Distributed Information-Theoretic Biclustering

Georg Pichler, Student Member, IEEE, Pablo Piantanida, Member, IEEE, and Gerald Matz, Senior Member, IEEE

Abstract

We study a novel multi-terminal source coding setup motivated by the biclustering problem. Two separate encoders observe two i.i.d. sources $X^n$ and $Z^n$, respectively. The goal is to find rate-limited encodings $f(x^n)$ and $g(z^n)$ that maximize the mutual information $I(f(X^n); g(Z^n))/n$. There are strong ties to a number of information theoretic problems, including hypothesis testing against independence, pattern recognition, the information bottleneck method and lossy source coding with logarithmic-loss distortion. Improving previous cardinality bounds allows us to thoroughly study the example of a binary symmetric source and quantifying the gap between the inner and the outer bound in this special case. Furthermore, we generalize our results to the case of more than two i.i.d. sources. As a special case of this generalization we investigate a Multiple Description extension of the CEO problem with log-loss distortion. Surprisingly this MD-CEO problem permits a tight single-letter characterization of the achievable region, which has the remarkable feature that it allows exploiting rates that are in general insufficient to guarantee successful typicality decoding.

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 social media bandwagon. A significant part of this data is unstructured and it is thus hard to extract the relevant information. This results in a growing need for a fundamental understanding and efficient methods for analyzing data and discovering valuable and relevant knowledge from it in the form of structured information.

When specifying certain hidden (unobserved) features of interest, the problem then consists of extracting those relevant features from a measurement, while neglecting other, irrelevant features. Formulating this idea in terms of lossy source compression [1], we can quantify the complexity of the encoded data via its rate and the quality via the information provided about specific (unobserved) features.

In this paper, we introduce and study the distributed biclustering problem from a formal information-theoretic perspective. Given correlated samples $X^n_1, X^n_2, \ldots, X^n_K$ observed at different encoders, the aim is to extract a description from each sample, such that the descriptions are maximally informative about each other. In other words, the $k$-th encoder tries to find a (lossy) description $U_k$ of its observation $X^n_k$ subject to complexity requirements (coding rate), such that the mutual information between two disjoint subsets of descriptions $(U_k)_{k \in A}$ and $(U_k)_{k \in B}$ is maximized. The goal is to characterize the optimal tradeoff between relevance (mutual information between the descriptions) and complexity (encoding rate).

A. Distributed Biclustering

As a clustering technique, biclustering (or co-clustering) was first explicitly considered by Hartigan [2] in 1972. A historical overview of biclustering including additional background can be found in [3, Section 3.2.4]. In general, given an $N \times M$ data matrix $(a_{nm})$, the goal of a biclustering algorithm [4] is to find partitions $B_k \subseteq \{1, \ldots, N\}$ and $C_l \subseteq \{1, \ldots, M\}$, $k = 1 \ldots K$, $l = 1 \ldots L$ such that all the “biclusters” $(a_{nm})_{n \in B_k, m \in C_l}$ are in a certain sense homogeneous. The measure of homogeneity of the biclusters depends on the specific application. The method received renewed attention when Cheng and Church [5] applied it to gene expression data. Many biclustering

Different parts of the material in this paper were presented at the 53rd Annual Allerton Conference on Communications, Control, and Computing (2015), the 2016 IEEE International Symposium on Information Theory, and submitted to the 2017 IEEE International Symposium on Information Theory.

G. Pichler and G. Matz are with the Institute of Telecommunications, TU Wien, Vienna, Austria. P. Piantanida is with Laboratoire des 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).
algorithms have been developed since (e.g., see [6] and references therein). An introductory overview of clustering algorithms for gene expression data can be found in the lecture notes [7]. The Information Bottleneck (IB) method, which can be viewed as a uni-directional information-theoretic variant of biclustering, was successfully applied to gene expression data as well [8].

In 2003, Dhillon et al. [9] adopted an information-theoretic approach to biclustering. They used mutual information to characterize the quality of a biclustering. Specifically, for the special case when the underlying matrix represents the joint probability distribution of two discrete random variables $X$ and $Y$, i.e., $a_{nm} = P\{X = n, Y = m\}$, their goal was to find functions $f : \{1, \ldots, N\} \to \{1, \ldots, K\}$ and $g : \{1, \ldots, M\} \to \{1, \ldots, L\}$ that maximize $I(f(X); g(Y))$ for specific $K$ and $L$. In information-theoretic terms this is a single-letter problem. But by using a stationary, memoryless process and defining achievability in the usual Shannon sense, the approach of Dhillon et al. can be addressed via information-theoretic tools.

The aim of the present paper is to characterize the achievable region of this information-theoretic biclustering problem and connect it to known problems in network information theory. Furthermore we provide an extension to more than two stationary, memoryless sources. This distributed information-theoretic biclustering problem offers a formidable mathematical complexity. It is fundamentally different from “classical” distributed source coding problems like distributed lossy compression [10] Chapter 12. Usually, one aims at reducing redundant information, i.e., information that is transmitted by multiple encoders, as much as possible, while still guaranteeing correct decoding. But in the biclustering problem, we are interested in maximizing precisely this redundant information. In this sense, the biclustering problem is complementary to distributed source coding. We also point out that the extension to multiple variables contains the Körner-Marton problem [11], which implies that in general Berger-Tung coding is suboptimal.

B. Contributions

We first study the case of two sources in Section II. This problem turns out to be closely related to a number of information-theoretic problems, including hypothesis testing against independence [12], pattern recognition [13], the information bottleneck method [14] and lossy source coding with logarithmic-loss distortion [15]. We study these equivalences in detail in Section II-B and exploit them to provide an inner bound on the achievable region. The outer bound follows from standard information-theoretic manipulations. In Section II-E we extensively study the doubly symmetric binary source as an example. In order to perform this analysis, we require stronger cardinality bounds, than the ones usually obtained using the convex cover method [10] Appendix C. To achieve this, we combine the convex cover method, the perturbation method and leverage ideas similar to [16], allowing us to deal only with the extreme points of the achievable region. The resulting bounds, proved in Appendix A2, allow for using binary auxiliaries in the case of binary sources. Based on a conjecture (which previously appeared in [13]), we then argue that there is indeed gap between the outer and the inner bound for a doubly symmetric binary source. This disproves another conjecture from [13]. In Section III we extend the inner and the outer bound to the case of multiple sources, which additionally requires a binning strategy for the achievability part. In Section III-B we investigate the Chief Executive Officer (CEO) problem under a mutual information constraint, a special case of the information-theoretic biclustering problem with multiple sources. We show that it is equivalent to classical multiterminal lossy source coding under logarithmic loss distortion. By leveraging this equivalence, we obtain tight bounds for a special case using results from [15]. In III-D we extend the CEO problem to a Multiple Description (MD) coding problem. Using tools from submodularity theory and convex analysis, we are able to provide a complete single-letter characterization of the resulting achievable region, which has the remarkable feature that it allows to exploit rate which is in general insufficient to guarantee successful typicality decoding.

C. Notation and Conventions

1) Non-Standard Conventions: We will use the shorthand $[l:k] := \{l, l+1, \ldots, k-1, k\}$. For convenience, we define the sets $K := [1:K]$, $J := [1:J]$, and $L := [1:L]$ for $K,J,L \in \mathbb{N}$. Furthermore let $\Omega$ denote the set of all pairs $(A,B)$, where $A,B \subseteq K$ are nonempty and disjoint. Note that $\Omega$ contains $3^K - 2^{K+1} + 1$ different pairs $(A,B)$. We also define $\Omega$ as the set of all pairs $(A,B)$, where $A \subseteq J$ and $B \subseteq L$ are nonempty. Hence, the set $\Omega$ contains $2^{J+L} - 2^J - 2^L + 1$ elements. For a linear ordering $\sqsubset$ of a set $E$ and $e \in E$ we will use the notation $\sqsubset e := \{e' \in E : e' \sqsubset e\}$ and accordingly for $\sqsupset$, $\sqsubset$ and $\square$. E.g., given the linear ordering of $\{1,2,3\}$ with $3 \sqsubset 1 \sqsubset 2$, we have $\sqsubset 3 = \{1,2\}$, $\sqsubset 1 = \{2\}$ and $\sqsubset 2 = \emptyset$. 

Enc. 1 (rate $R_1$) \[ f(X) \]

Enc. 2 (rate $R_2$) \[ g(Z) \]

$X \xrightarrow{\text{Enc. 1}} f(X)$

$Z \xrightarrow{\text{Enc. 2}} g(Z)$

$n \mu \leq I( f(X); g(Z) )$

Figure 1: Biclustering of two memoryless sources.

2) Standard Conventions: We use $\overline{A}$ to denote the topological closure of a set $A$ and $A^c$ for the complement of a set (or event). The symbol $|A|$ is used for the number of elements in a finite set $A$. We use $1_A$ for the indicator function of a set (or event) $A$. When there is no possibility of confusion we identify a set with one element with that element, e.g., $\{1, 2, 3\} \setminus \{1\} = \{2, 3\}$. Let $\mathbb{R}_+$ be the set of non-negative reals and $\mathbb{R}_-$ the set of non-positive reals. We denote random quantities and their realizations by capital, sans-serif and lowercase letters, respectively. Furthermore, vectors are indicated by bold-face type and have length $n$, if not otherwise specified. Random variables are assumed to be supported on finite sets. We use the same letter for the random variable and for its support set, e.g., $Y$ takes values in $\mathcal{Y}$ and $X_3$ takes values in $\mathcal{X}_3$. Given a random variable $X$, we write $p_X \in \mathcal{P}(\mathcal{X})$ for its probability mass function (pmf), where $\mathcal{P}(\mathcal{X})$ is the set of all pmfs on $\mathcal{X}$. We use $X \sim p_X$ or $X \simeq X$ to indicate that $X$ and $X$ have the same distribution. We use $E[X]$ and $P\{A\}$ for the expectation of the random variable $X$ and the probability of an event $A$, respectively. Subscripts indicate parts of vectors, e.g., $x_A := (x_i)_{i \in A}$. We further use the common notation $x_i := x_{\{i\}}$, $x^j := : x_{\{i,...,j\}}$. If a vector is already carrying a subscript, it will be separated by a comma, e.g., $x_{3,1}^1 = (x_3^0)_{1} = (x_3^0)_{1}$. Additionally we use subscript sets to denote tuples of real values and their components, e.g., $x_A \in \mathbb{R}^K$ or $x_{\Omega} \in \mathbb{R}^{3^K-2^{K+1}+1}$. Naturally, slices of tuples are indexed by subsets, e.g., $x_A \subseteq K$ is a slice of $x_K$. This notation extends naturally to tuples of vectors, where the subscript indices are separated by a comma, e.g., for $x_K \in \mathbb{R}^{nK}$, we have $x_{A,l} \subseteq \mathbb{R}^{(k-l+1)\vert A\vert}$. We use the notation of [17, Chapter 2] for information-theoretic quantities. All logarithms in this paper are to base $e$ and therefore all information theoretic quantities are measured in nats. The notation $h_0(p) := -p \log p - (1 - p) \log(1 - p)$ is used for the binary entropy function and $a \star b := a(1 - b) + (1 - a)b$ is the binary convolution operation. The symbol $\oplus$ stands for binary addition and $X \sim B(p)$ is used to denote a Bernoulli distribution with parameter $p$. The notation $X \oplus Y \oplus Z$ indicates that $X$, $Y$, and $Z$ form a Markov chain in this order and $X \perp Y$ denotes that $X$ and $Y$ are independent random variables. When generating codebooks we will assume that the codebook size is an integer to keep the notation simple. We will use superscript to indicate that a relation follows from a specific equation, e.g., the inequality $a \leq b$ follows from equation (1). Our achievability proofs are based on the notion of robust typicality [18], also used in [12]. We will adopt the usual notation for types and typical sequences and make use of the $\delta$-convention [19, Convention 2.11]. For convenience, the necessary notation and relevant results are summarized in Appendix C.

II. BICLUSTERING WITH TWO SOURCES

A. Problem Statement

In this section we will introduce the information-theoretic biclustering problem (or biclustering problem for short) with two sources and provide bounds on its achievable region. A schematic overview of the problem is presented in Figure 1. Let $(X, Z)$ be two random variables. The random vectors $(X, Z)$ consist of $n$ i.i.d. copies of $(X, Z)$. Given a block length $n \in \mathbb{N}$ and coding rates $R_1, R_2 \in \mathbb{R}_+$, an $(n, R_1, R_2)$ code $(f, g)$ consists of two functions $f: \mathcal{X}^n \rightarrow \mathcal{M}_1$ and $g: \mathcal{Z}^n \rightarrow \mathcal{M}_2$ such that the finite sets $\mathcal{M}_k$ satisfy $\log \vert \mathcal{M}_k \vert \leq nR_k$, $k \in \{1, 2\}$. Thus, the coding rates $R_k$, $k \in \{1, 2\}$, limit the complexity of the encoders.

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

$$\Theta(f; g) := \frac{1}{n} I( f(X); g(Z) ).$$  (1)
This co-information serves as a measure of the mutual relevance of the two encodings \( f(X) \) and \( g(Z) \). In contrast to rate-distortion theory, we do not require a specific distortion measure; rather, we quantify the quality of a code in pure information-theoretic terms, namely via mutual information. The idea is to find functions \( f \) and \( g \) that extract a compressed version of the common randomness in the observed data \( X \) and \( Z \).

**Definition 2.** A triple \((\mu, R_1, R_2) \in \mathbb{R}^3\) is achievable if for some \( n \in \mathbb{N} \) there exists an \((n, R_1, R_2)\) code \((f, g)\) such that

\[
\Theta(f; g) \geq \mu.
\]  

(2)

The achievable region \( \overline{\mathcal{R}} \) is defined as the closure of the set \( \mathcal{R} \) of achievable triples.

We note that stochastic encodings cannot enlarge the achievable region. Assume that the triple \((\mu, R_1, R_2)\) can be achieved using a stochastic encoding. The fact that any stochastic encoding can be represented as a convex combination of deterministic encodings implies that at least one of these deterministic encodings also achieves \((\mu, R_1, R_2)\).

**Remark 1.** Note that \( \overline{\mathcal{R}} \) is a convex set. This can be shown by a standard time-sharing argument.

**B. Connection with other Problems**

The biclustering problem with two sources turns out to be equivalent to an hypothesis testing problem and a pattern recognition problem. In this section we will clarify these equivalences explicitly, using the “multi-letter” region \( \mathcal{R}_s \).

**Definition 3.** Let \( \mathcal{R}_s \) be the set of triples \((\mu, R_1, R_2)\) such that there exist \( n \in \mathbb{N} \) and random variables \( U, V \) satisfying \( U \leftrightarrow X \leftrightarrow Z \leftrightarrow V \) and

\[
nR_1 \geq I(X; U) \quad (3)
\]

\[
nR_2 \geq I(Z; V) \quad (4)
\]

\[
n\mu \leq I(U; V). \quad (5)
\]

We will now showcase the connection between the biclustering problem and several other information-theoretic problems. Leveraging these equivalences will provide us with the achievability of \( \mathcal{R}_s \).

1) **Hypothesis Testing:** The biclustering problem can be shown to be equivalent to the hypothesis testing problem with data compression [12] when testing against independence. For completeness sake we briefly describe the problem setup here.

Given a source \((X, Z) \sim p_{XZ}\) define the independent random variables \(X \sim p_X\) and \(Z \sim p_Z\). An \((n, R_1, R_2)\) hypothesis test consists of an \((n, R_1, R_2)\) code \((f_n, g_n)\) and a set \(A_n \subseteq M_1 \times M_2\). The type I and type II errors of \((f_n, g_n, A_n)\) are defined as

\[
\alpha_n := P\{f_n(X), g_n(Z)\} \in A_n\}
\]

(6)

\[
\beta_n := P\{f_n(X), g_n(Z)\} \notin A_n\}
\]

(7)

A triple \((\mu, R_1, R_2)\) is HT-achievable if, for every \( \varepsilon > 0 \), there is a sequence of \((n, R_1, R_2)\) hypothesis tests \((f_n, g_n, A_n)\), \(n \in \mathbb{N}\) such that

\[
\lim_{n \to \infty} \alpha_n \leq \varepsilon \quad (8)
\]

\[
\lim_{n \to \infty} -\frac{1}{n} \log \beta_n \geq \mu. \quad (9)
\]

Let \(\mathcal{R}_{HT}\) denote the set of all HT-achievable triples.

**Theorem 4.** \(\mathcal{R}_{HT} = \overline{\mathcal{R}}_s\).

**Proof.** Assume \((\mu, R_1, R_2) \in \mathcal{R}_{HT}\). For \( \varepsilon > 0 \), pick an \((n, R_1, R_2)\) hypothesis tests \((f_n, g_n, A_n)\) such that \(\alpha_n \leq \varepsilon\) and \(\log \beta_n \leq -n(\mu - \varepsilon)\). By applying the Log-Sum inequality one can show with \(U := f_n(X)\) and \(V := g_n(Z)\) that

\[
I(U; V) \geq (1 - \alpha_n) \log \frac{1 - \alpha_n}{\beta_n} + \alpha_n \log \frac{\alpha_n}{1 - \beta_n}. \quad (10)
\]
As (3) and (4) are also satisfied, this shows that
\[ \left( \mu - \varepsilon \mu - \varepsilon - \frac{\log(2)}{n}, R_1, R_2 \right) \in \mathcal{R}_* \] and as \( \varepsilon \) and \( n^{-1} \) can be arbitrarily small, \( (\mu, R_1, R_2) \in \mathcal{R}_* \).

Consider the bound [12, Corollary 6]. It shows that \( (n\mu, nR_1, nR_2) \) is asymptotically achievable in the hypothesis testing problem for the vector source \((X, Z)\) if \((\mu, R_1, R_2) \in \mathcal{R}_*\), i.e., for any \( \varepsilon, \varepsilon' > 0 \), there is a sequence of \((k, nR_1 + \varepsilon', nR_2 + \varepsilon')\) hypothesis tests \((f_k, g_k, A_k)\) for \((X, Z)\), \(k \in \mathbb{N}\) such that
\[
\lim_{k \to \infty} \frac{1}{k} \log \beta_k \geq n\mu - \varepsilon'.
\]
This shows that \( (\mu - \varepsilon' n, R_1 + \varepsilon', R_2 + \varepsilon' n) \in \overline{\mathcal{R}}_{\text{HT}} \) is achievable for the source \((X, Z)\) and as \( \varepsilon' \) was arbitrary, this completes the proof.

We can leverage this equivalence with hypothesis testing to show that \( \overline{\mathcal{R}} \) is indeed a multi-letter characterization of \( \mathcal{R}_* \).

**Corollary 5.** \( \overline{\mathcal{R}} = \mathcal{R}_* \).

**Proof.** To prove \( \mathcal{R} \subseteq \mathcal{R}_* \), assume \((\mu, R_1, R_2) \in \mathcal{R}\) and choose \( n, f, \) and \( g \) appropriately as in Definition 2. Defining \( U := f(X) \) and \( V := g(Z) \) yields inequalities (3) and (5) and satisfies the Markov chain.

