Distributed Information Bottleneck Method for Discrete and Gaussian Sources

Conference Paper

Author(s):
Estella Aguerri, Iñaki; Zaidi, Abdellatif

Distributed Information Bottleneck Method for Discrete and Gaussian Sources

Iñaki Estella Aguerri [†]   Abdellatif Zaidi [†‡]

[†] Mathematics and Algorithmic Sciences Lab. France Research Center, Huawei Technologies, Boulogne-Billancourt, 92100, France
[‡] Université Paris-Est, Champs-sur-Marne, 77454, France
{inaki.estella@huawei.com, abdellatif.zaidi@u-pem.fr}

Abstract—We study the problem of distributed information bottleneck, in which multiple encoders separately compress their observations in a manner such that, collectively, the compressed signals preserve as much information as possible about another signal. The model generalizes Tishby's centralized information bottleneck method to the setting of multiple distributed encoders. We establish single-letter characterizations of the information-rate region of this problem for both i) a class of discrete memoryless sources and ii) memoryless vector Gaussian sources. Furthermore, assuming a sum constraint on rate or complexity, for both models we develop Blahut-Arimoto type iterative algorithms that allow to compute optimal information-rate trade-offs, by iterating over a set of self-consistent equations.

I. INTRODUCTION

The information bottleneck (IB) method was introduced by Tishby [1] (see [2] for an earlier equivalent formulation of the IB problem) as an information-theoretic principle for extracting the relevant information that some signal $Y \in \mathcal{Y}$ provides about another one, $X \in \mathcal{X}$, that is of interest. The approach has found remarkable applications in supervised and unsupervised learning problems such as classification, clustering and prediction, wherein one is interested in extracting the relevant features, i.e., X, of the available data Y [3], [4]. Perhaps key to the analysis, and development, of the IB method is its elegant connection with information-theoretic rate-distortion problems. Recent works show that this connection turns out to be useful also for a better understanding of deep neural networks [5]. Other connections, that are more intriguing, exist also with seemingly unrelated problems such as hypothesis testing [6] or systems with privacy constraints [7].

Motivated by applications of the IB method to settings in which the relevant features about X are to be extracted from separately encoded signals, we study the model shown in Figure 1. Here, X is the signal to be predicted and $(Y_1, \ldots, Y_K)$ are correlated signals that could each be relevant to extract one or more features of X. The features could be distinct or redundant. We make the assumption that the signals $(Y_1, \ldots, Y_K)$ are independent given X. This assumption holds in many practical scenarios. For example, the reader may think of $(Y_1, ..., Y_K)$ as being the results of K clinical tests that are performed independently at different clinics and are used to diagnose a disease X. A third party (decoder or detector) has to decide without access to the original data. In general, at every encoder k there is a tension among the *complexity* of the encoding, measured by the minimum description length or rate R_k at which the observation is compressed, and the information that the produced description, say U_k, provides about the signal X. The *relevance* of $(U_1, \ldots, U_K)$ is measured in terms of the information that the descriptions collectively preserve about X; and is captured by Shannon's mutual information $I(U_1, \ldots, U_K; X)$. Thus, the performance of the entire system
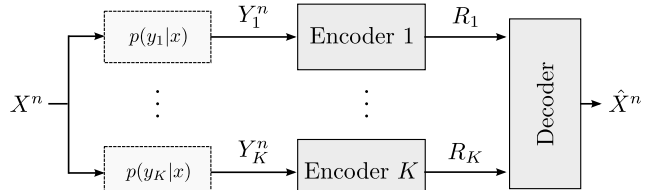

Fig. 1. A model for distributed information bottleneck (D-IB).

can be evaluated in terms of the tradeoff between the vector $(R_1, \ldots, R_K)$ of minimum description lengths and the mutual information $I(U_1, \ldots, U_K; X)$.

In this paper, we study the aforementioned tradeoff among relevant information and complexity for the model shown in Figure 1. First, we establish a single-letter characterization of the information-rate region of this model for discrete memoryless sources. In doing so, we exploit its connection with the distributed Chief Executive Officer (CEO) source coding problem under logarithmic-loss distortion measure studied in [8]. Next, we extend this result to memoryless vector Gaussian sources. Here, we prove that Gaussian test channels are optimal, thereby generalizing a similar result of [9] and [10] for the case of a single encoder IB setup.

In a second part of this paper, assuming a sum constraint on rate or complexity, we develop Blahut-Arimoto [11] type iterative algorithms that allow to compute optimal tradeoffs between information and rate, for both discrete and vector Gaussian models. We do so through a variational formulation that allows the determination of the set of self-consistent equations satisfied by the stationary solutions. In the Gaussian case, the algorithm reduces to an appropriate updating of the parameters of noisy linear projections. Here as well, our algorithms can be seen as generalizations of those developed for the single-encoder IB method, for discrete sources in [1] and for Gaussian sources in [10]; as well as a generalization of the Blahut-Arimoto algorithm proposed in [12] for the CEO source coding problem for $K = 2$ and discrete sources, to $K \geq 2$ encoders and for both discrete and Gaussian sources.

Notation: Upper case letters denote random variables, e.g., X; lower case letters denote realizations of random variables, e.g., x; and calligraphic letters denote sets, e.g., $\mathcal{X}$. The cardinality of a set is denoted by $|\mathcal{X}|$. For a random variable X with probability mass function (pmf) P_X, we use $P_X(x) = p(x)$, $x \in \mathcal{X}$ for short. Boldface upper case letters denote vectors or matrices, e.g., $\mathbf{X}$, where context makes the distinction clear. For an integer $n \in \mathbb{N}$, we denote the set $[1, n] := \{1, 2, \ldots, n\}$. We denote by $D_{\mathrm{KL}}(P, Q)$ the Kullback-Leibler divergence between the pmfs P and Q. For a set of integers $\mathcal{K} \subseteq \mathbb{N}$, $X_{\mathcal{K}}$ denotes the set $X_{\mathcal{K}} = \{X_k : k \in \mathcal{K}\}$. For a zero-mean vector $\mathbf{X}$ we define the matrices $\mathbf{\Sigma_x} := \mathrm{E}[\mathbf{X}\mathbf{X}^H]$; $\mathbf{\Sigma_{x,y}} := \mathrm{E}[\mathbf{X}\mathbf{Y}^H]$, and $\mathbf{\Sigma_{x|y}} := \mathbf{\Sigma_x} - \mathbf{\Sigma_{x,y}}\mathbf{\Sigma_y^{-1}}\mathbf{\Sigma_{y,x}}$.

