Received 14 June 2024; accepted 7 July 2024. Date of publication 10 July 2024; date of current version 25 July 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3426049

Multi-Source Distributed Data Compression Based on Information Bottleneck Principle

SHAYAN HASSANPOUR (Member, IEEE), ALIREZA DANAEE (Member, IEEE), DIRK WÜBBEN (Senior Member, IEEE), AND ARMIN DEKORSY (Senior Member, IEEE)

Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany

CORRESPONDING AUTHOR: S. HASSANPOUR (e-mail: hassanpour@ant.uni-bremen.de)

This work was supported in part by the German Ministry of Education and Research (BMBF) under Grant 16KISK109 (6G-ANNA), Grant 16KISK016 (Open6GHub), and Grant 16KISK068 (6G-TakeOff).

ABSTRACT In this article, we focus on a generic multiterminal (remote) source coding scenario in which, via a *joint* design, several intermediate nodes must *locally* compress their *noisy* observations from various sets of user / source signals ahead of forwarding them through multiple *error-free* and *rate-limited* channels to a (remote) processing unit. Although different local compressors might receive noisy observations from a / several common source signal(s), each local quantizer should also compress noisy observations from its own, i.e., uncommon source signal(s). This, in turn, yields a highly generalized scheme with most flexibility w.r.t. the assignment of users to the serving nodes, compared to the State-of-the-Art techniques designed exclusively for a common source signal. Following the *Information Bottleneck (IB)* philosophy, we choose the *Mutual Information* as the fidelity criterion here, and, by taking advantage of the *Variational Calculus*, we characterize the form of stationary solutions for two different types of processing flow / strategy. We utilize the derived solutions as the core of our devised algorithmic approach, the *GEneralized Multivariate IB (GEMIB)*, to (efficiently) address the corresponding design problems. We further provide the respective convergence proofs of GEMIB to a stationary point of the pertinent objective functionals and substantiate its effectiveness by means of numerical investigations over a couple of (typical) digital transmission scenarios.

INDEX TERMS 6G, distributed remote source coding, information bottleneck, multi-user data compression.

I. INTRODUCTION

THE *Information Bottleneck (IB)* method for data compression was first introduced in [1]. Its original formulation was based upon the seminal work of Shannon on *lossy* source coding [2] (specifically, the *single-letter* characterization of the *Rate-Distortion* function) with a twist stemming from a fresh and intuitive idea: rather than upper-bounding an expected distortion term, pinpoint a relevant / target variable and lower-bound a *Mutual Information* term. Why? Simply since in plenty of real-world applications in which the data compression must be performed, it is much easier to identify a relevant / target variable (whose information should be retained) than figuring out the proper distortion function. Later on, it was realized (see, e.g., [3]) that the basic constrained optimization problem which the IB framework is established upon, determines the boundary of achievable rate-distortion region for a *remote / indirect* source coding problem with the certain choice of *Logarithmic Loss* distortion [4]. Interested readers are referred to [5] for further in-depth discussions on different aspects of the IB principle (from the standpoints of both the Learning and Information Theory), together with its connections to several other interesting problems, including (but not confined to) the Wyner-Ahlswede-Körner problem [6], [7], the efficiency of investment information [8], and the privacy funnel [9], [10]. Moreover, to get a better view on the more recent works regarding both the theory and applications of the IB method, interested readers are referred to [11].

The IB principle has also been leveraged as a theoretical framework to better understand the underlying dynamics of deep learning models [12], [13], [14]. Further, it has been applied to various aspects of deep learning, from the

optimization of neural network parameters to the design of novel network architectures, and even as an effective means to reduce the problem of overfitting in complex inference tasks [15], [16], [17], [18].

Aside from purely theoretical studies, the IB method has already found its place and been implemented in (various parts of) modern digital data transmission schemes. To mention a few examples, one can list miscellaneous applications, i.a., in Analog-to-Digital converters for receiver front ends [19], in (efficient) construction of Polar Codes [20], [21], in discrete channel decoding schemes [22], [23], [24], in forward-aware vector quantization [25], [26], [27], and, last but not least, in Semantic / Task-Oriented Communications [28], [29], [30].

In the pertinent literature on this subject, it has also been considered how the original IB framework can be generalized to the multiterminal / distributed scenarios (see, e.g., [31], [32], [33], [34], [35]). Such extended schemes mainly focus on a particular setup where multiple *noisy* observations from one *common* source signal are compressed by several intermediate nodes (at potentially different rates, and also according to various strategies) to preserve (collectively) as much information as possible about that source signal. In practice, it happens (rather) frequently that alongside the common source signals, local compression units should serve *uncommon* source signals as well. In this context, a *common* source is referred to the one getting served by at least two intermediate nodes, and an *uncommon* source is the one getting served by only one intermediate node. Extending the IB framework to a hybrid case with both types of sources is the main focus here.

A. CONTRIBUTIONS

Within the scope of this article, we develop novel distributed (remote) source coding schemes for a (generic) multiterminal setup in which several intermediate nodes compress various sets of *noisy* observations from (potentially) common and uncommon source signals, before forwarding their signals via several *error-free* and *rate-limited* channels to a (remote) processing unit. Specifically, based upon the *Information Bottleneck* framework, we choose the *Mutual Information* as the fidelity criterion, and, by means of *Variational Calculus*, we derive the stationary solutions of the challenging design problems. Thereupon, we present an iterative algorithm, the *GEneralized Multivariate IB (GEMIB)*, to efficiently address the design problems, and, via a detailed analysis, we also provide the convergence proofs to a stationary point of the objective functionals. As the final puzzle piece of our comprehensive theoretical support, via an in-depth mathematical analysis, we further justify the behavior of GEMIB over the whole gamut of its main input parameters.

To get a crisp feeling about the generality of our results in this article, it should be noted that the considered distributed scenario appears in a broad variety of applications regarding the fifth (5G) and sixth (6G) generations of wireless network technologies, i.a., in (distributed) inference sensor networks with rate-limited channels to the fusion center [36], in cooperative relaying schemes with the *Quantize-and-Forward* strategy [37], [38], and also in *Cloud-based Radio Access Networks (Cloud-RANs)* [39], [40], as well as *(User-Centric) Cell-Free massive Multiple-Input Multiple-Output ((UC) CF-mMIMO) systems* with limited fronthaul rates [41], [42], [43].

B. OUTLINE

The centralized IB-based noisy source coding is briefly discussed in Section II as a prelude to the distributed extensions. The considered system model is then presented in Section III, together with two distinct design problems for the *parallel* and *successive* processing. In Section IV, two theorems are provided to fully characterize the form of stationary solutions for both processing strategies. Subsequently, in Section V, an iterative algorithm, namely, the *GEneralized Multivariate IB (GEMIB)*, is presented to address both design problems, together with its proof of convergence as well as an in-depth mathematical discussion on the behavior of this algorithm over the entire range of its main parameters. In Section VI, several numerical investigations are presented to corroborate the effectiveness of the proposed approach. Finally, a brief wrap-up in Section VII concludes this article. The detailed proofs of two main theorems are relegated to the Appendix.

C. NOTATIONS

According to the distribution, $p(\mathsf{a})$, the realizations, $a \in \mathcal{A}$, of the (discrete) random variable, a, happen. With boldface counterparts, the same holds true for the (discrete) random vector, $\mathbf{a}_{1:J} = \{\mathsf{a}_1, \ldots, \mathsf{a}_J\}$. Furthermore, $\mathbf{a}_{1:J}^{-j} = \mathbf{a}_{1:J} \setminus \{\mathsf{a}_j\}$, and $\mathbf{Pa}_{\mathsf{a}}^{\mathcal{G}_\bullet}$ denotes the parent nodes of random variable, a, in the Bayesian Network, $\mathcal{G}_\bullet$. Moreover, $D_{\mathrm{KL}}(\cdot\|\cdot)$, $H(\cdot)$, $D_{\mathrm{JS}}^{\{\cdot,\cdot\}}(\cdot\|\cdot)$, and $I(\cdot;\cdot)$, stand for the *Kullback-Leibler (KL)* divergence, the Shannon's entropy, the *Jensen-Shannon (JS)* divergence, and the Mutual Information [44], respectively. Also, the expectation operator is denoted by $\mathbb{E}_\bullet$.

II. IB-BASED NOISY SOURCE CODING: POINT-TO-POINT CASE IN A NUTSHELL

Consider the system model in Fig. 1. Specifically, the noisy observation, y, from a single source, x, shall be compressed at an Intermediate Node (IN) to the signal, z, before getting forwarded through an Ideal Rate-limited Channel (IRC) with the capacity, R, to the Remote Processing Unit (RPU). The interrelation between x and y is established via a Discrete Memoryless Channel (DMC) whose transition probabilities, $p(\mathsf{y}|\mathsf{x})$, as well as its input distribution, $p(\mathsf{x})$, are presumed to be known. By following the *Information Bottleneck (IB)* framework [1], the design problem is then formulated as a basic trade-off between two *Mutual Information* terms.

The first term, $I(\mathsf{x};\mathsf{z})$, is called the *relevant information*, and quantifies the *informativity* of outcome, and the second term, $I(\mathsf{y};\mathsf{z})$, is called the *compression rate*, and quantifies its *compactness*. The goal is then to maximize the relevant

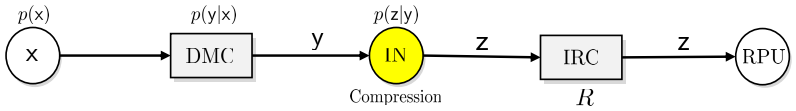


FIGURE 1. System model for IB-based point-to-point noisy source coding. DMC, IN, IRC, and RPU stand for Discrete Memoryless Channel, Intermediate Node, Ideal Rate-limited Channel, and Remote Processing Unit, respectively.

information such that the compression rate does not exceed the capacity, R, of the forward link to RPU. For the design of compressor, $p(\mathsf{z}|\mathsf{y})$, it holds[1]

$$p^*(\mathsf{z}|\mathsf{y}) = \underset{p(\mathsf{z}|\mathsf{y}):\, I(\mathsf{y};\,\mathsf{z})\le R}{\operatorname{argmax}} \; I(\mathsf{x};\,\mathsf{z}), \tag{1}$$

wherein, $0 \le R \le \log_2 |\mathcal{Z}|$ bits, sets an upper-bound on the compression rate. By making use of the method of *Lagrange Multipliers* [45], this design problem can be recast into an unconstrained optimization (up to the validity of compressor mapping)

$$p^*(\mathsf{z}|\mathsf{y}) = \underset{p(\mathsf{z}|\mathsf{y})}{\operatorname{argmax}} \; I(\mathsf{x};\,\mathsf{z}) - \lambda I(\mathsf{y};\,\mathsf{z}), \tag{2}$$

wherein, $\lambda \ge 0$, denotes the counterpart of upper-bound, R, in (1). Given R, the corresponding λ value can be found, e.g., by performing a bi-section search over a finite range. The form of stationary solution for the (non-convex) design problem (2) has been characterized in [1] as

$$p(z|y) = \frac{p(z)}{\psi(y,\beta)} \exp(-\beta D_{\mathrm{KL}}(p(\mathsf{x}|y)\|p(\mathsf{x}|z))), \tag{3}$$

for each pair $(y, z) \in \mathcal{Y} \times \mathcal{Z}$, wherein $\beta = \frac{1}{\lambda}$, and $\psi(y, \beta)$, is a partition function to ensure the compressor mapping's validity. Specifically, for each realization $y \in \mathcal{Y}$, the sum of calculated terms in (3) (ignoring ψ) over all output clusters $z \in \mathcal{Z}$ acts as the partition function. Furthermore, an iterative algorithm has been presented in [1] to address the design problem, carrying out the *Fixed-Point Iterations* [46] on the derived *implicit* solution (3).

III. DISTRIBUTED EXTENSIONS: SYSTEM MODEL AND PROBLEM FORMULATION

We consider the system model illustrated in Fig. 2. A total number, $N_{\text{tot.}}$, of source / user signals must be served by J intermediate nodes. Every node, IN_j with $j \in \{1, \ldots, J\}$, receives (non-interfering) *noisy* observations, $\{\mathsf{y}_{m\ell}^{(j)}\}$, from the set of source signals, $\{\mathsf{x}_{m\ell}\}$, that it must serve, and then quantizes them to a (compressed) representative, z_j, ahead of a forward transmission via an *error-free* link with capacity, R_j, to RPU. In the presented notations, $m \in \{1, \ldots, J\}$, is the index of the intermediate node to which a user is allocated. Further, $\ell \in \{1, \ldots, N_m\}$, is the index of the user within the allocated user set, and, N_m, is the number of users allocated to the m-th intermediate node. Moreover, the interrelation between $\mathsf{x}_{m\ell}$ and $\mathsf{y}_{m\ell}^{(j)}$ is modeled through a DMC whose transition probabilities, $p(\mathsf{y}_{m\ell}^{(j)}|\mathsf{x}_{m\ell})$, and input distribution, $p(\mathsf{x}_{m\ell})$, are presumed to be known.

