\ge \mathrm{I}(\mathsf{X};\mathsf{U}) \\ R_2 &\ge \mathrm{I}(\mathsf{Z};\mathsf{V}) \\ \mu &\le \mathrm{I}(\mathsf{U};\mathsf{V}) \ . \end{aligned}$$

Let the region $\mathcal{R}_b$ be the set of all points (μ, R_1, R_2) such that there exist parameters $0 \le \alpha, \beta \le \frac{1}{2}$ with

$$\begin{aligned} R_1 &\ge \log 2 - \mathrm{h_b}(\alpha) \\ R_2 &\ge \log 2 - \mathrm{h_b}(\beta) \\ \mu &\le \log 2 - \mathrm{h_b}(\alpha * p * \beta) \ . \end{aligned}$$

We obtain $\mathcal{R}_b$ from $\mathcal{R}'$ by forcing the channels to be BSCs and therefore have $\mathcal{R}_b \subseteq \mathcal{R}'$. Based on numerical evaluation we conjecture the following result, which implies $\mathrm{conv}(\mathcal{R}_b) = \mathcal{R}'$.

Conjecture 10. *Given two stochastic binary channels $\mathsf{X} \to \mathsf{U}$ and $\mathsf{Z} \to \mathsf{V}$, there exist parameters $0 \le \alpha, \beta \le \frac{1}{2}$ with*

$$\begin{aligned} \mathrm{I}(\mathsf{X};\mathsf{U}) &\ge \log 2 - \mathrm{h_b}(\alpha) \\ \mathrm{I}(\mathsf{Z};\mathsf{V}) &\ge \log 2 - \mathrm{h_b}(\beta) \\ \mathrm{I}(\mathsf{U};\mathsf{V}) &\le \log 2 - \mathrm{h_b}(\alpha * p * \beta) \ . \end{aligned}$$

To illustrate the tradeoff between complexity (R_1, R_2) and relevance (μ), the upper boundary of $\mathcal{R}_b$ is depicted in Figure 2 for $p = \frac{1}{4}$.

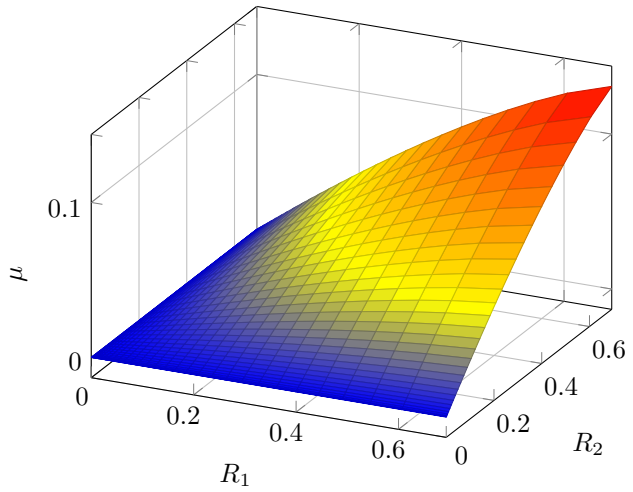

Fig. 2: Boundary of $\mathcal{R}_b$ for $p = \frac{1}{4}$

IV. Noisy Lossy Source Coding and Information Bottleneck

The biclustering problem introduced in Section II can be solved exactly in terms of a single-letter characterization in the special case when one rate is "large". If $R_2 \ge \log|\mathcal{Z}|$, Encoder 2 can transmit $\mathbf{Z}$ directly, which clearly maximizes $\mathrm{I}(f(\mathbf{X}); g(\mathbf{Z})) = \mathrm{I}(f(\mathbf{X}); \mathbf{Z})$. The problem then reduces to the IB, with the following two-dimensional achievable region.