II. System Model

Consider the discrete memoryless D-IB model shown in Figure 1. Let $\{X_i, Y_{1,i}, \ldots, Y_{K,i}\}_{i=1}^n = (X^n, Y_1^n, \ldots, Y_K^n)$ be a sequence of n independent, identically distributed (i.i.d.) random variables with finite alphabets $\mathcal{X}, \mathcal{Y}_k$, $k \in \mathcal{K} := \{1, \ldots, K\}$ and joint pmf $P_{X, Y_1, \ldots, Y_K}$. Throughout this paper, we make the assumption that the observations at the encoders are independent conditionally on X, i.e.,

$$Y_{k,i} \multimap X_i \multimap Y_{\mathcal{K}/k,i} \quad \text{for } k \in \mathcal{K} \text{ and } i \in [1, n]. \quad (1)$$

Encoder $k \in \mathcal{K}$ maps the observed sequence Y_k^n to an index $J_k := \phi_k(Y_k^n)$, where $\phi_k : \mathcal{Y}_k^n \to \mathcal{M}_k$ is a given map and $\mathcal{M}_k := [1, M_k^{(n)}]$. The index J_k is sent error-free to the decoder. The decoder collects all indices $J_\mathcal{K} := (J_1, \ldots, J_K)$ and then estimates the source X^n as $\hat{X}^n = g^{(n)}(J_\mathcal{J})$, where $g^{(n)} : \mathcal{M}_1 \times \cdots \times \mathcal{M}_L \to \hat{\mathcal{X}}^n$ is some decoder map and $\hat{\mathcal{X}}^n$ is the reconstruction alphabet of the source.

The quality of the reconstruction is measured in terms of the n-letter *relevant information* between the unobserved source X^n and its reconstruction at the decoder $\hat{X}^n$, given by

$$\Delta^{(n)} := \frac{1}{n} I(X^n; g^{(n)}(\phi_1^{(n)}(Y_1^n), \ldots, \phi_K^{(n)}(Y_K^n))). \quad (2)$$

Definition 1. *A tuple $(\Delta, R_1, \ldots, R_K)$ is said to be achievable for the D-IB model if there exists a blocklength n, encoder maps $\phi_k^{(n)}$ for $k \in \mathcal{K}$, and a decoder map $g^{(n)}$, such that*

$$R_k \geq \frac{1}{n} \log M_k^{(n)}, \ k \in \mathcal{K}, \ \text{and} \ \ \Delta \leq \frac{1}{n} I(X^n; \hat{X}^n). \quad (3)$$

where $\hat{X}^n = g^{(n)}(\phi_1^{(n)}(Y_1^n), \ldots, \phi_K^{(n)}(Y_K^n))$. The information-rate region $\mathcal{R}_{\text{IB}}$ is given by the closure of all achievable rates tuples $(\Delta, R_1, \ldots, R_K)$.

We are interested in characterizing the region $\mathcal{R}_{\text{IB}}$. Due to space limitations, some results are only outlined or provided without proof. We refer to [13] for a detailed version.

III. Information-Rate Region Characterization

In this section we characterize the information-rate region $\mathcal{R}_{\text{IB}}$ for a discrete memoryless D-IB model. It is well known that the IB problem is essentially a source-coding problem where the distortion measure is of logarithmic loss type [14]. Likewise, the D-IB model of Figure 1 is essentially a K-encoder CEO source coding problem under logarithmic loss (log-loss) distortion measure. The log-loss distortion between sequences is defined as

$$d_{\text{LL}}(x^n, \hat{x}^n) := -\frac{1}{n} \log \left(\frac{1}{\hat{x}^n(x^n)} \right), \quad (4)$$

where $\hat{x}^n = s(x^n | j_\mathcal{K})$ and s is a pmf on $\mathcal{X}^n$.

The rate-distortion region of the K-encoder CEO source coding problem under log-loss, with $K \geq 2$ which we denote hereafter as $\mathcal{RD}_{\text{CEO}}$, has been established recently in [8, Theorem 10] for the case in which the Markov chain (1) holds.

We first state the following proposition, the proof of which is easy and omitted for brevity.

Proposition 1. *A tuple $(\Delta, R_1, \ldots, R_K) \in \mathcal{R}_{\text{IB}}$ if and only if $(H(X) - \Delta, R_1, \ldots, R_K) \in \mathcal{RD}_{\text{CEO}}$.*

Proposition 1 implies that [8, Theorem 10] can be applied to characterize the information-rate region $\mathcal{R}_{\text{IB}}$ as given next.

Theorem 1. *In the case in which the Markov chain* (1) *holds, the rate-information region $\mathcal{R}_{\text{IB}}$ of the D-IB model is given by the set of all tuples $(\Delta, R_1, \ldots, R_K)$ which satisfy for $\mathcal{S} \subseteq \mathcal{K}$*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k - I(Y_k; U_k | X, Q)] + I(X; U_{\mathcal{S}^c} | Q), \quad (5)$$

for some joint pmf $p(q)p(x) \prod_{k=1}^K p(y_k|x) \prod_{k=1}^K p(u_k|y_k, q)$.

A. Memoryless Vector Gaussian D-IB

Consider now the following memoryless vector Gaussian D-IB problem. In this model, the source vector $\mathbf{X} \in \mathbb{C}^N$ is Gaussian and has zero mean and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$, i.e., $\mathbf{X} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{x}})$. Encoder k, $k \in \mathcal{K}$, observes a noisy observation $\mathbf{Y}_k \in \mathbb{C}^{M_k}$, that is given by

$$\mathbf{Y}_k = \mathbf{H}_k \mathbf{X} + \mathbf{N}_k, \quad (6)$$

where $\mathbf{H}_k \in \mathbb{C}^{M_k \times N}$ is the channel connecting the source to encoder k, and $\mathbf{N}_k \in \mathbb{C}^{M_k}$, $k \in \mathcal{K}$, is the noise vector at encoder k, assumed to be Gaussian, with zero-mean and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{n}_k}$, and independent from all other noises and the source vector $\mathbf{X}$.

