Received 8 August 2025; accepted 30 August 2025. Date of publication 3 September 2025; date of current version 18 September 2025.
Digital Object Identifier 10.1109/OJCOMS.2025.3605737

Forward-Aware Multi-Source Distributed Vector Quantization for Noisy Channels via 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 by the German Federal Ministry of Research, Technology and Space (BMFTR) under Grant 16KISK016 (Open6GHub), Grant 16KISK068 (6G-TakeOff), and Grant 16KISK109 (6G-ANNA).

ABSTRACT A generic distributed / multiterminal setup is considered here wherein, through a *joint* design, several intermediate nodes must *locally* quantize their *noisy* observations from various sets of potentially common and uncommon user / source signals, ahead of a forward transmission over multiple *error-prone* and *rate-limited* channels to a remote processing unit. The local compressors should be designed in such a fashion that the impacts of the *noisy* forward channels are taken into account as well. Fully aligned with the principal idea of the *Information Bottleneck (IB)* method, the *Mutual Information* is then selected here as the fidelity criterion, and by means of *Variational Calculus*, the corresponding stationary solutions are derived for two various types of processing flow / strategy. Thereupon, an iterative algorithm, the *Forward-Aware GEneralized Multivariate IB (FAGEMIB)*, is introduced as well to efficiently tackle the challenging (non-convex) design problems. The pertinent convergence proofs to a stationary point of the objective functionals are also provided, together with a couple of numerical results on typical digital data transmission scenarios, corroborating the effectiveness of these joint source-channel coding techniques. The presented compression schemes in this article, with the most flexibility w.r.t. the assignment of users to the serving nodes, highly generalize the State-of-the-Art techniques designed exclusively for a single (common) user / source signal.

INDEX TERMS 6G, information bottleneck method, joint source-channel coding, multi-user compression.

I. INTRODUCTION

THE ORIGINAL formulation of the *Information Bottleneck (IB)* method [1] for data compression was built upon the *single-letter* characterization of the *Rate-Distortion* function in Shannon's seminal work on *lossy* source coding [2]. The major modification there was coming from an intuitive idea: first, pinpoint your *relevant / target variable* and then try to retain its information in compressed data by lower-bounding a *Mutual Information* term, rather than upper-bounding an expected distortion term. In this manner, one does not need to be concerned about how to figure out the suitable distortion function for any application of interest. Instead, all it should be done is to identify the target variable that is a way simpler task to accomplish in many practical scenarios. Years later, the connections of the IB method to a *remote / indirect* source coding problem [3], [4], [5], [6], [7] were established by the particular choice of *Logarithmic Loss* distortion [8]. More specifically, it was shown that the boundary of achievable rate-distortion region is obtained by solving the central optimization in the IB framework. We refer the interested readers to [9], [10], [11] for further details on various aspects of the IB principle from the standpoints of Information and Learning Theory. Important to note is the connections to several other (classic) problems including (but not confined to) the Wyner-Ahlswede-Körner problem [12], [13], the efficiency of investment information [14], the privacy funnel [15], [16], [17], and also the (distributed) functional compression [18], [19], [20], [21].

The IB method has also been used to better understand the dynamics of deep learning models [22], [23], [24]. Furthermore, it has been applied to various aspects of deep learning, ranging from the generation of novel network architectures to the optimization of neural network parameters, and even as an efficient method towards reducing the issue of overfitting in intricate inference tasks [25], [26], [27].

Besides purely theoretical considerations, the IB method has already gained practical acceptance and has been utilized in a wide variety of real-world applications. Some examples of that include, i.a., semantic / task-oriented communications [28], [29], [30], discrete (channel) decoding techniques [31], [32], [33], [34], (efficient) construction of the Polar Codes [35], [36], and also in Analog-to-Digital converters for receiver front ends [37], [38]. Interested readers are referred to [39] to acquire a better view on recent studies concerning both the theoretical as well as the practical aspects of the IB method.