[1]Refer to [3] for the (asymptotic) remote source coding formulation with the *Logarithmic Loss* distortion function [4].

Following a certain processing flow / strategy (which will be specified in the design formulation), all individual source signals should be retrieved in RPU. One should note that, in this description, a *common* user that gets served by more than one intermediate node, will be allocated to only one of them, such that $\sum_m N_m = N_{\text{tot.}}$, which is the total number of users. This is done just for the sake of a clear enumeration and to simplify the mathematical formulation of the design problems. As another assisting tool in that regard, based upon the utilized formalism in [47], we take advantage of two Bayesian Networks (BNs) as graphical models to portray various aspects of the design problems. More specifically, the input BN, $\mathcal{G}_{\text{in}}$, in Fig. 2 depicts "*what is compressing what*" in a fashion that every compression variable, z_j, quantizes its parents in $\mathcal{G}_{\text{in}}$. Further, the statistical (in)dependencies between all involved random variables are encoded by the structure of $\mathcal{G}_{\text{in}}$ (through the basic rule that every variable is *conditionally* independent of all its non-descendants, given the value of its parents). The output BN, $\mathcal{G}_{\text{out}}$, in Fig. 2, on the other hand, illustrates "*what must retain information about what*" in a sense that each compression variable, z_j, appears as a parent for any source signal it should be informative of.

The design problem is then formulated as a (constrained) optimization, setting a fundamental trade-off between the *informativity* and *compactness* of resultant outcomes. The informativity is naturally quantified by the sum of *Mutual Information* terms between each source signal, $\mathsf{x}_{m\ell}$ with $m \in \{1, \ldots, J\}, \ell \in \{1, \ldots, N_m\}$, and its parents in $\mathcal{G}_{\text{out}}$ (denoted by $\mathbf{v}_{\mathsf{x}_{m\ell}}$), i.e., the set of all compression variables that must preserve information about $\mathsf{x}_{m\ell}$. However, the other side of the trade-off offers no natural, unique choice. Hence, different meaningful expressions can be employed. In the following part, two distinct constraint sets are considered to quantify the compactness of outcomes. Subsequently, by taking advantage of the *Variational Calculus*, we derive the stationary solutions for all (local) compressors and utilize them as the core of our (iterative) algorithm to efficiently address both design problems. The convergence proofs to a stationary point of the pertinent objective functionals will be presented as well.

A. PARALLEL SCHEME: IGNORING SIDE-INFORMATION

As the first choice of the processing strategy, we consider a *parallel* scheme in which no side-information is utilized at RPU, when retrieving the source signals. In this case, the design problem is formulated as the following optimization (with $P^* = \{p^*(\mathsf{z}_1|\mathbf{y}_1), \ldots, p^*(\mathsf{z}_J|\mathbf{y}_J)\}$ and $\mathbf{y}_m = \mathbf{Pa}_{\mathsf{z}_m}^{\mathcal{G}_{\text{in}}}$, i.e.,

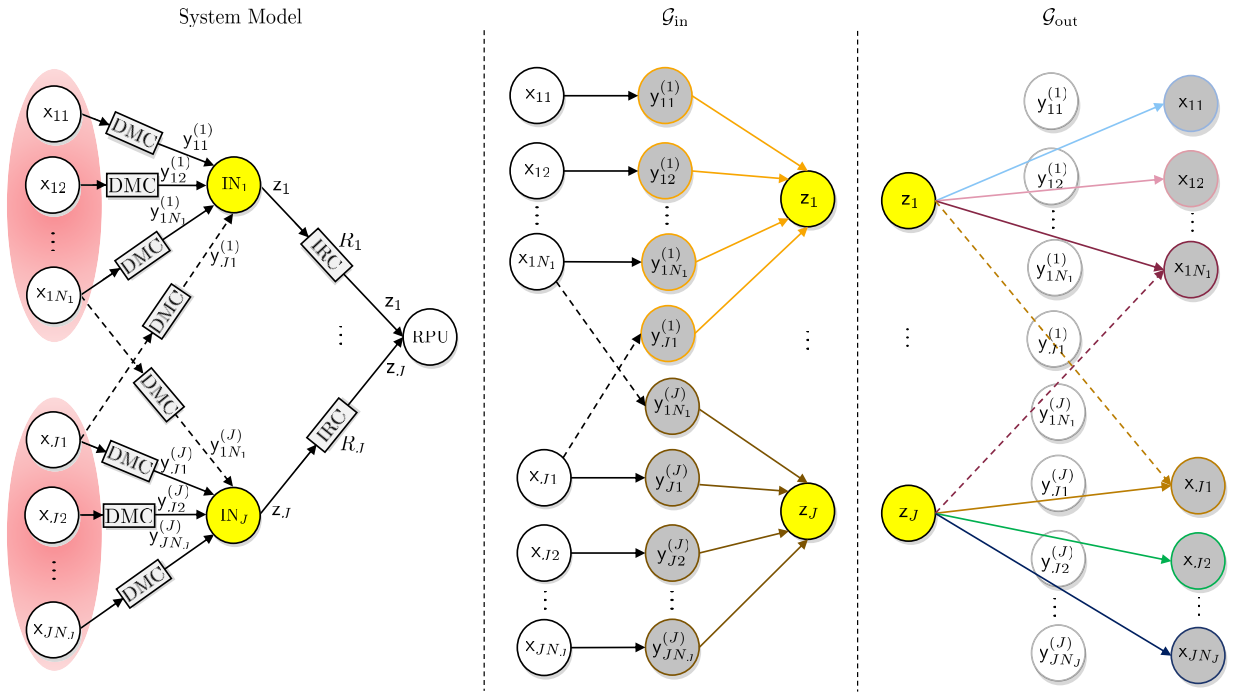


FIGURE 2. Considered system model for distributed noisy source coding, together with the corresponding input/output Bayesian Networks (BNs). The input BN, $\mathcal{G}_{\text{in}}$, portrays the *compression* side and the output BN, $\mathcal{G}_{\text{out}}$, portrays the *information preservation* side. For each common source, the connections to other (i.e., not allocated) serving INs are depicted by dashed lines.

the set of variables to be compressed by the m-th IN)

$$P^* = \underset{P:\,\forall m\, I(\mathbf{y}_m;\,\mathsf{z}_m)\le R_m}{\operatorname{argmax}} \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell};\,\mathbf{v}_{\mathsf{x}_{m\ell}}\big), \tag{4}$$

wherein, $0 \le R_m \le \log_2|\mathcal{Z}_m|$ bits, sets an upper-bound on the m-th compression rate, $I(\mathbf{y}_m;\,\mathsf{z}_m)$.[2] By making use of the method of *Lagrange Multipliers* [45], the design problem (4) can be recast into the unconstrained optimization (up to the validity of all compressor mappings)

$$P^* = \underset{P}{\operatorname{argmax}} \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell};\,\mathbf{v}_{\mathsf{x}_{m\ell}}\big) - \sum_{m=1}^{J}\lambda_m I\big(\mathbf{y}_m;\,\mathsf{z}_m\big), \tag{5}$$

with $\lambda_m \ge 0$, as the counterpart of the rate, R_m, in (4).

B. SUCCESSIVE SCHEME: USING SIDE-INFORMATION

As the second choice of the processing flow, we consider a *successive* scheme wherein the side-information from the already retrieved signals is utilized at RPU, when recovering a particular source signal. Compared to the previous approach, generally, this leads to a superior "*informativity-compactness*" trade-off, but at the expense of (processing) complexity. Essentially, this scheme follows the Wyner-Ziv setup for source coding [48] in which, statistically correlated signals are utilized as side-information at the decoder. The respective design problem is then formulated as

$$P^* = \underset{P:\,\forall m\, I(\mathbf{y}_m;\,\mathsf{z}_m|\mathbf{z}_{1:m-1})\le R_m}{\operatorname{argmax}} \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell};\,\mathbf{v}_{\mathsf{x}_{m\ell}}\big), \tag{6}$$

[2]Besides one-shot formulations, a multi-letter description is required for the (asymptotic) coding problems in Section III.

where, $0 \le R_m \le \log_2|\mathcal{Z}_m|$ bits, sets an upper-bound on the m-th *conditional* compression rate, $I(\mathbf{y}_m;\,\mathsf{z}_m|\mathbf{z}_{1:m-1})$. Note that, here, there is an extra degree of freedom, that is the processing order. Generally, it affects the performance and should be optimized (e.g., through a brute-force search). Henceforth, we continue our discussion with a fixed choice of ordering. Like the parallel processing, by making use of the method of *Lagrange Multipliers* [45], the design problem (6) can be recast into the unconstrained optimization (up to the validity of all compressor mappings)

$$P^* = \underset{P}{\operatorname{argmax}} \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell};\,\mathbf{v}_{\mathsf{x}_{m\ell}}\big) - \sum_{m=1}^{J}\lambda_m I\big(\mathbf{y}_m;\,\mathsf{z}_m|\mathbf{z}_{1:m-1}\big), \tag{7}$$

with $\lambda_m \ge 0$, as the counterpart the rate, R_m, in (6).

Please note that, for the special case of *full-informativity*, corresponding to letting $\lambda_m \to 0$ for $m = 1$ to J, the objective functionals of parallel and successive processing schemes coincide. To clearly perceive this, note that the difference in the objective functionals in (5) and (7) is in their second term that vanishes, when letting $\lambda_m \to 0$.

IV. CHARACTERIZATION OF STATIONARY SOLUTIONS

This part of the article is dedicated to the characterization of stationary solutions of the design problems for parallel and successive processing schemes. As will be perceived later on, these solutions become the core components of our devised iterative algorithm to efficiently tackle both design problems by solving a *Multivariate Fixed-Point System* [46].

A. PARALLEL PROCESSING

The following theorem delivers the stationary solutions for local compressors when addressing the design problem for the parallel processing scheme.

Theorem 1 (Parallel Scheme): Presume that the joint distribution of input variables (i.e., all nodes in $\mathcal{G}_{\text{in}}$ except the leaves) and λ_m are given for all $m \in \{1, \ldots, J\}$. The set of local compressors, $\{p(\mathsf{z}_j|\mathbf{y}_j) \; | j \in \{1, \ldots, J\}\}$, is a stationary point of the Lagrangian for parallel scheme

$$\mathcal{L}_{\text{Par.}} = \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}\big) - \sum_{m=1}^{J} \lambda_m I\big(\mathbf{y}_m; \mathsf{z}_m\big) \tag{8}$$

iff for each pair, $(\boldsymbol{y}_j, z_j) \in \boldsymbol{\mathcal{Y}}_j \times \mathcal{Z}_j$, it holds true that

$$p\big(z_j|\boldsymbol{y}_j\big) = \frac{p\big(z_j\big)}{\psi^{\text{Par.}}_{\mathsf{z}_j}\big(\boldsymbol{y}_j, \beta_j\big)} \exp\big(-d_{\text{Par.}}\big(\boldsymbol{y}_j, z_j\big)\big), \tag{9}$$

where, $\psi^{\text{Par.}}_{\mathsf{z}_j}(\boldsymbol{y}_j, \beta_j)$, is a normalization function that ensures the validity of pertinent quantizer mapping, and the relevant distortion, $d_{\text{Par.}}(\boldsymbol{y}_j, z_j)$, is calculated as

$$\begin{aligned} &d_{\text{Par.}}\big(\boldsymbol{y}_j, z_j\big) \\ &= \beta_j \sum_{(m,\ell):\; \mathsf{z}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}}|\boldsymbol{y}_j\right)} \Big\{ D_{\text{KL}}\Big(p\Big(\mathsf{x}_{m\ell}|\boldsymbol{y}_j, \mathbf{v}^{-j}_{\mathsf{x}_{m\ell}}\Big) \| p\big(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}}\big)\Big)\Big\}, \end{aligned} \tag{10}$$

with $\beta_j = \frac{1}{\lambda_j}$, $\mathbf{y}_j = \mathbf{Pa}^{\mathcal{G}_{\text{in}}}_{\mathsf{z}_j}$, $\mathbf{v}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$, and $\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{v}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$.

The proof has been presented in Appendix-A.

It is noteworthy that Theorem 1 generalizes the obtained results in [49]. There, the input to every local compressor was a noisy observation of a single *common* user. In contrast, here, alongside the potentially common users, different local compressors quantize various sets of noisy observations from different (i.e., uncommon) users as well.