The studied Gaussian model satisfies the Markov chain (1); and thus, the result of Theorem 1, which can be extended to continuous sources using standard techniques, characterizes the information-rate region of this model, denoted by $\mathcal{R}_{\text{IB}}^{\text{G}}$. Next theorem characterizes $\mathcal{R}_{\text{IB}}^{\text{G}}$, shows that the optimal test channels $P_{U_k|Y_k}$, $k \in \mathcal{K}$, are Gaussian and that there is not need for time-sharing, i.e., $Q = \emptyset$.

Theorem 2. *If $(\mathbf{X}, \mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ are jointly Gaussian as in* (6)*, the information-rate region $\mathcal{R}_{\text{IB}}^{\text{G}}$ is given by the set of all tuples $(\Delta, R_1, \ldots, R_L)$ satisfying that for all $\mathcal{S} \subseteq \mathcal{K}$*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k + \log|\mathbf{I} - \mathbf{B}_k|] + \log \left| \sum_{k \in \mathcal{S}^c} \bar{\mathbf{H}}_k^H \mathbf{B}_k \bar{\mathbf{H}}_k + \mathbf{I} \right|,$$

for some $\mathbf{0} \preceq \mathbf{B}_k \preceq \mathbf{I}$ and where $\bar{\mathbf{H}}_k = \boldsymbol{\Sigma}_{\mathbf{n}_k}^{-1/2} \mathbf{H}_k \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}$. In addition, the information-rate tuples in $\mathcal{R}_{\text{IB}}^{\text{G}}$ are achievable with $Q = \emptyset$ and $p^*(\mathbf{u}_k|\mathbf{y}_k, q) = \mathcal{CN}(\mathbf{y}_k, \boldsymbol{\Sigma}_{\mathbf{n}_k}^{1/2}(\mathbf{B}_k - \mathbf{I})\boldsymbol{\Sigma}_{\mathbf{n}_k}^{1/2})$.

Proof. An outline of the proof is given in Appendix A. □

IV. Computation of the Information Rate Region under Sum-Rate Constraint

In this section, we describe an iterative Blahut-Arimoto (BA)-type algorithm to compute the pmfs $P_{U_k|Y_k}$, $k \in \mathcal{K}$, that maximize information Δ under sum-rate constraint, i.e., $R_{\text{sum}} := \sum_{k=1}^K R_k$, for tuples $(\Delta, R_1, \ldots, R_K)$ in $\mathcal{R}_{\text{IB}}$. From Theorem 1 we have:

$$\mathcal{R}_{\text{sum}} := \text{convex-hull}\{(\Delta, R_{\text{sum}}) : \Delta \leq \Delta_{\text{sum}}(R_{\text{sum}})\}, \quad (7)$$

where we define the information-rate function

$$\Delta_{\text{sum}}(R) := \max_{\mathbf{P}} \min \left\{ I(X; U_\mathcal{K}), R - \sum_{k=1}^K I(Y_k; U_k | X) \right\},$$

and where the optimization is over the set of K conditional pmfs $P_{U_k|Y_k}$, $k \in \mathcal{K}$, which, for short, we define as

$$\mathbf{P} := \{P_{U_1|Y_1}, \ldots, P_{U_K|Y_K}\}. \quad (8)$$

Next proposition provides a characterization of the pairs $(\Delta, R_{\text{sum}}) \in \mathcal{R}_{\text{sum}}$ in terms of a parameter $s \geq 0$.

Proposition 2. *Each tuple* (Δ, R_{sum}) *on the information rate curve* $\Delta = \Delta_{\text{sum}}(R_{\text{sum}})$, *can be obtained for some* $s \geq 0$, *as* (Δ_s, R_s), *parametrically defined by*

$$(1+s)\Delta_s = (1+sK)H(X) + sR_s - \min_{\mathbf{P}} F_s(\mathbf{P}), \quad (9)$$

$$R_s = I(Y_{\mathcal{K}}; U_{\mathcal{K}}^*) + \sum_{k=1}^{K}[I(Y_k; U_k^*) - I(X; U_k^*)], \quad (10)$$

where $\mathbf{P}^*$ *are the pmfs yielding the minimum in* (9) *and*

$$F_s(\mathbf{P}) := H(X|U_{\mathcal{K}}) + s\sum_{k=1}^{K}[I(Y_k; U_k) + H(X|U_k)]. \quad (11)$$

Proof. The proof of Proposition 2 follows along the lines of [12, Theorem 2] and is omitted for brevity. Note that the rate expression in Theorem 1 is different to that in [12]. □

From Proposition 2, the information-rate function can be computed by solving (9) and evaluating (10) for all $s \geq 0$. Inspired by the standard Blahut-Arimoto (BA) method [11], and following similar steps as for the BA-type algorithm proposed in [12] for the CEO problem with $K = 2$ encoders, we show that problem (9) can be solved with an alternate optimization procedure, with respect to $\mathbf{P}$ and some appropriate auxiliary pmfs Q_{U_k}, $Q_{X|U_k}$, $k \in \mathcal{K}$ and $Q_{X|U_1,\ldots,U_K}$, denoted as

$$\mathbf{Q} := \{Q_{U_1}, \ldots, Q_{U_K}, Q_{X|U_1}, \ldots, Q_{X|U_K}, Q_{X|U_1,\ldots,U_K}\}.$$