The extensions of the IB principle to multiterminal setups have also been considered in the pertinent literature (see, e.g., [40], [41], [42], [43], [44], [45], [46], [47]). Typically, such distributed schemes focus on a certain setup wherein various *noisy* observations of a single *common* user signal should be locally compressed at some intermediate nodes, ahead of a forward transmission through a couple of *error-free* and *rate-limited* channels to a (remote) processing unit. In practice, however, it happens quite often that, every intermediate node should further serve its own (i.e., uncommon) source / user signals as well, alongside the common users. Furthermore, the forward channels to the remote processing unit are usually *error-prone*. In [48], two multiterminal / distributed IB-based joint source-channel coding schemes have been developed for the simpler case of dealing exclusively with a single common source signal. Herein, we generalize those results to the more complicated scenario described above, i.e., incorporating (in the design) uncommon source signals alongside the common ones, and as a direct consequence, enabling the full flexibility w.r.t. the assignment of users to the serving nodes. It should be noted that, in this context, by an uncommon user, it is meant the one getting served by only one intermediate node, and by a common user, it is meant the one getting served by at least two intermediate nodes. The devised schemes in this article directly extend the presented (distributed multi-user) *remote source coding* schemes in [49] to the context of (distributed multi-user) *joint source-channel coding*.

A. CONTRIBUTIONS

In this work, two novel joint source-channel coding schemes are developed for a (generic) multiterminal / distributed setup wherein a couple of intermediate nodes quantize different sets of *noisy* observations from uncommon and (potentially) common user / source signals, before a forward transmission through multiple *rate-limited* and *error-prone* channels to a remote processing unit. Explicitly, pursuing the *Information Bottleneck* philosophy, the *Mutual Information* is chosen as the fidelity criterion for designing the compression schemes, and taking advantage of *Variational Calculus*, the stationary solutions are derived for the challenging design problems. Subsequently, we introduce the *Forward-Aware GEneralized Multivariate IB (FAGEMIB)*, which is an (iterative) algorithm using the derived stationary solutions as its core component. To deliver a full-package support, via in-depth mathematical analyses, we further provide both the convergence proofs as well as the cogent justifications of the behavior of FAGEMIB over the entire range of its main (input) parameters.

The distributed scenario which we focused on here, shows up in a wide variety of applications regarding the fifth (5G) and sixth (6G) generations of wireless network technologies, i.a., in the Cell-Free massive Multiple-Input Multiple-Output systems (CF-mMIMO) [50], [51], [52], [53], [54], and in Cloud-based Radio Access Networks (Cloud-RANs) [55], [56] with error-prone fronthaul links, in inference sensor networks with imperfect channels to the fusion center [57], and in relaying schemes with the "*Quantize-and-Forward*" strategy [58], [59].

The main practical relevance of the proposed compression schemes in this article reveals itself in applications wherein, due to some stringent (end-to-end) delay / latency constraints, modern (iterative) forward error-correcting techniques might not be applicable as a feasible solution to tackle the problem of imperfect / error-prone forwarding [60], or when dealing with hardware imperfections wherein, the separate usage of those schemes may lead to substantial overheads in terms of energy efficiency [61]. Moreover, the theoretical importance of the proposed schemes in this article is realized by recalling that the source and channel coding separation's optimality breaks down in certain occasions, including the relevant case of working in a finite / non-asymptotic block-length regime.

B. OUTLINE

We begin our (technical) discussions by presenting the point-to-point joint source-channel coding scheme in Section II. Thereupon, in Section III, we introduce our system model for distributed data compression and formulate the design problems for the *parallel* and *successive* processing schemes. After that, in Section IV, we derive the stationary solutions of the design problems. These solutions are then utilized as the core of our devised (iterative) algorithm, the FAGEMIB, which is presented in Section V. There, we provide the convergence proofs to a stationary point of the objective functionals as well, together with a detailed discussion on the way this algorithm behaves over the whole range of its input parameters. To substantiate the effectiveness of our novel compression schemes, multiple numerical results are presented in Section VI. Eventually, we conclude this article by providing a succinct wrap-up in Section VII. The detailed proofs of two main theorems are relegated to the Appendix.