The derived relevant distortion in (10) quantifies the degree of proximity / closeness of the conditional distributions in which $\boldsymbol{y}_j$ is involved to those where $\boldsymbol{y}_j$ is substituted by its compressed representative, z_j. For example, if a certain cluster $z_j \in \mathcal{Z}_j$ behaves more similarly to $\boldsymbol{y}_j \in \boldsymbol{\mathcal{Y}}_j$ compared to another cluster, $z'_j \in \mathcal{Z}_j$, it holds $d_{\text{Par.}}(\boldsymbol{y}_j, z_j) < d_{\text{Par.}}(\boldsymbol{y}_j, z'_j)$, which implies $p(z_j|\boldsymbol{y}_j) > p(z'_j|\boldsymbol{y}_j)$. In other words, if z_j is a good representative of $\boldsymbol{y}_j$ the corresponding membership probability, $p(z_j|\boldsymbol{y}_j)$, is increased accordingly.

B. SUCCESSIVE PROCESSING

The following theorem delivers the stationary solutions for local compressors when addressing the design problem for successive processing.

Theorem 2 (Successive Scheme): Presume that the joint distribution of input variables (i.e., all nodes in $\mathcal{G}_{\text{in}}$ except the leaves) and λ_m are given for all $m \in \{1, \ldots, J\}$. The set of local compressors, $\{p(\mathsf{z}_j|\mathbf{y}_j) \; | j \in \{1, \ldots, J\}\}$, is a stationary point of the Lagrangian for successive scheme

$$\mathcal{L}_{\text{Suc.}} = \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}\big) - \sum_{m=1}^{J} \lambda_m I\big(\mathbf{y}_m; \mathsf{z}_m|\mathbf{z}_{1:m-1}\big) \tag{11}$$

iff for each pair, $(\boldsymbol{y}_j, z_j) \in \boldsymbol{\mathcal{Y}}_j \times \mathcal{Z}_j$, it holds true that

$$p\big(z_j|\boldsymbol{y}_j\big) = \frac{p\big(z_j\big)}{\psi^{\text{Suc.}}_{\mathsf{z}_j}\big(\boldsymbol{y}_j, \beta_j\big)} \exp\big(-d_{\text{Suc.}}\big(\boldsymbol{y}_j, z_j\big)\big), \tag{12}$$

where, $\psi^{\text{Suc.}}_{\mathsf{z}_j}(\boldsymbol{y}_j, \beta_j)$, is a normalization function that ensures the validity of pertinent quantizer mapping, and the relevant distortion, $d_{\text{Suc.}}(\boldsymbol{y}_j, z_j)$, is calculated as

$$\begin{aligned} &d_{\text{Suc.}}\big(\boldsymbol{y}_j, z_j\big) \\ &= \beta_j \sum_{(m,\ell):\; \mathsf{z}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}}|\boldsymbol{y}_j\right)} \Big\{ D_{\text{KL}}\Big(p\Big(\mathsf{x}_{m\ell}|\boldsymbol{y}_j, \mathbf{v}^{-j}_{\mathsf{x}_{m\ell}}\Big) \| p\big(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}}\big)\Big)\Big\} \\ &\quad - \sum_{z_{1:j-1}} p\big(z_{1:j-1}|\boldsymbol{y}_j\big) \log p\big(z_{1:j-1}|z_j\big) \\ &\quad - \beta_j \sum_{k=j+1}^{J} \frac{1}{\beta_k} \sum_{z^{-j}_{1:k}} p\Big(z^{-j}_{1:k}|\boldsymbol{y}_j\Big) \log p\big(z_k|z_{1:k-1}\big), \end{aligned} \tag{13}$$

with $\beta_j = \frac{1}{\lambda_j}$, $\mathbf{y}_j = \mathbf{Pa}^{\mathcal{G}_{\text{in}}}_{\mathsf{z}_j}$, $\mathbf{v}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$, and $\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{v}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$.

The proof has been presented in Appendix-B.

It is noteworthy that Theorem 2 generalizes the presented results in [50]. There, like in [49], the input signal to every local compressor was a noisy observation of a single *common* source signal. In contrast, here, analogous to the setup for the parallel processing, alongside the potentially common source signals, different (local) compressors quantize various sets of noisy observations from different (i.e., uncommon) source signals as well. The main difference, compared to the parallel scheme, lies in the consideration of *side-information* at the compression rates to further leverage the (potentially present) correlations in the output signals of (local) compressors.

This extra level of complexity in the design formulation reflects itself in the obtained stationary solutions. Comparing the derived relevant distortion (13) for successive processing with its counterpart (10) for parallel scheme reveals that it extends it by two extra terms appearing due to conditioning the compression rates which translates into the consideration of side-information in the respective design problem.

V. AN ITERATIVE DESIGN ROUTINE

In this part, we present an iterative algorithm to address the challenging design problems for both parallel and successive processing schemes. To that end, the common structure in the derived stationary solutions will be leveraged. Furthermore, through an in-depth analysis, we provide the proof of convergence (to a stationary point of the objective functionals) for both parallel and successive processing schemes. We further delve into the behavior of our proposed algorithmic approach over the entire range of its main input parameters.

A. THE DEVISED ALGORITHM

To achieve a generic algorithm, note that irrespective of the selected strategy, i.e., the parallel or successive scheme, the calculated stationary solution for each pair, $(\mathbf{y}_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$, of the (local) compressors takes the *implicit* form below

$$p(z_j|\mathbf{y}_j) = \frac{p(z_j)}{\psi^r_{\mathsf{z}_j}(\mathbf{y}_j, \beta_j)} \exp\big(-d_r(\mathbf{y}_j, z_j)\big), \tag{14}$$

with $r \in \{\text{Par.}, \text{Suc.}\}$. All the (local) compressor mappings come into play in the calculation of the relevant distortion, $d_r(\mathbf{y}_j, z_j)$. Therefore, we can interpret the right side of (14) as a functional, featuring all (local) compressors as its input arguments. As a direct result, by sweeping through all (local) compressors, we achieve a non-linear system of equations, extending the structure of *Multivariate Fixed-Point Systems* [46]. Particularly, here, the functionals of multiple mappings replace the functions of multiple variables:

$$\begin{cases} p(\mathsf{z}_1|\mathbf{y}_1) = \mathcal{F}_1\big(p(\mathsf{z}_1|\mathbf{y}_1), \ldots, p(\mathsf{z}_J|\mathbf{y}_J)\big) \\ p(\mathsf{z}_2|\mathbf{y}_2) = \mathcal{F}_2\big(p(\mathsf{z}_1|\mathbf{y}_1), \ldots, p(\mathsf{z}_J|\mathbf{y}_J)\big) \\ \vdots \\ p(\mathsf{z}_J|\mathbf{y}_J) = \mathcal{F}_J\big(p(\mathsf{z}_1|\mathbf{y}_1), \ldots, p(\mathsf{z}_J|\mathbf{y}_J)\big), \end{cases} \tag{15}$$

with $\mathcal{F}_j$, denoting a certain functional. To solve this system, the conventional (iterative) methods can be leveraged. In this article, we propose an algorithm, featuring the *synchronous* updating rule, which is reminiscent of the *Jacobi* method for linear systems [46].

Particularly, we initialize with a set of random mappings, $\{p^{(0)}(\mathsf{z}_j|\mathbf{y}_j)\ \ |j\}$. Subsequently, (till convergence by $\varepsilon \ll 1$) we update the mappings for local compressors iteratively via

$$p^{(i+1)}(z_j|\mathbf{y}_j) = \frac{p^{(i)}(z_j)}{\psi^{r\,(i+1)}_{\mathsf{z}_j}(\mathbf{y}_j, \beta_j)} \exp\Big(-d^{(i)}_r(\mathbf{y}_j, z_j)\Big), \tag{16}$$

wherein, i, denotes the iteration counter. By the synchronous updating, it is meant that at each iteration, all local compressor mappings, $\{p^{(i+1)}(\mathsf{z}_j|\mathbf{y}_j)\ \ |j\}$, are updated based upon the previous configuration of the same set, i.e., $\{p^{(i)}(\mathsf{z}_j|\mathbf{y}_j)\ \ |j\}$. The pertinent pseudo-code of this algorithm, the *GEneralized Multivariate IB (GEMIB)*, has been presented in Alg. 1.

In the following section, we show that GEMIB converges to a stationary point of the functionals for both parallel and successive schemes. Since the quality of outcome depends on the choice of initialization, as a popular workaround to avoid poor local optima, one can repeat GEMIB with different starting points, $\{p^{(0)}(\mathsf{z}_j|\mathbf{y}_j)\ \ |j\}$, and retain the best outcome.

B. CONVERGENCE PROOFS

Herein, first we reformulate the design problems for both parallel and successive processing schemes as an alternating minimization w.r.t. the set of all compressors, P, and another set of (auxiliary) distributions, Q. To do so, we introduce a (tight) *variational upper-bound*, $\bar{\mathcal{F}}_r(P, Q)$, on the pertinent objective functional, $\mathcal{F}_r(P)$. Then, with an unfolding trick, we show that the main update step of our proposed iterative algorithm is obtained by merging the respective updates for P and Q. By this, it is inferred that, principally, GEMIB lies within the class of *successive upper-bound minimization*[3] [51]. Consequently, its convergence to a stationary point is immediately ensured. It is noteworthy that similar road maps have been followed in [33] and [34] to verify the convergence of their devised algorithms.

Algorithm 1 GEneralized Multivariate IB (GEMIB)

Input: Joint input distribution, convergence parameter $\varepsilon > 0$, $\beta_j = \frac{1}{\lambda_j} > 0$, $r \in \{\text{Par.}, \text{Suc.}\}$

Output: A (generally soft) partition z_j of $\mathcal{Y}_j$ into $|\mathcal{Z}_j|$ bins

Initialization: $i = 0$, random mappings $\{p^{(0)}(\mathsf{z}_j|\mathbf{y}_j)\ \ |j\}$

while True **do**

for $j = 1{:}J$ **do**

- $p^{(i)}(z_j) \leftarrow \sum_{\mathbf{y}_j} p^{(i)}(z_j|\mathbf{y}_j)p(\mathbf{y}_j) \quad \forall z_j \in \mathcal{Z}_j$
- Calculate the i-th update for all distributions involved in the relevant distortion $d_r(\mathbf{y}_j, \mathsf{z}_j)$
- $p^{(i+1)}(z_j|\mathbf{y}_j) \leftarrow \frac{p^{(i)}(z_j)}{\psi^{r\,(i+1)}_{\mathsf{z}_j}(\mathbf{y}_j, \beta_j)} \exp\Big(-d^{(i)}_r(\mathbf{y}_j, z_j)\Big)$

end for

if $\forall j, \forall \mathbf{y}_j{:}D_{\text{JS}}^{\{\frac{1}{2}, \frac{1}{2}\}}\big(p^{(i+1)}(\mathsf{z}_j|\mathbf{y}_j) \| p^{(i)}(\mathsf{z}_j|\mathbf{y}_j)\big) \le \varepsilon$ **then**

Break

else

$i \leftarrow i + 1$

end if

end while

1) PARALLEL PROCESSING

We can recast the design problem for parallel processing into minimizing the following functional over P

$$\begin{aligned} \mathcal{F}_{\text{Par.}}(P) &= \sum_{m=1}^{J} \sum_{\ell=1}^{N_m} H(\mathsf{x}_{m\ell}) - \mathcal{L}_{\text{Par.}} \\ &= \sum_{m=1}^{J} \lambda_m I\big(\mathbf{y}_m; \mathsf{z}_m\big) + \sum_{m=1}^{J} \sum_{\ell=1}^{N_m} H\big(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}}\big). \end{aligned} \tag{17}$$

By defining a set of auxiliary conditional probability distributions $Q = \{q(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}})\ \ |m \in \{1, \ldots, J\}, \ell \in \{1, \ldots, N_m\}\}$ and the functional

$$\begin{aligned} \bar{\mathcal{F}}_{\text{Par.}}(P, Q) = &\sum_{m=1}^{J} \lambda_m I\big(\mathbf{y}_m; \mathsf{z}_m\big) \\ &- \sum_{m=1}^{J} \sum_{\ell=1}^{N_m} \mathbb{E}_{\mathsf{x}_{m\ell}, \mathbf{v}_{\mathsf{x}_{m\ell}}} \big\{\log q\big(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}}\big)\big\}, \end{aligned} \tag{18}$$

the next four Lemmas hold:

[3]The underlying idea is to optimize a sequence of approximate objective function(al)s (which satisfy some mild assumptions), instead of directly optimizing the original *non-convex* and/or *non-smooth* objective function(al).