To this end, we define the function $\bar{F}_s(\cdot)$ and write (9) as a minimization over the pmfs $\mathbf{P}$ and pmfs $\mathbf{Q}$, where

$$\bar{F}_s(\mathbf{P}, \mathbf{Q}) := -s\sum_{k=1}^{K} H(U_k|Y_k) - s\sum_{k=1}^{K} \mathrm{E}_{U_k}[\log q(U_k)]$$
$$- s\sum_{k=1}^{K} \mathrm{E}_{X,U_k}[\log q(X|U_k)] - \mathrm{E}_{X,U_{\mathcal{K}}}[\log q(X|U_{\mathcal{K}})]. \quad (12)$$

Lemma 1. *We have*

$$F^* := \min_{\mathbf{P}} F_s(\mathbf{P}) = \min_{\mathbf{P}} \min_{\mathbf{Q}} \bar{F}_s(\mathbf{P}, \mathbf{Q}). \quad (13)$$

Algorithm 1 describes the steps to successively minimize $\bar{F}_s(\mathbf{P}, \mathbf{Q})$ by optimizing a convex problem over $\mathbf{P}$ and over $\mathbf{Q}$ at each iteration. The proof of Lemma 1 and the steps of the proposed algorithm are justified with the following lemmas, whose proofs are along the lines of Lemma 1, Lemma 2, Lemma 3 in [12], and are omitted due to space limitations.

Lemma 2. $\bar{F}_s(\mathbf{P}, \mathbf{Q})$ *is convex in* $\mathbf{P}$ *and convex in* $\mathbf{Q}$.

Lemma 3. *For fixed pmfs* $\mathbf{P}$, $\bar{F}_s(\mathbf{P}, \mathbf{Q}) \geq F_s(\mathbf{P})$ *for all pmfs* $\mathbf{Q}$, *and there exists a unique* $\mathbf{Q}$ *that achieves the minimum* $\min_{\mathbf{Q}} \bar{F}_s(\mathbf{P}, \mathbf{Q}) = F_s(\mathbf{P})$, *given by*

$$Q_{U_k}^* = P_{U_k}, \quad Q_{X|U_k}^* = P_{X|U_k}, \quad k \in \mathcal{K}, \quad (14)$$

$$Q_{X|U_1,\ldots,U_K}^* = P_{X|U_1,\ldots,U_K}, \quad (15)$$

where P_{U_k}, $P_{X|U_k}$ *and* $P_{X|U_1,\ldots,U_K}$ *are computed from* $\mathbf{P}$.

Lemma 4. *For fixed* $\mathbf{Q}$, *there exists a* $\mathbf{P}$ *that achieves the minimum* $\min_{\mathbf{P}} \bar{F}_s(\mathbf{P}, \mathbf{Q})$, *where* $P_{U_k|Y_k}$ *is given by*

$$p^*(u_k|y_k) = q(u_k)\frac{\exp\left(-\psi_s(u_k, y_k)\right)}{\sum_{u_k \in \mathcal{U}_k} q(u_k)\exp(-\psi_s(u_k, y_k))}, \quad (16)$$

for $u_k \in \mathcal{U}_k$ *and* $y_k \in \mathcal{Y}_k$, $k \in \mathcal{K}$, *and where we define*

$$\psi_s(u_k, y_k) := D_{\text{KL}}(P_{X|y_k} || Q_{X|u_k}) \quad (17)$$
$$+ \frac{1}{s}\mathrm{E}_{U_{\mathcal{K}\backslash k}|y_k}[D_{\text{KL}}(P_{X|U_{\mathcal{K}\backslash k}, y_k} || Q_{X|U_{\mathcal{K}\backslash k}, u_k})].$$

Algorithm 1 BA-type algorithm for the Discrete D-IB

1: **input:** pmf $P_{X,Y_1,\ldots,Y_k}$, parameter $s \geq 0$.
2: **output:** optimal $P_{U_k|Y_k}^*$, pair (Δ_s, R_s).
3: **initialization** Set $t = 0$ and set $\mathbf{P}^{(0)}$ with $p(u_k|y_k) = \frac{1}{|\mathcal{U}_k|}$
$\quad$ for $u_k \in \mathcal{U}_k$, $y_k \in \mathcal{Y}_k$, $k = 1, \ldots, K$.
4: **repeat**
5: $\quad$ Compute $\mathbf{Q}^{(t+1)}$ as (14) and (15) from $\mathbf{P}^{(t)}$.
6: $\quad$ Compute $\mathbf{P}^{(t+1)}$ as (16) from $\mathbf{Q}^{(t+1)}$ and $\mathbf{P}^{(t)}$.
7: $\quad$ $t \leftarrow t + 1$.
8: **until** convergence.

Algorithm 1 essentially falls in the Successive Upper-Bound Minimization (SUM) framework [15] in which $\bar{F}_s(\mathbf{P}, \mathbf{Q})$ acts as a globally tight upper bound on $F_s(\mathbf{P})$. Algorithm 1 provides a sequence $\mathbf{P}^{(t)}$ for each iteration t, which converges to a stationary point of the optimization problem (13).

Proposition 3. *Every limit point of the sequence* $\mathbf{P}^{(t)}$ *generated by Algorithm 1 converges to a stationary point of* (13).

Proof. Let $\mathbf{Q}^*(\mathbf{P}) := \arg\min_{\mathbf{Q}} \bar{F}_s(\mathbf{P}, \mathbf{Q})$. From Lemma 3, $\bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P}')) \geq \bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P})) = F_s(\mathbf{P})$ for $\mathbf{P}' \neq \mathbf{P}$. It follows that $F_s(\mathbf{P})$ and $\bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P}'))$ satisfy [15, Proposition 1] and thus $\bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P}'))$ satisfies A1-A4 in [15]. Convergence to a stationary point of (13) follows from [15, Theorem 1]. □

Remark 1. *The resulting set of self consistent equations* (14), (15) *and* (17) *satisfied by any stationary point of the D-IB problem, remind that of the original IB problem [9]. Note the additional divergence term in* (17) *for encoder* k *averaged over the descriptions at the other* $\mathcal{K} \setminus k$ *encoders.*

V. COMPUTATION OF THE INFORMATION RATE REGION FOR THE VECTOR GAUSSIAN D-IB

Computing the maximum information under sum-rate constraint from Theorem 2 is a convex optimization problem on $\mathbf{B}_k$, which can be efficiently solved with generic tools. Alternatively, next we extend Algorithm 1 for Gaussian sources.