Definition 11. *A point (μ, R) is* IB-achievable, *if for some $n \in \mathbb{N}$, there exists an encoding function $f\colon \mathcal{X}^n \to \mathcal{M}$ with $\log|\mathcal{M}| \le nR$ such that*

$$\text{co-in}(f; \mathrm{id}_{\mathcal{Z}^n}) = \frac{1}{n}\mathrm{I}(f(\mathbf{X}); \mathbf{Z}) \ge \mu \ , \tag{16}$$

where $\mathrm{id}_{\mathcal{Z}^n}$ is the identity function on $\mathcal{Z}^n$. Denote the set of all IB-achievable points $\mathcal{R}_{IB}$.

As mentioned before, we can obtain $\mathcal{R}_{\mathrm{IB}}$ as a special case from $\mathcal{R}$ because $(\mu, R) \in \mathcal{R}_{\mathrm{IB}}$ if and only if $(\mu, R, \log|\mathcal{Z}|) \in \mathcal{R}$.

Remark 3. As the inner bound $\mathrm{conv}(\mathcal{R}^i)$ and the outer bound $\mathcal{R}^o$ coincide for large R_2, we have that $(\mu, R) \in \overline{\mathcal{R}_{\mathrm{IB}}}$ if and only if there exist random variables Q and U such that $\mathrm{p}_{\mathsf{Q},\mathsf{U},\mathsf{X},\mathsf{Z}} = \mathrm{p}_{\mathsf{Q}}\mathrm{p}_{\mathsf{X},\mathsf{Z}}\mathrm{p}_{\mathsf{U}|\mathsf{X},\mathsf{Q}}$ with

$$\begin{aligned} \mathrm{I}(\mathsf{U};\mathsf{X}|\mathsf{Q}) &\le R \\ \mathrm{I}(\mathsf{U};\mathsf{Z}|\mathsf{Q}) &\ge \mu \ . \end{aligned}$$

We will now formulate a noisy lossy source coding problem [7] with a distortion constraint, which turns out to be equivalent to Definition 11, when choosing logarithmic loss distortion [8, Section II].

Definition 12 (Noisy lossy source coding)**.** *Given random variables (X,Z) and for each $n \in \mathbb{N}$, a reconstruction alphabet $\widetilde{\mathcal{Z}}_n$ with a distortion function $d\colon \widetilde{\mathcal{Z}}_n \times \mathcal{Z}^n \to \mathbb{R}_+$, a point (ρ, R) is d-*achievable *if for some $n \in \mathbb{N}$, there exists an encoding function $f\colon \mathcal{X}^n \to \mathcal{M}$ and a decoding function $g\colon \mathcal{M} \to \widetilde{\mathcal{Z}}$ such that $\log|\mathcal{M}| \le nR$ and*

$\mathbb{E}[d(g(f(\mathbf{X})),\mathbf{Z})] \leq \rho$. *Denote the set of all d-achievable points* $\mathcal{R}_d$.

Choose the reconstruction alphabet as $\widetilde{\mathcal{Z}}_n = \mathcal{P}(\mathcal{Z}^n)$, the set of all pmfs on $\mathcal{Z}^n$ and the logarithmic loss distortion [8, Section II] d_{LL}, defined as

$$d_{\mathrm{LL}} \colon \mathcal{P}(\mathcal{Z}^n) \times \mathcal{Z}^n \to \mathbb{R}_+ \ , \qquad (\mathrm{p}, \boldsymbol{z}) \mapsto -\frac{1}{n} \log \mathrm{p}(\boldsymbol{z}) \ .$$

We next argue that $\mathcal{R}_{\mathrm{IB}}$ and $\mathcal{R}_{d_{\mathrm{LL}}}$ are equivalent.

Lemma 13. *A point* (μ, R) *is in* $\mathcal{R}_{IB}$ *if and only if* $(\mathrm{H}(\mathsf{Z})-\mu, R) \in \mathcal{R}_{d_{LL}}$.

In order to show Lemma 13, we need the following result from [8].

Lemma 14 ([8, Lemma 1])**.** *For any encoding function* f *and decoding function* g*,*

$$\mathbb{E}[d_{LL}(g(f(\mathbf{X})),\mathbf{Z})] \geq \frac{1}{n}\mathrm{H}(\mathbf{Z}|f(\mathbf{X})) \ , \tag{17}$$

with equality iff $g(m) = \mathrm{P}\{\mathbf{Z} = \,\cdot\, | f(\mathbf{X}) = m\}$.