C. NOTATIONS

For the (discrete) random variable, a, each realization, $a \in \mathcal{A}$ occurs according to the probability mass function, $p(\mathsf{a})$. The

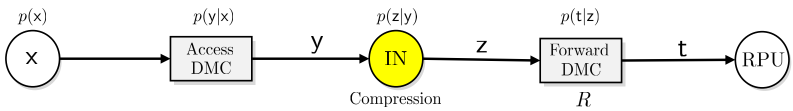


FIGURE 1. The presumed system model for point-to-point IB-based joint source-channel coding. DMC, IN, and RPU stand for Discrete Memoryless Channel, Intermediate Node, and Remote Processing Unit, respectively. The Markov chain $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z} \leftrightarrow \mathsf{t}$ applies.

TABLE 1. List of denotations used for the main variables in this article.

Denotation	Description
$\mathsf{x}_{\circ}$	source signal
$\mathsf{y}_{\circ}^{\circ}$	noisy observation of a source signal
$\mathsf{z}_{\circ}$	compression variable
$\mathsf{t}_{\circ}$	forward variable
$\mathbf{y}_{\circ}$	set of noisy observations to be compressed at $\text{IN}_{\circ}$
$\mathbf{v}_{\mathsf{x}_{\circ}}$	set of forward variables preserving information about $\mathsf{x}_{\circ}$
$\mathbf{w}_{\mathsf{x}_{\circ}}$	set of compression variables preserving information about $\mathsf{x}_{\circ}$

same holds true for the vector, $\mathbf{a}_{1:J} = \{\mathsf{a}_1, \ldots, \mathsf{a}_J\}$, with the boldface counterparts. Moreover, $\mathbf{a}_{1:J}^{-j} = \mathbf{a}_{1:J} \setminus \{\mathsf{a}_j\}$, and the parent nodes of the random variable a, in the Bayesian Network, $\mathcal{G}_{\circ}$ are denoted by $\mathbf{Pa}_{\mathsf{a}}^{\mathcal{G}_{\circ}}$. Further, $D_{\text{KL}}(\cdot\|\cdot)$, $I(\cdot;\cdot)$, $H(\cdot)$, $D_{\text{JS}}^{\{\cdot,\cdot\}}(\cdot\|\cdot)$, denote *Kullback-Leibler (KL)* divergence, Mutual Information, Shannon's entropy, and *Jensen-Shannon (JS)* divergence [62]. $\mathbb{E}_{\circ}$ represents the expectation operator. Table 1 lists the denotations used for the main variables:

II. IB-BASED JOINT SOURCE-CHANNEL CODING: POINT-TO-POINT SETUP IN A NUTSHELL

Consider the illustrated system model in Fig. 1. The noisy signal,[1] y, from a (single) source signal, x, is observed by an Intermediate Node (IN) via an access link modeled by a Discrete Memoryless Channel (DMC). The goal is then to compress the signal, y, into another signal, namely, z, ahead of a forward transmission via an *error-prone* channel with the rate-limit, R, to a Remote Processing Unit (RPU). The impacts of the noisy forward link must be taken into account within the design formulation. The source statistics, $p(\mathsf{x})$, as well as both the access, $p(\mathsf{y}|\mathsf{x})$, and the forward transition probabilities, $p(\mathsf{t}|\mathsf{z})$, are presumed to be known. By pursuing the IB philosophy [1], the design problem has been formulated in [63] as a trade-off between two *Mutual Information* terms.