We will show \( \mathcal{R}_{\text{HT}} \subseteq \mathcal{R} \), which is equivalent to \( \mathcal{R}_* \subseteq \mathcal{R} \) by Theorem 4. Assume \((\mu, R_1, R_2) \in \mathcal{R}_{\text{HT}}\) and for \( \varepsilon > 0 \) choose a sequence of \((n, R_1, R_2)\) hypothesis tests \((f_n, g_n, A_n)\), \(n \in \mathbb{N}\) such that (6) and (9) are satisfied. Pick \( n_0 \in \mathbb{N} \) such that \( \alpha_{n_0} \leq 2\varepsilon \) and \( -\log \beta_{n_0} \geq n(\mu - \varepsilon) \). By the same reasoning as in (13), the \((n_0, R_1, R_2)\), the code \((f_{n_0}, g_{n_0})\) achieves \((\mu - \varepsilon\mu - \varepsilon - \frac{\log(2)}{n_0}, R_1, R_2)\), implying \((\mu, R_1, R_2) \in \mathcal{R}\).

2) **Pattern Recognition:** Consider the pattern recognition problem introduced in [13]. For completeness sake we restate the problem here.

A triple \((\mu, R_1, R_2)\) is said to be PR-achievable if for any \( \varepsilon > 0 \) there exists an \((n, R_1, R_2)\) code \((f, g)\) in the sense defined in Section II-A such that the following conditions are met. Let \((X(i), Z(i))\) be \( n \) i.i.d. copies of \((X, Z)\), independently generated for each \( i \in [1 : e^{n\mu}] \). Pick an index \( W \sim U([1 : e^{n\mu}]) \) uniformly at random, independent of \((X(i), Z(i))\), \(i \in [1 : e^{n\mu}]\). There needs to exist a function \( \phi : [1 : e^{nR_1}] \times [1 : e^{nR_2}] \to [1 : e^{n\mu}]\) such that,
\[
P\{ W = \phi(C_u, g(Z(W))) \} \geq 1 - \varepsilon,
\]
where \( C_u := (f(X(i)))_{i \in [1 : e^{n\mu}]} \). Let \( \mathcal{R}_{\text{PR}} \) denote the set of all PR-achievable points.

**Remark 2.** We want to point out that the variant of the inner bound for the pattern recognition problem stated in [13, Theorem 1] is flawed. To see this, note that (using the notation of [13]) the point \((R_x = 0, R_y = b, R_c = b)\) is contained in \( R_{\text{in}} \) (choose \( U = V = \emptyset \)) for any \( b > 0 \) even if the random variables \( X \) and \( Y \) are independent. As a consequence, this point is clearly not achievable in general. However, the region \( R_{\text{in}}' \) defined in the right column of [13, p. 303], coincides with our findings and the proof given in [13, Appendix A] holds for this region.

**Proposition 6.** \( \mathcal{R}_{\text{PR}} = \overline{\mathcal{R}}_* \).

**Proof.** Assume \((\mu, R_1, R_2) \in \mathcal{R}_{\text{PR}}\) and for an arbitrary \( \varepsilon > 0 \) and suitably large \( n \in \mathbb{N} \) choose appropriate functions \( f, g, \phi \). By defining \( U := f(X) \) and \( V := g(Z) \), we have (3) and (4). Furthermore,
\[
I(U; V) = I(f(X); g(Z)) = I(C_u; g(Z(W))) | W) = I(C_u; g(Z(W))), W) \geq I(C_u; W | g(Z(W)))
\]
\[
= \mathbb{H}(W | g(Z(W))) - \mathbb{H}(W | C_u, g(Z(W))) \quad (21) \\
\geq n\mu - \mathbb{H}(W | \phi(C_u, g(Z(W)))) \quad (22) \\
\geq n\mu - \log(2) - \varepsilon n\mu. \quad (23)
\]

The equality in (18) holds as \(X(i) \perp Z(i)\) for \(i \neq j\), (19) follows from \(W \perp C_u\), (22) follows from \(W \perp Z(W)\), the fact that \(\mathbb{H}(W) = n\mu\) and the data processing inequality. Fano’s inequality was used in (23). This shows \(\mathcal{R}_{PR} \subseteq \overline{\mathcal{R}_*}\) as \(\varepsilon\) and \(n^{-1}\) can be arbitrarily small.

To show the other direction, we apply the achievability result \cite[Theorem 1]{13} to the multi-letter source \((X, Z)\). Assuming \((\mu, R_1, R_2) \in \mathcal{R}_*\), we know that for some \(n \in \mathbb{N}\) there are random variables \((U, V)\) satisfying the Markov chain \(U \rightarrow X \rightarrow Z \rightarrow V\) and (3) \cite{5}. By \cite[Theorem 1]{13}, the triple \((n\mu, nR_1, nR_2)\) is asymptotically achievable for the pattern recognition problem for the source \((X, Z)\) with an arbitrary error probability \(\varepsilon > 0\). Select \(k \in \mathbb{N}\) such that the functions \(f, g, \phi\) asymptotically achieve this point, i.e., for \(\varepsilon' > 0\), \(\log|f| \leq k(nR_1 + \varepsilon')\), \(\log|g| \leq k(nR_2 + \varepsilon')\), \(\log|\phi| \geq k(n\mu - \varepsilon')\) and (16) are satisfied. Thus, the triple \((\mu - \varepsilon'_n, R_1 + \varepsilon'_n, R_2 + \varepsilon'_n)\) is achievable for the source \((X, Z)\) and as \(\varepsilon'\) can be arbitrarily small, this completes the proof.

\[\square\]

\section{Bounds on the Achievable Region}

We first provide outer bounds on the set of achievable triples for biclustering with two sources.

\textbf{Theorem 7.} We have \(\mathcal{R} \subseteq \mathcal{R}_0 \subseteq \mathcal{R}'_0\), where the two regions \(\mathcal{R}_0\) and \(\mathcal{R}'_0\) are given by

\[
\mathcal{R}_0 := \{ (\mu, R_1, R_2) : R_1 \geq I(U; X), R_2 \geq I(V; Z), \mu \leq I(V; Z) + I(U; X) - I(UV; ZX) \},
\]

\[
\mathcal{R}'_0 := \{ (\mu, R_1, R_2) : R_1 \geq I(U; X), R_2 \geq I(V; Z), \mu \leq \min(I(U; Z), I(V; X)) \},
\]

with \(U\) and \(V\) any pair of random variables satisfying \(U \rightarrow X \rightarrow Z\) and \(X \rightarrow Z \rightarrow V\).

Although Theorem 7 follows from the outer bound for the pattern recognition problem \cite[Appendix B]{13} via the equivalence shown in II-B2, we provide a short proof in Appendix A1 for the sake of completeness.

The regions \(\mathcal{R}_0\) and \(\mathcal{R}'_0\) are both convex since a time-sharing variable can be incorporated into \(U\) and \(V\). Furthermore, \(\mathcal{R}'_0\) remains unchanged when \(U\) and \(V\) are required to satisfy the complete Markov chain \(U \rightarrow X \rightarrow Z \rightarrow V\).

The numerical computation of the outer bounds requires the cardinalities of the auxiliary random variables to be bounded. We therefore complement Theorem 7 with the following result, whose proof is provided in Appendix A2.

\textbf{Proposition 8.} We have \(\mathcal{R}_0 = \text{conv}(S_0)\) and \(\mathcal{R}'_0 = \text{conv}(S'_0)\), where the regions \(S_0\) and \(S'_0\) are defined as \(\mathcal{R}_0\) and \(\mathcal{R}'_0\), respectively, but with the additional cardinality bounds \(|U| \leq |X|\) and \(|V| \leq |Z|\).

The cardinality bounds in this result are tighter than the usual bounds obtained with the convex cover method \cite[Appendix C]{10}, where the cardinality has to be increased by one. Thus, when dealing with the binary case in Section II-B, binary auxiliaries will be sufficient. The smaller cardinalities come at the price of convexification in Proposition 8 since the regions \(S_0\) and \(S'_0\) are not necessarily convex.

The following inner bound follows from the inner bounds of the two equivalent problems, detailed in Section II-B.

\textbf{A more general inner bound will be proved in Theorem 20 in Section III-A for an extension of the biclustering problem to more than two source.}

\textbf{Theorem 9.} We have \(\mathcal{R}_1 \subseteq \overline{\mathcal{R}}\) where

\[
\mathcal{R}_1 := \{ (\mu, R_1, R_2) : R_1 \geq I(U; X), R_2 \geq I(V; Z), \mu \leq I(U; V) \},
\]

with auxiliary random variables \(U, V\) satisfying \(U \rightarrow X \rightarrow Z \rightarrow V\).

Theorem 9 follows from the achievability result of Han on the hypothesis testing problem \cite[Corollary 6]{12}, leveraging the equivalence detailed in Theorem 4 and Corollary 5. Alternatively, it also follows from the inner bound for the pattern recognition problem \cite[Appendix A]{13} using Corollary 5 and Proposition 6. Interestingly, the outer bound \(\mathcal{R}_0\) and the inner bound \(\mathcal{R}_1\) would coincide if the Markov condition \(U \rightarrow X \rightarrow Z \rightarrow V\) were imposed in the definition of \(\mathcal{R}_0\) since then \(I(V; Z) + I(U; X) - I(UV; ZX) = I(U; V) - I(U; V|XZ) = I(U; V)\).

Employing a binning scheme does not increase the achievable region. The intuition is that binning reduces redundant information transmitted by both encoders, whereas in information-theoretic biclustering this quantity
should actually be maximized. A tight bound on the achievable region can be obtained if \( \mu \) is not greater than the common information (cf. \([20]−[22]\)) of \( X \) and \( Z \), as stated in the following corollary.

**Corollary 10.** If \( Y = \zeta_1(X) = \zeta_2(Z) \) is common to \( X \) and \( Z \) in the sense of \([21]\) and \( 0 \leq \mu \leq H(Y) \) then 
\((\mu, R_1, R_2) \in \overline{\mathcal{R}} \) if and only if \( \mu \leq \min(R_1, R_2) \).

**Proof.** Theorem\( [7] \) entails \( \mu \leq \min(R_1, R_2) \) for any \((\mu, R_1, R_2) \in \overline{\mathcal{R}} \). With \( U = V = Y \), Theorem\( [9] \) implies \((H(Y), H(Y), H(Y)) \in \overline{\mathcal{R}} \). Using time-sharing with \((0, 0, 0) \in \overline{\mathcal{R}} \) we obtain \((\mu, \mu, \mu) \in \overline{\mathcal{R}} \) for \( \mu \leq H(Y) \) and hence 
\((\mu, R_1, R_2) \in \overline{\mathcal{R}} \) if \( \mu \leq \min(R_1, R_2) \).

We next improve the inner bound \( \mathcal{R}_i \) via convexification. Furthermore, we incorporate the same cardinality bounds as for the outer bound in Proposition\( [8] \) thereby enabling us to use binary auxiliaries when dealing with binary sources in Section II-E.

**Proposition 11.** We have \( S_i' := \text{conv}(S_i) = \text{conv}(\mathcal{R}_i) \subseteq \overline{\mathcal{R}} \) where \( S_i \) is defined as \( \mathcal{R}_i \), but with the additional cardinality bounds \(|U| \leq |X|\), and \(|V| \leq |Z|\). Furthermore, \( S_i' \) can be explicitly expressed as
\[
S_i' = \{ (\mu, R_1, R_2) : R_1 \geq I(U; X|Q), R_2 \geq I(V; Z|Q), \mu \leq I(U; V|Q) \},
\]
where \( U, V, \) and \( Q \) are random variables such that \( p_{X,Z,U,V,Q} = p_Q p_{X,Z} p_{U|X,Q} p_{V|Z,Q} \), \(|U| \leq |X|\), \(|V| \leq |Z|\), and \(|Q| \leq 3\).

The proof of this result is given in Appendix\( A3 \).

**D. The Information Bottleneck Method**

The information-theoretic problem posed by the Information Bottleneck (IB) method\( [14] \) can be obtained as a special case from the biclustering problem. We will introduce the problem setup and subsequently show how it can be derived as a special case of Definition\( [2] \). Note that the definition slightly differs from\( [9] \) Definition 1], where the IB problem was first studied formally. However, the achievable region is identical.

**Definition 12.** A pair \((\mu, R_1)\) is IB-achievable if, for some \( n \in \mathbb{N} \), there exists \( f : X^n \to M_1 \) with \( \log |M_1| \leq nR_1 \) and
\[
\mu \leq \frac{1}{n} I(f(X) ; Z).
\]
Let \( \mathcal{R}_{IB} \) be the set of all IB-achievable pairs.

Furthermore, Definition\( [12] \) is equivalent to lossy source coding with logarithmic loss distortion\( [15] \).

**Definition 13.** A pair \((\mu, R_1)\) is LL-achievable if, for some \( n \in \mathbb{N} \), there exist \( f : X^n \to M_1 \) and \( \phi : M_1 \to \mathcal{P}(Z^n) \) with \( \log |M_1| \leq nR_1 \) such that
\[
\mu \leq \frac{1}{n} E[\text{d}_{LL}(\phi \circ f(X) , Z)],
\]
where \( \text{d}_{LL}(p, z) := - \log p(z) \). Let \( \mathcal{R}_{LL} \) be the set of all LL-achievable pairs.

Using results from\( [15] \) we can show the following equivalences.

**Proposition 14.** For a pair \((\mu, R_1)\), the following are equivalent:

1. \((\mu, R_1) \in \overline{\mathcal{R}_{IB}} \).
2. \((H(Z) - \mu, R_1) \in \overline{\mathcal{R}_{LL}} \).
3. \((\mu, R_1, \log |Z|) \in \overline{\mathcal{R}} \).
4. There exists a random variable \( U \) such that \( U \rightleftharpoons X \rightleftharpoons Z \), \( I(X; U) \leq R \) and \( I(Z; U) \geq \mu \).

**Proof.** To show \( [1] \Leftrightarrow [2] \), apply\( [15] \), Lemma 1], noting that equality can be achieved when selecting \( \phi(m) = P(Z = f(X) = m) \) (see Lemma\( [26] \) for a proof).

\( [3] \Leftrightarrow [4] \) is shown in\( [15] \), Section III.F.

To show \( [3] \Leftrightarrow [4] \) apply the outer bound \( \mathcal{R}'_o \) of Theorem\( [7] \) and Theorem\( [9] \)(identifying \( V = Z \)).

\( \blacksquare \)
Thus, this tradeoff between “relevance” and “complexity” can be characterized by
\[
\mu_{IB}(R) := \max_{U : I(U; X) \leq R} I(U; Z),
\]
the IB function (cf. \cite{15, 23}). Interestingly, the function \( (30) \) is the solution to a variety of different problems in information theory. As mentioned in \cite{23}, \( (30) \) is the solution to the problem of loss-less source coding with one helper \cite{24, 25}. Witsenhausen and Wyner \cite{26} investigated a lower bound for a conditional entropy when simultaneously requiring another conditional entropy to fall below a threshold. Their work was a generalization of an earlier result \cite{27} and furthermore related to \cite{24, 28–30}. The conditional entropy bound in \cite{26} turns out to be an equivalent characterization of \( (30) \). Also in the context of gambling in the horse race market, \( (30) \) occurs as the maximum incremental growth in wealth when rate-limited side-information is available to the gambler \cite{31, Theorem 3}.

\textbf{E. Example: Binary Symmetric Sources}

Let \( (X, Z) \sim \text{DSBS}(p) \) be a doubly symmetric binary source \cite[Example 10.1]{10} with parameter \( p \), i.e., \( X \sim B(\frac{1}{2}) \) is a Bernoulli random variable with parameter \( \frac{1}{2} \), \( N \sim B(p) \), \( N \perp X \), and \( Z := X \oplus N \). We first show that the inner bound \( S_1' \) and the outer bound \( R_0' \) do not coincide.

\textbf{Proposition 15.} For the doubly symmetric binary source, \( S_1' \neq R_0' \).

The proof of this proposition is given in Appendix A4.

Let the region \( S_b \) be defined as
\[
S_b := \bigcup_{0 \leq \alpha, \beta \leq \frac{1}{2}} \{ (\mu, R_1, R_2) : R_1 \geq \log 2 - h_0(\alpha), R_2 \geq \log 2 - h_0(\beta), \mu \leq \log 2 - h_0(\alpha \ast p \ast \beta) \}.
\]

By choosing \( U = X \oplus N_1 \) and \( V = Z \oplus N_2 \), where \( N_1 \sim B(\alpha) \) and \( N_2 \sim B(\beta) \) are independent of \( (X, Z) \) and of each other, it follows that \( S_b \subseteq S_i \). To illustrate the tradeoff between complexity \( (R_1, R_2) \) and relevance \( (\mu) \), the boundary of \( S_b \) is depicted for \( p = 0.1 \) in Figure 2.

Based on numerical experiments, we formulate the following conjecture.

\textbf{Conjecture 16.} Given two binary variables \( U \) and \( V \) satisfying \( U \leftrightsquigarrow X \leftrightsquigarrow Z \leftrightsquigarrow V \), there exist parameters \( 0 \leq \alpha, \beta \leq \frac{1}{2} \) such that
\[
I(X; U) \geq \log 2 - h_0(\alpha)
\]

Figure 2: Boundary of \( S_b \) for \( p = 0.1 \).
This conjecture is equivalent to $S_b = S_i$ and already appeared in [13] as [13, Eq. (14), Conj. 1]. However, the second part of this conjecture, [13, Eq. (15), Conj. 1], does not hold. It claims that also $\Conv(S_b) = \mathcal{R}_o$. In what follows, we will construct a counterexample to that claim. Assuming Conjecture 16, this counterexample also implies that $S_i' \neq \mathcal{R}_o$. For the sake of simplicity we will restrict our discussion to the case $R_1 = R_2 = R$, which leads to a two-dimensional subset of the achievable region. Define the concave functions $\hat{\mu}_b(R) := \max\{\mu : (\mu, R, R) \in \Conv(S_b)\}$ and $\hat{\mu}_o(R) := \max\{\mu : (\mu, R, R) \in \mathcal{R}_o\}$ for $R \in [0, \log(2)]$. In order to show $\Conv(S_b) \neq \mathcal{R}_o$, it suffices to find $\bar{R} \in [0, \log(2)]$ with $\hat{\mu}_b(\bar{R}) < \hat{\mu}_o(\bar{R})$. It is straightforward to compute the function $\hat{\mu}_b$ numerically. On the other hand, we can obtain a lower bound on $\hat{\mu}_o$, by randomly sampling the binary probability mass functions that satisfy the Markov constraints in Theorem 7. In doing so, we encounter points strictly above the graph of $\hat{\mu}_b$.

For $\alpha, \beta \in [0, \frac{1}{2}]$, we compute

$$\tilde{R}_1 := \log(2) - h_0(\alpha)$$

(35)

$$\tilde{R}_2 := \log(2) - h_0(\beta)$$

(36)

$$\tilde{\mu} := \log(2) - h_0(\alpha \ast p \ast \beta)$$

(37)

on a suitably fine grid and numerically bound the upper concave hull of $\tilde{\mu}(\tilde{R}_1, \tilde{R}_2)$. Evaluating it at $R = \tilde{R}_1 = \tilde{R}_2$ yields an upper bound on $\hat{\mu}_b(\bar{R})$.

Based on the cardinality bounds in Propositions 8 and 11 we restrict the auxiliaries $U$ and $V$ to be binary.

We thus obtain the parameterized outer boundary of $S_i$ by evaluating for each (symmetric) rate $\tilde{R}(\alpha) = \log(2) - h_0(\alpha)$ the respective relevance bound

$$\tilde{\mu}_i(\alpha) = \log(2) - h_0(\alpha \ast p \ast \alpha).$$

(38)

To obtain the upper boundary of $\Conv(S_i)$ we numerically compute the upper concave envelope of $\tilde{\mu}(\bar{R})$. We then approximate the boundary of $\Conv(S_i)$ from below. This is achieved by randomly sampling the binary probability mass functions that satisfy the Markov constraints in Theorem 7 (but not necessarily the long Markov chain $U \rightarrow X \rightarrow Z \rightarrow V$). It turned out that the optimal choice also in this case are doubly symmetric binary marginals $(U, X)$ and $(V, Z)$. Again, the upper concave envelope was computed numerically to obtain an inner approximation of the boundary of $\Conv(S_o)$.