Proof of Lemma 13: To show the first part assume $(\mathrm{H}(\mathsf{Z}) - \mu, R) \in \mathcal{R}_{d_{\mathrm{LL}}}$. We obtain f and g as given in Definition 12. With Lemma 14 we have

$$\begin{aligned} \frac{1}{n}\mathrm{I}(\mathbf{Z}; f(\mathbf{X})) &= \mathrm{H}(\mathsf{Z}) - \frac{1}{n}\mathrm{H}(\mathbf{Z}|f(\mathbf{X})) \\ &\overset{(17)}{\geq} \mathrm{H}(\mathsf{Z}) - \mathbb{E}[d_{\mathrm{LL}}(g(f(\mathbf{X})),\mathbf{Z})] \\ &\geq \mu \ . \end{aligned}$$

This shows $(\mu, R) \in \mathcal{R}_{\mathrm{IB}}$.

For the second part assume $(\mu, R) \in \mathcal{R}_{\mathrm{IB}}$ and obtain an encoding function f such that (16) holds. Choosing g to obtain equality in (17), we obtain from Lemma 14,

$$\begin{aligned} \mathbb{E}[d_{\mathrm{LL}}(g(f(\mathbf{X})),\mathbf{Z})] &= \frac{1}{n}\mathrm{H}(\mathbf{Z}|(f(\mathbf{X})) \\ &= \mathrm{H}(\mathsf{Z}) - \frac{1}{n}\mathrm{I}(\mathbf{Z}; f(\mathbf{X})) \\ &\overset{(16)}{\leq} \mathrm{H}(\mathsf{Z}) - \mu \ . \end{aligned}$$

Therefore, we have $(\mathrm{H}(\mathsf{Z}) - \mu, R) \in \mathcal{R}_{d_{\mathrm{LL}}}$. ∎

Thus, we could have also obtained the statement in Remark 3 by applying Dobrushin's result [7] to logarithmic loss distortion.

V. Summary and Discussion

The biclustering problem was introduced as a multiterminal source coding problem. We provided and outer and an inner bound on the achievable region. In general these bounds do not meet. However, in case one rate is "large", we obtain a tight bound. This case was also shown to be equivalent to a rate-distortion problem. Furthermore we provided a closed form expression of an inner bound for a binary example.

While the inner bound seems sufficient, the outer bound appears to be severely lacking. Obtaining a good upper bound for the mutual information between two arbitrary encodings solely based on their rates is a difficult task. Standard information-theoretic manipulations appear incapable of handling this dependence well. However, we conjecture that the key for solving this problem lies in improving the outer bound.

Appendix

A. Proof of Theorem 4

For $(\mu, R_1, R_2) \in \mathcal{R}$, let (f, g) be an (n, R_1, R_2) code for some $n \in \mathbb{N}$ such that co-in$(f; g) \geq \mu$. We define the random variables $\mathsf{U}_l \triangleq \big(\mathbf{X}^{l-1}, f(\mathbf{X})\big)$ and $\mathsf{V}_l \triangleq \big(\mathbf{Z}^{l-1}, g(\mathbf{Z})\big)$ and obtain

$$\begin{aligned} nR_1 \geq \mathrm{H}\big(f(\mathbf{X})\big) &= \mathrm{I}\big(f(\mathbf{X}); \mathbf{X}\big) \\ &= \sum_{l=1}^{n} \mathrm{I}\Big(\mathbf{X}_l; f(\mathbf{X})\Big|\mathbf{X}^{l-1}\Big) \\ &= \sum_{l=1}^{n} \mathrm{I}(\mathbf{X}_l; \mathsf{U}_l) \end{aligned}$$