Lemma 1: The following equivalence holds

$$\min_{P} \mathcal{F}_{\text{Par.}}(P) = \min_{P} \min_{Q} \bar{\mathcal{F}}_{\text{Par.}}(P, Q) \ . \tag{19}$$

Proof: The difference of $\mathcal{F}_{\text{Par.}}(P)$ and $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$ equals

$$\begin{aligned} &\bar{\mathcal{F}}_{\text{Par.}}(P, Q) - \mathcal{F}_{\text{Par.}}(P) \\ &= \sum_{m=1}^{J} \sum_{\ell=1}^{N_m} \sum_{\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}} p\big(\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big) D_{\text{KL}}\big(p\big(\mathbf{x}_{m\ell}|\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big) \| q\big(\mathbf{x}_{m\ell}|\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big)\big) \geq 0, \end{aligned} \tag{20}$$

with the equality iff $q(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}) = p(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}})$ for every $m \in \{1, \ldots, J\}$ and $\ell \in \{1, \ldots, N_m\}$.

Lemma 2: The functional $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$ is separately convex in P and Q.

Proof: It is immediately deduced from the application of *log-sum* inequality [44].

Lemma 3: Given P, a unique Q minimizes $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$, namely,

$$q^*\big(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}\big) = p\big(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}\big), \tag{21}$$

with $p(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}})$, getting calculated from P.

Proof: It is directly inferred from the proof of Lemma 1.

Lemma 4: Given Q, there exists a P which minimizes $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$, namely,

$$p^*\big(z_j|\mathbf{y}_j\big) = \frac{p\big(z_j\big)}{\bar{\psi}^{\text{Par.}}_{\mathbf{z}_j}\big(\mathbf{y}_j, \beta_j\big)} \exp\big(-\bar{d}_{\text{Par.}}\big(\mathbf{y}_j, z_j\big)\big), \tag{22}$$

for each $(\mathbf{y}_j, z_j) \in \boldsymbol{\mathcal{Y}}_j \times \mathcal{Z}_j$, with $\bar{\psi}^{\text{Par.}}_{\mathbf{z}_j}(\mathbf{y}_j, \beta_j)$, acting as the partition function which ensures the validity of compressor mapping. The distortion, $\bar{d}_{\text{Par.}}(\mathbf{y}_j, z_j)$, is calculated by

$$\begin{aligned} &\bar{d}_{\text{Par.}}\big(\mathbf{y}_j, z_j\big) \\ &= \beta_j \sum_{(m,\ell):\, \mathbf{z}_j \in \mathbf{v}_{\mathbf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{\nu}^{-j}_{\mathbf{x}_{m\ell}}|\mathbf{y}_j\right)} \Big\{ D_{\text{KL}}\Big(p\Big(\mathbf{x}_{m\ell}|\mathbf{y}_j, \boldsymbol{\nu}^{-j}_{\mathbf{x}_{m\ell}}\Big) \| q\big(\mathbf{x}_{m\ell}|\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big)\Big)\Big\}. \end{aligned} \tag{23}$$

Proof: The derivation follows the same road map as in the proof of Theorem 1, noting that

$$\begin{aligned} &\frac{\delta\Big(\sum_{m=1}^{J}\sum_{\ell=1}^{N_m} \mathbb{E}_{\mathbf{x}_{m\ell}, \mathbf{v}_{\mathbf{x}_{m\ell}}} \big\{\log q\big(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}\big)\big\}\Big)}{\delta p\big(\mathbf{z}_j|\mathbf{y}_j\big)} = p\big(\mathbf{y}_j\big) \\ &\times \sum_{(m,\ell):\, \mathbf{z}_j \in \mathbf{v}_{\mathbf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{\nu}^{-j}_{\mathbf{x}_{m\ell}}|\mathbf{y}_j\right)} \Big\{ \sum_{x_{m\ell}} p\Big(x_{m\ell}|\mathbf{y}_j, \boldsymbol{\nu}^{-j}_{\mathbf{x}_{m\ell}}\Big) \log q\big(x_{m\ell}|\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big) \Big\}. \end{aligned} \tag{24}$$

When we merge the results for P and Q from the last two Lemmas, we obtain the main update step of our proposed algorithm for every (local) compressor. Hence, it is ensured by [51, Thm 1] that GEMIB converges to a stationary point since $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$ and $\mathcal{F}_{\text{Par.}}(P)$ satisfy [51, Proposition 1]. ■

2) SUCCESSIVE PROCESSING

Analogous to the previous case, we can recast the respective design problem for successive processing into minimizing the following functional over P

$$\begin{aligned} \mathcal{F}_{\text{Suc.}}(P) = \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} H(\mathbf{x}_{m\ell}) - \mathcal{L}_{\text{Suc.}} &= \sum_{m=1}^{J} \lambda_m I\big(\mathbf{y}_m; \mathbf{z}_m\big) \\ &- \sum_{m=2}^{J} \lambda_m I\big(\mathbf{z}_m; \mathbf{z}_{1:m-1}\big) + \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} H\big(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}\big). \end{aligned} \tag{25}$$

By defining $Q = \{q(\mathbf{z}_2|\mathbf{z}_1), \ldots, q(\mathbf{z}_J|\mathbf{z}_{1:J-1}), q(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}})\}$ for $m \in \{1, \ldots, J\}$, $\ell \in \{1, \ldots, N_m\}$ and the functional

$$\begin{aligned} \bar{\mathcal{F}}_{\text{Suc.}}(P, Q) = &\sum_{m=1}^{J} \lambda_m I\big(\mathbf{y}_m; \mathbf{z}_m\big) \\ &- \sum_{m=2}^{J} \lambda_m \, \mathbb{E}_{\mathbf{z}_{1:m}} \left\{ \log \frac{q(\mathbf{z}_m|\mathbf{z}_{1:m-1})}{p(\mathbf{z}_m)} \right\} \\ &- \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} \mathbb{E}_{\mathbf{x}_{m\ell}, \mathbf{v}_{\mathbf{x}_{m\ell}}} \big\{\log q\big(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}\big)\big\}, \end{aligned} \tag{26}$$

the next four Lemmas hold:

Lemma 5: The following equivalence holds

$$\min_{P} \mathcal{F}_{\text{Suc.}}(P) = \min_{P} \min_{Q} \bar{\mathcal{F}}_{\text{Suc.}}(P, Q) \ . \tag{27}$$

Proof: The difference of $\mathcal{F}_{\text{Suc.}}(P)$ and $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$ equals

$$\begin{aligned} &\bar{\mathcal{F}}_{\text{Suc.}}(P, Q) - \mathcal{F}_{\text{Suc.}}(P) \\ &= \sum_{m=1}^{J}\sum_{\ell=1}^{N_m}\sum_{\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}} p\big(\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big) D_{\text{KL}}\big(p\big(\mathbf{x}_{m\ell}|\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big) \| q\big(\mathbf{x}_{m\ell}|\boldsymbol{\nu}_{\mathbf{x}_{m\ell}}\big)\big) \\ &\quad + \sum_{m=2}^{J} \lambda_m \sum_{z_{1:m-1}} p(z_{1:m-1}) D_{\text{KL}}(p(\mathbf{z}_m|z_{1:m-1}) \| q(\mathbf{z}_m|z_{1:m-1})) \\ &\geq 0, \end{aligned} \tag{28}$$

where the equality holds true iff $q(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}) = p(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}})$ and $q(\mathbf{z}_m|\mathbf{z}_{1:m-1}) = p(\mathbf{z}_m|\mathbf{z}_{1:m-1})$ for $m \in \{1, \ldots, J\}$ and $\ell \in \{1, \ldots, N_m\}$.

Lemma 6: $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$ is separately convex in P and Q.

Proof: It is immediately deduced from the application of *log-sum* inequality [44].

Lemma 7: Given P, a unique Q minimizes $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$, namely,

$$\begin{aligned} q^*\big(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}\big) &= p\big(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}}\big) \\ q^*(\mathbf{z}_m|\mathbf{z}_{1:m-1}) &= p(\mathbf{z}_m|\mathbf{z}_{1:m-1}), \end{aligned} \tag{29}$$

for $m \in \{1, \ldots, J\}$, $\ell \in \{1, \ldots, N_m\}$, with $p(\mathbf{x}_{m\ell}|\mathbf{v}_{\mathbf{x}_{m\ell}})$ and $p(\mathbf{z}_m|\mathbf{z}_{1:m-1})$, calculated from P.

Proof: It is directly deduced from the proof of Lemma 5.

Lemma 8: Given Q, there exists a P which maximizes $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$, namely,

$$p^*\big(z_j|\mathbf{y}_j\big) = \frac{p\big(z_j\big)}{\bar{\psi}^{\text{Suc.}}_{\mathbf{z}_j}\big(\mathbf{y}_j, \beta_j\big)} \exp\big(-\bar{d}_{\text{Suc.}}\big(\mathbf{y}_j, z_j\big)\big), \tag{30}$$

for each $(\boldsymbol{y}_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$, with $\bar{\psi}^{\text{Suc.}}_{\mathsf{z}_j}(\boldsymbol{y}_j, \beta_j)$, acting as a partition function which ensures the validity of compressor mapping. The distortion, $\bar{d}_{\text{Suc.}}(\boldsymbol{y}_j, z_j)$, is calculated by

$$\begin{aligned}
&\bar{d}_{\text{Suc.}}(\boldsymbol{y}_j, z_j) \\
&= \beta_j \sum_{(m,\ell):\, \mathsf{z}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}} | \boldsymbol{y}_j\right)} \left\{ D_{\text{KL}}\left(p\left(\mathsf{x}_{m\ell} | \boldsymbol{y}_j, \boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}\right) \| q\left(\mathsf{x}_{m\ell} | \boldsymbol{v}_{\mathsf{x}_{m\ell}}\right)\right)\right\} \\
&\quad - \sum_{z_{1:j-1}} p\left(z_{1:j-1} | \boldsymbol{y}_j\right) \log \frac{q\left(z_j | z_{1:j-1}\right)}{p\left(z_j\right)} \\
&\quad - \beta_j \sum_{k=j+1}^{J} \frac{1}{\beta_k} \sum_{z^{-j}_{1:k}} p\left(z^{-j}_{1:k} | \boldsymbol{y}_j\right) \log q\left(z_k | z_{1:k-1}\right). \qquad (31)
\end{aligned}$$

Proof: The derivation follows the same road map as in the proof of Theorem 2, noting that

$$\begin{aligned}
&\frac{\delta\left(\mathbb{E}_{\mathbf{z}_{1:j}}\left\{\log \frac{q\left(\mathsf{z}_j | \mathbf{z}_{1:j-1}\right)}{p\left(\mathsf{z}_j\right)}\right\}\right)}{\delta p\left(\mathsf{z}_j | \mathbf{y}_j\right)} \\
&= p\left(\boldsymbol{y}_j\right) \sum_{z_{1:j-1}} p\left(z_{1:j-1} | \boldsymbol{y}_j\right) \log \frac{q\left(z_j | z_{1:j-1}\right)}{p\left(z_j\right)}, \qquad (32)
\end{aligned}$$

and for $j < m \leq J$

$$\begin{aligned}
&\frac{\delta\left(\mathbb{E}_{\mathbf{z}_{1:m}}\left\{\log \frac{q\left(\mathsf{z}_m | \mathbf{z}_{1:m-1}\right)}{p\left(\mathsf{z}_m\right)}\right\}\right)}{\delta p\left(\mathsf{z}_j | \mathbf{y}_j\right)} \\
&= p\left(\boldsymbol{y}_j\right) \sum_{z^{-j}_{1:m}} p\left(z^{-j}_{1:m} | \boldsymbol{y}_j\right) \log \frac{q\left(z_m | z_{1:m-1}\right)}{p\left(z_m\right)}. \qquad (33)
\end{aligned}$$

Like the previous case, when we merge the results for P and Q from the last two Lemmas, we obtain the main update step of our proposed algorithm for every (local) compressor. Consequently, it is ensured by [51, Thm 1] that GEMIB converges to a stationary solution since $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$ and $\mathcal{F}_{\text{Suc.}}(P)$ satisfy [51, Proposition 1] as well. ■

C. MATHEMATICAL DISCUSSION

In this part, we provide some insights into the behavior of the GEMIB algorithm over the whole range of its main input parameters, thereby answering the important questions of what to expect from the algorithm and, more importantly, how to justify its observed behavior when working in various regimes of the input parameters. Further, we shortly discuss an important extension of the GEMIB algorithm to the case of uneven user signal recovery preferences.

1) Presuming fixed $p(\mathsf{z}_m|\mathbf{y}_m)$ and (finite) λ_m for all $m = 1 \text{ to } J$ and $m \neq j$, and by letting $\beta_j \to 0$, the design problem for parallel scheme boils down to the minimization of j-th compression rate, $I(\mathbf{y}_j; \mathsf{z}_j)$, w.r.t. the j-th local compressor, $p(\mathsf{z}_j|\mathbf{y}_j)$. This extreme case leads to the state of *full diffusion* in which, each realization, $\boldsymbol{y}_j \in \mathcal{Y}_j$, is allocated equiprobably to all output clusters, $z_j \in \mathcal{Z}_j$. In this fashion, the input and output of the j-th (local) compressor become statistically independent, and, consequently, the pertinent compression rate, $I(\mathbf{y}_j; \mathsf{z}_j)$, becomes zero that is its global minimum.