For finite alphabet sources the updates of $\mathbf{Q}^{(t+1)}$ and $\mathbf{P}^{(t+1)}$ in Algorithm 1 are simple, but become unfeasible for continuous alphabet sources. We leverage on the optimality of Gaussian descriptions, shown in Theorem 2, to restrict the optimization of $\mathbf{P}$ to Gaussian distributions, which are easily represented by a finite set of parameters, namely its mean and covariance. We show that if $\mathbf{P}^{(t)}$ are Gaussian pmfs, then $\mathbf{P}^{(t+1)}$ are also Gaussian pmfs, which can be computed with an efficient update algorithm of its representing parameters. In particular, if at time t, the k-th pmf $P_{\mathbf{U}_k|\mathbf{Y}_k}^{(t)}$ is given by

$$\mathbf{U}_k^t = \mathbf{A}_k^t \mathbf{Y}_k + \mathbf{Z}_k^t, \quad (18)$$

where $\mathbf{Z}_k^t \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}_k^t})$; we show that for $\mathbf{P}^{(t+1)}$ updated as in (16), $P_{\mathbf{U}_k|\mathbf{Y}_k}^{(t+1)}$ corresponds to $\mathbf{U}_k^{t+1} = \mathbf{A}_k^{t+1}\mathbf{Y}_k + \mathbf{Z}_k^{t+1}$, where $\mathbf{Z}_k^{t+1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}})$ and $\mathbf{A}_k^{t+1}$, $\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}$ are updated as

$$\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}} = \left(\left(1 + \frac{1}{s}\right)\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1} - \frac{1}{s}\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\backslash k}^t}^{-1}\right)^{-1}, \quad (19)$$

$$\mathbf{A}_k^{t+1} = \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1}\left(\left(1 + \frac{1}{s}\right)\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1})\right.$$
$$\left. - \frac{1}{s}\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\backslash k}^t}^{-1}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{u}_{\mathcal{K}\backslash k}^t}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1})\right). \quad (20)$$

The detailed update procedure is given in Algorithm 2.

Remark 2. *Algorithm 2 generalizes the iterative algorithm for single encoder Gaussian D-IB in [10] to the Gaussian D-IB with K encoders and sum-rate constraint. Similarly to the solution in [10], the optimal description at each encoder is given by a noisy linear projection of the observation, whose dimensionality is determined by the parameter s and the second order moments between the observed data and the source of interest, as well as a term depending on the observed data with respect to the descriptions at the other encoders.*

A. Derivation of Algorithm 2

In this section, we derive the update rules in Algorithm 2 and show that the Gaussian distribution is invariant to the update rules in Algorithm 1, in line with Theorem 2.

First, we recall that if $(\mathbf{X}_1, \mathbf{X}_2)$ are jointly Gaussian, then

$$P_{\mathbf{X}_2|\mathbf{X}_1=\mathbf{x}_1} = \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}_2|\mathbf{x}_1}, \boldsymbol{\Sigma}_{\mathbf{x}_2|\mathbf{x}_1}), \tag{21}$$

where $\boldsymbol{\mu}_{\mathbf{x}_2|\mathbf{x}_1} := \mathbf{K}_{\mathbf{x}_2|\mathbf{x}_1}\mathbf{x}_1$, with $\mathbf{K}_{\mathbf{x}_2|\mathbf{x}_1} := \boldsymbol{\Sigma}_{\mathbf{x}_2,\mathbf{x}_1}\boldsymbol{\Sigma}_{\mathbf{x}_1}^{-1}$.

Then, for $\mathbf{Q}^{(t+1)}$ computed as in (14) and (15) from $\mathbf{P}^{(t)}$, which is a set of Gaussian distributions, we have

$$Q_{\mathbf{X}|\mathbf{u}_k}^{(t+1)} = \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{u}_k^t}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{u}_k^t}), Q_{\mathbf{X}|\mathbf{u}_{\mathcal{K}}}^{(t+1)} = \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{u}_{\mathcal{K}}^t}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{u}_{\mathcal{K}}^t}).$$

Next, we look at the update $\mathbf{P}^{(t+1)}$ as in (16) from given $\mathbf{Q}^{(t+1)}$. First, we have that $p(\mathbf{u}_k^t)$ is the marginal of $\mathbf{U}_k^t$, given by $\mathbf{U}_k^t \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{u}_k^t})$ where $\boldsymbol{\Sigma}_{\mathbf{u}_k^t} = \mathbf{A}_k^t \boldsymbol{\Sigma}_{\mathbf{y}_k} \mathbf{A}_k^{t,H} + \boldsymbol{\Sigma}_{\mathbf{z}_k^t}$.

Then, to compute $\psi_s(\mathbf{u}_k^t, \mathbf{y}_k)$, first, we note that

$$E_{U_{\mathcal{K}\backslash k}|y_k}[D_{\mathrm{KL}}(P_{X|U_{\mathcal{K}\backslash k},y_k} || Q_{X|U_{\mathcal{K}\backslash k},u_k})] \tag{22}$$
$$=D_{\mathrm{KL}}(P_{X,U_{\mathcal{K}\backslash k}|y_k} || Q_{X,U_{\mathcal{K}\backslash k}|u_k}) - D_{\mathrm{KL}}(P_{U_{\mathcal{K}\backslash k}|y_k} || Q_{U_{\mathcal{K}\backslash k}|u_k}),$$

and that for two generic multivariate Gaussian distributions $P_1 \sim \mathcal{CN}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $P_2 \sim \mathcal{CN}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ in $\mathbb{C}^N$,

$$D_{\mathrm{KL}}(P_1, P_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^H \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$
$$+ \log|\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1}| - N + \mathrm{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\}. \tag{23}$$