and accordingly

$$nR_2 \geq \sum_{l=1}^{n} \mathrm{I}(\mathbf{Z}_l; \mathsf{V}_l) \ .$$

We also have

$$\begin{aligned} n\mu &\leq \mathrm{I}\big(f(\mathbf{X}); g(\mathbf{Z})\big) \\ &= \mathrm{I}\big(f(\mathbf{X}); \mathbf{X}\big) - \mathrm{I}\big(f(\mathbf{X}); \mathbf{X}\big|g(\mathbf{Z})\big) \\ &= \mathrm{I}\big(f(\mathbf{X}); \mathbf{X}\big) + \mathrm{I}\big(g(\mathbf{Z}); \mathbf{Z}\big) \\ &\quad - \mathrm{I}\big(f(\mathbf{X}); \mathbf{X}\big|g(\mathbf{Z})\big) - \mathrm{I}\big(g(\mathbf{Z}); \mathbf{Z}\big) \\ &= \mathrm{I}\big(f(\mathbf{X}); \mathbf{X}\big) + \mathrm{I}\big(g(\mathbf{Z}); \mathbf{Z}\big) \\ &\quad - \mathrm{I}\big(f(\mathbf{X}); \mathbf{XZ}\big|g(\mathbf{Z})\big) - \mathrm{I}\big(g(\mathbf{Z}); \mathbf{XZ}\big) \\ &= \mathrm{I}\big(f(\mathbf{X}); \mathbf{X}\big) + \mathrm{I}\big(g(\mathbf{Z}); \mathbf{Z}\big) \\ &\quad - \mathrm{I}\big(f(\mathbf{X}), g(\mathbf{Z}); \mathbf{XZ}\big) \\ &= \sum_{l=1}^{n} \Big[\mathrm{I}(\mathsf{U}_l; \mathbf{X}_l) + \mathrm{I}(\mathsf{V}_l; \mathbf{Z}_l) - \mathrm{I}(\mathsf{U}_l\mathsf{V}_l; \mathbf{X}_l\mathbf{Z}_l)\Big] \ . \end{aligned}$$

Now a standard time-sharing argument shows $\mathcal{R} \subseteq \mathcal{R}^o$. ∎

B. Proof of Lemma 7

Fix $\varepsilon' > 0$ and $n \in \mathbb{N}$. For n sufficiently large we find $M_1, M_2 \in \mathbb{N}$ satisfying (6) and (7). We can thus apply [31, Lemma 3.4] (with $\{1,2\} \to \Sigma$, $\mathsf{U} \to U_1$, $\mathsf{V} \to U_2$, $\mathsf{X} \to X_1$, $\mathsf{Z} \to X_2$, $\{1\} \to \Psi$, and $\{1,2\} \to \Sigma_1$). Denote the codebooks $\mathcal{C}_{\mathsf{U}} \triangleq (\hat{\mathbf{U}}_i)_{i=[1:M_1]}$ and $\mathcal{C}_{\mathsf{V}} \triangleq (\hat{\mathbf{V}}_j)_{j=[1:M_2]}$, which are drawn independently uniform from $\mathcal{T}^n_{[\mathsf{U}]\delta}$ and from $\mathcal{T}^n_{[\mathsf{V}]\delta}$, respectively, where $\delta > 0$ is suitably small. Denoting the resulting randomized coding functions as $\mathbf{U}^* = f(\mathbf{X}, \mathcal{C}_{\mathsf{U}})$ and $\mathbf{V}^* = g(\mathbf{Z}, \mathcal{C}_{\mathsf{V}})$, we have

$$P_e \triangleq \mathrm{P}\Big\{(\mathbf{U}^*, \mathbf{X}, \mathbf{Z}, \mathbf{V}^*) \notin \mathcal{T}^n_{[\mathsf{UXZV}]\delta}\Big\} \leq \varepsilon' \ .$$

if n is chosen large enough.