Figure 3 shows the resulting boundaries for $p = 0.1$ in the vicinity of $R = \log(2)$. Albeit small, there is clearly a gap between $\Conv(S_b)$ and $\Conv(S_o)$, outside the margin of numerical error. This shows that the bounds are not tight and [13, Eq. (15), Conj. 1] does not hold.

Figure 3: Outer boundaries of bounds on $\mathcal{R}$ for $p = 0.1$.
| u | v | x | z | \( P\{U = u, V = v | X = x, Z = z \} \) |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0.9952741162497373 |
| 0 | 1 | 0 | 0 | 0.0021933512177303 |
| 1 | 0 | 0 | 0 | 0.0021933512177303 |
| 1 | 1 | 0 | 0 | 0.0003391813148023 |
| 0 | 0 | 0 | 1 | 0.002529431127998 |
| 0 | 1 | 0 | 1 | 0.9949380363394698 |
| 1 | 0 | 0 | 1 | 0.0000031014045345 |
| 1 | 1 | 0 | 1 | 0.002529431127998 |
| 0 | 0 | 1 | 0 | 0.002532000611774 |
| 0 | 1 | 1 | 0 | 0.9949354668556935 |
| 1 | 0 | 0 | 1 | 0.0003751446975464 |
| 1 | 1 | 0 | 1 | 0.002532000611774 |
| 0 | 0 | 1 | 1 | 0.0000031014045345 |
| 0 | 1 | 1 | 1 | 0.0021573878349861 |
| 1 | 0 | 0 | 1 | 0.0021573878349861 |
| 1 | 1 | 0 | 1 | 0.9953100796324815 |

Table I: Distribution with largest gap

particular, we observed the largest gap between the two boundaries at a rate of about \( R = R_1 = R_2 \approx 0.675477 \). The particular distribution of \((U, V)\) at this rate, that resulted from randomly optimizing over the distribution that satisfy the Markov constraints in Theorem 7 is given in Table I. It achieves \( I(U; V) \approx 0.359613 \) which is \( \Delta \approx 1.977334 \cdot 10^{-4} \) above the inner bound. Thus, this distribution corresponds to a point \( x \in conv(S_o) \) with \( x \notin conv(S_b) \), and therefore provides a counterexample to [13, Eq. (15), Conj. 1]. We firmly believe that a tight characterization of the achievable region requires an improved outer bound. However, it appears very difficult to find a manageable outer bound based on the full Markov chain \( U \rightarrow X \rightarrow Z \rightarrow V \).

**Remark 3.** Recently, Kumar and Courtade introduced a conjecture [32], [33] concerning Boolean functions that maximize mutual information. Their work was inspired by a similar problem in computational biology [34]. A weaker form of their conjecture [33, Section IV, 2)], which was solved in [35], corresponds to a zero-rate variant of the binary example studied here.

### III. Biclustering with Multiple Sources

#### A. Problem Statement and Results

In this section we extend the information-theoretic biclustering problem introduced in Section II-A to the case of multiple sources and we provide bounds on the associated achievable region. A schematic illustration of the problem is shown in Figure 4.

Let \( X_K \) be \( K \) random variables, taking values in the finite sets \( X_k \). The random vectors \( X_K \) consist of \( n \) i.i.d. copies of \( X_k \). For \( n \in \mathbb{N} \) and \( R_K \in \mathbb{R}^{K} \), an \((n, R_K)\) code \( f_K \) consists of \( K \) functions \( f_k: X^n_k \rightarrow M_k \), where \( M_k \) is an arbitrary finite set with \( \log |M_k| \leq nR_k \) for each \( k \in K \). Recall that the symbol \( \mu_\Omega \) refers to a tuple \( \mu_\Omega \in \mathbb{R}^{3K - 2^K + 1} \), where \( \Omega \) is the set of all pairs \((A, B)\) with \( A, B \subset K \) nonempty and disjoint and \( |\Omega| = 3^K - 2^K + 1 \).

**Definition 17.** Consider an \((n, R_K)\) code \( f_K \) with \( U_k := f_k(X_k) \); for any \((A, B) \in \Omega \) we define the co-information of \( f_A \) and \( f_B \) as

\[
\Theta(f_A; f_B) := \frac{1}{n} I(U_A; U_B).
\]

**Definition 18.** A point \((\mu_\Omega, R_K)\) is achievable if, for some \( n \in \mathbb{N} \), there exists an \((n, R_K)\) code \( f_K \) such that for any \((A, B) \in \Omega \)

\[
\Theta(f_A; f_B) \geq \mu_{A,B}.
\]

\[
\text{Table I: Distribution with largest gap}
\]
The set of all achievable points is denoted $\mathcal{R}$ and we refer to its closure $\overline{\mathcal{R}}$ as achievable region.

Remark 4. A standard time sharing argument can be used to show that $\overline{\mathcal{R}}$ is a convex set.

We first state an outer bound for the achievable region whose proof is provided in Appendix A5.

Theorem 19. We have the outer bounds $\mathcal{R} \subseteq \mathcal{R}_o \subseteq \mathcal{R}'$. Here, the region $\mathcal{R}'$ is defined as

$$\mathcal{R}' := \{ (\mu_\Omega, R_K) : \sum_{k \in A} R_k \geq I(U_A; X_k | U_C) \text{ for } A, C \subseteq K, \text{ and } \mu_{A,B} \leq I(U_A; X_B), (A, B) \in \Omega \}$$

for some random variables $U_C$ with $U_A \rightarrow X_A \rightarrow X_{K\setminus A}$ for any $A \subseteq K$. The region $\mathcal{R}_o$ is defined similarly as $\mathcal{R}'$ only that the inequality for the relevance $\mu_{A,B}$ is replaced with

$$\mu_{A,B} \leq I(U_A; X_A) + I(U_B; X_B) - I(U_A U_B; X_A X_B). \quad (41)$$

The next result, whose proof is detailed in Section III-C, provides an achievable region.

Theorem 20. An inner bound for the achievable region is given by $\mathcal{R}_i \subseteq \overline{\mathcal{R}}$ where the region $\mathcal{R}_i$ consists of all points $(\mu_\Omega, R_K)$ for which there exist random variables $U_K$ satisfying $U_k \rightarrow X_k \rightarrow (X_{K\setminus k}, U_{K\setminus k})$ for all $k \in K$ and for all $(A, B) \in \Omega$ there exist subsets $\tilde{A} \subseteq A$ and $\tilde{B} \subseteq B$ such that

$$\sum_{k \in A'} R_k \geq I(X_{A'}; U_{A'} | U_{A' \setminus A'}) \text{ for all } A' \subseteq A \text{ with } A' \cap \tilde{A} \neq \emptyset, \quad (42)$$

$$\sum_{k \in B'} R_k \geq I(X_{B'}; U_{B'} | U_{B' \setminus B'}) \text{ for all } B' \subseteq B \text{ with } B' \cap \tilde{B} \neq \emptyset, \quad (43)$$

$$\mu_{A,B} \leq I(U_{\tilde{A}}; U_{\tilde{B}}). \quad (44)$$

In contrast to the case of two sources binning helps for $K > 2$ sources. For illustration, consider the case $K = 3$ and assume we are only interested in maximizing $\Theta(f_{\{1,2\}}; f_3)$. Then any information encoded by both $f_1$ and $f_2$ is redundant as it does not increase $I(f_1(X_1), f_2(X_2); f_3(X_3))$. The corresponding rate loss can be reduced by a quantize-and-bin scheme (see [24], [36], [37]).

The proof that $\mathcal{R}_i$ is indeed achievable uses typicality coding and binning. The conditions (42) and (43) ensure that $U_{\tilde{A}}$ and $U_{\tilde{B}}$ can be correctly decoded from the output of the encoders $A$ and $B$, respectively. By (44), $U_{\tilde{A}}$ and
Remark 5. The convexity of $\mathcal{R}$ (Remark 4) and Proposition 23 imply that $\overline{\mathcal{R}}_{\text{MI}}$ is a convex set.

To shorten notation we will introduce the set of random variables

$$
\mathcal{P}_* := \{U_j, Q : Q \perp (X_j, Y), U_j \rightarrow (X_j, Q) \rightarrow (X_j \cup_j Y, U_j \cup_j Y) \text{ for all } j \in \mathcal{J} \}.
$$

We can obtain an inner bound on $\overline{\mathcal{R}}_{\text{MI}}$ directly from Theorem 20 as stated in the following corollary.
Corollary 24. An inner bound for the achievable region is given by $R_{1}^{\text{MI}} \subseteq \overline{R_{\text{MI}}}$ where the region $R_{1}^{\text{MI}}$ consists of all points $(\nu_{\Omega}, R_{\mathcal{F}})$ for which there exist random variables $(U_{\mathcal{J}}, \emptyset) \in \mathcal{P}$ and for all $(A, B) \in \Omega$ there exists a subset $\tilde{A} \subseteq A$ such that
\[
\sum_{k \in \mathcal{A}'} R_{k} \geq I(X_{\mathcal{A}'}; U_{A}\mid U_{\mathcal{A}'\setminus A}) \quad \text{for all } A' \subseteq A \text{ with } A' \cap \tilde{A} \neq \emptyset, \quad (49)
\]
\[
\mu_{A,B} \leq I(U_{\tilde{A}}; Y_{B}). \quad (50)
\]

We next argue that the CEO problem introduced in Definition [22] is equivalent to the log-loss distortion approach in [15]. For $A \subseteq \mathcal{J}$ and $B \subseteq \mathcal{L}$ we consider a decoding function $g_{A,B} : \mathcal{M}_{A} \rightarrow \mathcal{P}^{Y_{B}}$ that produces a probabilistic estimate of $Y_{B}$ given the output of the encoders $A$. The quality of this probabilistic estimate is measured by log-loss distortion.

Definition 25. A point $(\nu_{\Omega}, R_{\mathcal{F}})$ is LL-achievable if, for some $n \in \mathbb{N}$, there exists an $(n, R_{\mathcal{F}})$ code $f_{\mathcal{F}}$ such that for all $(A, B) \in \tilde{\Omega}$ there exists a decoding function $g_{A,B} : \mathcal{M}_{A} \rightarrow \mathcal{P}^{Y_{B}}$ with
\[
\mathbb{E}[d_{\text{LL}}(g_{A,B}(U_{A}), Y_{B})] \leq \nu_{A,B}, \quad (51)
\]
where $U_{j} := f_{j}(X_{j})$, $j \in \mathcal{J}$. Let $R_{\text{LL}}$ be the set of all LL-achievable points.

We note that [15] considers the case where $\nu_{A,B} = 0$ only if $A = \mathcal{J}$ and $B = \mathcal{L} = \{1\}$. To show the equivalence of $R_{\text{MI}}$ and $R_{\text{LL}}$, we first state an auxiliary lemma which is essentially [15] Lemma 1] and provided just for the sake of completeness (see Appendix B2 for the proof).

Lemma 26. For any decoding function $g_{A,B}$ and code $f_{\mathcal{F}}$, we have
\[
\mathbb{E}[d_{\text{LL}}(g_{A,B}(U_{A}), Y_{B})] \geq \frac{1}{n} H(Y_{B}\mid U_{A}), \quad (52)
\]
where $U_{j} := f_{j}(X_{j})$, $j \in \mathcal{J}$ and with equality if and only if $g_{A,B}(u_{A}) = p_{Y_{B}\mid U_{A}}(\cdot\mid u_{A})$.

According to Lemma 26 the MI performance of an encoder-decoder pair is at least as good as its log-loss performance. On the other hand, the optimum can always be achieved when equality holds in (52). This is the basis for the next result.

Lemma 27. $(\nu_{\Omega}, R_{\mathcal{F}}) \in R_{\text{LL}}$ if and only if $(\nu'_{\Omega}, R_{\mathcal{F}}) \in R_{\text{MI}}$, where $\nu'_{A,B} := H(Y_{B}) - \nu_{A,B}$.

Proof. Assume that $(\nu'_{\Omega}, R_{\mathcal{F}}) \in R_{\text{MI}}$ is achieved by the $(n, R_{\mathcal{F}})$-code $f_{\mathcal{F}}$, i.e., (45) holds for all $(A, B) \in \tilde{\Omega}$. By Lemma 26, choosing the decoding functions $g_{A,B}(u_{A}) := p_{Y_{B}\mid U_{A}}(\cdot \mid u_{A})$ implies $\mathbb{E}[d_{\text{LL}}(g_{A,B}(U_{A}), Y_{B})] \leq \nu_{A,B}$ and thus $(\nu'_{\Omega}, R_{\mathcal{F}}) \in R_{\text{LL}}$. To show $R_{\text{LL}} \subseteq R_{\text{MI}}$, assume that $(\nu_{\Omega}, R_{\mathcal{F}}) \in R_{\text{LL}}$ is achieved by $f_{\mathcal{F}}$ and $g_{A,B}$, i.e., (51) holds for all $(A, B) \in \Omega$. Lemma 26 then implies $I(U_{A}; Y_{B}) \geq n\nu'_{A,B}$ and hence $(\nu'_{\Omega}, R_{\mathcal{F}}) \in R_{\text{LL}}$.

For the rest of this section, assume $L = 1$ and $X_{j} \rightarrow Y_{j} \rightarrow X_{\mathcal{J}\setminus j}$ for all $j \in \mathcal{J}$. For brevity we will write $Y := Y_{1}$ and $\nu_{A} := \nu_{A,1}$ in the following.

As a consequence of Lemma 27, the results in [15] directly apply to the CEO problem with a mutual information constraint.

Lemma 28. Assume $\nu_{A} = 0$, whenever $A \neq \mathcal{J}$. Then $(\nu_{\Omega}, R_{\mathcal{F}}) \in R_{\text{MI}}$ if and only if there exist random variables $(U_{\mathcal{J}}, Q) \in \mathcal{P}$, and the following inequalities hold.
\[
\sum_{k \in \mathcal{A}'} R_{k} \geq I(X_{\mathcal{A}'}; U_{\mathcal{A}'\setminus A}, Q) \quad \text{for all } A' \subseteq \mathcal{J}, \quad (53)
\]
\[
\mu_{\mathcal{J}} \leq I(U_{\mathcal{J}}; Y\mid Q). \quad (54)
\]

Proof. $(\nu_{\Omega}, R_{\mathcal{F}}) \in R_{\text{MI}}$ follows by applying Corollary 24 with $\tilde{A} = \emptyset$ for $A \neq \mathcal{J}$ and $\tilde{A} = \mathcal{J}$ for $A = \mathcal{J}$, taking into account the convexity of $R_{\text{MI}}$ (Remark 5). The converse follows from [15, Lemma 5] and Lemma 27.
Remark 6. The achievable region of the multiterminal source coding problem with logarithmic-loss distortion, introduced in [15, Section II], can be obtained as a special case of $\mathcal{R}_{\text{MI}}$ as well. Choose $J = L = 2$ and set $Y_j = X_j$, $j \in \{1, 2\}$. The inner bound $\mathcal{R}_{\text{MI}}$ is also tight (up to convexification) due to the results in [15].

C. Proof of Theorem 20

The proof of Theorem 20 extends the methods developed in [12] for the hypothesis testing problem (cf. Section II-B1) to a setup with multiple sources. We begin by extending [12, Lemma 8] and incorporating a binning strategy.

Lemma 29 (Existence of a code). Let $\varepsilon > 0$, $U_k \leftrightarrow X_k \leftrightarrow (X_K \setminus k, U_K \setminus k)$ for all $k \in K$, and $R_K \in \mathbb{R}_+^K$. Then, for sufficiently large $n \in \mathbb{N}$ we can obtain an $(n, R_K + \varepsilon)$ code $f_K$ with $W_k := f_k(X_k)$ and decoding functions $g_{A,\tilde{A}} : \mathcal{M}_{\tilde{A}} \rightarrow U^n_{\tilde{A}}$ for each $A, \tilde{A} \subseteq K$ with $\emptyset \neq \tilde{A} \subseteq A$ such that the following two properties hold.

For every $(A, \tilde{A}) \in \Omega$ and $\emptyset \neq \tilde{A} \subseteq A$ as well as $\emptyset \neq \tilde{B} \subseteq B$:

1) If [42] and [43] hold, then
$$P\left\{ (g_{A,\tilde{A}}(W_A), X_A, X_B, g_{B,\tilde{B}}(W_B)) \notin \mathcal{T}_{[U_{\tilde{A}}]} \right\} \leq \varepsilon. \tag{55}$$

2) $$\left| (g_{A,\tilde{A}}(\mathcal{M}_A) \times g_{B,\tilde{B}}(\mathcal{M}_B)) \cap \mathcal{T}_{[U_{\tilde{A}}]} \right| \leq \exp \left( n \left( I(U_{\tilde{A}}; X_{\tilde{A}}; X_B) + \varepsilon \right) \right). \tag{56}$$

The proof of Lemma 29 is provided in Appendix B3.

Furthermore we will need the following set of random variables.

Definition 30. For random variables $(A, B, C, D)$ and $\delta \geq 0$, define the set of random variables
$$S_{\delta}(A, B, C, D) := \{ A, \tilde{A}, B, \tilde{B} : (\tilde{A}, \tilde{B}) \in \mathcal{T}_{[AB]} \setminus \Omega, (\tilde{A}, \tilde{B}) \in \mathcal{T}_{[CD]} \setminus \Omega, (\tilde{A}, \tilde{B}) \in \mathcal{T}_{[AD]} \setminus \Omega \}. \tag{57}$$

Consider $(\mu_{\Omega}, R_K) \in \mathcal{R}_i$ and choose $U_K$ as in Theorem 20. Fix $\varepsilon > 0$ and apply Lemma 29 to obtain encoding functions $f_K$ and decoding functions $g_{A,\tilde{A}}$. For any pair $(A, B) \in \Omega$, find the nonempty subsets $\tilde{A} \subseteq A \subseteq A$ and $\tilde{B} \subseteq B \subseteq B$ such that (42)–(44) hold. Consider $(A, \tilde{A}) = (\emptyset, \emptyset)$ can be ignored since due to (44) $\tilde{A} = \emptyset \neq \tilde{B} = \emptyset$ it leads to $\mu_{A,B} \leq 0$, which is achieved by any code.) Define the functions $h_1 := g_{A,\tilde{A}} \circ f_A$ and $h_2 := g_{B,\tilde{B}} \circ f_B$. To analyze $\Theta(f_A; f_B)$, we define $D_1 := h_1(X_{\tilde{A}})$ and partition $X_{\tilde{A}}$ as $X_{\tilde{A}} = \bigcup_{u_{\tilde{A}} \in D_1} h_1^{-1}(u_{\tilde{A}})$. We may assume without loss of generality that $h_1^{-1}(u_{\tilde{A}}) \subseteq \mathcal{T}_{[X_{\tilde{A}}]}(u_{\tilde{A}})$ whenever $u_{\tilde{A}} \in \mathcal{T}_{[U_{\tilde{A}}]}$ as this does not interfere with the properties (55) and (56) of the code.