When β_j takes finite (non-zero) values, usually the compressor mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$, becomes *soft / stochastic*. On the other hand, by letting $\beta_j \to \infty$, a *hard / deterministic* mapping is generated, corresponding to the state of *full concentration* for this other extreme case. To justify this behavior, note that by letting $\beta_j \to \infty$, the design problem for parallel processing boils down to the following optimization

$$p^*\left(\mathsf{z}_j|\mathbf{y}_j\right) = \underset{p(\mathsf{z}_j|\mathbf{y}_j)}{\text{argmax}} \sum_{(m,\ell):\, \mathsf{z}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} I\left(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}\right), \qquad (34)$$

that is a *convex maximization* problem. To realize that, first it must be noted that $I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}})$ is convex w.r.t. $p(\mathbf{v}_{\mathsf{x}_{m\ell}}|\mathsf{x}_{m\ell})$, when $p(\mathsf{x}_{m\ell})$ is given [44]. Since the interrelation between $p(\mathbf{v}_{\mathsf{x}_{m\ell}}|\mathsf{x}_{m\ell})$ and $p(\mathsf{z}_j|\mathbf{y}_j)$ is established through

$$\begin{aligned}
p\left(\mathbf{v}_{\mathsf{x}_{m\ell}}|\mathsf{x}_{m\ell}\right) &= p\left(\mathsf{z}_j|\mathsf{x}_{m\ell}\right) p\left(\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathsf{x}_{m\ell}\right) \\
&= \left(\sum_{\boldsymbol{y}_j} p\left(\mathsf{z}_j|\boldsymbol{y}_j\right) p\left(\boldsymbol{y}_j|\mathsf{x}_{m\ell}\right)\right) p\left(\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathsf{x}_{m\ell}\right), \qquad (35)
\end{aligned}$$

which is an *affine* transform preserving convexity [52, Sec. 3.2], it is deduced that $I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}})$ is also convex w.r.t. $p(\mathsf{z}_j|\mathbf{y}_j)$. Further, as the sum of several convex functions is also a convex function, it is immediately inferred that the objective in (34) is a convex function of the j-th (local) compressor, $p(\mathsf{z}_j|\mathbf{y}_j)$. A (quite well-known) theorem from *convex maximization* theory asserts that a convex function that is defined over a closed and convex set reaches its global maximum at an *extreme point* of that set [53, Ch. 4]. This theorem implies that it is sufficient to solely focus on the *hard / deterministic* mappings in this case (i.e., when letting $\beta_j \to \infty$). To clearly discern this, note that the search space, i.e., the space of (valid) probability mappings, $p(\mathsf{z}_j|\mathbf{y}_j)$, is a closed and convex polytope which is generated from the Cartesian product of $|\mathcal{Y}_j|$ probability simplices [54]. The corners / vertices of this polytope are its extreme points, and each corner corresponds to the Cartesian product of some corners of the constituent probability simplices, yielding a *hard / deterministic* mapping in the end.

2) Regarding the successive processing, a similar behavior is observed, and hence, an analogous justification is made. Particularly, once again, by presuming fixed $p(\mathsf{z}_m|\mathbf{y}_m)$ and (finite) λ_m for all $m = 1 \text{ to } J$ and $m \neq j$, by letting $\beta_j \to 0$, the design problem for successive processing boils down to the minimization of the j-th conditional compression rate, $I(\mathbf{y}_j; \mathsf{z}_j|\mathbf{z}_{1:j-1})$, w.r.t. the j-th (local) compressor, $p(\mathsf{z}_j|\mathbf{y}_j)$. Like the parallel processing scheme, this extreme case also leads to the state of *full diffusion* in which, every realization, $\boldsymbol{y}_j \in \mathcal{Y}_j$, is allocated equiprobably to all output bins / clusters, $z_j \in \mathcal{Z}_j$. In this manner, the input and output of the j-th local compressor become statistically independent. Thus, the conditional compression rate, $I(\mathbf{y}_j; \mathsf{z}_j|\mathbf{z}_{1:j-1})$, becomes zero, since it is non-negative and upper-bounded by $I(\mathbf{y}_j; \mathsf{z}_j) = 0$.

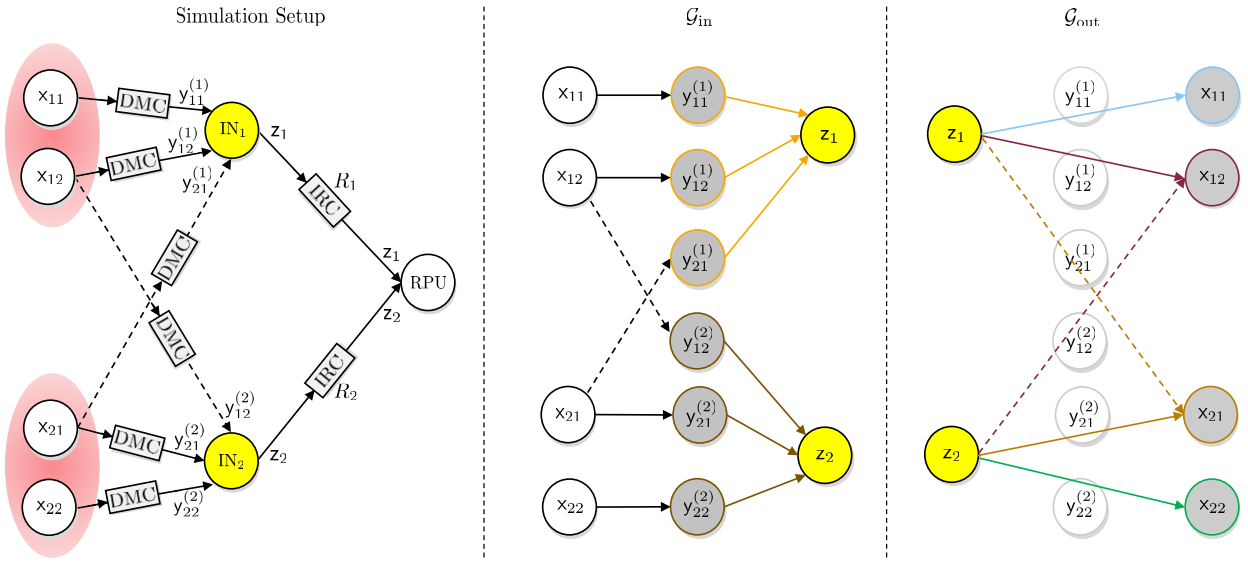


FIGURE 3. Considered setup for numerical simulations regarding distributed noisy source coding with two common (x_{12} and x_{21}) and two uncommon (x_{11} and x_{22}) sources, along with the corresponding input/output BNs.

Similarly, when β_j takes finite (non-zero) values, usually the compressor mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$, becomes *soft/stochastic*. On the other hand, by letting $\beta_j \to \infty$, the state of *full concentration* is reached wherein *hard/deterministic* mappings are obtained. An analogous line of argumentation as in the parallel processing applies here, too. By letting $\beta_j \to \infty$, the design problem for successive scheme boils down to[4]

$$p^*(\mathsf{z}_j|\mathbf{y}_j) = \underset{p(\mathsf{z}_j|\mathbf{y}_j)}{\operatorname{argmax}} \sum_{(m,\ell):\, \mathsf{z}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} I\big(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}\big) + \sum_{m=j+1}^{J} \lambda_m I(\mathsf{z}_m; \mathbf{z}_{1:m-1}). \tag{36}$$

This, again, is a *convex maximization* problem. To perceive that, it totally suffices to show that $I(\mathsf{z}_m; \mathbf{z}_{1:m-1})$ is a convex function of the j-th (local) compressor mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$, as if so, the second term on the right side of (36) is also convex w.r.t. $p(\mathsf{z}_j|\mathbf{y}_j)$, since λ_m is non-negative and the sum of several convex functions is also a convex function. To conclude the claimed proposition's proof, note that $I(\mathsf{z}_m; \mathbf{z}_{1:m-1})$ is a convex function of $p(\mathbf{z}_{1:m-1}|\mathsf{z}_m)$, when $p(\mathsf{z}_m)$ is given [44]. Since the relation between $p(\mathbf{z}_{1:m-1}|\mathsf{z}_m)$ and $p(\mathsf{z}_j|\mathbf{y}_j)$ is established by

$$p(\mathbf{z}_{1:m-1}|\mathsf{z}_m) = \frac{\sum_{\mathbf{y}_{1:m}} p(\mathbf{y}_{1:m}) p\big(\mathbf{z}_{1:m}^{-j}|\mathbf{y}_{1:m}^{-j}\big) p(\mathsf{z}_j|\mathbf{y}_j)}{p(\mathsf{z}_m)}, \tag{37}$$

which is an *affine* transform that preserves convexity, it is inferred that $I(\mathsf{z}_m; \mathbf{z}_{1:m-1})$ is also convex w.r.t. $p(\mathsf{z}_j|\mathbf{y}_j)$.

[4]To clearly see this, note that, from the Markovian relations, it is deduced that $I(\mathbf{y}_m; \mathsf{z}_m|\mathbf{z}_{1:m-1}) = I(\mathbf{y}_m; \mathsf{z}_m) - I(\mathsf{z}_m; \mathbf{z}_{1:m-1})$.

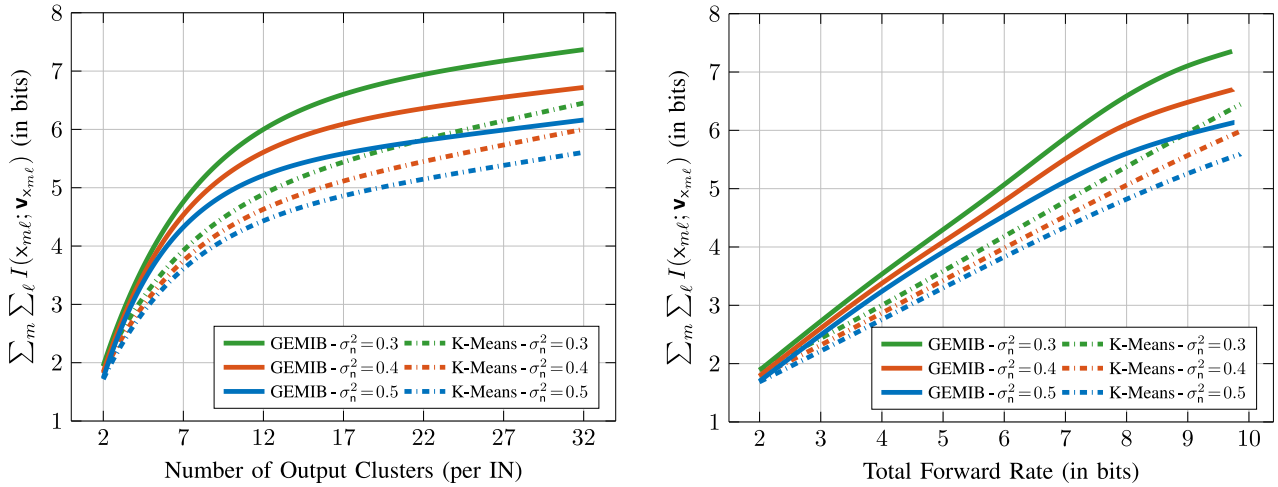
3) Finally, it is worth mentioning that GEMIB can be easily adapted to the case in which, via a "normalized preference" set, $\{0 \le \alpha_{m\ell} \le 1 \ | m \in \{1, \ldots, J\}, \ell \in \{1, \ldots, N_m\}\}$, the user signal recoveries are prioritized based upon their importance. In this case, the *informativity* is quantified by the *weighted* sum of relevant information terms, i.e.,

$$\sum_{m=1}^{J} \sum_{\ell=1}^{N_m} \alpha_{m\ell}\, I\big(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}\big). \tag{38}$$

This does not change the form of derived stationary solutions for both parallel and successive processing schemes in (10) and (13), up to a multiplicative prefactor, $\alpha_{m\ell}$, in front of the expected values in the first term of both solutions. Hence, in case of *uneven* recovery preferences, the corresponding set, $\{0 \le \alpha_{m\ell} \le 1 \ | m \in \{1, \ldots, J\}, \ell \in \{1, \ldots, N_m\}\}$, is fed as another input to the GEMIB algorithm.