To quantify / measure the informativity of outcome, the *relevant information*, $I(\mathsf{x};\mathsf{t})$, is utilized, and to quantify its compactness, the *compression rate*, $I(\mathsf{y};\mathsf{z})$, is applied. The design problem of the compressor, $p(\mathsf{z}|\mathsf{y})$, is then formulated as a constrained maximization, namely,

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

with $0 \le R \le \log_2|\mathcal{Z}|$ bits, bounding the compression rate from above. This problem can then be reformulated as an unconstrained optimization (up to the compressor mapping's validity) by exploiting the method of *Lagrange Multipliers* [64]. Explicitly, it applies

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

with $\lambda \ge 0$, being associated with the upper-bound, R, in (1). The corresponding λ value for a given R can be found, e.g., by performing a bisection search over a finite range. The form of stationary solutions for (2) has been derived in [63] for each $(y,z) \in \mathcal{Y} \times \mathcal{Z}$ as

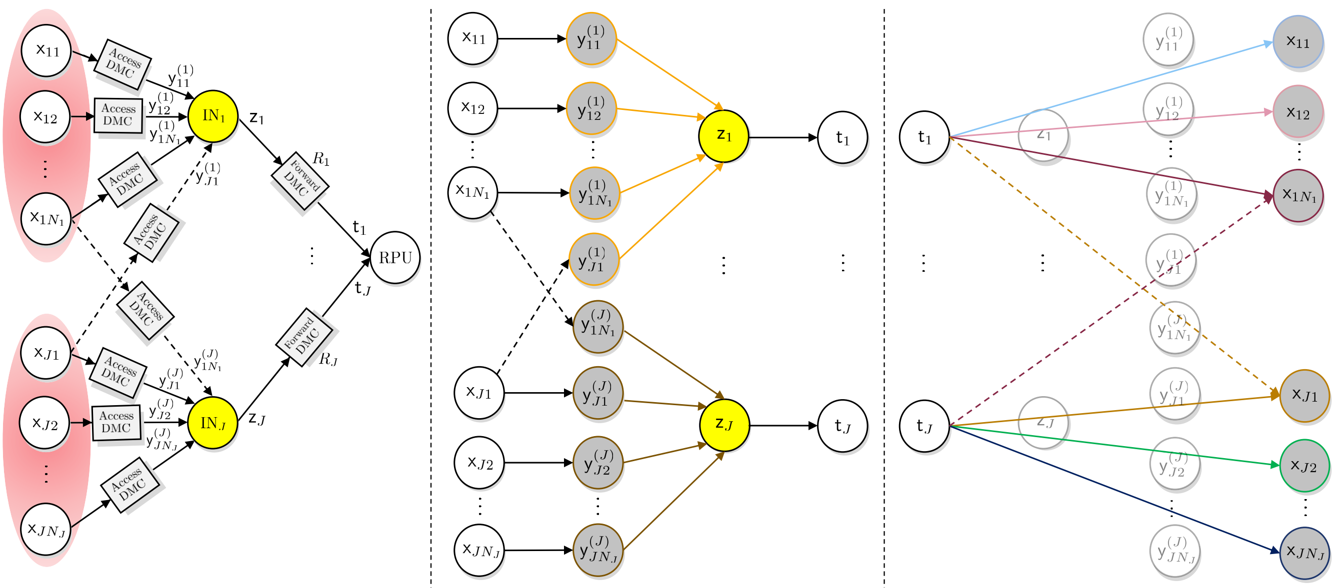
$$p(z|y) = \frac{p(z)}{\psi(y,\beta)} \exp\Big(-\beta \sum_{t \in \mathcal{T}} p(t|z) D_{\text{KL}}\big(p(\mathsf{x}|y) \| p(\mathsf{x}|t)\big)\Big), \tag{3}$$

with $\beta = \frac{1}{\lambda}$ and $\psi(y,\beta)$, a normalization function, ensuring the compressor mapping's validity. Explicitly, the sum of all calculated terms in (3) (ignoring ψ) over all output clusters, $z \in \mathcal{Z}$, for every realization, $y \in \mathcal{Y}$, acts as the normalization function. An iterative algorithm has also been proposed in [63] to address the design problem which, in essence, applies the *Fixed-Point Iterations* [65] on (3). The convergence proof of this iterative algorithm to a stationary point of the pertinent objective functional has been presented as well, together with several numerical results, substantiating the fact that this data compression scheme enjoys an inherent *error-protection* capability. Interested readers are referred to [66] for the data-driven counterpart.