Defining $D_2$ accordingly, we set $\mathcal{F} := (D_1 \times D_2) \cap \mathcal{T}_{[U_{\tilde{A}}]}$. Using the shorthand notation $U_1 := h_1(X_{\tilde{A}})$ and $U_2 := h_2(X_B)$, we have
$$n \Theta(f_A; f_B) \geq n \Theta(h_1; h_2) = I(h_1(X_{\tilde{A}}); h_2(X_B)) \tag{58}$$

$$= \sum_{u_{\tilde{A}} \in D_1} \sum_{u_B \in D_2} P\{ U_1 = u_{\tilde{A}}, U_2 = u_B \} \log \frac{P\{ U_1 = u_{\tilde{A}}, U_2 = u_B \}}{P\{ U_1 = u_{\tilde{A}} \} P\{ U_2 = u_B \} } \tag{59}$$

$$= \sum_{(u_{\tilde{A}}, u_B) \in \mathcal{F}} P\{ U_1 = u_{\tilde{A}}, U_2 = u_B \} \log \frac{P\{ U_1 = u_{\tilde{A}}, U_2 = u_B \}}{P\{ U_1 = u_{\tilde{A}} \} P\{ U_2 = u_B \} } \tag{60}$$

where (58) follows from the data processing inequality [17, Theorem 2.8.1] and (59) is a consequence of the log-sum inequality [17, Theorem 2.7.1]. Furthermore, we defined $p_{\mathcal{F}} := P\{ (U_1, U_2) \in \mathcal{F} \}$ and $\bar{p}_{\mathcal{F}} := P\{ (U_1, U_2) \notin \mathcal{F} \}$.
with $\overline{U}_1 := h_1(\overline{X}_A)$, $\overline{U}_2 := h_2(\overline{X}_B)$, where $(\overline{X}_A, \overline{X}_B)$ are i.i.d. copies of $(\overline{X}_A, \overline{X}_B) \sim p_{X_A} p_{X_B}$. The expression (61) can be further bounded as
\begin{equation}
p_F \log \frac{p_F}{\bar{p}_F} + (1-p_F) \log \frac{1-p_F}{1-\bar{p}_F} = -h_0(p_F) - p_F \log \bar{p}_F - (1-p_F) \log (1-\bar{p}_F) \geq -h_0(p_F) - p_F \log \bar{p}_F \geq -h_0(p_F) - (1-\varepsilon) \log \bar{p}_F 
\end{equation}

where (d) follows from (55). For each $u_\tilde{A} \in D_1$ and $u_\tilde{B} \in D_2$ define
\begin{equation}
S(u_\tilde{A}, u_\tilde{B}) := \{ u_\tilde{A} \} \times h_1^{-1}(u_\tilde{A}) \times h_2^{-1}(u_\tilde{B}) \times \{ u_\tilde{B} \}
\end{equation}

and
\begin{equation}
S := \bigcup_{(u_\tilde{A}, u_\tilde{B}) \in F} S(u_\tilde{A}, u_\tilde{B}).
\end{equation}

Now, pick any $(u_\tilde{A}, \hat{x}_A, \hat{x}_B, \tilde{u}_B) \in S$. Let $\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B$, and $\hat{U}_\tilde{B}$ be the type variables corresponding to $u_\tilde{A}$, $\hat{x}_A$, $\hat{x}_B$, and $\tilde{u}_B$, respectively. From part [1] of Lemma 47 we know
\begin{equation}
P\{ \overline{X}_A = \hat{x}_A, \overline{X}_B = \hat{x}_B \} = \exp( -n \{ H(\hat{X}_A \hat{X}_B) + D_{KL}(\hat{X}_A \hat{X}_B || \overline{X}_A \overline{X}_B) \} )
\end{equation}

Let $\kappa(u_\tilde{A}, u_\tilde{B}; \hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B})$ denote the number of elements in $S(u_\tilde{A}, u_\tilde{B})$ that have type $(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B})$. Then, by part [2] of Lemma 47
\begin{equation}
\kappa(u_\tilde{A}, u_\tilde{B}; \hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) \leq \exp(n(\kappa(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) + \varepsilon)).
\end{equation}

Letting $\kappa(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B})$ be the number of elements of $S$ with type $(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B})$, we have
\begin{equation}
\kappa(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) = \sum_{(u_\tilde{A}, u_\tilde{B}) \in F} \kappa(u_\tilde{A}, u_\tilde{B}; \hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) \leq \sum_{(u_\tilde{A}, u_\tilde{B}) \in F} \exp(n(\kappa(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) + \varepsilon)).
\end{equation}

Thus,
\begin{equation}
P\{ (U_\tilde{A}, U_\tilde{B}) \in F \} \leq \sum_{\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}} \kappa(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) \cdot \exp( -n(\kappa(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) + D_{KL}(\overline{X}_A \overline{X}_B || \overline{X}_A \overline{X}_B) ) ) \leq \sum_{\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}} \exp( -n(\kappa(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) + \varepsilon) )
\end{equation}

where the sum is over all types that occur in $S$.
\begin{equation}
k(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) := I(\hat{U}_\tilde{A}; \hat{X}_A, \hat{X}_B) - I(\hat{U}_\tilde{A}; \hat{X}_A \overline{X}_B) + D_{KL}(\overline{X}_A \overline{X}_B || \overline{X}_A \overline{X}_B).
\end{equation}

Using a type counting argument (Lemma 46) we can further bound
\begin{equation}
P\{ (U_\tilde{A}, U_\tilde{B}) \in F \} \leq \exp( -n(k(\hat{U}_\tilde{A}, \hat{X}_A, \hat{X}_B, \hat{U}_\tilde{B}) + \varepsilon) )
\end{equation}
where the maximum is over all types occurring in $\mathcal{S}$. For any type $(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B)$ in $\mathcal{S}$, we have by construction $(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) \in \mathcal{S}_\delta(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B)$ (following the $\delta$-convention, Remark 10) and we can thus conclude
\[
P\{ (\tilde{U}_A, \tilde{U}_B) \in \mathcal{F} \} \leq (n + 1)^{\lfloor k \mu C \rfloor |X_k|} \cdot \max_{(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) \in \mathcal{S}_\delta(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B)} \exp \left( -n(k(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) - \varepsilon) \right).
\]

Combining (64) and (76) we have shown that for $n$ large enough
\[
\Theta(f_A; f_B) \geq -\frac{\log(2)}{n} - \frac{1}{n} \log P\{ (\tilde{U}_A, \tilde{U}_B) \in \mathcal{F} \}
\]
\[
\geq -\varepsilon + (1 - \varepsilon) \min \left( k(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) - \varepsilon \right)
\]
\[
\geq -2\varepsilon + (1 - \varepsilon) \min \left( k(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) \right)
\]
\[
\geq \min \left( k(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) \right) - (2 + I(X_A; X_B))\varepsilon,
\]
where the minimum is over all $(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) \in \mathcal{S}_\delta(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B)$. As $k(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B)$ is continuous as a function of $p_{X_A, X_B, U_B}$ it follows that for $n$ large enough,
\[
\Theta(f_A; f_B) \geq \min_{(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) \in \mathcal{S}_\delta(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B)} k(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) - C\varepsilon
\]
for some constant $C$. Observe that for $(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) \in \mathcal{S}_\delta(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B)$ we have
\[
k(\tilde{U}_A, \tilde{X}_A, \tilde{X}_B, \tilde{U}_B) = I(\tilde{U}_A; \tilde{X}_A) - I(\tilde{U}_A; \tilde{X}_B)
\]
\[
\geq I(\tilde{U}_A; \tilde{U}_B) = I(U_A; U_B).
\]
Combining (81) and (83) we have
\[
\Theta(f_A; f_B) \geq I(U_A; U_B) - C\varepsilon \geq \mu_{A,B} - C\varepsilon,
\]
where the second inequality follows from assumption (44). We hence obtain $(\mu_{\Omega} - C\varepsilon, R_K + \varepsilon) \in \mathcal{R}$; since $\varepsilon$ was arbitrary, this completes the proof.

D. Multiple Description CEO Problem

We will continue the discussion of the CEO problem of Section III-B and assume $L = 1$ as well as $X_j \rightarrow Y := Y_1 \rightarrow X_j, j \in \mathcal{J}$. To shorten notation we will again use $\nu_A := \nu_{A,1}$. In contrast to the discussion in Section III-B we will allow for $\nu_j > 0$ for any $j \in \mathcal{J}$. Loosely speaking this requires a Multiple Description code for the CEO problem, enabling the CEO to obtain valuable information from the message of the $j$-th agent alone.

Surprisingly, this extension also permits a single-letter characterization. In particular for the case $J = 2$ this allows us to give a full single-letter characterization of the achievable region, which will be explicitly stated in Corollary 33.

**Definition 31.** For a linear ordering $\sqsupseteq$ of $\mathcal{J}$ and set $\mathcal{I} \subseteq \mathcal{J}$, let the region $\mathcal{R}^{(\mathcal{I} \sqsupseteq \mathcal{J})}_{\mathcal{M}}$ be the set of tuples $(\nu_{\Omega}, R_{\mathcal{J}})$ such that there exist random variables $(U_{\mathcal{J}}, \emptyset) \in \mathcal{P}_+^*$ with
\[
R_j \geq I(U_j; X_{j} | \bigcup_{j \in \mathcal{I}}), \quad \text{for} \quad j \in \mathcal{J},
\]
\[
R_j \geq I(U_j; X_{j}), \quad \text{for} \quad j \in \mathcal{I},
\]
\[
\nu_j \leq I(U_j; Y | \bigcup_{j \in \mathcal{I}}), \quad \text{for} \quad j \notin \mathcal{I},
\]
\[
\nu_j \leq I(U_j; Y), \quad \text{for} \quad j \in \mathcal{I},
\]
\[
\nu_K \leq I(U_{\mathcal{J}}; Y),
\]
\[
\nu_A \leq 0 \quad \text{for} \quad |A| \neq 1 \land A \neq \mathcal{J}.
\]

\(^1\)For the notation regarding linear orderings refer to Section I-C1.
Remark 7. The purpose of the ordering ⊑ is to determine the order of the messages for successive decoding. Equivalently, Definition [31] could be rephrased in terms of a permutation of \( J \) in place of the linear ordering \( ⊑ \).

We are now able to state the single-letter characterization of \( \mathcal{R}_{\text{MI}} \), given that (90) holds.

**Theorem 32.** We have \( \{(ν_\Omega, R_J) \in \mathcal{R}_{\text{MI}}^{(1)} \mid (90) \text{ holds}\} = \text{conv}\left( \bigcup_{\mathcal{I} \subseteq \mathcal{J}} \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})}\right) \), where the union is over all linear orderings \( ⊑ \) of \( J \) and sets \( \mathcal{I} \subseteq \mathcal{J} \).

The proof of Theorem 32 is provided at the end of this section.

In particular, we found a single-letter characterization of the full achievable region of the Multiple Description CEO problem for the case of \( J = 2 \) agents. We state this special case separately in the following corollary to showcase some interesting features of this single-letter region.

**Corollary 33.** For \( J = 2 \), we have \( \mathcal{R}_{\text{MI}} = \text{conv}\left( \mathcal{R}_{\text{MI}}^{(1)} \cup \mathcal{R}_{\text{MI}}^{(2)} \cup \mathcal{R}_{\text{MI}}^{(3)}\right) \), where \( (ν_\Omega, R_J) \in \mathcal{R}_{\text{MI}}^{(i)} \) if, for some \( (U_J, \emptyset) \in \mathcal{P}_* \), the following inequalities are satisfied

\[
\begin{align*}
\mathcal{R}_{\text{MI}}^{(1)}: & \quad ν_1 ≤ I(Y; U_1) \\
\mathcal{R}_{\text{MI}}^{(2)}: & \quad ν_1 ≤ I(Y; U_1 | U_2) \\
\mathcal{R}_{\text{MI}}^{(3)}: & \quad ν_1 ≤ I(Y; U_1) \\
& \quad ν_2 ≤ I(Y; U_2) \\
& \quad ν_{1,2} ≤ I(Y; U_1 U_2) \\
& \quad R_1 ≥ I(U_1; X_1) \\
& \quad R_2 ≥ I(U_2; X_1) \\
& \quad R_1 ≥ I(U_1; X_1) \\
& \quad R_2 ≥ I(U_2; X_2).
\end{align*}
\]

**Proof.** Assuming \( 1 ⊑ 2 \), we obtain \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} = \mathcal{R}_{\text{MI}}^{(2)} \) if \( 1 \notin \mathcal{I} \) and otherwise \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} = \mathcal{R}_{\text{MI}}^{(3)} \). On the other hand, if \( 2 ⊑ 1 \), we obtain \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} = \mathcal{R}_{\text{MI}}^{(1)} \) if \( 2 \notin \mathcal{I} \) and otherwise also \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} = \mathcal{R}_{\text{MI}}^{(3)} \).

**Remark 8.** Note that the total available rate of encoder 2 is \( R_2 = I(X_2; U_2 | U_1) \) to achieve a point in \( \mathcal{R}_{\text{MI}}^{(1)} \). Interestingly, this rate is in general less than the rate required to ensure successful typicality decoding of \( U_2 \). However, \( ν_2 = I(Y; U_2 | U_1) \) can still be achieved.

**Remark 9.** On the other hand, fixing the random variables \( U_1, U_2 \) in the definition of \( \mathcal{R}_{\text{MI}}^{(i)} \) shows another interesting feature of this region. The achievable values for \( ν_1 \) and \( ν_2 \) vary across \( i \in \{1, 2, 3\} \) and hence do not only depend on the chosen random variables \( U_1 \) and \( U_2 \), but also on the specific rates \( R_1 \) and \( R_2 \).

It is worth mentioning that by setting \( ν_1 = ν_2 = 0 \) the region \( \mathcal{R}_{\text{MI}} \) reduces to the rate region of the CEO problem with log-loss distortion derived in [15].

The following proposition shows that \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} \) is computable, at least in principle. The given cardinality bounds are certainly not optimal, but they serve to show that \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} \) is topologically closed.

**Proposition 34.** The region \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} \) remains unchanged if the cardinality bound \( |U_j| ≤ |X_j| + 4^J \) is imposed for every \( j \in \mathcal{J} \).

The proof of Proposition 34 is given in Appendix B6.

The following two theorems provide an inner and an outer bound for \( \mathcal{R}_{\text{MI}} \). In order to show that Theorem 32 holds, we subsequently show that these bounds are indeed tight.

**Theorem 35.** We have \( \mathcal{R}_{\text{MI}}^{(\mathcal{J}, \mathcal{I})} \subseteq \mathcal{R}_{\text{MI}} \) for any \( \mathcal{I} \subseteq \mathcal{J} \) and linear ordering \( ⊑ \) of \( \mathcal{J} \).

**Theorem 36.** If \( (ν_\Omega, R_J) \in \mathcal{R}_{\text{MI}} \) then

\[
\sum_{j \in B} R_j - ν_A ≥ I(X_B; U_B | Y_Q) - I(Y; U_A \setminus B | Q) \quad \text{for } A, B \subseteq \mathcal{J}
\]

for some random variables \( (U_J, Q) \in \mathcal{P}_* \).

The proof of Theorems 35 and 36 is given in Appendices B4 and B5, respectively.

We will, however, only require the following corollary of Theorem 36.
Corollary 37. For any \((\nu_\Omega^*, R_J) \in \mathcal{R}_{MI}\) there are random variables \((U_J, Q) \in \mathcal{P}_*\) with
\[
R_j \geq 0 \quad \text{for all } j \in J,
\]
\[
\sum_{j \in A} R_j - \nu_J \geq I(X_A; U_A|Q) - I(Y; U_{J\setminus A}|Q) \quad \text{for all } A \subseteq J,
\]
\[
R_j - \nu_j \geq I(X_j; U_j|Q) \quad \text{for all } j \in J,
\]
\[
\nu_j \leq I(Y; U_j|Q) \quad \text{for all } j \in J.
\]

Proof of Theorem 33 We will make use of some rather technical results on convex polyhedra, given in Appendix B8.

Assume \((\nu_\Omega^*, R_J) \in \mathcal{R}_{MI}\). We can then find \((U_J, Q) \in \mathcal{P}_*\) such that \((97)-(100)\) hold. We define \(\nu_\Omega^* := -\nu_\Omega^*\) to simplify notation. It is straightforward to check that the inequalities \((97)-(100)\) define a sequence of closed convex polyhedra \(\mathcal{S}^{(j)}\) in the variables \((R_J, \tilde{\nu}_\Omega)\) that satisfy assumptions 1 and 2 of Lemma 44. \(\mathcal{S}^{(0)}\) is defined by \((97)\) and \((98)\) alone and for \(j \in [0:J]\), the polyhedron \(\mathcal{S}^{(j)}\) is given in the \(K + j\) variables \((R_J, \tilde{\nu}_J, \tilde{\nu}_{[1:j]}^*)\) by adding constraints \((99)\) and \((100)\) for each \(j_0 \in [1:j]\). The set \(\mathcal{S}^{(0)}\) is a supermodular polyhedron [39, Section 2.3] in the \(K\) variables \((R_J, \tilde{\nu}_J)\) on \((\mathcal{K}, 2^\mathcal{K})\) with rank function
\[
f(A) = \begin{cases} 0, & K \notin A, \\ I(X_{A\setminus K}; U_{A\setminus K}|Q) - I(Y; U_{J\setminus A}|Q), & K \in A, \end{cases}
\]
where supermodularity follows via standard information-theoretic arguments. By the extreme point theorem [39, Theorem 3.22], every extreme point of \(\mathcal{S}^{(0)}\) is associated with a linear ordering \(\subseteq\) of \(\mathcal{K}\). Such an extreme point is given by
\[
R_j^{(\subseteq)} = 0 \quad \text{for } j \subset K,
\]
\[
R_j^{(\subseteq)} = I(U_j; X_j|U_{\supseteq j}Q) \quad \text{for } j \subset K,
\]
\[
\nu_j^{(\subseteq)} = I(Y; U_{\supseteq jK}|Q) - I(Y; U_{\subset K}|Q).
\]
Assumption 3 of Lemma 44 can now be verified by
\[
R_j^{(\subseteq)} \leq I(X_j; U_j|Q) + I(Y; U_j|Q) = I(X_j; U_j|Q).
\]
By applying Lemma 44 we find that every extreme point of \(\mathcal{S}^{(J)}\) is given by a subset \(I \subseteq J\) and an ordering \(\subseteq\) of \(\mathcal{K}\) as
\[
R_j^{(\subseteq, I)} = I(X_j; U_j|Q), \quad j \in I
\]
\[
R_j^{(\subseteq, I)} = 0, \quad j \notin I \land j \subset K
\]
\[
R_j^{(\subseteq, I)} = I(U_j; X_j|U_{\supseteq j}Q), \quad j \notin I \land j \subset K
\]
\[
\nu_j^{(\subseteq, I)} = I(Y; U_{\supseteq jK}|Q) - I(Y; U_{\subset K}|Q), \quad j \in I
\]
\[
\nu_j^{(\subseteq, I)} = I(U_j; Y|Q), \quad j \notin I \land j \subset K
\]
\[
\nu_j^{(\subseteq, I)} = -I(U_j; X_j|Y), \quad j \notin I \land j \subset K
\]
\[
\nu_j^{(\subseteq, I)} = I(U_j; Y|U_{\supseteq j}Q), \quad j \notin I \land j \subset K.
\]
For each \(q \in Q\) with \(P\{Q = q\} > 0\) let the point \((\nu_{\Omega}^{(\subseteq, I,q)} R_J^{(\subseteq, I,q)})\) be defined by \((106)-(112)\) but given \(\{Q = q\}\). By substituting \(U_j \rightarrow \emptyset\) if \(j \notin I \land j \subset K\), we see that \((\nu_{\Omega}^{(\subseteq, I,q)}, R_J^{(\subseteq, I,q)}) \in \mathcal{R}_{MI}^{(\subseteq, I)}\) and consequently \((\nu_{\Omega}^{(\subseteq, I)}, R_J^{(\subseteq, I)}) \in \mathcal{R}_{MI}^{(\subseteq, I)}\). Defining the orthant \(O' := \{\nu_{\Omega}, R_J : \nu_{\Omega} \leq 0, R_J \geq 0\}\), this implies \((\nu_K, R_J) \in \mathcal{R}_{MI}^{(\subseteq, I)}\). Together with Theorem 35 and the convexity of \(\mathcal{R}_{MI}^{(\subseteq, I)}\)
(Remark 5) we obtain
\[ \mathcal{R}_{\text{MI}} \subseteq \text{conv}\left( \bigcup_{\subset \mathcal{I}} \mathcal{R}_{\text{MI}}^{(\subset, \mathcal{I})} \right) \subseteq \overline{\mathcal{R}_{\text{MI}}}. \] (113)