We next analyze the random quantity $\mathsf{L} \triangleq \sum_{i,j=1}^{M_1,M_2} \mathbb{1}_{\mathcal{T}^n_{[\mathsf{UV}]\delta}}(\hat{\mathbf{U}}_i, \hat{\mathbf{V}}_j)$. For n large enough, we have

$$\begin{aligned}
\mathbb{E}[\mathsf{L}] &= \sum_{i,j=1}^{M_1,M_2} \mathbb{E}\Big[\mathbb{1}_{\mathcal{T}^n_{[\mathsf{UV}]\delta}}(\hat{\mathbf{U}}_i, \hat{\mathbf{V}}_j)\Big] \\
&= \sum_{i,j=1}^{M_1,M_2} \frac{\left|\mathcal{T}^n_{[\mathsf{UV}]\delta}\right|}{\left|\mathcal{T}^n_{[\mathsf{U}]\delta}\right|\left|\mathcal{T}^n_{[\mathsf{V}]\delta}\right|} \\
&\leq M_1 M_2 \frac{\mathrm{e}^{n(\mathrm{H}(\mathsf{UV})+\varepsilon_1(\delta))}}{\mathrm{e}^{n(\mathrm{H}(\mathsf{U})+\mathrm{H}(\mathsf{V})-\varepsilon_2(\delta))}} \qquad (18) \\
&\leq M_1 M_2 \mathrm{e}^{-n(\mathrm{I}(\mathsf{U};\mathsf{V})-\varepsilon_3(\delta))} \qquad (19)
\end{aligned}$$

where $\varepsilon_1(\delta), \varepsilon_2(\delta), \varepsilon_3(\delta) \to 0$ as $\delta \to 0$. Here (18) follows from the properties of typical sets [30, Section 2.4]. If δ is suitably small, we can choose $\varepsilon_1, \varepsilon_2 < \varepsilon$ such that $\varepsilon_1 + \varepsilon_2 + \varepsilon_3(\delta) < \varepsilon$. Requiring $M_1 \leq \mathrm{e}^{n(\mathrm{I}(\mathsf{U};\mathsf{X})+\varepsilon_1)}$ and $M_2 \leq \mathrm{e}^{n(\mathrm{I}(\mathsf{V};\mathsf{Z})+\varepsilon_2)}$, we have from (19)

$$\mathbb{E}[\mathsf{L}] \leq \exp\big(n(\mathrm{I}(\mathsf{UV};\mathsf{XZ}) + \varepsilon_1 + \varepsilon_2 + \varepsilon_3(\delta))\big)$$

and know from Markov's inequality for n large enough

$$\begin{aligned}
&\mathrm{P}\big\{\mathsf{L} \geq \exp\big(n(\mathrm{I}(\mathsf{UV};\mathsf{XZ}) + \varepsilon)\big)\big\} \\
&\quad \leq \exp\big(n(\varepsilon_1 + \varepsilon_2 + \varepsilon_3(\delta) - \varepsilon)\big) \\
&\quad \leq \varepsilon' \ .
\end{aligned}$$

Define the error events $\mathcal{E}_1 = \{(\mathbf{U}^*, \mathbf{X}, \mathbf{Z}, \mathbf{V}^*) \notin \mathcal{T}^n_{[\mathsf{UXZV}]\delta}\}$ and $\mathcal{E}_2 = \{\mathsf{L} > \exp\big(n(\mathrm{I}(\mathsf{UV};\mathsf{XZ}) + \varepsilon)\big)\}$. From Markov's inequality we know

$$\begin{aligned}
\mathrm{P}\Big\{\mathrm{P}\{\mathcal{E}_1|\mathcal{C}_\mathsf{U}, \mathcal{C}_\mathsf{V}\} \geq \sqrt{\varepsilon'}\Big\} &\leq \sqrt{\varepsilon'} \\
\mathrm{P}\Big\{\mathrm{P}\{\mathcal{E}_2|\mathcal{C}_\mathsf{U}, \mathcal{C}_\mathsf{V}\} \geq \sqrt{\varepsilon'}\Big\} &\leq \sqrt{\varepsilon'} \ .
\end{aligned}$$