VI. NUMERICAL RESULTS

In this section, by performing several numerical simulations, we intend to corroborate the effectiveness of our proposed (distributed) multi-source compression schemes. Throughout our investigations, we focus on a particular setup, illustrated in Fig. 3, including four user signals and two intermediate nodes. Specifically, two of these user signals, x_{11} and x_{22}, are uncommon as they get served by a single intermediate node, namely, IN_1 and IN_2, respectively. The other two source signals are common since they get served by both intermediate nodes. We (arbitrarily) assign x_{12} to IN_1, and x_{21} to IN_2. To formalize both *compression* and *informativity* sides of the design problem(s), the pertinent input/output BNs have also been presented in Fig. 3, indicating that each intermediate node should jointly compress its incoming *noisy*



(a) Total relevant information vs. compressors' output cardinality

(b) Total relevant information vs. total required forward rate

FIGURE 4. Total relevant information of GEMIB (Parallel) and K-Means vs. a) number of output clusters and b) total forward rate. Equiprobable bipolar 4-ASK source signaling ($\sigma^2_{\mathsf{x}_{m\ell}} = 5$) over AWGN access channels ($\sigma_n^2 = 0.3, 0.4, 0.5$), with $\lambda_m = 0.01$ for $m = 1, 2$, and convergence parameter $\epsilon = 10^{-3}$.

observations, while preserving information about the original source signals (being served by it).

In the following part, first we present the *overall* dynamics of our devised IB-based compression scheme and compare it with one of the most popular methods regarding the vector quantization, i.e., the K-Means algorithm [55]. Subsequently, we present the *individual* dynamics of all four source signals. For these investigations, we focus on the parallel processing scheme. Finally, we compare the obtained performance from both parallel and successive processing schemes, confirming the fact that by leveraging the potential correlations in signals of different intermediate nodes, the latter can yield a superior "*information-compression*" trade-off.

A. OVERALL SYSTEM DYNAMICS

For this investigation, we consider a standard bipolar 4-ASK (Amplitude Shift Keying) source signaling for all four users. Moreover, for all access connections from the users to the intermediate nodes, we consider a DMC that approximates a discrete-time, discrete-input, and continuous-output AWGN (Additive White Gaussian Noise) channel, characterized by the same noise variance, σ_n^2. To get the transition probability matrices, rather than prequantizing the output signals and by following a purely "*Monte Carlo*" approach, 40 samples are generated per access link. We set both trade-off parameters, λ_1 and λ_2 to 0.01, indicating that through a hard clustering, the focus will be mainly on the information preservation.

To perform vector quantization at each intermediate node, we applied the following two algorithms: GEMIB (parallel) and standard K-Means [55]. To avoid poor local optima, we initialized both algorithms 100 times and retained the best outcome. We repeated this procedure for 100 regenerations of access transition matrices and averaged the outcomes. The obtained results have been illustrated in Fig. 4.

Specifically, in Fig. 4a, we illustrated the obtained total relevant information, namely, $\sum_m \sum_\ell I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}})$, when varying the number of output bins / clusters (per intermediate node). In Fig. 4b, we presented the respective overall system dynamics in the "*information-compression*" plane. As the main takeaway from both results, it is clearly observed that the GEMIB (parallel) algorithm yields superior performance compared to the standard K-Means routine. This performance superiority reflects itself in both aspects of the *informativity* and *compactness*. As a concrete example, by considering the depicted results for $\sigma_n^2 = 0.3$, it is observed that, to support 5 bits of total relevant information, the obtained solution with the GEMIB algorithm requires around 6 bits of total forward rate, that is notably less than that of the K-Means algorithm, which requires around 7.4 bits. If, on the other hand, the total forward rate is fixed to 6 bits, it is observed that the GEMIB algorithm can support up to around 5 bits of total relevant information, that is notably more than the K-Means algorithm, which can only support up to 4.2 bits.

Expectedly, by loosening the compression bottleneck via increasing the number of output clusters, the performance (in terms of the overall relevant information) enhances. Naturally, the same holds true for better access link qualities, corresponding to lower σ_n^2 values, since by increasing the capacity of access channels, more information about user signals will be flown into the system. This, in turn, leads to a *nested* sequence of achievable rate-information regions (note that the union of all the points on and below each curve in Fig. 4b constitutes the achievable region for a given σ_n^2).

B. INDIVIDUAL SYSTEM DYNAMICS

In this part, we intend to investigate the individual dynamics regarding various source signals. To that end, we apply the same procedure described for the overall system dynamics in the previous part (which we do not repeat for the sake of

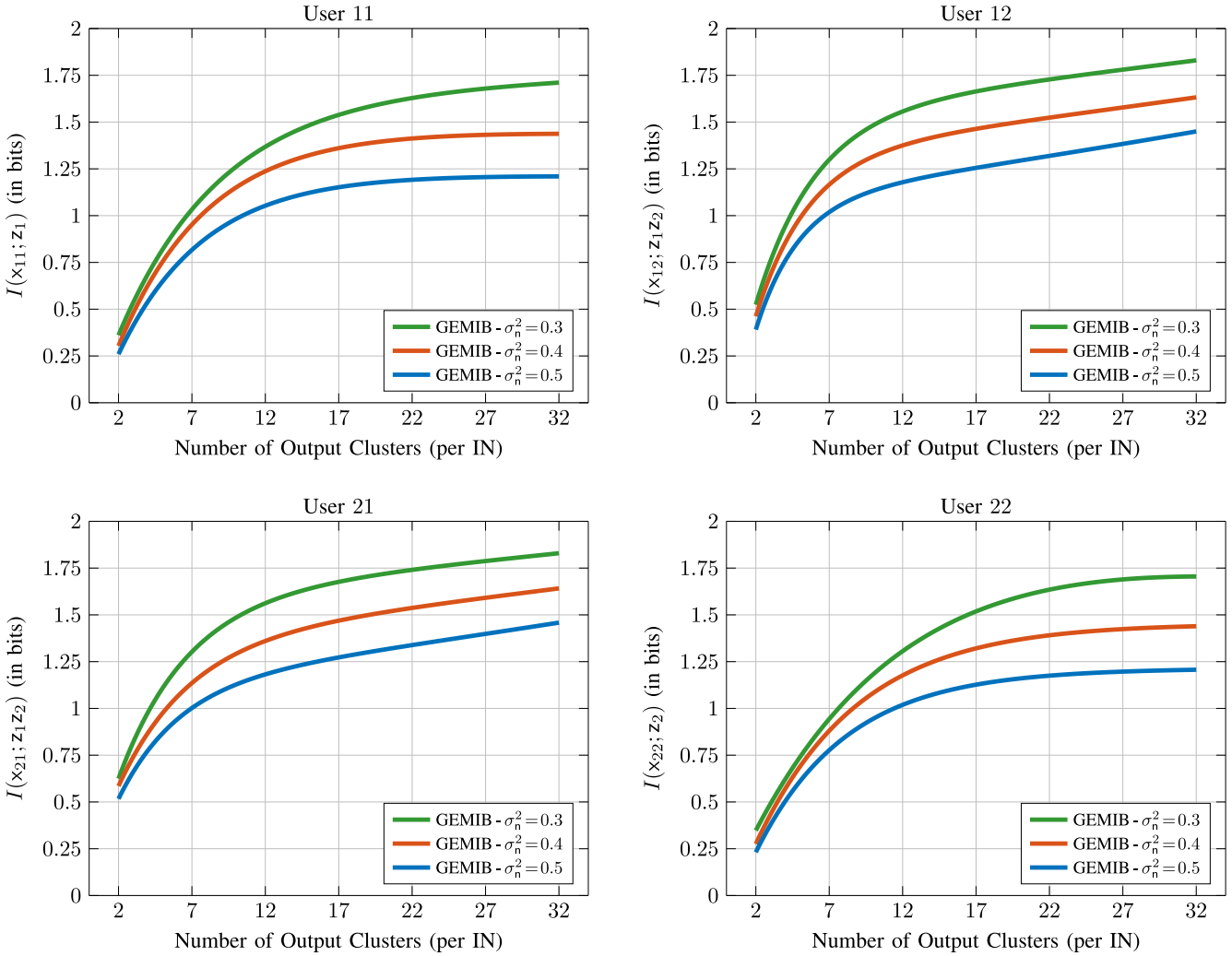


FIGURE 5. Individual relevant information vs. number of output clusters for GEMIB (Parallel). Equiprobable QPSK source signaling ($\sigma^2_{\mathsf{x}_{m\ell}}=1$) over AWGN access channels ($\sigma_n^2=0.3, 0.4, 0.5$), with $\lambda_m=0.01$ for $m=1,2$, and convergence parameter $\epsilon=10^{-3}$.

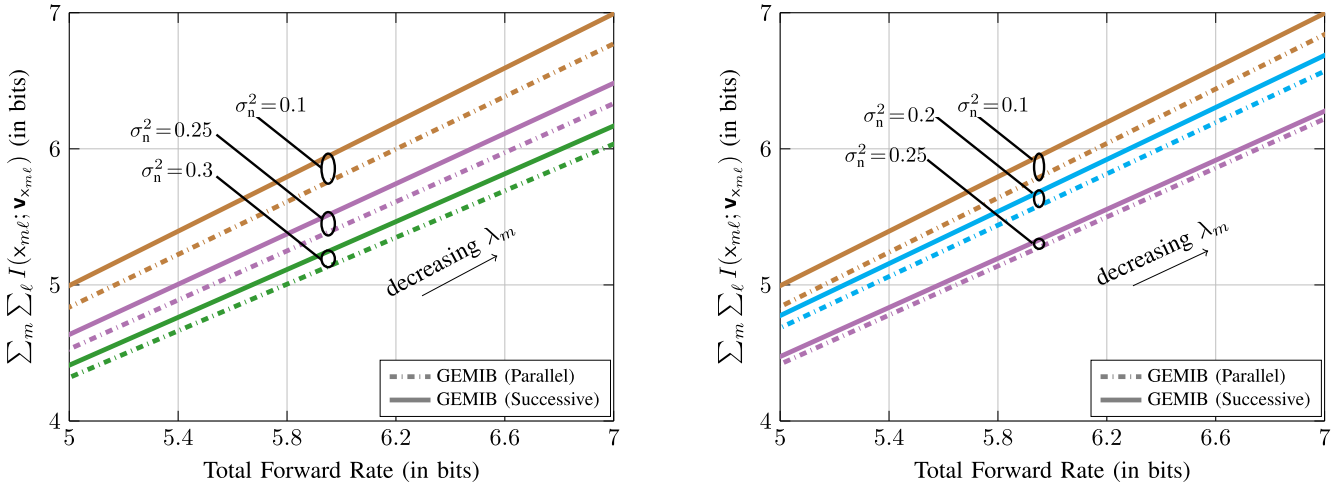
brevity), except for the choice of source signaling, where we consider a standard QPSK (Quadrature Phase Shift Keying) constellation. In Fig. 5, we illustrated the individual (i.e., per user) relevant information, $I(\mathsf{x}_{m\ell};\mathbf{v}_{\mathsf{x}_{m\ell}})$, when varying the number of output bins / clusters (per intermediate node).

Specifically, two distinct types of results are observed. The top-right and the bottom-left results belong to the users being served by both intermediate nodes. Further, the top-left and the bottom-right results belong to the users being served by only one intermediate node. Due to the present symmetry in our considered setup the obtained results in each of these two groups are quite similar. Principally, the same trend as in the overall dynamics is observed here as well, that is, the looser the compression bottleneck and the better the access channel qualities, the higher the relevant information. It must be noted that the users served by both nodes exhibit a better performance compared to those served by a single node.

C. PARALLEL VS. SUCCESSIVE PROCESSING

In the last part of our numerical investigations, we intend to compare the obtained performance of parallel and successive processing schemes. Like previous case, we apply the same procedure described for the overall system dynamics. Here, we consider both (equiprobable) 4-ASK and QPSK source signaling. Furthermore, we fix the number of output clusters (per intermediate node) to 32 and vary the trade-off parameters λ_m for $m=1,2$ over a certain range. We then calculate the obtained total relevant information and the required total forward rate for supporting it. The corresponding results have been illustrated in Fig. 6.

First, it should be noted that, by varying the values of λ_m for $m=1,2$, one can sweep through different points in the "*information-compression*" plane. Specifically, small values of λ_m correspond to the solutions with more focus on the informativity, and conversely, large values of λ_m correspond to the solutions with more focus on the compactness.



(a) 4-ASK source signaling ($\sigma^2_{\mathsf{x}_{m\ell}} = 5$)

(b) QPSK source signaling ($\sigma^2_{\mathsf{x}_{m\ell}} = 1$)

FIGURE 6. Total relevant information vs total forward rate for a) 4-ASK and b) QPSK source signaling over AWGN access channels ($\sigma_n^2 = 0.1$ to 0.3), with 32 output clusters per intermediate node and convergence parameter $\epsilon = 10^{-3}$, a) $0.25 \le \lambda_m \le 0.30$ and b) $0.25 \le \lambda_m \le 0.33$ for $m = 1, 2$.