It remains to show that \( \text{conv}\left( \bigcup_{\subset \mathcal{I}} \mathcal{R}_{\text{MI}}^{(\subset, \mathcal{I})} \right) \) is topologically closed. Using Proposition 34 we can write
\[ \mathcal{R}_{\text{MI}}^{(\subset, \mathcal{I})} = F^{(\subset, \mathcal{I})}(\mathcal{P}^*_s) + \mathcal{O}', \] where \( \mathcal{P}^*_s := \{p_{Y, X_j, U_j} : (U_j, \emptyset) \in \mathcal{P}_s, |U_j| \leq |X_j| + 4^J, j \in \mathcal{J} \} \) is a compact subset of the probability simplex and \( F^{(\subset, \mathcal{I})} \) is a continuous function, given by the definition of \( \mathcal{R}_{\text{MI}}^{(\subset, \mathcal{I})}, \) (85)–(90). We can thus write
\[ \text{conv}\left( \bigcup_{\subset \mathcal{I}} \mathcal{R}_{\text{MI}}^{(\subset, \mathcal{I})} \right) = \text{conv}\left( \bigcup_{\subset \mathcal{I}} F^{(\subset, \mathcal{I})}(\mathcal{P}^*_s) + \mathcal{O}' \right) = \text{conv}\left( \bigcup_{\subset \mathcal{I}} F^{(\subset, \mathcal{I})}(\mathcal{P}^*_s) \right) + \mathcal{O}', \] (114)
which follows from [40, Theorem 1.1.2]. Note that \( F^{(\subset, \mathcal{I})}(\mathcal{P}^*_s) \) is compact as the continuous image of a compact set and thus, \( \text{conv}\left( \bigcup_{\subset \mathcal{I}} F^{(\subset, \mathcal{I})}(\mathcal{P}^*_s) \right) \) is compact as the convex hull of a compact set (finite union of compact sets) [41, Theorem 2.3.4]. This finishes the proof, as \( \mathcal{O}' \) is closed and the sum of a compact and a closed set is closed [42, Exercise 1.3(e)].

IV. SUMMARY AND DISCUSSION

We introduced a novel multi-terminal source coding problem termed information-theoretic biclustering. Interestingly, this problem is related to several other problems at the frontier of statistics and information theory and offers a formidable mathematical complexity. Indeed, it is fundamentally different from “classical” distributed source coding problems where the encoders usually aim at reducing, as much as possible, redundant information among the sources while still satisfying a fidelity criterion. Whereas in the considered problem, the encoders are interested in maximizing precisely such redundant information.

Although an exact characterization of the achievable region is mathematically very challenging and still remains elusive, we provided outer and inner bounds to the set of achievable rates. We thoroughly studied the special case of two symmetric binary sources for which novel cardinality bounding techniques were developed. Based on numerical evidence we formulated a conjecture that entails an explicit expression for the inner bound. This conjecture provides strong evidence that our inner and outer bound do not meet in general. We firmly believe that an improved outer bound, satisfying the adequate Markov chains, is required for a tight characterization of the achievable region.

We further established analogous bounds to the achievable rate region of information-theoretic biclustering with more than two sources. However, these bounds cannot be tight since the famous Körner-Marton problem constitutes a counterexample. For an analogue of the well-known CEO problem we showed that our bounds are tight in a special case, leveraging existing results from multiterminal lossy source coding. Furthermore we considered a Multiple Description CEO problem which surprisingly also permits a single-letter characterization of the achievable region. The resulting region has the remarkable feature that it allows to exploit rate that is in general insufficient to guarantee successful typicality encoding.

The interesting challenge of the biclustering problem lies in the fact that one needs to bound the mutual information between two arbitrary encodings solely based on their rates. Available information-theoretic manipulations seem incapable of handling this requirement well.

ACKNOWLEDGMENT

The authors would like to thank Shlomo Shamai (Shitz) and Emre Telatar for insightful discussions regarding the binary example.

APPENDIX

A. Proofs from Section [11]
1) Proof of Theorem \[\theta\] 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 \(\Theta(f; g) \geq \mu\). We define the random variables \(U_l := (X^{l-1}, f(X))\) and \(V_l := (Z^{l-1}, g(Z))\) and obtain

\[nR_1 \geq H(f(X)) = I(f(X); X)\]  \hspace{1cm} (115)

\[= \sum_{l=1}^{n} I(X_l; f(X) | X^{l-1})\]  \hspace{1cm} (116)

\[= \sum_{l=1}^{n} I(X_l; U_l)\]  \hspace{1cm} (117)

and accordingly

\[nR_2 \geq \sum_{l=1}^{n} I(Z_l; V_l)\]  \hspace{1cm} (118)

We also have

\[n\mu \leq I(f(X); g(Z))\]  \hspace{1cm} (119)

\[= I(f(X); X) - I(f(X); X | g(Z))\]  \hspace{1cm} (120)

\[= I(f(X); X) + I(g(Z); Z) - I(f(X); X | g(Z)) - I(g(Z); Z)\]  \hspace{1cm} (121)

\[= I(f(X); X) + I(g(Z); Z) - I(f(X); X | g(Z)) - I(g(Z); XZ)\]  \hspace{1cm} (122)

\[= I(f(X); X) + I(g(Z); Z) - I(f(X), g(Z); XZ)\]  \hspace{1cm} (123)

\[= \sum_{l=1}^{n} \left[ I(U_l; X_l) + I(V_l; Z_l) - I(U_lV_l; X_lZ_l) \right].\]  \hspace{1cm} (124)

Now a standard time-sharing argument shows \(\mathcal{R} \subseteq \mathcal{R}_o\). To see \(\mathcal{R}_o \subseteq \mathcal{R}_o'\), note that

\[I(U; X) + I(V; Z) - I(UV; XZ)\]  \hspace{1cm} (125)

\[= I(V; Z) - I(V; XZ | U)\]  \hspace{1cm} (126)

\[= I(U; Z) + I(V; Z) - I(U; Z) - I(V; XZ | U)\]  \hspace{1cm} (127)

\[= I(U; Z) + I(V; Z) - I(U; Z) - I(V; Z | U) - I(V; X | ZU)\]  \hspace{1cm} (128)

\[= I(U; Z) + I(V; Z) - I(UV; Z) - I(V; X | ZU)\]  \hspace{1cm} (129)

\[= I(U; Z) - I(U; Z | V) - I(V; X | ZU)\]  \hspace{1cm} (130)

\[\leq I(U; Z)\]  \hspace{1cm} (131)

and by a symmetric argument

\[I(U; X) + I(V; Z) - I(UV; XZ) \leq I(V; X).\]  \hspace{1cm} (132)

2) Proof of Proposition \[\hat{\nu}\] Most steps in the proof apply to both \(\mathcal{R}_o\) and \(\mathcal{R}_o'\). We thus state the proof for \(\mathcal{R}_o\) and point out the required modifications where appropriate.

Define the set of pmfs (with finite, but arbitrarily large support)

\[Q := \{p_{U', V', X', Z'} : X' \sim X, Z' \sim Z', U' \sim U, V' \sim V'\}, \text{ and } (X', Z') \sim (X, Z)\} \hspace{1cm} (133)\]

and the compact set of pmfs with fixed alphabet size

\[Q(a, b) := \{p_{U', V', X', Z'} \in Q : |U'| = a, |V'| = b\}.\]  \hspace{1cm} (134)

Define the continuous vector valued function \(F := (F_1, F_2, F_3)\) as

\[F_1(p_{U,V,X,Z}) := I(X; U) + I(Z; V) - I(UV; XZ),\]  \hspace{1cm} (135)

\[F_2(p_{U,V,X,Z}) := I(U; X),\]  \hspace{1cm} (136)

\[F_3(p_{U,V,X,Z}) := I(V; Z).\]  \hspace{1cm} (137)
In the proof of $R'_o = \text{conv}(S'_o)$, define $F_1$ as $F_1(p_\mathcal{U}; \mathcal{V}; \mathcal{X}; \mathcal{Z}) := \min\{I(U;Z), I(V;X)\}$. We can now write $R_o = F(Q) + O'$ and $S_o = F(Q(\mathcal{X}|, |\mathcal{Z}|)) + O'$ where $O' := (\mathbb{R}_- \times \mathbb{R}_+ \times \mathbb{R}_+)$. Thus, we need to show
$$R_o = \text{conv}\left(F(Q(\mathcal{X}|, |\mathcal{Z}|)) + O'\right).$$
(138)

Since $R_o$ is convex, by defining the extended real function $\bar{\psi}(\lambda) := \sup_{\lambda \in R_o} \lambda \cdot x$ we obtain [41] Theorem 2.2, 3.
$$\text{conv}(R_o) = R_o = \bigcap_{\lambda \in R^3} \{ x \in \mathbb{R}^3 : x \cdot \lambda \leq \bar{\psi}(\lambda) \}.$$  
(139)

From the definition of $R_o$, we clearly have $\bar{\psi}(\lambda_1, \lambda_2, \lambda_3) = +\infty$ whenever $\lambda_1 < 0$, $\lambda_2 > 0$, or $\lambda_3 > 0$, and $\bar{\psi}(\lambda) = \sup_{p \in Q} \lambda \cdot F(p)$ everywhere else. Thus,
$$R_o = \bigcap_{\lambda \in \Lambda} \{ x \in \mathbb{R}^3 : x \cdot \lambda \leq \psi(\lambda) \},$$  
(140)
where $\Lambda$ is the quadrant defined by $\lambda_1 \geq 0$, $\lambda_2 \leq 0$, and $\lambda_3 \leq 0$. We next show that for any $\lambda \in \Lambda$,
$$\psi(\lambda) = \max_{p \in Q(\mathcal{X}|, |\mathcal{Z}|)} \lambda \cdot F(p).$$  
(141)

Choose arbitrary $\lambda \in \Lambda$ and $\delta > 0$. We can find random variables $(\bar{\mathcal{U}}, \bar{\mathcal{X}}, \bar{\mathcal{Z}}, \bar{\mathcal{V}})$ such that $\bar{\psi}(\lambda) = \psi(\lambda) + \delta$. By compactness of $Q(a, b)$ and continuity of $F$, there are random variables $(U, X, Z, V) \sim p \in Q(\mathcal{U}|, |\mathcal{V})$ with
$$\lambda \cdot F(p) = \max_{p' \in Q(\mathcal{U}|, |\mathcal{V})} \lambda \cdot F(p') \geq \lambda \cdot F(\bar{p}) \geq \psi(\lambda) - \delta.$$  
(142)

We now show that there exists $p'' \in Q(\mathcal{X}|, |\mathcal{Z}|)$ with
$$\lambda \cdot F(p'') = \lambda \cdot F(p).$$  
(143)

By (142), $\lambda \cdot F(p) = \max_{p' \in Q(\mathcal{U}|, |\mathcal{V})} \lambda \cdot F(p') = 0$ for any $\lambda \in \Lambda$ with $\lambda_1 + \min(\lambda_2, \lambda_3) \leq 0$ as a consequence of the inequalities $F_1 \leq F_2$ and $F_1 \leq F_3$. Thus, we only need to show (143) for $\lambda \in \Lambda$ with $\lambda_1 + \min(\lambda_2, \lambda_3) > 0$. To this end we use the perturbation method [43], [44] and perturb $p$, obtaining
$$U', X', Z', V' \sim p'(u, x, z, v) = p(u, x, z, v)(1 + \varepsilon \phi(u)).$$  
(144)

We require
$$1 + \varepsilon \phi(u) \geq 0,$$  
(145)
$$E[\phi(U)] = 0,$$  
(146)
$$E[\phi(U)|X = x, Z = z] = 0.$$  
(147)

The conditions (145) and (146) ensure that $p'$ is a valid pmf and (147) implies $p' \in Q$. Observe that for any $\phi$, there is an $\varepsilon_0 > 0$ such that (145) is satisfied for $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$. Furthermore, (147) is equivalent to
$$E[\phi(U)|X = x] = 0$$  
(148)
because of the Markov chain $U \rightarrow X \rightarrow Z$. If $|\mathcal{U}| \geq |\mathcal{X}| + 1$ there is a non-trivial solution to (146) and (148), which means there exists $\phi \neq 0$ such that (145) and (147) are satisfied. We have
$$\lambda \cdot F(p') = \lambda_1 [I(X;U) - I(UV;XZ) + H(Z) + \varepsilon H(U) - \varepsilon H(UX) \phi]$$
$$- \varepsilon H(UV) + \varepsilon H(UXZV) \phi + H(V') - H(\mathcal{Z}'\mathcal{V}')$$
$$+ \lambda_2 [I(X;U) + \varepsilon H(U) - \varepsilon H(UX) \phi]$$
$$+ \lambda_3 [H(Z) + H(V') - H(\mathcal{Z}'\mathcal{V}')].$$  
(149)

We used the shorthand $H(UX) \phi := -\sum_{u,x} p(u,x)\phi(u) \log p(u,x)$ and analogous for other combinations of random
variables. By (142) we have \( \frac{\partial^2}{\partial \varepsilon^2} \mathbf{F}(p') \big|_{\varepsilon=0} \leq 0 \). Observe that
\[
\frac{\partial}{\partial \varepsilon} (H(V') - H(Z'V')) = \frac{\partial}{\partial \varepsilon} \sum_{z,v} p'(z,v) \log \frac{p'(z,v)}{p'(v)}
\]
\[
= \sum_{z,v} \frac{\partial p'(z,v)}{\partial \varepsilon} \log \frac{p'(z,v)}{p'(v)} + p'(z,v) \frac{\partial^2}{\partial \varepsilon^2} \log \frac{p'(z,v)}{p'(v)}
\]
\[
= \sum_{z,v} \frac{\partial p'(z,v)}{\partial \varepsilon} \log \frac{p'(z,v)}{p'(v)} + \left( \frac{\partial p'(z,v)}{\partial \varepsilon} \right)^2 - \frac{p'(z,v) \partial^2 p'(v)}{p'(v)^2}
\]
and consequently
\[
\frac{\partial^2}{\partial \varepsilon^2} \lambda \cdot \mathbf{F}(p') = (\lambda_1 + \lambda_3) \frac{\partial^2}{\partial \varepsilon^2} (H(V') - H(Z'V'))
\]
\[
= (\lambda_1 + \lambda_3) \sum_{z,v} \left( \frac{\partial p'(z,v)}{\partial \varepsilon} \right)^2 \frac{1}{p'(z,v)}
\]
\[
- 2 \frac{\partial p'(z,v)}{\partial \varepsilon} \frac{\partial p'(v)}{\partial \varepsilon} \frac{1}{p'(v)} + \left( \frac{\partial p'(v)}{\partial \varepsilon} \right)^2.
\]
Here we already used that \( \frac{\partial^2 p'(v)}{\partial \varepsilon^2} = \frac{\partial^2 p'(z,v)}{\partial \varepsilon^2} = 0 \). It is straightforward to calculate
\[
\frac{\partial p'(v)}{\partial \varepsilon} = p_V(v) E[\phi(U) | V = v]
\]
\[
\frac{\partial p'(z,v)}{\partial \varepsilon} = p_{Z,V}(z,v) E[\phi(U) | V = v, Z = z]
\]
\[
p'(z,v) \big|_{\varepsilon=0} = p_{Z,V}(z,v)
\]
\[
p'(v) \big|_{\varepsilon=0} = p_V(v)
\]
and, thus, taking into account that \( \lambda_1 + \lambda_3 > 0 \),
\[
0 \geq \sum_{z,v} p(z,v) \left( E[\phi(U) | V = v, Z = z] - E[\phi(U) | V = v] \right)^2,
\]
which implies for any \((z,v)\) with \( p(z,v) > 0 \),
\[
\sum_u p_{U|Z,V}(u|z,v) \phi(u) = \sum_u p_{U|V}(u|v) \phi(u).
\]
From (161) we can conclude
\[
H(V') - H(Z'V') = \sum_{z,v} p'(z,v) \log \frac{p'(z,v)}{p'(v)}
\]
\[
= \sum_{z,v,u} p(u,z,v)(1 + \varepsilon \phi(u)) \log \frac{\sum_{u'} p(u',z,v)(1 + \varepsilon \phi(u'))}{\sum_{u'} p(u',v)(1 + \varepsilon \phi(u'))}
\]
\[
= \sum_{z,v,u} p(u,z,v)(1 + \varepsilon \phi(u)) \log \frac{p(z,v) \sum_{u'} p(u'|z,v)(1 + \varepsilon \phi(u'))}{p(v) \sum_{u'} p(u'|v)(1 + \varepsilon \phi(u'))}
\]
\[
= \sum_{z,v,u} p(u,z,v)(1 + \varepsilon \phi(u)) \log \frac{p(z,v)(1 + \varepsilon \sum_{u'} p(u'|z,v) \phi(u'))}{p(v)(1 + \varepsilon \sum_{u'} p(u'|v) \phi(u'))}
\]
Define the continuous vector-valued function
\[ |U| \leq |X| \]
\[ \text{hull of a connected set in } \mathbb{R} \]
follows directly from the strengthened Carathéodory theorem \[45, \text{Theorem 18(ii)} \] because \[ \text{conv} \]
carried out for \[ V \]
reduces the cardinality of \[ U \]
may now choose \[ \lambda \]
and yields \[ F \]
maximal, i.e., such that there is at least one \[ p \]
by one and may be repeated until \[ \phi \equiv 0 \], i.e. \[ |U'| = |X| \]. The same process can be carried out for \[ V \] and yields \[ p'' \in \mathcal{Q}(|X|, |Z|) \], such that \[ (143) \] holds.

In the proof of \[ R'_o \), we apply the support lemma \[ [10] \text{ Appendix C} \] with \[ |X| - 1 \] test functions \[ f_x(p_{X'}) := p_{X'}(x) \] \( x \in X \) and with the function
\[ f(p_{X'}) := \lambda_1 \min \{ I(V; X), H(Z) - H(Z') \} + \lambda_2 (H(X) - H(X')) + \lambda_3 I(Z; V), \]
where \( (Z', X') \sim p_{X'} p_{Z|x} \). Consequently there exists a random variable \( U' \) with \( (U', X, Z, V) \sim p' \in \mathcal{Q}(|X|, |Z|) \) and \[ \lambda \cdot F(p') = \lambda \cdot F(p) \]. By applying the same argument to \( V \), we obtain \[ p'' \in \mathcal{Q}(|X|, |Z|) \] such that \[ (143) \] holds.