Choosing $\varepsilon' = \varepsilon^2$, by the union bound we have that our random coding scheme with probability at least $1 - 2\varepsilon$ yields a code $\mathcal{C}_u = (\boldsymbol{u}_i)_{i=[1:M_1]}$, $\mathcal{C}_v = (\boldsymbol{v}_j)_{j=[1:M_2]}$ and deterministic encoding functions $f\colon \mathcal{X}^n \to \mathcal{C}_u$, $g\colon \mathcal{Z}^n \to \mathcal{C}_v$, such that

$$\begin{aligned}
&\mathrm{P}\{\mathcal{E}_1|\mathcal{C}_\mathsf{V} = \mathcal{C}_v, \mathcal{C}_\mathsf{U} = \mathcal{C}_u\} \\
&\quad = \mathrm{P}\Big\{(f(\mathbf{X}), \mathbf{X}, \mathbf{Z}, g(\mathbf{Z})) \notin \mathcal{T}^n_{[\mathsf{UXZV}]\delta}\Big\} \leq \varepsilon \qquad (20)
\end{aligned}$$

and

$$\sum_{i,j=1}^{M_1,M_2} \mathbb{1}_{\mathcal{T}^n_{[\mathsf{UV}]\delta}}(\boldsymbol{u}_i, \boldsymbol{v}_j) \leq \exp\big(n(\mathrm{I}(\mathsf{UV};\mathsf{XZ}) + \varepsilon)\big) \ . \qquad (21)$$

Pick any such code and define $\mathcal{C}_i = f^{-1}(\{\boldsymbol{u}_i\}) \cap \mathcal{T}^n_{[\mathsf{X}|\mathsf{U}]\delta}(\boldsymbol{u}_i)$ if $\boldsymbol{u}_i \neq \boldsymbol{u}_{i'}$ for all $i' < i$ and $\mathcal{C}_i = \varnothing$ otherwise. $\mathcal{D}_j$ is defined accordingly. The conditions (8) and (9) now follow directly from (20) and (21). ∎

C. Proof of Proposition 9

As $\overline{\mathcal{R}}$ is convex by definition, we only need to show $\mathcal{R}' = \mathrm{conv}(\mathcal{R}^i)$. The cardinality bound $|\mathcal{Q}| \leq 3$ follows directly from the strengthened Carathéodory theorem [32, Theorem 18(ii)] as $\mathrm{conv}(\mathcal{R}^i)$ is the convex hull of a connected set in $\mathbb{R}^3$.

One can use the convex cover method [30, Appendix C] directly to show the weaker bounds $|\mathcal{U}| \leq |\mathcal{X}| + 1$ and $|\mathcal{V}| \leq |\mathcal{Z}| + 1$, also given by Han [22, Corrolary 6]. By only dealing with the extreme points of the achievable region on the upper concave envelope [33] we are able to improve the cardinality bounds on the auxiliaries. We will only show the cardinality bound $|\mathcal{U}| \leq |\mathcal{X}|$ as the bound for $|\mathcal{V}|$ follows analogously.