Applying (22) and (23) in (17) and noting that all involved distributions are Gaussian, it follows that $\psi_s(\mathbf{u}_k^t, \mathbf{y}_k)$ is a quadratic form. Then, since $p(\mathbf{u}_k^t)$ is Gaussian, the product $\log(p(\mathbf{u}_k^t)\exp(-\psi_s(\mathbf{u}_k^t, \mathbf{y}_k)))$ is also a quadratic form, and identifying constant, first and second order terms, we can write

$$\log p^{(t+1)}(\mathbf{u}_k|\mathbf{y}_k) = Z(\mathbf{y}_k) + (\mathbf{u}_k - \boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k})^H \boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1}$$
$$\cdot (\mathbf{u}_k - \boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k}), \tag{24}$$

where $Z(\mathbf{y}_k)$ is a normalization term independent of $\mathbf{u}_k$, and

$$\boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{u}_k^t}^{-1} + \mathbf{K}_{\mathbf{x}|\mathbf{u}_k^t}^H \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{u}_k^t}^{-1} \mathbf{K}_{\mathbf{x}|\mathbf{u}_k^t}$$
$$+ \frac{1}{s} \mathbf{K}_{\mathbf{xu}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^H \boldsymbol{\Sigma}_{\mathbf{xu}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^{-1} \mathbf{K}_{\mathbf{xu}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}$$
$$- \frac{1}{s} \mathbf{K}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^H \boldsymbol{\Sigma}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^{-1} \mathbf{K}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}, \tag{25}$$

$$\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k} = \boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}} \Big(\mathbf{K}_{\mathbf{x}|\mathbf{u}_k^t}^H \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{u}_k^t}^{-1} \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_k}$$
$$+ \frac{1}{s} \mathbf{K}_{\mathbf{x},\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t} \boldsymbol{\Sigma}_{\mathbf{x},\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^{-1} \boldsymbol{\mu}_{\mathbf{x},\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{y}_k}$$
$$- \frac{1}{s} \mathbf{K}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t} \boldsymbol{\Sigma}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^{-1} \boldsymbol{\mu}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{y}_k} \Big). \tag{26}$$

This shows that $p^{(t+1)}(\mathbf{u}_k|\mathbf{y}_k)$ is a Gaussian distribution and that $\mathbf{U}_k^{t+1}|\{\mathbf{Y}_k = \mathbf{y}_k\}$ is distributed as $\mathcal{CN}(\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k}, \boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}})$.

Algorithm 2 BA-type algorithm for the Gaussin Vector D-IB

1: **input:** covariance $\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k}$, parameter $s \geq 0$.
2: **output:** optimal pairs $(\mathbf{A}_k^*, \boldsymbol{\Sigma}_{\mathbf{z}_k^*})$, $k = 1, \ldots, K$.
3: **initialization** Set randomly $\mathbf{A}_k^0$ and $\boldsymbol{\Sigma}_{\mathbf{z}_k^0} \succeq 0$, $k \in \mathcal{K}$.
4: **repeat**
5:      Compute $\boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{u}_{\mathcal{K}\backslash k}^t}$ and update for $k \in \mathcal{K}$

$$\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}} = \mathbf{A}_k^t \boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{x}} \mathbf{A}_k^{t,H} + \boldsymbol{\Sigma}_{\mathbf{z}_k^t} \tag{27}$$

$$\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\backslash k}^t} = \mathbf{A}_k^t \boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{u}_{\mathcal{K}\backslash k}^t} \mathbf{A}_k^{t,H} + \boldsymbol{\Sigma}_{\mathbf{z}_k^t}, \tag{28}$$

6:      Compute $\boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}$ as in (19) for $k \in \mathcal{K}$.
7:      Compute $\mathbf{A}_k^{t+1}$ as (20), $k \in \mathcal{K}$.
8:      $t \leftarrow t + 1$.
9: **until** convergence.

Next, we simplify (25) and (26) to obtain the update rules (19) and (20). From the matrix inversion lemma, similarly to [10], for $(\mathbf{X}_1, \mathbf{X}_2)$ jointly Gaussian we have

$$\boldsymbol{\Sigma}_{\mathbf{x}_2|\mathbf{x}_1}^{-1} = \boldsymbol{\Sigma}_{\mathbf{x}_2}^{-1} + \mathbf{K}_{\mathbf{x}_1|\mathbf{x}_2}^H \boldsymbol{\Sigma}_{\mathbf{x}_1|\mathbf{x}_2}^{-1} \mathbf{K}_{\mathbf{x}_1|\mathbf{x}_2}. \tag{29}$$

Applying (29), in (25) we have

$$\boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1} + \frac{1}{s}\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{xu}_{\mathcal{K}\backslash k}}^{-1} - \frac{1}{s}\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\backslash k}}^{-1}, \tag{30}$$

$$= \left(1 + \frac{1}{s}\right)\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1} - \frac{1}{s}\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\backslash k}}^{-1}, \tag{31}$$

where (31) is due to the Markov chain $\mathbf{U}_k \multimap \mathbf{X} \multimap \mathbf{U}_{\mathcal{K}\backslash k}$.

Then, also from the matrix inversion lemma, we have for jointly Gaussian $(\mathbf{X}_1, \mathbf{X}_2)$,

$$\boldsymbol{\Sigma}_{\mathbf{x}_2|\mathbf{x}_1}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}_1,\mathbf{x}_2}\boldsymbol{\Sigma}_{\mathbf{x}_1}^{-1} = \boldsymbol{\Sigma}_{\mathbf{x}_2}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}_1,\mathbf{x}_2}\boldsymbol{\Sigma}_{\mathbf{x}_1|\mathbf{x}_2}^{-1}. \tag{32}$$