As the main takeaway, it is clearly seen from both results that, irrespective of the certain choice of model parameters, the successive scheme outperforms the parallel one, yielding superior *information-compression* trade-offs. This, indeed, is obtained by leveraging the present *correlations* in the signals of intermediate nodes (as both *commonly* serve x_{12} and x_{21}). Note that the Markovian relations imply that the conditioning on previous signals can help to reduce the forward rate since the *conditional* compression rate can be decomposed as

$$I\big(\mathbf{y}_m; \mathsf{z}_m|\mathbf{z}_{1:m-1}\big) = I\big(\mathbf{y}_m; \mathsf{z}_m\big) - \underbrace{I(\mathsf{z}_m; \mathbf{z}_{1:m-1})}_{\ge 0}. \quad (39)$$

The larger the access links' Signal-to-Noise Ratios (SNRs), the higher becomes the correlations between signals of two intermediate nodes. Consequently, the higher becomes the gain of utilizing the side-information, and, hence, the more widens the gap between the performance curves of parallel and successive processing schemes.

VII. SUMMARY

In this article, the *Information Bottleneck (IB)* principle was fully generalized to design multiterminal / distributed remote source coding schemes for a (generic) scenario appearing in a variety of real-world applications. The considered scenario, with highest flexibility w.r.t. the assignment of users to the serving nodes, went beyond the State-of-the-Art techniques designed exclusively for a single (common) source signal. Specifically, the *Mutual Information* was selected here as the fidelity criterion, and, by taking advantage of the *Variational Calculus*, the stationary solutions for two different types of processing were derived and exploited later on as the core of our devised (iterative) algorithm, the GEMIB, to efficiently address both challenging design problems. Based upon an in-depth analysis, it was further proven that GEMIB converges to a stationary point of the pertinent objective functionals. At the end, the effectiveness of the introduced compression schemes was substantiated as well by a couple of numerical simulations over typical digital data transmission scenarios.

APPENDIX

This part of the article details the proof of two main theorems for the stationary solutions of parallel and successive schemes. To obtain these solutions, the concept of functional derivative in Variational Calculus plays the key role, since it generalizes the concept of gradient to the case in which functions of multiple functions, known as functionals, should be optimized over their input functions.

A. PROOF OF THEOREM 1 (PARALLEL PROCESSING)

The Lagrangian for parallel scheme, $\mathcal{L}_{\text{Par.}}$, is a functional that features all the individual local compressors, $\{p(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$, as its input arguments. Hence, to obtain a stationary solution, all functional derivatives w.r.t. local compressors should be equated to zero. To incorporate the validity conditions into the analysis, we further associate a Lagrange multiplier, $\lambda_{\boldsymbol{y}_m}$, to every realization, $\boldsymbol{y}_m \in \boldsymbol{\mathcal{Y}}_m$, of the m-th compressor's input set of variables, $\mathbf{y}_m = \mathbf{Pa}_{\mathsf{z}_m}^{\mathcal{G}_{\text{in}}}$, and introduce the overall Lagrangian for parallel processing, $\mathcal{L}^{\text{Ov.}}_{\text{Par.}}$, as

$$\mathcal{L}^{\text{Ov.}}_{\text{Par.}} = \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}\big) - \sum_{m=1}^{J} \lambda_m I\big(\mathbf{y}_m; \mathsf{z}_m\big) + \sum_{m=1}^{J}\sum_{\boldsymbol{y}_m} \lambda_{\boldsymbol{y}_m}\left(\sum_{z_m} p\big(z_m|\boldsymbol{y}_m\big) - 1\right), \quad (40)$$

wherein, $\mathbf{v}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}_{\mathsf{x}_{m\ell}}^{\mathcal{G}_{\text{out}}}$. Next, the functional derivative of the overall Lagrangian, $\mathcal{L}^{\text{Ov.}}_{\text{Par.}}$, w.r.t. the j-th compressor, $p(\mathsf{z}_j|\mathbf{y}_j)$,

is calculated in some steps. To that end, it should be noted that the following applies

$$\frac{\delta\Big(\sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell};\mathbf{v}_{\mathsf{x}_{m\ell}}\big)\Big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = p\big(\mathbf{y}_j\big) \times \sum_{(m,\ell):\,\mathsf{z}_j\in\mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathbf{y}_j\right)}\left\{\sum_{x_{m\ell}} p\Big(x_{m\ell}|\mathbf{y}_j,\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}\Big)\log\frac{p\big(x_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big)}{p\Big(x_{m\ell}|\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}\Big)}\right\}. \tag{41}$$

Furthermore, the following holds true

$$\frac{\delta\Big(\sum_{m=1}^{J}\lambda_m I\big(\mathbf{y}_m;\mathsf{z}_m\big)\Big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = \lambda_j\frac{\delta I\big(\mathbf{y}_j;\mathsf{z}_j\big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = \lambda_j\, p\big(\mathbf{y}_j\big)\log\frac{p\big(z_j|\mathbf{y}_j\big)}{p\big(z_j\big)}, \tag{42}$$

and

$$\frac{\delta\left(\sum_{m=1}^{J}\sum_{\mathbf{y}_m}\lambda_{\mathbf{y}_m}\left(\sum_{z_m}p\big(z_m|\mathbf{y}_m\big)-1\right)\right)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = \lambda_{\mathbf{y}_j}. \tag{43}$$

Applying the stationary condition, namely, $\frac{\delta\mathcal{L}^{\text{Ov.}}_{\text{Par.}}}{\delta p(\mathsf{z}_j|\mathbf{y}_j)}=0$, and by noting that $p(\mathbf{y}_j)>0$, from (41), (42) and (43), and by definition of the KL divergence, it is deduced that

$$-\sum_{(m,\ell):\,\mathsf{z}_j\in\mathbf{v}_{\mathsf{x}_{m\ell}}}\mathbb{E}_{p\left(\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathbf{y}_j\right)}\Big\{D_{\text{KL}}\Big(p\Big(\mathsf{x}_{m\ell}|\mathbf{y}_j,\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}\Big)\|p\big(\mathsf{x}_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big)\Big)\Big\} - \lambda_j\log\frac{p\big(z_j|\mathbf{y}_j\big)}{p\big(z_j\big)}+\tilde{\lambda}^{\text{Par.}}_{\mathbf{y}_j}=0, \tag{44}$$

in which, it applies

$$\tilde{\lambda}^{\text{Par.}}_{\mathbf{y}_j} = \frac{\lambda_{\mathbf{y}_j}}{p\big(\mathbf{y}_j\big)} + \sum_{(m,\ell):\,\mathsf{z}_j\in\mathbf{v}_{\mathsf{x}_{m\ell}}}\mathbb{E}_{p\left(\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathbf{y}_j\right)}\Big\{D_{\text{KL}}\Big(p\Big(\mathsf{x}_{m\ell}|\mathbf{y}_j,\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}\Big)\|p\Big(\mathsf{x}_{m\ell}|\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}\Big)\Big)\Big\}. \tag{45}$$

Next, the following steps are made: the second term in (44) is brought into the other side of equality, both sides are first multiplied by $\beta_j=\frac{1}{\lambda_j}$ and then exponentiated, and, finally, both sides are again multiplied by $p(z_j)$. Then, it applies

$$p\big(z_j|\mathbf{y}_j\big) = p\big(z_j\big)\exp\Big(-d_{\text{Par.}}\big(\mathbf{y}_j,z_j\big)+\beta_j\tilde{\lambda}^{\text{Par.}}_{\mathbf{y}_j}\Big). \tag{46}$$

By enforcing the quantizer mapping's validity condition, i.e., $\sum_{z_j}p(z_j|\mathbf{y}_j)=1$, we can simply treat $\exp(-\beta_j\tilde{\lambda}^{\text{Par.}}_{\mathbf{y}_j})$ as the partition function, $\psi^{\text{Par.}}_{\mathbf{z}_j}$, to obtain the presented solution in the statement of Theorem 1. ■

B. PROOF OF THEOREM 2 (SUCCESSIVE PROCESSING)

Similar to the previous case, the Lagrangian for successive processing, $\mathcal{L}_{\text{Suc.}}$, is a functional that features all individual local compressors, $\{p(\mathsf{z}_j|\mathbf{y}_j)\ \forall j\}$, as its input arguments. Thus, to have a stationary solution, all functional derivatives w.r.t. local compressors should be equated to zero. Like before, first we associate a Lagrange multiplier, $\lambda_{\mathbf{y}_m}$, to each realization, $\mathbf{y}_m\in\boldsymbol{\mathcal{Y}}_m$, of the m-th compressor's input set of variables, $\mathbf{y}_m=\mathbf{Pa}^{\mathcal{G}_{\text{in}}}_{\mathsf{z}_m}$, to bring the validity conditions into the analysis. Then, we introduce the overall Lagrangian for successive processing, $\mathcal{L}^{\text{Ov.}}_{\text{Suc.}}$, as

$$\mathcal{L}^{\text{Ov.}}_{\text{Suc.}} = \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I\big(\mathsf{x}_{m\ell};\mathbf{v}_{\mathsf{x}_{m\ell}}\big) - \sum_{m=1}^{J}\lambda_m I\big(\mathbf{y}_m;\mathsf{z}_m|\mathbf{z}_{1:m-1}\big) + \sum_{m=1}^{J}\sum_{\mathbf{y}_m}\lambda_{\mathbf{y}_m}\left(\sum_{z_m}p\big(z_m|\mathbf{y}_m\big)-1\right), \tag{47}$$

wherein, $\mathbf{v}_{\mathsf{x}_{m\ell}}=\mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$. Then, the functional derivative of the overall Lagrangian, $\mathcal{L}^{\text{Ov.}}_{\text{Suc.}}$, w.r.t. the j-th (local) compressor, $p(\mathsf{z}_j|\mathbf{y}_j)$, is calculated. Note that, we should only calculate the derivative of the second term on the right side of (47) since the derivatives of the other terms have already been calculated in (41) and (43). To do so, first it should be noted that the following holds true

$$\frac{\delta\Big(\sum_{m=1}^{J}\lambda_m I\big(\mathbf{y}_m;\mathsf{z}_m|\mathbf{z}_{1:m-1}\big)\Big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = \frac{\delta\Big(\sum_{m=j}^{J}\lambda_m I\big(\mathbf{y}_m;\mathsf{z}_m|\mathbf{z}_{1:m-1}\big)\Big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)}, \tag{48}$$

as the first $j-1$ terms in the summation do not depend on the mapping $p(\mathsf{z}_j|\mathbf{y}_j)$. Further, the following relations apply

$$\frac{\delta I\big(\mathbf{y}_j;\mathsf{z}_j|\mathbf{z}_{1:j-1}\big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = \frac{\delta\big(I\big(\mathbf{y}_j;\mathsf{z}_j\big)-I\big(\mathsf{z}_j;\mathbf{z}_{1:j-1}\big)\big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = p\big(\mathbf{y}_j\big)\left[\log\frac{p\big(z_j|\mathbf{y}_j\big)}{p\big(z_j\big)}-\sum_{z_{1:j-1}}p\big(z_{1:j-1}|\mathbf{y}_j\big)\log\frac{p\big(z_{1:j-1}|z_j\big)}{p\big(z_{1:j-1}\big)}\right], \tag{49}$$

and for $j<m\le J$

$$\frac{\delta I\big(\mathbf{y}_m;\mathsf{z}_m|\mathbf{z}_{1:m-1}\big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = \frac{\delta H\big(\mathsf{z}_m|\mathbf{z}_{1:m-1}\big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} - \underbrace{\frac{\delta H\big(\mathsf{z}_m|\mathbf{y}_m\big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)}}_{0} = \frac{\delta H\big(\mathsf{z}_m|\mathbf{z}_{1:m-1}\big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = p\big(\mathbf{y}_j\big)\sum_{z^{-j}_{1:m}}p\Big(z^{-j}_{1:m}|\mathbf{y}_j\Big)\log\frac{1}{p\big(z_m|z_{1:m-1}\big)}. \tag{50}$$