Define the continuous function $F(\mathrm{p}_{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}}})$ that maps a pmf onto the vector $(\mathrm{I}(\widetilde{\mathsf{X}};\widetilde{\mathsf{U}}), \mathrm{I}(\widetilde{\mathsf{Z}};\widetilde{\mathsf{V}}), \mathrm{I}(\widetilde{\mathsf{U}};\widetilde{\mathsf{V}}))$. Defining the compact, connected set of pmfs $\mathcal{P} \triangleq \{\mathrm{p}_{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}}} : \mathrm{p}_{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}}} = \mathrm{p}_{\widetilde{\mathsf{U}}|\widetilde{\mathsf{X}}}\mathrm{p}_{\mathsf{X},\mathsf{Z}}\mathrm{p}_{\widetilde{\mathsf{V}}|\widetilde{\mathsf{Z}}}, \widetilde{\mathcal{U}} = \{0, \ldots, |\mathcal{X}|\}, \widetilde{\mathcal{V}} = \{0, \ldots, |\mathcal{Z}|\}\}$, we obtain the compact, connected region $\mathcal{S} \triangleq F(\mathcal{P}) \subseteq \mathbb{R}^3$. As $\mathcal{S}$ is compact, its convex hull $\mathrm{conv}(\mathcal{S})$ is compact and can be represented as an intersection of halfspaces in the following manner. For $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3$ define $V(\boldsymbol{\lambda}) \triangleq \max_{\boldsymbol{x} \in \mathcal{S}} \boldsymbol{\lambda} \cdot \boldsymbol{x}$. Then $\mathrm{conv}(\mathcal{S}) = \bigcap_{\boldsymbol{\lambda} \in \mathbb{R}^3} \{\boldsymbol{x} \in \mathbb{R}^3 : \boldsymbol{\lambda} \cdot \boldsymbol{x} \leq V(\boldsymbol{\lambda})\}$. Taking the cardinality bound on U into account, defining $\mathcal{P}' \triangleq \{\mathrm{p}_{\widetilde{\mathsf{U}},\widetilde{\mathsf{X}},\widetilde{\mathsf{Z}},\widetilde{\mathsf{V}}} \in \mathcal{P} : |\widetilde{\mathcal{U}}| = |\mathcal{X}|\}$ and $\mathcal{S}' \triangleq F(\mathcal{P}')$, with the same reasoning we obtain $\mathrm{conv}(\mathcal{S}') = \bigcap_{\boldsymbol{\lambda} \in \mathbb{R}^3} \{\boldsymbol{x} \in \mathbb{R}^3 : \boldsymbol{\lambda} \cdot \boldsymbol{x} \leq V'(\boldsymbol{\lambda})\}$ where $V'(\boldsymbol{\lambda}) \triangleq \max_{\boldsymbol{x} \in \mathcal{S}'} \boldsymbol{\lambda} \cdot \boldsymbol{x}$. We next show $V'(\boldsymbol{\lambda}) \geq V(\boldsymbol{\lambda})$ which proves $\mathrm{conv}(\mathcal{S}') = \mathrm{conv}(\mathcal{S})$. Let $\mathcal{X}' \triangleq \mathcal{X} \setminus \{x\}$ where $x \in \mathcal{X}$ is arbitrary. Define the test function $t_x(\mathrm{p}_{\widetilde{\mathsf{X}}}) \triangleq \mathrm{p}_{\widetilde{\mathsf{X}}}(x)$ for $x \in \mathcal{X}$ and abbreviate $\boldsymbol{t} = (t_x)_{x \in \mathcal{X}'}$. Choose any $\boldsymbol{\lambda} \in \mathbb{R}^3$ and fix $(\mathsf{U}, \mathsf{X}, \mathsf{Z}, \mathsf{V})$ that achieve $\boldsymbol{\lambda} \cdot F(\mathrm{p}_{\mathsf{U},\mathsf{X},\mathsf{Z},\mathsf{V}}) = V(\boldsymbol{\lambda})$. Define the function

$$\begin{aligned}
f(\mathrm{p}_{\widetilde{\mathsf{X}}}) \triangleq\ & \lambda_1(\mathrm{H}(\mathsf{X}) - \mathrm{H}(\widetilde{\mathsf{X}})) \\
&+ \lambda_2 \mathrm{I}(\mathsf{Z};\mathsf{V}) + \lambda_3(\mathrm{H}(\mathsf{V}) - \mathrm{H}(\widetilde{\mathsf{V}}))
\end{aligned}$$