Due to \[ (142) \] and \[ (143) \] we now have
\[ \lambda \cdot F(p'') = \lambda \cdot F(p) \geq \psi(\lambda) - \delta. \] (173)

Since this holds for arbitrary \( \delta > 0 \) and since \( \mathcal{Q}(|X|, |Z|) \) is compact, \[ (141) \] holds. Now \[ (140) \] implies
\[ R'_o = \text{conv} \mathcal{Q}(|X|, |Z|) + O' \]
\[ = \text{conv} \mathcal{Q}(|X|, |Z|) + O' \]
\[ = \text{conv} (F(\mathcal{Q}(|X|, |Z|)) + O' \]
\[ \subseteq F(Q) + O' \]
\[ = R'_o, \]
where \[ (175) \] follows from \[ (40) \text{ Theorem 1.1.2} \] and \[ (176) \] is a consequence of \[ (42) \text{ Exercise 1.3(e)} \], considering that \( F(\mathcal{Q}(|X|, |Z|)) \) and therefore also its convex hull is compact \[ (41) \text{ Theorem 2.3.4} \]. The relation \[ (177) \] is a consequence of \( \mathcal{Q}(|X|, |Z|) \subseteq Q \) and the convexity of \( F(Q) \).

3) Proof of Proposition \[ [11] \]
We only need to show \[ \text{conv}(S_i) = \text{conv}(R_i) \] as the cardinality bound \( |Q| \leq 3 \) follows directly from the strengthened Carathéodory theorem \[ [45] \text{ Theorem 18(ii)} \] because \( \text{conv}(R_i) \) is the convex hull of a connected set in \( \mathbb{R}^3 \). We will only show the cardinality bound \( |U| \leq |X| + 1 \) and \( |V| \leq |Z| + 1 \) can be shown directly using the convex cover method \[ [10] \text{ Appendix C}, [24], [46] \]. Define the continuous vector-valued function
\[ F(p_{U', X', Z,V}) := (I(\tilde{U}; \tilde{V}), I(\tilde{X}; \tilde{U}), I(\tilde{Z}; \tilde{V})). \] (179)

Define the compact, connected sets of pmfs
\[ Q := \{ p_{U', X', Z,V} : p_{U', X', Z,V} = p_{U'|X} p_{X} z p_{V|Z}, U = \{0, \ldots, |X|\}, V = \{0, \ldots, |Z|\} \}, \] (180)
\( Q' := \{ p_{\tilde{u},x,z,v} \in Q : \tilde{u} = \{ 1, \ldots, |\mathcal{X}| \} \}. \tag{181} \)

To complete the proof of the proposition, it suffices to show

\[
\text{conv}(\mathcal{F}(Q)) \subseteq \text{conv}(\mathcal{F}(Q')) \tag{182}
\]

since we then have

\[
\begin{align*}
\text{conv}(\mathcal{R}_i) &= \text{conv}(\mathcal{F}(Q) + \mathcal{O}') \\
&= \text{conv}(\mathcal{F}(Q)) + \text{conv}(\mathcal{O}') \\
&= \text{conv}(\mathcal{F}(Q)) + \mathcal{O}' \\
&\subseteq \text{conv}(\mathcal{F}(Q')) + \mathcal{O}' \tag{183} \\
&= \text{conv}(\mathcal{F}(Q') + \mathcal{O}') \tag{184} \\
&= \text{conv}(\mathcal{F}(Q')) \tag{185}
\end{align*}
\]

where \( (184) \) and \( (187) \) follow from \cite[Theorem 1.1.2]{40}, and we used \( \mathcal{O}' = (\mathbb{R}_- \times \mathbb{R}_+ \times \mathbb{R}_+) \). The region \( \mathcal{F}(Q) \subseteq \mathbb{R}^3 \) is compact and connected [\cite[Theorem 26.5]{47}, \cite[Theorem 4.22]{48}]. Therefore, its convex hull \( \text{conv}(\mathcal{F}(Q)) \) is compact [\cite[Corollary 5.33]{49}] and can be represented as an intersection of halfspaces in the following manner [\cite[Proposition 2.2, 3.]{41}]: For \( \lambda = (\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3 \), define \( V(\lambda) := \max_{x \in \mathbb{R}^3} \lambda \cdot x \), then

\[
\text{conv}(\mathcal{F}(Q)) = \bigcap_{\lambda \in \mathbb{R}^3} \{ x \in \mathbb{R}^3 : \lambda \cdot x \leq V(\lambda) \}. \tag{189}
\]

With the same reasoning we obtain

\[
\text{conv}(\mathcal{F}(Q')) = \bigcap_{\lambda \in \mathbb{R}^3} \{ x \in \mathbb{R}^3 : \lambda \cdot x \leq V'(\lambda) \}, \tag{190}
\]

where \( V'(\lambda) := \max_{x \in \mathbb{F}(Q')} \lambda \cdot x \). We next show \( V'(\lambda) \geq V(\lambda) \) which already implies (182) due to (189) and (190).

Let \( \lambda' := \lambda \setminus \{ x \in \mathcal{X} \} \) where \( x \in \mathcal{X} \) is arbitrary. Define the test function \( t_x(p_{\tilde{x}}) := p_{\tilde{x}}(x) \) for \( x \in \mathcal{X} \) and abbreviate \( t = (t_x)_{x \in \mathcal{X}'} \). Choose any \( \lambda \in \mathbb{R}^3 \) and fix \( (U, X, Z, V) \) that achieve \( \lambda \cdot \mathcal{F}(p_{U,X,Z,V}) = V(\lambda) \). Define the continuous function

\[
f(p_{\tilde{x}}) := \lambda_1(H(X) - H(\tilde{X})) + \lambda_2I(Z;V) + \lambda_3(H(V) - H(\tilde{V})) \tag{191}
\]

where \( (\tilde{V}, \tilde{Z}, \tilde{X}) \sim p_{V/Z/P}\mathbb{X}\). Obviously \( (p_{x}(x))_{x \in \mathcal{X}}, V(\lambda) \) lies in the convex hull of the compact, connected set \( (t, f)(Q(\mathcal{X}')) \). Therefore, by the strengthened Carathéodory theorem [\cite[Theorem 18(ii)]{45}, \cite[Theorem 18(ii)]{45}], \( |\mathcal{X}| \) points suffice, i.e., there exists a random variable \( U' \) with \( |U'| = 24 \mathcal{X} \) and \( p_{U',X,Z,V} \in Q' \), such that \( E_{U'} [f(p_{X,(\cdot)|U'})] = \lambda \cdot \mathcal{F}(p_{U',X,Z,V}) = V(\lambda) \). This shows \( V'(\lambda) \geq V(\lambda) \).

By applying the same reasoning to \( V \), one can show that \( |V| = |Z| \) is sufficient.

4) Proof of Proposition [15]. With \( U = X \oplus N_1 \) and \( V = Z \oplus N_2 \), where \( N_1, N_2 \sim B(\alpha) \) are independent of \( (X, Z) \) and of each other, it follows that \( (\mu, R, R) := \{ \log 2 - h_0(\alpha * p), \log 2 - h_0(\alpha), \log 2 - h_0(\alpha) \} \subseteq \mathcal{R}_\mathcal{X} \) for some \( \alpha \in (0, \frac{1}{2}) \). Assuming \( (\mu, R, R) \in \mathcal{S}_\mathcal{X} \) and choosing \( U, V, \) and \( Q \) according to Proposition [11] implies that

\[
\begin{align*}
H(X|UQ) &\geq h_0(\alpha), \\
H(Z|VQ) &\geq h_0(\alpha), \\
I(U;V|Q) &\geq \log 2 - h_0(\alpha * p). \tag{194}
\end{align*}
\]

Applying Mrs. Gerber’s Lemma (MGL) [\cite[Theorem 1]{27}] yields

\[
H(X|VQ) \geq h_0^{-1}(H(Z|VQ)) * p \geq h_0(\alpha * p), \tag{195}
\]

where the second inequality follows from \cite[Theorem 1]{193}. Thus, \( I(X;V|Q) \leq \log 2 - h_0(\alpha * p) \) and furthermore \( I(X;V|Q) \geq I(U;V|Q) \) due to the Markov chain property. These two inequalities in combination with (194) imply \( I(X;V|Q) = I(U;V|Q) \), or equivalently \( I(X;V|UQ) = 0 \), which corresponds to the Markov chain property.
Assume without loss of generality that $p_Q(q) > 0$ for all $q \in Q$.
If $p_{U|X,Q}(u|x,q) > 0$ for $(u,x,q) \in \{0,1\} \times \{0,1\} \times Q$, then (197) necessitates

$$p_{V|U,Q}(v|u,q) = p_{V|X,Q}(v|x,q)$$

(198)

for $v \in \{0,1\}$. Next, we partition $Q$ into the three disjoint subsets

$$Q_1 := \{ q \in Q : U = X|Q = q \text{ or } U = 1 \oplus X|Q = q \}$$

(199)

$$Q_2 := \{ q \in Q : U = 0|Q = q \text{ or } U = 1|Q = q \}$$

(200)

$$Q_3 := \{ q \in Q : p_{U|X,Q}(0|x,q) > 0 \text{ and } p_{U|X,Q}(1|x,q) > 0 \text{ for some } x \in \{0,1\} \}.$$  

(201)

Given $q \in Q_3$, we apply (198) twice and obtain

$$p_{V|U,Q}(v|0,q) = p_{V|X,Q}(v|x,q) = p_{V|U,Q}(v|1,q).$$

(202)

i.e., $I(U;V|Q = q) = 0$, which is also true for $q \in Q_2$. We can then develop (194) as

$$\log 2 - h_0(\alpha * p) \leq I(U;V|Q) = P\{Q \in Q_1\} I(X;Z) = P\{Q \in Q_1\} (\log 2 - h_0(p)).$$

(203)

On the other hand we obtain from (192) that

$$h_0(\alpha) \leq H(X|UQ) \leq P\{Q \in (Q_2 \cup Q_3)\} \log 2 = (1 - P\{Q \in Q_1\}) \log 2.$$  

(204)

Combination of the previous two inequalities leads to

$$\frac{\log 2 - h_0(\alpha * p)}{\log 2 - h_0(p)} \leq P\{Q \in Q_1\} \leq 1 - \frac{h_0(\alpha)}{\log 2},$$

(205)

which is a contradiction since $\frac{\log 2 - h_0(\alpha * p)}{\log 2 - h_0(p)} > 1 - \frac{h_0(\alpha)}{\log 2}$.

5) **Proof of Theorem 19** If $(\mu_{\Omega}, R_K) \in \mathcal{R}$ we obtain an $(n, R_K)$ code $f_K$ for some $n \in \mathbb{N}$ such that (40) holds. Define $U_k := f_k(X_k)$ and the auxiliary random variables $U_{k,l} := (U_k, X_{k,l}^{l-1})$ for $k \in K$ and $l \in [1:n]$. For any two sets $\mathcal{A}, \mathcal{C} \subseteq K$ we have

$$n \sum_{k \in \mathcal{A}} R_k \geq H(U_{\mathcal{A}})$$

(206)

$$= I(U_{\mathcal{A}}; X_K)$$

(207)

$$\geq I(U_{\mathcal{A}}; X_{K\mid \mathcal{C}})$$

(208)

$$= \sum_{l=1}^{n} I(U_{\mathcal{A}}; X_{K,l}|U_{\mathcal{C}}X_{K,l}^{l-1})$$

(209)

$$= \sum_{l=1}^{n} I(U_{\mathcal{A},l}; X_{K,l}|U_{\mathcal{C},l})$$

(210)

Furthermore, for any pair $(\mathcal{A}, \mathcal{B}) \in \Omega$

$$n\mu_{\mathcal{A},\mathcal{B}} \leq I(U_{\mathcal{A}}; U_{\mathcal{B}})$$

(211)

$$= I(U_{\mathcal{A}}; X_{\mathcal{A}}) - I(U_{\mathcal{A}}; X_{\mathcal{A}}|U_{\mathcal{B}})$$

(212)

$$= I(U_{\mathcal{A}}; X_{\mathcal{A}}) + I(U_{\mathcal{B}}; X_{\mathcal{B}}) - I(U_{\mathcal{A}}; X_{\mathcal{A}}|U_{\mathcal{B}}) - I(U_{\mathcal{B}}; X_{\mathcal{B}})$$

(213)

$$= I(U_{\mathcal{A}}; X_{\mathcal{A}}) + I(U_{\mathcal{B}}; X_{\mathcal{B}}) - I(U_{\mathcal{A}}; X_{\mathcal{A}}X_{\mathcal{B}}|U_{\mathcal{B}}) - I(U_{\mathcal{B}}; X_{\mathcal{A}}X_{\mathcal{B}})$$

(214)

$$= I(U_{\mathcal{A}}; X_{\mathcal{A}}) + I(U_{\mathcal{B}}; X_{\mathcal{B}}) - I(U_{\mathcal{A}}U_{\mathcal{B}}; X_{\mathcal{A}}X_{\mathcal{B}})$$

(215)
\begin{align*}
&= \sum_{l=1}^{n} \left[ I(U_{A,l}; X_{A,l}) + I(U_{B,l}; X_{B,l}) - I(U_{A,l}U_{B,l}; X_{A,l}X_{B,l}) \right].
\end{align*}

Now a standard time-sharing argument shows \( \mathcal{R} \subseteq \mathcal{R}_o \). To see \( \mathcal{R}_o \subseteq \mathcal{R}'_o \) note that

\begin{align*}
I(U_{A}; X_{A}) + I(U_{B}; X_{B}) &- I(U_{A}U_{B}; X_{A}X_{B}) \\
&= I(U_{A}; X_{B}) - I(U_{B}; X_{A}X_{B}|U_{A}) \\
&= I(U_{A}; X_{B}) + I(U_{B}; X_{B}) - I(U_{A}; X_{B}) - I(U_{B}; X_{A}X_{B}|U_{A}) \\
&= I(U_{A}; X_{B}) + I(U_{B}; X_{B}) - I(U_{A}U_{B}; X_{B}) - I(U_{B}; X_{A}X_{B}U_{A}) \\
&= I(U_{A}; X_{B}) - I(U_{A}; X_{B}|U_{B}) - I(U_{B}; X_{A}X_{B}U_{A}) \\
&\leq I(U_{A}; X_{B}).
\end{align*}

B. Proofs from Section III

1) Proof of Proposition 21

Pick an arbitrary \( k \in K \). For nonempty \( A, B \subseteq K \) with \( k \in B \) we can write

\( H(X_A|U_B) = E_{U_k} \left[ f_{A,B}(p_{X_k}|u_k(\cdot|u_k)) \right] \)

where

\( f_{A,B}(p_{X_k}|u_k(\cdot|u_k)) := H(X_A|U_B|k, U_k = u_k) \).

Furthermore, \( H(U_A|U_B) = E_{U_{k'}} g_{A,B}(p_{X_k}|u_k(\cdot|u_k)) \)

where

\( g_{A,B}(p_{X_k}|u_k(\cdot|u_k)) := H(U_A|U_B|k, U_k = u_k) \).

Observe that both \( f_{A,B} \) and \( g_{A,B} \) are continuous functions of \( p_{X_k}|u_k(\cdot|u_k) \). Apply the support lemma [10] Appendix C] with the functions \( f_{A,B} \) and \( g_{A,B} \) for all nonempty \( A, B \subseteq K \) such that \( k \in B \), and \( |X_k| - 1 \) test functions, which guarantee that the marginal distribution \( p_{X_k} \) does not change. We obtain a new random variable \( U_{k'} \) with

\( H(X_A|U_B|k, U_{k'}) = H(X_A|U_B|k) \) and \( H(U_A|U_B|k, U_{k'}) = H(U_A|U_B|k) \). By rewriting (42) and (44) in terms of conditional entropies, it is evident that the defining inequalities for \( \mathcal{R}_i \) remain the same when replacing \( U_k \) by \( U_{k'} \). \( U_{k'} \) satisfies the required cardinality bound

\( |U_{k'}| \leq |X_k| - 1 + 2(2^K - 1)2^{K-1} \\
= |X_k| - 1 + 2^{2^K} - 2^K \\
\leq |X_k| + 4^K. \)

The same process is repeated for every \( k \in K \).

2) Proof of Lemma 26

We can interpret a reproduction \( p_{Y_B|U_A}(\cdot|u_A) := g_{A,B}(u_A) \) as the conditional probability distribution of \( Y_B \) given \( U_A = u_A \). Note that this is in general not a product distribution. We calculate

\( E[d_{LL}(Y_B, g_{A,B}(U_A))|U_A = u_A] = -\frac{1}{n} \sum_{y_b \in Y_B} p_{Y_B|U_A}(y_B|u_A) \log p_{Y_B|U_A}(y_B|u_A) \)

\( = \frac{1}{n} \sum_{y_b \in Y_B} p_{Y_B|U_A}(y_B|u_A) \log \frac{p_{Y_B|U_A}(y_B|u_A)}{p_{Y_B|U_A}(y_B|u_A)} + \frac{1}{n} H(Y_B|U_A = u_A) \)

\( = \frac{1}{n} D_{KL}(p_{Y_B|U_A}(\cdot|u_A)||p_{Y_B|U_A}(\cdot|u_A)) + \frac{1}{n} H(Y_B|U_A = u_A) \)

\( \geq \frac{1}{n} H(Y_B|U_A = u_A) \),

where in the last step we used the non-negativity of Kullback-Leibler divergence. Equality in (232) is obtained with \( g_{A,B}(u_A) = p_{Y_B|U_A}(\cdot|u_A) \). By averaging over \( u_A \) we obtain the desired result.

3) Proof of Lemma 29

Fix \( 0 < \varepsilon', \varepsilon'' < \varepsilon \) and \( \bar{R}_K \subseteq \mathbb{R}^K_+ \) as \( \bar{R}_k = I(X_k; U_k) + \varepsilon''/2 \) for each \( k \in K \).

\(^2\)There are \((2^K-1)\) ways to choose \( A \) and \( 2^{K-1} \) ways to choose \( B \).
• **Encoding:** For \( n \in \mathbb{N} \) define \( \tilde{M}_k := e^nR_k \) and \( \tilde{M}_k := [1: \tilde{M}_k] \). We apply Lemma 52 and consider the random codebooks \( C_k := (\mathbf{V}^{(k)})_{i \in \tilde{M}_k} \), which are drawn independently uniform from \( T^n_{[U]} \) for each \( k \in K \). Denote the resulting randomized coding functions as \( \tilde{W}_k = \tilde{f}_k(X_k, C_k) \) and the corresponding decoded value as \( \tilde{U}_k := \mathbf{V}^{(k)}_{\tilde{W}_k} \). If \( n \) is chosen large enough we have therefore
\[
P_e := P\left\{ (\tilde{U}_K, X_K) \notin T^n_{[U]} \right\} \leq \varepsilon'.
\] (233)

Next, we introduce (deterministic) binning. If \( R_k < I(X_k; U_k) \), partition \( \tilde{M}_k \) into \( M_k := e^n(R_k + \varepsilon'') \) equally sized, consecutive bins, each of size \( e^n\Delta_k \) with
\[
\Delta_k := \tilde{R}_k - R_k - \varepsilon'' = I(X_k; U_k) - R_k - \varepsilon''/2.
\] (234)

The deterministic function \( \beta_k : \tilde{M}_k \rightarrow M_k \) maps a codeword to the index of the bin in \( M_k := [1:M_k] \) to which it belongs. Now use the randomized encoding function \( f_k := \beta_k \circ \tilde{f}_k \). If \( R_k \geq I(X_k; U_k) \), we do not require binning and let \( \beta_k \) be the identity on \( \tilde{M}_k \) and hence \( f_k := \tilde{f}_k \).