Applying (32) in (26), for the first term, we have

$$\mathbf{K}_{\mathbf{x}|\mathbf{u}_k^t}^H \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{u}_k^t}^{-1}\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_k} = \boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{u}_k^t}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_k} \tag{33}$$
$$= \boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1}\mathbf{A}_k^t\boldsymbol{\Sigma}_{\mathbf{y}_k,\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{y}_k}\boldsymbol{\Sigma}_{\mathbf{y}_k}^{-1}\mathbf{y}_k$$
$$= \boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1}\mathbf{A}_k^t(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{y}_k}^{-1})\mathbf{y}_k, \tag{34}$$

where $\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{u}_k^t} = \mathbf{A}_k^t\boldsymbol{\Sigma}_{\mathbf{y}_k,\mathbf{x}}$; and (34) is due to the definition of $\boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{x}}$. Similarly, for the second term, we have

$$\mathbf{K}_{\mathbf{xu}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^H \boldsymbol{\Sigma}_{\mathbf{xu}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^{-1}\boldsymbol{\mu}_{\mathbf{x},\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{y}_k}$$
$$= \boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{xu}_{\mathcal{K}\backslash k}}^{-1}\mathbf{A}_k^t(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{xu}_{\mathcal{K}\backslash k}^t}\boldsymbol{\Sigma}_{\mathbf{y}_k}^{-1})\mathbf{y}_k, \tag{35}$$
$$= \boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1}\mathbf{A}_k^t(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{y}_k}^{-1})\mathbf{y}_k, \tag{36}$$

where we use $\boldsymbol{\Sigma}_{\mathbf{u}_k^t,\mathbf{xu}_{\mathcal{K}\backslash k}} = \mathbf{A}_k^t\boldsymbol{\Sigma}_{\mathbf{y}_k,\mathbf{xu}_{\mathcal{K}\backslash k}^t}$; and (36) is due to the Markov chain $\mathbf{U}_k \multimap \mathbf{X} \multimap \mathbf{U}_{\mathcal{K}\backslash k}$. For the third term,

$$\mathbf{K}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}\boldsymbol{\Sigma}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{u}_k^t}^{-1}\boldsymbol{\mu}_{\mathbf{u}_{\mathcal{K}\backslash k}^t|\mathbf{y}_k}$$
$$= \boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\backslash k}}^{-1}\mathbf{A}_k^t(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{u}_{\mathcal{K}\backslash k}^t}\boldsymbol{\Sigma}_{\mathbf{y}_k}^{-1})\mathbf{y}_k. \tag{37}$$

Equation (20) follows by noting that $\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k} = \mathbf{A}_k^{t+1}\mathbf{y}_k$, and that from (26) $\mathbf{A}_k^{t+1}$ can be identified as given in (20).

Finally, we note that due to (18), $\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\backslash k}^t}$ are given as in (27) and (28), where $\boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{n}_k}$ and $\boldsymbol{\Sigma}_{\mathbf{y}_k|\mathbf{u}_{\mathcal{K}\backslash k}^t}$ can be computed from its definition.

VI. NUMERICAL RESULTS

In this section, we consider the numerical evaluation of Algorithm 2, and compare the resulting relevant information to two upper bounds on the performance for the D-IB: i) the information-rate pairs achievable under centralized IB encoding, i.e., if $(Y_1, \ldots, Y_K)$ are encoded jointly at a rate equal to the total rate $R_{\text{sum}} = R_1 + \cdots + R_K$, characterized in [10]; ii) the information-rate pairs achievable under centralized IB encoding when $R_{\text{sum}} \to \infty$, i.e., $\Delta = I(X; Y_1, \ldots, Y_K)$.

Figure 2 shows the resulting (Δ, R_{sum}) tuples for a Gaussian vector model with $K = 2$ encoders, source dimension $N = 4$, and observations dimension $M_1 = M_2 = 2$ for different values of s calculated as in Proposition 2 using Algorithm 2, and its upper convex envelope. As it can be seen, the distributed IB encoding of sources performs close to the Tishby's centralized IB method, particularly for low R_{sum} values.

APPENDIX A

Let $(\mathbf{X}, \mathbf{U})$ be two complex random vectors. The conditional Fischer information is defined as $\mathbf{J}(\mathbf{X}|\mathbf{U}) := \mathrm{E}[\nabla \log p(\mathbf{X}|\mathbf{U}) \nabla \log p(\mathbf{X}|\mathbf{U})^H]$, and the MMSE is given by $\mathrm{mmse}(\mathbf{X}|\mathbf{U}) := \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}|\mathbf{U}])(\mathbf{X} - \mathrm{E}[\mathbf{X}|\mathbf{U}])^H]$. Then [16]

$$\log |(\pi e)\mathbf{J}^{-1}(\mathbf{X}|\mathbf{U})| \leq h(\mathbf{X}|\mathbf{U}) \leq \log |(\pi e)\mathrm{mmse}(\mathbf{X}|\mathbf{U})|. \quad (38)$$

We outer bound the information-rate region in Theorem 1 for $(\mathbf{X}, \mathbf{Y}_{\mathcal{K}})$ as in (6). For $q \in \mathcal{Q}$ and fixed $\prod_{k=1}^{K} p(\mathbf{u}_k|\mathbf{y}_k, q)$, choose $\mathbf{B}_{k,q}$, $k \in \mathcal{K}$ satisfying $\mathbf{0} \preceq \mathbf{B}_{k,q} \preceq \mathbf{\Sigma}_{\mathbf{n}_k}^{-1}$ such that

$$\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, \mathbf{U}_{k,q}, q) = \mathbf{\Sigma}_{\mathbf{n}_k} - \mathbf{\Sigma}_{\mathbf{n}_k} \mathbf{B}_{k,q} \mathbf{\Sigma}_{\mathbf{n}_k}. \quad (39)$$