Applying the stationary condition, namely, $\frac{\delta\mathcal{L}^{\text{Ov.}}_{\text{Suc.}}}{\delta p(\mathsf{z}_j|\mathbf{y}_j)}=0$, and by noting that $p(\mathbf{y}_j)>0$, from (41), (43), (49), (50), and by definition of the KL divergence, the following is inferred

$$-\sum_{(m,\ell):\,\mathsf{z}_j\in\mathbf{v}_{\mathsf{x}_{m\ell}}}\mathbb{E}_{p\left(\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathbf{y}_j\right)}\Big\{D_{\text{KL}}\Big(p\Big(\mathsf{x}_{m\ell}|\mathbf{y}_j,\boldsymbol{v}^{-j}_{\mathsf{x}_{m\ell}}\Big)\|p\big(\mathsf{x}_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big)\Big)\Big\} - \lambda_j\log\frac{p\big(z_j|\mathbf{y}_j\big)}{p\big(z_j\big)}+\lambda_j\sum_{z_{1:j-1}}p\big(z_{1:j-1}|\mathbf{y}_j\big)\log p\big(z_{1:j-1}|z_j\big)$$

$$+\sum_{k=j+1}^{J}\lambda_k\sum_{z_{1:k}^{-j}}p\left(z_{1:k}^{-j}|\boldsymbol{y}_j\right)\log p(z_k|z_{1:k-1})+\tilde{\lambda}_{\boldsymbol{y}_j}^{\text{Suc.}}=0, \tag{51}$$

wherein, it applies

$$\tilde{\lambda}_{\boldsymbol{y}_j}^{\text{Suc.}}=-\lambda_j\sum_{z_{1:j-1}}p(z_{1:j-1}|\boldsymbol{y}_j)\log p(z_{1:j-1})+\frac{\lambda_{\boldsymbol{y}_j}}{p(\boldsymbol{y}_j)}$$
$$+\sum_{(m,\ell):\,\mathbf{z}_j\in\mathbf{v}_{\mathbf{x}_{m\ell}}}\mathbb{E}_{p\left(v_{\mathbf{x}_{m\ell}}^{-j}|\boldsymbol{y}_j\right)}\left\{D_{\text{KL}}\left(p\left(\mathbf{x}_{m\ell}|\boldsymbol{y}_j,\boldsymbol{v}_{\mathbf{x}_{m\ell}}^{-j}\right)\|p\left(\mathbf{x}_{m\ell}|\boldsymbol{v}_{\mathbf{x}_{m\ell}}^{-j}\right)\right)\right\}. \tag{52}$$

Next, analogous to the previous case, the following steps are made: the second term in (51) is brought into the other side of equality, both sides are first multiplied by $\beta_j=\frac{1}{\lambda_j}$ and then exponentiated, and, finally, both sides are again multiplied by $p(z_j)$. Subsequently, it applies

$$p(z_j|\boldsymbol{y}_j)=p(z_j)\exp\left(-d_{\text{Suc.}}(\boldsymbol{y}_j,z_j)+\beta_j\tilde{\lambda}_{\boldsymbol{y}_j}^{\text{Suc.}}\right). \tag{53}$$

By enforcing the quantizer mapping's validity condition, i.e., $\sum_{z_j}p(z_j|\boldsymbol{y}_j)=1$, we can simply treat $\exp(-\beta_j\tilde{\lambda}_{\boldsymbol{y}_j}^{\text{Suc.}})$ as the partition function, $\psi_{\mathbf{z}_j}^{\text{Suc.}}$, to obtain the presented solution in the statement of Theorem 2. ■

REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 2000, pp. 1–16.

[2] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Int. Conv. Rec.*, vol. 4, pp. 142–163, Mar. 1959.

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

[4] 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.

[5] A. Zaidi, I. Estella-Aguerri, and S. Shamai, "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 2, p. 151, Jan. 2020.

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

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

[8] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[9] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop*, 2014, pp. 501–505.

[10] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, p. 15, Mar. 2016.

[11] J. Lewandowsky and G. Bauch, "Theory and application of the information bottleneck method," *Entropy*, vol. 26, no. 3, p. 187, Feb. 2024.

[12] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop*, 2015, pp. 1–5.

[13] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.

[14] A. M. Saxe et al., "On the information bottleneck theory of deep learning," *J. Stat. Mechan. Theory Expe.*, vol. 2019, no. 12, Dec. 2019, Art. no. 124020.

[15] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–19.

[16] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[17] H. Hafez-Kolahi and S. Kasaei, "Information bottleneck and its applications in deep learning," 2019, *arXiv:1904.03743*.

[18] B. C. Geiger and G. Kubin, "Information bottleneck: Theory and applications in deep learning," *Entropy*, vol. 22, no. 12, p. 1408, Dec. 2020.

[19] G. Zeitler, A. C. Singer, and G. Kramer, "Low-precision A/D conversion for maximum information rate in channels with memory," *IEEE Trans. Commun.*, vol. 60, no. 9, pp. 2511–2521, Sep. 2012.

[20] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6562–6582, Oct. 2013.

[21] M. Stark, A. Shah, and G. Bauch, "Polar code construction using the information bottleneck method," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops*, 2018, pp. 7–12.

[22] F. J. C. Romero and B. M. Kurkoski, "LDPC decoding mappings that maximize mutual information," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 9, pp. 2391–2401, Sep. 2016.

[23] M. Stark, L. Wang, G. Bauch, and R. D. Wesel, "Decoding rate-compatible 5G-LDPC codes with coarse Quantization using the information bottleneck method," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 646–660, 2020.

[24] T. Monsees, O. Griebel, M. Herrmann, D. Wübben, A. Dekorsy, and N. Wehn, "Minimum-integer computation finite alphabet message passing decoder: From theory to decoder implementations towards 1 Tb/s," *Entropy*, vol. 24, no. 10, p. 1452, Oct. 2022.

[25] A. Winkelbauer, G. Matz, and A. Burg, "Channel-Optimized vector Quantization with mutual information as fidelity criterion," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2013, pp. 851–855.

[26] S. Hassanpour, D. Wübben, and A. Dekorsy, "A graph-based message passing approach for joint source-channel coding via information bottleneck principle," in *Proc. IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process.*, 2018, pp. 1–5.

[27] S. Hassanpour, T. Monsees, D. Wübben, and A. Dekorsy, "Forward-aware information bottleneck-based vector Quantization for noisy channels," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7911–7926, Dec. 2020.

[28] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[29] D. Gündüz et al., "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[30] E. Beck, C. Bockelmann, and A. Dekorsy, "Semantic information recovery in wireless networks," *Sensors*, vol. 23, no. 14, p. 6347, Jul. 2023.

[31] I. Estella-Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and gaussian sources," in *Proc. Int. Zurich Semin. Inf. Commun.*, 2018, pp. 1–6.

[32] I. Estella-Aguerri, A. Zaidi, G. Caire, and S. S. Shitz, "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4575–4596, Jul. 2019.

[33] Y. Uğur, I. Estella-Aguerri, and A. Zaidi, "Vector Gaussian CEO problem under logarithmic loss and applications," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4183–4202, Jul. 2020.

[34] I. Estella-Aguerri and A. Zaidi, "Distributed variational representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 120–138, Jan. 2021.

[35] S. Hassanpour, D. Wübben, and A. Dekorsy, "Forward-aware information bottleneck-based vector Quantization: Multiterminal extensions for parallel and successive retrieval," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6633–6646, Oct. 2021.

[36] S. Movaghati and M. Ardakani, "Distributed channel-aware quantization based on maximum mutual information," *Int. J. Distrib. Sens. Netw.*, vol. 12, no. 5, May 2016, Art. no. 3595389.

[37] G. Zeitler, G. Bauch, and J. Widmer, "Quantize-and-forward schemes for the orthogonal multiple-access relay channel," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 1148–1158, Apr. 2012.

[38] I. Avram, N. Aerts, H. Bruneel, and M. Moeneclaey, "Quantize and forward cooperative communication: Channel parameter estimation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1167–1179, Mar. 2012.

[39] D. Wübben et al., "Benefits and impact of cloud computing on 5G signal processing: Flexible Centralization through cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.
[40] S.-H. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.
[41] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.
[42] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.
[43] M. Bashar, P. Xiao, R. Tafazolli, K. Cumanan, A. G. Burr, and E. Björnson, "Limited-fronthaul cell-free massive MIMO with local MMSE receiver under Rician fading and phase shifts," *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1934–1938, Sep. 2021.
[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[45] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Cambridge, MA, USA: Academic, 1982.
[46] J. H. Mathews and K. D. Fink, *Numerical Methods Using MATLAB*, 4th ed. Hoboken, NJ, USA: Pearson Prentice-Hall, 2004.
[47] N. Slonim, N. Friedman, and N. Tishby, "Multivariate information bottleneck," *Neural Comput.*, vol. 18, no. 8, pp. 1739–1789, Aug. 2006.
[48] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.
[49] S. Hassanpour, D. Wübben, and A. Dekorsy, "A novel approach to distributed Quantization via multivariate information bottleneck method," in *Proc. IEEE Glob. Commun. Conf.*, 2019, pp. 1–6.
[50] S. Hassanpour, D. Wübben, and A. Dekorsy, "Generalized distributed information bottleneck for fronthaul rate reduction at the cloud-RANs uplink," in *Proc. IEEE Glob. Commun. Conf.*, 2020, pp. 1–6.
[51] M. Razaviyan, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, Jun. 2013.
[52] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[53] R. Horst, P. M. Pardalos, and N. Van Thoai, *Introduction to Global Optimization*, 2nd ed. New York, NY, USA: Springer, 2000.
[54] S. Hassanpour, D. Wübben, A. Dekorsy, and B. M. Kurkoski, "On the relation between the asymptotic performance of different algorithms for information bottleneck framework," in *Proc. IEEE Int. Conf. Commun.*, 2017, pp. 1–6.
[55] A. K. Jain, "Data clustering: 50 Years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

SHAYAN HASSANPOUR (Member, IEEE) received the B.Sc. degree in electrical engineering (electronics) from the University of Mazandaran, Iran, in 2011, the M.Sc. degree (with outstanding performance) in communications engineering from the University of Ulm, Germany, in 2014, and the Dr.-Ing. degree (summa cum laude) in electrical engineering from the University of Bremen, Germany, in 2022, where he is currently working as a Postdoctoral Researcher with the Department of Communications Engineering. Over the last couple of years, he has been prolifically contributing to the top-level international journals and IEEE flagship conferences on his Ph.D. topic, i.e., the Information Bottleneck method. His other research interests include information theory, MU/MIMO systems, wireless/mobile communications, statistical signal processing, and the application of machine learning in the design of communication systems. He won the 2021's VDE ITG Prize and the 2023's OHB Ph.D. Prize for an outstanding scientific publication and his doctoral dissertation, respectively.

ALIREZA DANAEE (Member, IEEE) received the bachelor's degree in electronic engineering from the University of Kurdistan, Sanandaj, Iran, in 2008, the master's degree in electrical engineering from Shahid Rajaee Teacher Training University, Tehran, Iran, in 2013, and the Ph.D. degree in electrical engineering, in the area of signal processing, automation and robotics from the Pontifical Catholic University of Rio de Janeiro, Brazil, in 2022. He is currently a Postdoctoral Researcher with the Department of Communications Engineering, University of Bremen, Germany. His current research interests center around statistical signal processing, information theory, and machine learning with applications to wireless communications and distributed data processing.

DIRK WÜBBEN (Senior Member, IEEE) received the Dipl.-Ing. (FH) degree in electrical engineering from the University of Applied Science Münster, Germany, in 1998, and the Dipl.-Ing. (Uni.) and Dr.-Ing. degrees in electrical engineering from the University of Bremen, Germany, in 2000 and 2005, respectively, where he is currently a Senior Research Group Leader and a Lecturer with the Department of Communications Engineering. He has published more than 140 papers in international journals and conference proceedings. His research interests include wireless communications, signal processing, multiple antenna systems, cooperative communication systems, channel coding, information theory, and machine learning. He has been an Editor of IEEE WIRELESS COMMUNICATIONS LETTERS. He is a Board Member of the Germany Chapter of the IEEE Information Theory Society and a member of VDE/ITG Expert Committee "Information and System Theory."

ARMIN DEKORSY (Senior Member, IEEE) is currently the Head of the Department of Communications Engineering, University of Bremen. He has more than ten years of industrial experience in leading research positions, such as an DMTS with Bell Labs Europe and the Head of Research Europe Qualcomm, Nuremberg, and by conducting international research projects (more than 25 BMBF/BMWI/EU Projects) in affiliation with his scientific expertise shown by more than 200 journals and conference publications and holds more than 19 patents. He investigates new lines of research in wireless communication and signal processing for the baseband of transceivers of future communication systems, the results of which are transferred to the pre-development of industry through political and strategic activities. His current research focuses on distributed signal processing, compressed sampling, information bottleneck method, and machine learning leading to the further development of communication technologies for 5G/6G, industrial wireless communications, and NewSpace satellite communications. He is a Senior Member of the IEEE Communications and Signal Processing Society, the Head of VDE/ITG Expert Committee "Information and System Theory," and a member of Executive Board of the Technologie-Zentrum Informatik und Informationstechnik, University of Bremen.