• **Decoding:** Given the codebooks, the decoding procedure \( g_{A,A'} : M_{\tilde{A}} \rightarrow U^n_{\tilde{A}} \) for each \( \emptyset \neq \tilde{A} \subseteq A \subseteq K \) is carried out as follows: Given \( m_A \in M_{A} \), let \( \tilde{m}_A := \beta^{-1}_A(m_A) \subseteq \tilde{M}_A \) be all indices that are in the bins \( m_A \). Consider only the typical sequences \( \mathbf{V}^{(A)} \cap T^n_{[U]} := \Phi \subseteq U^n_{A} \). Denote the restriction of \( \Phi \) to the coordinates \( \tilde{A} \) as \( [\Phi]_{\tilde{A}} \). If \( \Phi \neq \emptyset \), choose the lexicographically smallest element of \( [\Phi]_{\tilde{A}} \) otherwise choose the lexicographically smallest element of \( \mathbf{V}^{(A)} \).

Let \( A, \tilde{A}, B, \tilde{B} \subseteq K \) be sets of indices, such that the conditions of part \( [\Box] \) are satisfied. Using \( W_k := f_k(X_k, C_k) \) and the randomized codings \( \tilde{U}_1 := g_{A,A}(W_A, C_A) \) and \( \tilde{U}_2 := g_{B,B}(W_B, C_B) \), consider the error event \( E_0 := \{(\tilde{U}_1, X_A, X_B, \tilde{U}_2) \notin T^n_{[U_A X_A X_B U_B]} \} \). Defining the other events
\[
E_1 := \{ (\tilde{U}_A, X_A, X_B, \tilde{U}_B) \notin T^n_{[U_A X_A X_B U_B]} \},
\] (235)
\[
E_2 := \left\{ \left| \left[ V^{(A)}_{\tilde{W}_A} \cap T^n_{[U_A]} \right]_{\tilde{A}} \right| > 1 \right\},
\] (236)
\[
E_3 := \left\{ \left| \left[ V^{(B)}_{\tilde{W}_B} \cap T^n_{[U_B]} \right]_{\tilde{B}} \right| > 1 \right\},
\] (237)

where we used the random set of indices \( \tilde{W}_A := \beta^{-1}_A(W_A) \). We clearly have \( E_0 \subseteq E_1 \cup E_2 \cup E_3 \) and thus
\[
P\{E_0\} \leq P\{E_1\} + P\{E_2|E_1^c\} + P\{E_3|E_1^c\} \leq P\{E_2|E_1^c\} + P\{E_3|E_1^c\} + \varepsilon'.
\] (238)

(239)

We can partition \( \tilde{W}_A \) into (random) subsets \( D_{A'} \), indexed by \( A' \subseteq A \) as
\[
D_{A'} := \{ \tilde{w}_A \in \tilde{W}_A : \tilde{w}_A^c = \tilde{W}_A^c \land \tilde{w}_k \neq \tilde{W}_k, \forall k \in A' \},
\] (240)

where we used \( A^c := A \setminus A' \). Observe that \( D_{\emptyset} = \{ \tilde{W}_A \} \). For each set \( \emptyset \neq A' \subseteq A \) we define the error event
\[
E_{A'} := \{ V^{(A')}_{D_{A'}} \cap T^n_{[U_A]} \neq \emptyset \}
\] (241)

and we have
\[
E_2 \subseteq \bigcup_{A' \subseteq A : A' \cap \tilde{A} \neq \emptyset} E_{A'}
\] (242)

which implies
\[
P\{E_2|E_1^c\} \leq \sum_{A' \subseteq A : A' \cap \tilde{A} \neq \emptyset} P\{E_{A'}|E_1^c\}.
\] (243)
By the construction of the codebook, $D_{A'}$ has $\prod_{k \in A'} (e^{n \Delta_k} - 1)$ elements. For $\tilde{w}_A \in D_{A'}$, we have that $V_{\tilde{w}_A}^{(A')}$ are uniformly distributed on $\prod_{k \in A'} T_{[u_k]}$ and $\tilde{w}_{A''} = \tilde{W}_{A''}$. Given $E_1^c$ we have in particular $U_A \in T_{[u_A]}$. Thus, for any $u_{A''} \in T_{[u_{A''}]}$, we can conclude,

\[
P\left\{ E_A | E_1^c, \tilde{U}_{A''} = u_{A''} \right\} = P\left\{ \bigcup_{\tilde{w}_A \in D_{A'}} \{ V_{\tilde{w}_A}^{(A)} \in T_{[u_A]} \} \big| E_1^c, \tilde{U}_{A''} = u_{A''} \right\}
\]
\[
\leq \sum_{\tilde{w}_A \in D_{A'}} P\left\{ V_{\tilde{w}_A}^{(A)} \in T_{[u_A]} \big| E_1^c, \tilde{U}_{A''} = u_{A''} \right\}
\]
\[
\leq \exp\left( n \left( \sum_{k \in A'} \Delta_k \right) \right) \frac{|T_{[u_A]}|}{\prod_{k \in A'} |T_{[u_k]}|}
\]
\[
\leq \exp\left( n \left( \sum_{k \in A'} \Delta_k \right) \right) \frac{\exp\left( n(H(U_A|U_{A''}) + \varepsilon_0(n))\right)}{\exp\left( n(\sum_{k \in A'} H(u_k) - \varepsilon_k(n))\right)}
\]
\[
\leq \exp\left( n \left( \varepsilon(n) + H(U_A|U_{A''}) + \sum_{k \in A'} (\Delta_k - H(U_k)) \right) \right),
\]

where $\varepsilon(n) = \sum_{k \in A' \cup \emptyset} \varepsilon_k(n)$ goes to zero as $n \to \infty$. Here, [42] follows from parts 2 and 3 of Lemma 51. We observe that the definition of $R_k$ and [42] imply for any $\emptyset \neq A' \subseteq A$ with $A' \cap \tilde{A} \neq \emptyset$ that

\[
\sum_{k \in A'} \Delta_k \leq -\frac{\varepsilon''}{2} - H(U_A|U_{A''}) + \sum_{k \in A'} H(u_k).
\]

Marginalize over $\tilde{U}_{A''}$ in (248) and use (249) to obtain

\[
P\left\{ E_A | E_1^c \right\} \leq \exp\left( n \left( \varepsilon(n) - \frac{\varepsilon''}{2} \right) \right) \leq \varepsilon'
\]

for $n$ large enough. Applying the same arguments to $P\left\{ E_0 | E_1^c \right\}$ and combining (239), (243) and (250) we have

\[
P\left\{ E_0 \right\} \leq \varepsilon' + 2|A|\varepsilon' + 2|B|\varepsilon' \leq 2^K \varepsilon'.
\]

For a set $\emptyset \neq A \subseteq K$, we next analyze the random quantity $L := \left| C_A \cap T_{[u_A]} \right|$. For $n$ large enough, we have

\[
E[L] \leq \sum_{\mathbf{V}_A \in C_A} E\left[ \mathbf{1}_{T_{[u_A]}^n} (\mathbf{V}_A) \right]
\]
\[
= \left( \prod_{k \in A} \tilde{M}_k \right) E\left[ \mathbf{1}_{T_{[u_A]}^n} (\mathbf{V}_A) \right] \quad \text{for any } \mathbf{V}_A \in C_A
\]
\[
= \left( \prod_{k \in A} \tilde{M}_k \right) \frac{|T_{[u_A]}^n|}{\prod_{k \in A} |T_{[u_k]}|^n}
\]
\[
\leq \left( \prod_{k \in A} \tilde{M}_k \right) \frac{e^{n(H(U_A) + \varepsilon_0(n))}}{e^{n(\sum_{k \in A} H(u_k) - \varepsilon_k(n))}}
\]
\[
\leq \left( \prod_{k \in A} \tilde{M}_k \right) e^{n(H(U_A) - \sum_{k \in A} H(u_k) + \varepsilon(n))}
\]
\[
= \exp\left( n \left( H(U_A) + \varepsilon(n) + \sum_{k \in A} I(U_k; X_k) + \frac{\varepsilon''}{2} - H(U_k) \right) \right)
\]
\[ P\{E_4\} \leq \exp\left( n\left( \varepsilon(n) - \varepsilon + \frac{|A| \varepsilon''}{2} \right) \right) \leq \varepsilon'. \]  

Using (251) and (260) we can apply Lemma 38 and obtain deterministic encoding functions \( f_k : \mathcal{X}_k^m \rightarrow \mathcal{M}_k \), and deterministic decoding functions \( g_{\mathcal{A},\mathcal{A}} : \mathcal{M}_\mathcal{A} \rightarrow \mathcal{U}_\mathcal{A}^m \), such that (55) holds whenever the conditions of part 1 are satisfied. Taking into account that \( g_{\mathcal{A},\mathcal{A}} : \mathcal{M}_\mathcal{A} \rightarrow \mathcal{U}_\mathcal{A}^m \subseteq \mathcal{C}_{\mathcal{A},\mathcal{B}} \), we also have (56). (Note that, given a specific code, \( P\{E_4\} < 1 \) already implies \( P\{E_4\} = 0 \) as the event \( E_4 \) is fully determined by the code \( C_K \) alone.)

**4) Proof of Theorem 35**

Pick a linear ordering \( \preceq \) of \( \mathcal{J} \), a set \( \mathcal{I} \subseteq \mathcal{J} \) and \( (U_j, \varnothing) \in \mathcal{P}_s \). To obtain a code we apply Lemma 29 with \( K = J + 1 \), \( X_K = U_K = Y \), \( B = \mathcal{B} = \{K\} \), \( \mathcal{A} = \mathcal{A} \) for all \( \varnothing \neq \mathcal{A} \subseteq \mathcal{J} \) and rates \( R_j = I(U_j; X_j | U_{\preceq j}, \varnothing) \), \( R_K = \log |\mathcal{Y}| \), as suggested by Proposition 23. As in the proof of Lemma 29 let \( f_j \) denote the encoding function without binning and with rate \( n^{-1} \log |f_j| \leq I(U_j; X_j) + \frac{\varepsilon}{2} \). Furthermore, let \( f_j' \) be the encoding function including binning, obtaining a rate of \( n^{-1} \log |f_j'| \leq I(U_j; X_j | U_{\preceq j}) + \varepsilon \). Finally we obtain the final \( (n, R_J + \varepsilon) \) code \( f_J \) by setting \( f_j := \tilde{f}_j \) for \( j \in \mathcal{I} \) and \( f_j := f_j' \) for \( j \notin \mathcal{I} \). Let the decoding functions be \( g_{\mathcal{A},\mathcal{A}} := g_{\mathcal{A},\mathcal{A}} \) for all \( \varnothing \neq \mathcal{A} \subseteq \mathcal{J} \). Furthermore, for each \( j \in \mathcal{J} \), we define the decoding function \( \tilde{g}_j \) which maps \( W_j := f_j(X_j) \) onto its codebook entry. Also let \( W_j := f_j(X_j) \) and \( W_j' := f_j'(X_j) \). Note that \( W_j' \) is a function of \( W_j \), which is in turn a function of \( \tilde{W}_j \).

Let the event \( S'_A \) be the success event that joint typicality \( (Y, X_{\mathcal{A}}, g_A(W_A')) \in \mathcal{T}_{[YX_{\mathcal{A}}U_{\mathcal{A}}]}^n \) holds. Also let \( \tilde{S}_j \) be the event that \( (Y, X_j, \tilde{g}_j(\tilde{W}_j)) \in \mathcal{T}_{[YX_{\mathcal{J}}u_{\mathcal{J}}]}^n \). For any \( \mathcal{A} = \bigcup_{j_0} A_j \), \( j_0 \in \mathcal{J} \), we have for \( \mathcal{A}' \subseteq \mathcal{A} \),

\[ \sum_{j \in \mathcal{A}'} R_j = \sum_{j \in \mathcal{A}'} I(U_j; X_j | U_{\preceq j}) \geq \sum_{j \in \mathcal{A}'} I(U_j; X_{\mathcal{A}'} | U_{\preceq j}, U_{\mathcal{A}' \setminus \mathcal{A}'} \cap A') = I(U_{\mathcal{A}' \setminus \mathcal{A}'}; X_{\mathcal{A}' \setminus \mathcal{A}' \setminus \mathcal{A}' \setminus \mathcal{A}' \setminus \mathcal{A}' \setminus \mathcal{A}' \setminus \mathcal{A}' \setminus \mathcal{A}') \]  

Thus, condition (42) is satisfied and for large enough \( n \) we have \( P\{S'_A\} \geq 1 - \varepsilon \) by Lemma 29. Clearly also \( P\{\tilde{S}_j\} \geq 1 - \varepsilon \) for each \( j \in \mathcal{J} \), using Lemma 49 and Lemma 50.

Let us pick an arbitrary \( \varepsilon' > 0 \). For any \( \mathcal{A} = \bigcup_{j_0} A_j \), provided that \( n \) is large enough and \( \varepsilon \) small enough, we have that

\[ \frac{1}{n} I(Y; W_A) \geq \frac{1}{n} I(Y; W_A') \geq \frac{1}{n} I(Y; g_A(W_A')) \geq H(Y) - \frac{1}{n} I(Y; g_A(W_A')) \geq H(Y) - \frac{1}{n} I(S_A') \geq H(Y) - \frac{1}{n} I(S_A') | g_A(W_A') \]  

\[ \geq H(Y) - \varepsilon' - \frac{1}{n} (1 - \varepsilon) H(Y | g_A(W_A'), S_A') \geq \varepsilon H(Y) \]

\[ \mu = \exp \left( n \left( \frac{H(U_A) + \varepsilon(n) + |A| \varepsilon''}{2} - \sum_{k \in \mathcal{A}} H(U_k | X_k) \right) \right) \]  

\[ = \exp \left( n \left( I(U_A; X_A) + \varepsilon(n) + |A| \varepsilon'' \right) \right) \]  

where \( \varepsilon(n) = \sum_{k \in \mathcal{A}} \varepsilon_k(n) \) goes to zero as \( n \to \infty \). Here, (255) follows from parts 1 and 2 of Lemma 51. Assume that \( \varepsilon'' \) is such that \( K \varepsilon'' / 2 < \varepsilon \). Defining the error event \( E_4 = \{ L \geq \exp \left( n \left( |I(U_A; X_A)| + \varepsilon \right) \right) \}, \) we know from Markov’s inequality that for large enough \( n \)

\[ P\{E_4\} \leq \exp \left( n \left( \varepsilon(n) - \varepsilon + |A| \varepsilon'' \right) \right) \]

In what follows, we will routinely merge expressions that can be made arbitrarily small (for \( n \) large and \( \varepsilon \) sufficiently small) and bound them by \( \varepsilon' \).
\[ \frac{1}{n} I(A_j; W_j) \geq I(A_j; Y_j) - \epsilon' \geq \nu_j - \epsilon'. \] (274)

Also, for \( j_0 \in J \) and \( A = \exists j_0 \),
\[ \frac{1}{n} I(Y; W_{j_0}) \geq \frac{1}{n} I(Y; W_{j_0}) - \frac{1}{n} I(Y; W_A) \]
\[ \geq I(U_A U_{j_0}; Y) - \epsilon' - \frac{1}{n} I(Y; W_A) \]
\[ \geq I(X_{j_0}; W_{j_0}) - \epsilon' - \frac{1}{n} I(X_A; W_A) = I(U_{j_0}; Y_{j_0}) \]
\[ \geq I(U_{j_0}; Y_{j_0}) - \epsilon' - I(X_A; U_A) + H(X_A|Y) - \frac{1}{n} H(X_A|W_A, Y) \]
\[ \geq I(U_{j_0}; Y_{j_0}) - \epsilon' - I(X_A; U_A) + H(X_A|Y) - H(X_A|U_A, Y) \]
\[ \geq I(U_{j_0}; Y|U_A) - \epsilon' \]
\[ \geq \nu_{j_0} - \epsilon'. \] (276)

Here, \( [278] \) follows from the Markov chain \( W_A \rightarrow X_A \rightarrow Y \). We need to justify \( [279] \) and \( [280] \). In \( [279] \) we used that for \( \epsilon \) small enough and \( n \) large enough,
\[ \frac{1}{n} I(X_A; W_A') = \frac{1}{n} H(W_A') \]
\[ \leq \frac{1}{n} \sum_{j \in A} H(W_j) \]
\[ \leq \sum_{j \in A} \left( I(U_j; X_j|U_{-j}) + \epsilon \right) \]
\[ \leq I(U_A; X_A) + \epsilon', \] (286)

where \( [284] \) follows from the chain rule for entropy \( [17] \) Theorem 2.2.1 and the data processing inequality and \( [285] \) follows from the entropy bound \( [17] \) Theorem 2.6.4 and the fact that \( n^{-1} \log |f_j| \leq I(U_j; X_j|U_{-j}) + \epsilon \). The inequality \( [280] \) follows similar to \( [275] \) as for \( n \) large enough and \( \epsilon \) small enough,
\[ \frac{1}{n} H(X_A|W_A', Y) \leq \frac{1}{n} H(X_A|g(A(W_A'), Y) \]
\[ \leq \frac{1}{n} H(X_A, |S'| A(W_A'), Y) \]
\[ \leq \epsilon' + \frac{1}{n} H(X_A|g(A(W_A'), Y, S')) \]
\[ \leq \epsilon' + \frac{1}{n} \sum_{u_A, y} P\{g(A(W_A') = u_A, Y = y'|S') \log |T_{[X_A|U_A,Y]}(u_A, y)| \]
\[ \leq \epsilon' + H(X_A|U_A, Y). \] (291)
For \( j_0 \in \mathcal{I} \), we have similar to (273) that
\[
\frac{1}{n} I(Y; W_{j_0}) = \frac{1}{n} I(Y; \tilde{W}_{j_0})
\]
\[
\geq \frac{1}{n} I(Y; g_{j_0}(\tilde{W}_{j_0}))
\]
\[
= H(Y) - \frac{1}{n} H(Y | g_{j_0}(\tilde{W}_{j_0}))
\]
\[
\geq H(Y) - \frac{1}{n} H(Y, 1_{\tilde{S}_{j_0}} | g_{j_0}(\tilde{W}_{j_0}))
\]
\[
\geq H(Y) - \frac{1}{n} H(Y | g_{j_0}(\tilde{W}_{j_0}), 1_{\tilde{S}_{j_0}})
\]
\[
\geq H(Y) - \frac{1}{n} H(Y | g_{j_0}(\tilde{W}_{j_0})) - \frac{1}{n} H(Y | g_{j_0}(\tilde{W}_{j_0}), 1_{\tilde{S}_{j_0}})
\]
\[
\geq H(Y) - \frac{1}{n} H(Y | g_{j_0}(\tilde{W}_{j_0}))
\]
\[
\leq H(Y) - \varepsilon' - \frac{1}{n} H(Y | U_{j_0})
\]
\[
= I(U_{j_0}; Y) - \varepsilon' \geq \nu_{j_0} - \varepsilon'.
\]

5) **Proof of Theorem 3.6** With \( U_{j,i} := (U_j, X_i^{i-1}) \), \( Q_i := (Y_i^{i-1}, Y_{i+1}) \).

\[
\sum_{j \in B} R_j \geq H(U_B)
\]
\[
= I(U_B; X_B)
\]
\[
= I(U_B; X_B | Y)
\]
\[
= I(U_B; Y) + I(U_B; X_B | Y)
\]
\[
= I(U_B; Y) - I(U_B; Y | U_B) + I(U_B; X_B | Y)
\]
\[
= I(U_B; Y) - I(U_B; Y | U_B) - I(U_B; Y | U_B) + I(U_B; X_B | Y)
\]
\[
\geq n\nu_A + I(U_B; Y | U_B) - I(U_B; Y | U_B) + I(U_B; X_B | Y)
\]
\[
\geq n\nu_A - I(U_B; X_B | Y)
\]
\[
= \sum_{i=1}^{n} [\nu_A - I(U_{A\setminus B}; Y_i | Y^{i-1}) + I(U_B; X_{B,i} | YX^{i-1})]
\]
\[
\geq \sum_{i=1}^{n} [\nu_A - I(U_{A\setminus B}; Y_i | Q_i) + I(U_B; X_{B,i} | YX^{i-1})]
\]
\[
= \sum_{i=1}^{n} [\nu_A - I(U_{A\setminus B}; Y_i | Q_i) + I(U_B; X_{B,i} | YQ_i)]
\]

The result follows by a standard time-sharing argument. Note that the required Markov chain and the independence are satisfied.