Such $\mathbf{B}_{k,q}$ always exists since $\mathbf{0} \preceq \mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, \mathbf{U}_{k,q}, q) \preceq \mathbf{\Sigma}_{\mathbf{n}_k}^{-1}$, for all $q \in \mathcal{Q}$, and $k \in \mathcal{K}$. We have from (5),

$$I(\mathbf{Y}_k; \mathbf{U}_k|\mathbf{X}, q) \geq \log |\mathbf{\Sigma}_{\mathbf{n}_k}| - \log |\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, \mathbf{U}_{k,q}, q)|$$
$$= -\log |\mathbf{I} - \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2} \mathbf{B}_{k,q} \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}|, \quad (40)$$

where the inequality is due to (38), and (40) is due to (39). Let $\bar{\mathbf{B}}_k := \sum_{q \in \mathcal{Q}} p(q) \mathbf{B}_{k,q}$. Then, we have from (40)

$$I(\mathbf{Y}_k; \mathbf{U}_k|\mathbf{X}, Q) \geq -\sum_{q \in \mathcal{Q}} p(q) \log |\mathbf{I} - \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2} \mathbf{B}_{k,q} \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}|$$
$$\geq -\log |\mathbf{I} - \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2} \bar{\mathbf{B}}_k \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}|, \quad (41)$$

where (41) follows from the concavity of the log-det function and Jensen's inequality. On the other hand, we have

$$I(\mathbf{X}; \mathbf{U}_{S^c, q}|q) \leq \log |\mathbf{\Sigma}_{\mathbf{x}}| - \log |\mathbf{J}^{-1}(\mathbf{X}|\mathbf{U}_{S^c, q}, q)| \quad (42)$$
$$= \log \left| \sum_{k \in S^c} \mathbf{\Sigma}_{\mathbf{x}}^{1/2} \mathbf{H}_k^H \mathbf{B}_{k,q} \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{x}}^{1/2} + \mathbf{I} \right|, \quad (43)$$

where (42) is due to (38); and (43) is due to to the following equality connecting the MMSE matrix (39) and the Fisher information as in [16]–[19] (We refer to [13] for details):

$$\mathbf{J}(\mathbf{X}|\mathbf{U}_{S^c, q}, q) = \sum_{k \in S^c} \mathbf{H}_k^H \mathbf{B}_{k,q} \mathbf{H}_k + \mathbf{\Sigma}_{\mathbf{x}}^{-1}. \quad (44)$$

Similarly to (41), from (43) and Jensen's Inequality we have

$$I(\mathbf{X}; \mathbf{U}_{S^c}|Q) \leq \log \left| \sum_{k \in S^c} \mathbf{\Sigma}_{\mathbf{x}}^{1/2} \mathbf{H}_k^H \bar{\mathbf{B}}_k \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{x}}^{1/2} + \mathbf{I} \right|. \quad (45)$$

Substituting (41) and (45) in (5) and letting $\mathbf{B}_k := \mathbf{\Sigma}_{\mathbf{n}_k}^{-1/2} \bar{\mathbf{B}}_k \mathbf{\Sigma}_{\mathbf{n}_k}^{-1/2}$ gives the outer bound. The proof is completed by noting that the outer bound is achieved with $Q = \emptyset$ and $p^*(\mathbf{u}_k|\mathbf{y}_k, q) = \mathcal{CN}(\mathbf{y}_k, \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}(\mathbf{B}_k - \mathbf{I})\mathbf{\Sigma}_{\mathbf{n}_k}^{1/2})$.
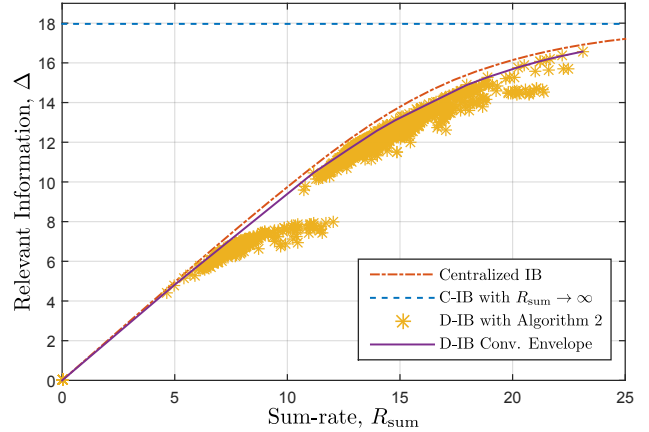

Fig. 2. Information vs. sum-rate for vector Gaussian D-IB with $K = 2$ encoders, source dimension $N = 4$, and observation dimension $M_1 = M_2 = 2$.

REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annual Allerton Conf. on Comm., Control, and Computing*, 1999, pp. 368–377.

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

[3] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. of 23rd Ann. Int'l ACM SIGIR Conf. on Res. and Dev. in Info. Retrieval*, 2000, pp. 208–215.

[4] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, "Information-based clustering," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18 297–18 302, 2005.

[5] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information." [Online]. Available: http://arxiv.org/abs/1703.00810

[6] C. Tian and J. Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Trans. Info. Theory*, vol. 54, no. 10, pp. 4666–4681, Oct. 2008.

[7] K. Kittichokechai and G. Caire, "Privacy-constrained remote source coding," in *IEEE Int. Symp. Inf. Th. (ISIT)*, Jul. 2016, pp. 1078–1082.

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

[9] A. Globerson and N. Tishby, "On the optimality of the Gaussian information bottleneck curve," *Hebrew University Tech. Report*, 2004.

[10] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables." *Journal of Machine Learning Research*, vol. 6, pp. 165–188, Feb. 2005.

[11] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul 1972.

[12] Y. Ugur, I. E. Aguerri, and A. Zaidi, "A generalization of Blahut-Arimoto algorithm to computing rate-distortion regions of multiterminal source coding under logarithmic loss," in *Proc. of IEEE Info. Theory Workshop, ITW*, Kaohsiung, Taiwan, Nov. 2017.

[13] I. Estella and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources." [Online]. Available: http://www-syscom.univ-mlv.fr/ zaidi/publications/proofs-dIB-izs2017.pdf

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

[15] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. on Opt.*, vol. 23, pp. 1126–1153, Jun. 2013.

[16] E. Ekrem and S. Ulukus, "An outer bound for the vector Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, Nov. 2014.

[17] Y. Zhou, Y. Xu, W. Yu, and J. Chen, "On the optimal fronthaul compression and decoding strategies for uplink cloud radio access networks," *IEEE Tr. Inf. Th.*, vol. 62, no. 12, pp. 7402–7418, Dec. 2016.

[18] I. Estella, A. Zaidi, G. Caire, and S. Shamai, "On the capacity of cloud radio access networks," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2017.

[19] I. E. Aguerri, A. Zaidi, G. Caire, and S. Shamai, "On the capacity of cloud radio access networks with oblivious relaying." [Online]. Available: http://arxiv.org/abs/1710.09275