where $(\widetilde{\mathsf{V}}, \widetilde{\mathsf{Z}}, \widetilde{\mathsf{X}}) \sim \mathrm{p}_{\mathsf{V}|\mathsf{Z}}\mathrm{p}_{\mathsf{Z}|\mathsf{X}}\mathrm{p}_{\widetilde{\mathsf{X}}}$. Obviously $((\mathrm{p}_\mathsf{X}(x))_{x \in \mathcal{X}'}, V(\boldsymbol{\lambda}))$ lies in the convex hull of the compact, connected set $(\boldsymbol{t}, f)(\mathcal{P}(\mathcal{X}))$. Therefore, by the strengthened Carathéodory theorem [32, Theorem 18(ii)], $|\mathcal{X}|$ points suffice, i.e., there exists a random variable U' with $|\mathcal{U}'| = |\mathcal{X}|$ and thus $\mathrm{p}_{\mathsf{U}',\mathsf{X},\mathsf{Z},\mathsf{V}} \in \mathcal{P}'$, such that $\mathbb{E}\big[f(\mathrm{p}_{\mathsf{X}|\mathsf{U}'}(\,\cdot\,|\mathsf{U}'))\big] = \boldsymbol{\lambda} \cdot F(\mathrm{p}_{\mathsf{U}',\mathsf{X},\mathsf{Z},\mathsf{V}}) = V(\boldsymbol{\lambda})$. This shows $V'(\boldsymbol{\lambda}) \geq V(\boldsymbol{\lambda})$. To obtain the full region observe that $\mathrm{conv}(\mathcal{R}^i) = \mathrm{conv}(\mathcal{S}) + (\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_-) = \mathrm{conv}(\mathcal{S}') + (\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_-)$.

By applying the same reasoning to V, one can show that $|\mathcal{V}| = |\mathcal{Z}|$ is sufficient. ∎

References

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Claude Elwood Shannon: collected papers*, N. J. A. Sloane and A. D. Wyner, Eds. IEEE Press, 1993, pp. 325–350.

[2] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

[3] Y. Seldin, N. Slonim, and N. Tishby, "Information bottleneck for non co-occurrence data," in *Proc. of the 20th Annu. Conf. on Neural Inform. Process. Syst.*, 2006, pp. 1241–1248.

[4] R. Bekkerman, R. El-Yaniv, N. Tishby, Y. Winter, I. Guyon, and A. Elisseeff, "Distributional word clusters vs. words for text categorization," *Machine Learning Research*, vol. 3, pp. 1183–1208, 2003.

[5] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[6] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 595–609.

[7] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *Information Theory, IRE Transactions on*, vol. 8, no. 5, pp. 293–304, September 1962.

[8] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan 2014.

[9] B. Mirkin, *Mathematical Classification and Clustering*. Kluwer Academic Publisher, 1996.

[10] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, Mar 1972.

[11] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[12] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proc. 8th Int. Conf. Intelligent Syst. for Molecular Biology*, vol. 8, 2000, pp. 93–103.

[13] A. Tanay, R. Sharan, and R. Shamir, "Biclustering algorithms: A survey," *Handbook of computational molecular biology*, vol. 9, no. 1-20, pp. 122–124, 2005.

[14] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proceedings of the* 9^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[15] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sep 1975.

[16] A. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications: Part I," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 769–772, Nov 1973.

[17] H. Witsenhausen, "Entropy inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 610–616, Sep 1974.

[18] A. Wyner, "A theorem on the entropy of certain binary sequences and applications: Part II," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 772–777, Nov 1973.

[19] R. Ahlswede and J. Körner, "On the connection between the entropies of input and output distributions of discrete memoryless channels," in *Proc. 5th Conf. Probability Theory*, 1977, pp. 13–23.

[20] ——, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, 1975.

[21] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communications," *Probability Theory and Related Fields*, vol. 34, no. 2, pp. 157–177, 1976.

[22] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[23] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, 1998.

[24] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[26] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, Mar 2001.

[27] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[28] S.-Y. Tung, "Multiterminal source coding," Ph.D. dissertation, Cornell University, May 1978.

[29] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, pp. 149–162, 1973.

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

[31] T. S. Han and K. Kobayashi, "A unified achievable rate region for a general class of multiterminal source coding systems," *IEEE Trans. Inf. Theory*, vol. 26, no. 3, pp. 277–288, May 1980.

[32] H. G. Eggleston, *Convexity*, P. Hall and F. Smithies, Eds. Cambridge University Press, 1958.

[33] C. Nair, "Upper concave envelopes and auxiliary random variables," *Int. J. of Advances in Eng. Sciences and Appl. Math.*, vol. 5, no. 1, pp. 12–20, 2013.