6) **Proof of Proposition 3.4** Pick arbitrary \( j, j_0 \in \mathcal{J} \). For nonempty \( B \subset \mathcal{J} \) with \( j \in B \) we can write
\[
H(X_{j_0} | U_B) = \mathbb{E}_{U_j} \left[ f_{j_0,B}(p_{X,j_0| U_j} \cdot | U_j) \right]
\]
and
\[
H(Y | U_B) = \mathbb{E}_{U_j} \left[ g_B(p_{X,j_0| U_j} \cdot | U_j) \right]
\]
where
\[
f_{j_0,B}(p_{X,i| U_k} \cdot | u_k) := H(X_{j_0} | U_B | k, U_k = u_k),
\]
\[
g_B(p_{X,i| U_k} \cdot | u_k) := H(Y | U_B | k, U_k = u_k).
\]

Observe that both \( f_{j_0,B} \) and \( g_B \) are continuous functions of \( p_{X,i| U_k} \cdot | u_k \). Apply the support lemma [10] Appendix C with the functions \( f_{j_0,B} \) and \( g_B \) for all \( j_0 \in \mathcal{J}, j \in B \subset \mathcal{J} \), and \( |X_j| - 1 \) test functions, which guarantee that the marginal distribution \( p_{X,i} \) does not change. We obtain a new random variable \( U'_j \) with \( H(X_{j_0} | U_B \setminus U'_j) = H(X_{j_0} | U_B) \) and \( H(Y | U_B \setminus U'_j) = H(Y | U_B) \). By rewriting (85)–(90) in terms of conditional entropies, it is evident that the
defining inequalities for $\mathcal{R}_{\text{MI}}^{(C)}$ remain the same when replacing $U_k$ by $U'_k$. $U'_k$ satisfies the required cardinality bound.

\[ |U'_j| \leq |X_j| - 1 + J2^J - 1 + 2^{J-1} \leq |X_j| + 4^J. \]  

The same process is repeated for every $j \in J$.

7) A Random Coding Lemma:

**Lemma 38.** Let $X$ be discrete random variables and for any $\delta > 0$ let $F_\delta$ be a random code (random vector-valued function) operating on $X^n$ ($n$ sufficiently large as a function of $\delta$). For finitely many error events $(E_i)_{i \in I}$ we have

\[ P\{E_i\} \leq \delta, \quad i \in I. \]  

Then, for any $\varepsilon > 0$ we can find $\delta > 0$, a sufficiently large $n \in \mathbb{N}$, and a code $f$, such that

\[ P\{E_i|F_\delta = f\} \leq \varepsilon, \quad i \in I. \]  

**Proof.** We can apply Markov’s inequality to the random variable $P\{E_i|F_\delta\}$ and obtain

\[ P \left\{ P\{E_i|F_\delta\} \geq \sqrt{\delta} \right\} \leq \frac{\delta}{\sqrt{\delta}} = \sqrt{\delta}. \]  

Applying the union bound yields

\[ P \left\{ \bigcup_{i \in I} \left\{ P\{E_i|F_\delta\} \geq \sqrt{\delta} \right\} \right\} \leq \sum_{i \in I} P\{P\{E_i|F_\delta\} \geq \sqrt{\delta}\} \leq |I|\sqrt{\delta}. \]

Thus, with a probability no smaller than $1 - |I|\sqrt{\delta}$ the random coding yields a specific code $f$ such that $P\{E_i|F_\delta\} < \sqrt{\delta}$ for all $i \in I$. Choosing $\delta = \min(\varepsilon^2, \sqrt{|I|}^{-2}/2)$ yields the desired result.

8) Results on Convex Polyhedra: Let $S$ be the convex polyhedron $S := \{x \in \mathbb{R}^n : Ax \geq b\}$ for an $m \times n$ matrix $A = (a_{(1)}, a_{(2)}, \ldots, a_{(m)})^T$ and $b \in \mathbb{R}^m$, where $a_{(j)}$ is the $j$th row of $A$. In this section we will use the notation of [41]. In particular, we shall call a closed convex set line-free if it does not contain a (straight) line. The characteristic cone of a closed convex set $C$ is defined as $ccC := \{y : x + \lambda y \in C \text{ for all } \lambda \geq 0\}$ ($x \in C$ arbitrary) and $\text{ext}C$ is the set of all extreme points of $C$, i.e., points $x \in C$ that cannot be written as $x = \lambda y + (1 - \lambda)z$ with $y, z \in C, y \neq z$ and $\lambda \in (0, 1)$.

**Lemma 39.** A point $y$ is in $ccS$ if and only if $Ay \geq 0$.

**Proof.** If $Ay \geq 0$, then $x \in S$ and $\lambda \geq 0$, $A(x + \lambda y) \geq Ax \geq b$. On the other hand, if we assume $a_{(j)}^Ty < 0$ without loss of generality, we have $a_{(j)}^T(x + \lambda y) < b_1$ for $\lambda > \frac{b_1 - a_{(j)}^Tx}{a_{(j)}^Ty} > 0$.

**Lemma 40.** If, for every $i \in [1:n]$, there exists $j \in [1:m]$ such that $e_i = a_{(j)}$ and for every $j \in [1:m]$, $a_{(j)} \geq 0$, then $S$ is line-free and $ccS = O = \{x : x \geq 0\}$.

**Proof.** For any $y \in O$, clearly $Ay \geq 0$ and hence $y \in ccS$ by Lemma [39]. If $y \notin O$ and without loss of generality $y_1 < 0$, then choose $j \in [1:m]$ such that $a_{(j)} = e_1$ and we have $a_{(j)}^Ty = y_1 < 0$. To show that $S$ is line-free assume for $y \neq 0$ that $x + \lambda y \in S$ for all $\lambda \in \mathbb{R}$. Without loss of generality we assume $y_1 \neq 0$ and obtain $x_1 + \lambda y_1 \geq b_j$, which cannot hold for every $\lambda \in \mathbb{R}$.

**Definition 41.** A point $x$ is on an extreme ray of the cone $ccS$ if $x = y + z$ with $y, z \in ccS$ implies that $y = \lambda z$ for some $\lambda \in \mathbb{R}$.

It is easy to see that the points on extreme rays of $O$ are given by $x = \lambda e_i$ for $\lambda \geq 0$ and $i \in [1:n]$.

\[ \text{There are } J \text{ ways to choose } j_0 \text{ and } 2^{J-1} \text{ ways to choose } B. \]
Define $A(x) := \{ j \in [1 : m] : a_j^T x = b_j \}$. We say that exactly $n_0$ linearly independent inequalities are satisfied with equality at $x$, if $Ax \geq b$ and $(a_{(j)})_{j \in A(x)}$ has rank $n_0$.

**Lemma 42.** $x \in \text{ext } S$ if and only if exactly $n$ linearly independent inequalities are satisfied with equality at $x$.

**Proof.** Assuming that less than $n$ linearly independent inequalities are satisfied with equality at $x$, we find $0 \neq c \perp (a_{(j)})_{j \in A(x)}$ and thus $x \pm \varepsilon c \in S$ for a small $\varepsilon > 0$, showing that $x \not\in \text{ext } S$.

Conversely assume $x \not\in \text{ext } S$, i.e., $x = \lambda x' + (1 - \lambda)x''$ for $\lambda \in (0, 1)$ and $x', x'' \in S$, $x' \neq x''$. For any $j \in A(x)$, we then have $\lambda a_{(j)}^T x' + (1 - \lambda)a_{(j)}^T x'' = b_j$, which implies $a_{(j)}^T x' = a_{(j)}^T x'' = b_j$ and therefore $0 \neq x' - x'' \perp (a_{(j)})_{j \in A(x)}$.

**Lemma 43.** Assuming that $S$ is line-free, and that exactly $n - 1$ linearly independent inequalities are satisfied with equality at $x$. Then either $x = \lambda c + (1 - \lambda)d$ where $\lambda \in (0, 1)$ and $c, d \in \text{ext } S$ or $x = c + d$ where $c \in \text{ext } S$ and $d \neq 0$ lies on an extreme ray of $ccS$.

**Proof.** We obtain $0 \neq r \perp (a_{(j)})_{j \in A(x)}$. Define $\lambda_1 := \inf \{ \lambda : x + \lambda r \in S \}$ and $\lambda_2 := \sup \{ \lambda : x + \lambda r \in S \}$. Clearly $0 \leq \lambda_1 \leq \lambda_2$. As $S$ is line-free, we may assume without loss of generality $\lambda_1 = -1$ (note that $x \not\in \text{ext } S$) and set $c = x - r$. We now have $c \in \text{ext } S$ as otherwise $c - \varepsilon r \in S$ for some small $\varepsilon > 0$.

If $\lambda_2 < \infty$, define $d = x + \lambda_2 r$ which yields $d \in \text{ext } S$ and $x = \lambda c + (1 - \lambda)d$ with $\lambda = \frac{\lambda_2}{\lambda_2 + 1}$. Note that $\lambda_2 \neq 0$ as $x \not\in \text{ext } S$.

If $\lambda_2 = \infty$ we have $x - c = r \in ccS$. We need to show that $r$ is also on an extreme ray of $ccS$. Assuming $r = r' + r''$ with $r', r'' \in ccS$ yields $a_{(j)}^T (r' + r'') = 0$, which implies $a_{(j)}^T r' = a_{(j)}^T r'' = 0$ for every $j \in A(x)$ by Lemma 39.

For each $j \in [0 : J]$, define the closed convex polyhedron $S^{(j)} := \{ x \in \mathbb{R}^{K + j} : A^{(j)} x \geq b^{(j)} \}$, where $A^{(j)}$ is a matrix and $b^{(j)}$ a vector of appropriate dimension. We make the following three assumptions:

1) $A^{(j)}$ and $b^{(j)}$ are defined recursively as

$$A^{(j)} := \begin{pmatrix} A^{(j-1)} & 0 \\ 0^T & 1 \\ e_j^T & 1 \end{pmatrix}, \quad b^{(j)} = \begin{pmatrix} b^{(j-1)} \\ c_1^{(j)} \\ c_2^{(j)} \end{pmatrix},$$

where $e_j$ is the $j$th unit vector of appropriate dimension and $c_1^{(j)}$ and $c_2^{(j)}$ are arbitrary reals.

2) Each entry of $A^{(0)}$ equals 0 or 1 and for all $k \in K$ at least one row of $A^{(0)}$ is equal to $e_k^T$. Due to assumption 1 this also implies that each entry of $A^{(j)}$ is in $\{0, 1\}$ and for all $k \in [1 : K + j]$ at least one row of $A^{(j)}$ is equal to $e_k^T$.

3) For any extreme point $x \in \text{ext } S^{(0)}$ and any $j \in J$, assume $x_j \leq c_2^{(j)} - c_1^{(j)}$.

**Lemma 44.** Under assumptions 1 to 3 for every $j_0 \in [0 : J]$ and every extreme point $y \in \text{ext } S^{(j_0)}$ there is an extreme point $x \in \text{ext } S^{(0)}$ and a subset $\mathcal{I}_{j_0} \subseteq [1 : j_0]$ such that $y_K = x_K$ and for every $j \in J$,

$$y_j = \begin{cases} x_j, & j \notin \mathcal{I}_{j_0}, \\ c_2^{(j)} - c_1^{(j)}, & j \in \mathcal{I}_{j_0}, \end{cases}$$

and for every $j \in [1 : j_0]$,

$$y_{K+j} = \begin{cases} c_2^{(j)} - x_j, & j \notin \mathcal{I}_{j_0}, \\ c_1^{(j)}, & j \in \mathcal{I}_{j_0}. \end{cases}$$

**Proof.** For every $j \in J$, $S^{(j)}$ is line-free by assumption 2 and Lemma 40 and can be written [41] Lemma 6, p. 25 as $S^{(j)} = ccS^{(j)} + \text{conv}(\text{ext } S^{(j)})$. Lemma 40 also implies $ccS^{(j)} = O$. Let us proceed inductively over $j_0 \in [0 : J]$. For $j_0 = 0$ the statement is trivial.

Given any $y \in \text{ext } S^{(j_0)}$, we need to obtain $x \in \text{ext } S^{(0)}$ and $\mathcal{I}_{j_0}$ such that $y$ is given according to (322) and (323). Let $z = y_{1}^{K+j_0-1}$ be the truncation of $y$. Exactly $K + j_0$ linear independent inequalities of $A^{(j_0)} y \geq b^{(j_0)}$ are satisfied with equality by Lemma 42, which is possible in only two different ways:
• **Construction I:** Exactly $K + j_0 - 1$ linear independent inequalities of $A_{(j_0-1)} z \geq b_{(j_0-1)}$ are satisfied with equality, i.e., $z \in \text{ext} \mathcal{S}_{(j_0-1)}$ by Lemma 42 and at least one of

$$y_{K+j_0} \geq c_1^{(j_0)},$$

$$y_{j_0} + y_{K+j_0} \geq c_2^{(j_0)},$$

is satisfied with equality.

As $z \in \text{ext} \mathcal{S}_{(j_0-1)}$, there exists $x \in \text{ext} \mathcal{S}^{(0)}$ and $I_{j_0-1}$ such that (322) holds for $j \in J$ and (323) holds for $j \in [1 : j_0 - 1]$ by the induction hypothesis. In particular $y_{j_0} = x_{j_0}$. Assuming that (325) holds with equality, we have $y_{K+j_0} = c_2^{(j_0)} - x_{j_0}$. Thus, the point $x$ together with $I_{j_0} = I_{j_0-1}$ yields $y$ from (322) and (323). Equality in (324) implies equality in (325) by assumption 3.

• **Construction II:** Exactly $K + j_0 - 2$ linear independent inequalities of $A_{(j_0-1)} z \geq b_{(j_0-1)}$ are satisfied with equality and (324) and (325) are both satisfied with equality as well. Additionally, these $K + j_0$ inequalities together need to be linearly independent. This can occur in two different ways by Lemma 43.

Assume $z = \lambda x + (1 - \lambda)x'$ for $x, x' \in \text{ext} \mathcal{S}^{(j_0-1)}$, $x \neq x'$ and $\lambda \in (0, 1)$. This implies $y_{K+j_0} = c_1^{(j_0)}$ and $y_{j_0} = \lambda x_{j_0} + (1 - \lambda)x'_{j_0} = c_2^{(j_0)} - c_1^{(j_0)}$, which by assumption 3 already implies $x_{j_0} = x'_{j_0} = c_2^{(j_0)} - c_1^{(j_0)}$. Thus, (324) and (325) are satisfied (with equality) for every $\lambda \in [0, 1]$ and $y$ cannot be an extreme point as it can be written as a non-trivial convex combination.

We can thus focus on the second option which is that $z$ is on an extreme ray of $\mathcal{S}^{(j_0-1)}$, i.e., $z = x + e_{j'}$ for some $x \in \text{ext} \mathcal{S}^{(j_0-1)}$, $\lambda > 0$ and $j' \in [1 : K + j_0 - 1]$. If $j' \neq j_0$, (324) and (325) are satisfied for all $\lambda > 0$ and thus $y$ cannot be an extreme point because it can be written as a non-trivial convex combination. As $j' = j_0$, the point $x$ with $I_{j_0} = I_{j_0-1} \cup j_0$ yields the desired extreme point.

C. Types, Typical Sequences and Related Results

In this section we introduce standard notions and results, as needed for the mathematical developments and proofs in this work. The results can be easily derived from the standard formulations provided in [10] and [19].

**Definition 45** (Type; [[19] Definition 2.1]). The type of a vector $x \in \mathcal{X}$ is the random variable $\hat{X} \sim p_X \in \mathcal{P}(\mathcal{X})$ defined by

$$p_X(x) = \frac{1}{n} N(x|\mathcal{X}), \text{ for every } x \in \mathcal{X},$$

(326)

where $N(x|\mathcal{X})$ denotes the number of occurrences of $x$ in $\mathcal{X}$. For a random variable $\hat{X}$, the set of $n$-sequences with type $\hat{X}$ is denoted $\mathcal{T}^n_{\hat{X}}$.

For a pair of random variables $(X, Y)$, we say that $y \in \mathcal{Y}$ has conditional type $Y$ given $x \in \mathcal{X}$ if $(x, y) \in \mathcal{T}^n_{XY}$. The set of all $n$-sequences $y \in \mathcal{Y}$ with conditional type $Y$ given $x$ will be denoted $\mathcal{T}^n_{Y|X}(x)$.

A key property of types is the following result, known as type counting.

**Lemma 46** (Type counting; [[19] Lemma 2.2]). The number of different types of sequences in $\mathcal{X}$ is less than $(n + 1)^{|\mathcal{X}|}$.

Some important properties of types are listed in the following lemma.

**Lemma 47** ([19] Lemmas 2.5 and 2.6).

1) For any two random variables $X, \tilde{X}$ on $\mathcal{X}$, and $x \in \mathcal{T}^n_{\tilde{X}}$

$$P\{\tilde{X} = x\} = \exp \left(-n(H(X) + D_{KL}(X||\tilde{X}))\right),$$

(327)

where $\tilde{X}$ is a sequence of $n$ i.i.d. copies of $\tilde{X}$.

2) For a pair of random variables $(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$ and $x \in \mathcal{X}$, such that $\mathcal{T}^n_{Y|X}(x) \neq \emptyset$

$$(n + 1)^{-|\mathcal{X}||\mathcal{Y}|} \exp \left(nH(Y|X)\right) \leq |\mathcal{T}^n_{Y|X}(x)| \leq \exp \left(nH(Y|X)\right).$$

(328)
Remark 10. We shall adopt the $\delta$-convention \[19\] Convention 2.11 and assume the existence of an adequate sequence $(\delta_n)_{n=1}^{\infty}$ approaching 0. We will omit $\delta$ in the notation, i.e., write $T_{[X]}^n$ and $T_{[X]}^n$.

REFERENCES

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

[2] J. A. Hartigan, “Direct clustering of a data matrix,” \textit{Journal of the American Statistical Association}, vol. 67, no. 337, pp. 123–129, Mar. 1972.
[43] A. A. Gohari and V. Anantharam, “Evaluation of Marton’s inner bound for the general broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 608–619, Feb. 2012.

[44] V. Jog and C. Nair, “An information inequality for the BSSC broadcast channel,” in *Inform. Theory and Applicat. Workshop (ITA)*, San Diego, CA, USA, Feb. 2010, pp. 1–8.

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

[46] A. D. 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.

[47] J. R. Munkres, *Topology*. Prentice Hall, 2000.

[48] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill, 1976.

[49] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker’s Guide*, 3rd ed. Springer, 2006.

[50] 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.