III. DISTRIBUTED EXTENSIONS: SYSTEM MODEL AND PROBLEM FORMULATION

The depicted system model in Fig. 2a is considered. A total number, $N_{\text{tot.}}$, of independent user signals must be served by J intermediate nodes. Each node, IN_j with $j \in \{1, \ldots, J\}$, receives non-interfering noisy observations, $\{\mathsf{y}_{m\ell}^{(j)}\}$, from its assigned set of user signals, $\{\mathsf{x}_{m\ell}\}$, and compresses them to a (compact) representative, z_j, prior to a forward transmission via an *error-prone* channel with the transition probabilities, $p(\mathsf{t}_j|\mathsf{z}_j)$, and the rate-limit, R_j, to the RPU. All user signals, $\mathsf{x}_{m\ell}$, should then be retrieved at the RPU by applying various processing strategies which get reflected in the design formulations. Through following a joint source-channel coding approach, the imperfections of the forward channels must be integrated into the design of (local) compressors. The access link between the user signal, $\mathsf{x}_{m\ell}$, and its noisy observation, $\mathsf{y}_{m\ell}^{(j)}$, at j-th intermediate node is modeled by a DMC whose transition probabilities, $p(\mathsf{y}_{m\ell}^{(j)}|\mathsf{x}_{m\ell})$, and input distribution, $p(\mathsf{x}_{m\ell})$, are assumed to

[1] Throughout this article, all considered random variables are discrete.



(a) System model (b) Input BN, $\mathcal{G}_{\text{in}}$ (c) Output BN, $\mathcal{G}_{\text{out}}$

FIGURE 2. a) System model for distributed joint source-channel coding, together with b) input BN and c) output BN. The input BN, $\mathcal{G}_{\text{in}}$, illustrates the *compression* side and the output BN, $\mathcal{G}_{\text{out}}$, reflects the *information preservation* side. For every common user, the connection to the allocated serving IN is depicted by the solid line and the connections to other (i.e., not allocated) serving INs are depicted by dashed lines.

be known. 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.

It is important to mention that, in this system description, a *common* source with more than one serving intermediate nodes, is allocated to only one of them, such that it holds $\sum_m N_m = N_{\text{tot.}}$, which is the total number of individual user signals. This yields a clear enumeration of all users and simplifies the mathematical formulation of design problems. Furthermore, totally aligned with the introduced formalism in [67], two Bayesian Networks (BNs) are used (as graphical models) to reflect different aspects (namely, the *compression* and *information preservation*) in the design problems. The input BN, $\mathcal{G}_{\text{in}}$, in Fig. 2b portrays "*what compresses what*". Specifically, every compression variable, z_j, quantizes its parents in $\mathcal{G}_{\text{in}}$. As another functionality, $\mathcal{G}_{\text{in}}$, also encodes (by its structure) the statistical relations between all random variables in the system model. How? Through the basic rule of *conditional* independence from all non-descendants, given the values of parent nodes. On the other hand, the output BN, $\mathcal{G}_{\text{out}}$, in Fig. 2c illustrates "*what should remain informative of what*". Particularly, the noisy version, t_j, of the compression variable, z_j, shows up as a parent for any user signal that z_j must preserve information about.

We state the design problem as a constrained optimization, establishing a basic trade-off between the *compactness* and *informativity* of resultant outcomes. Naturally, we quantify the informativity by summing up all the *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 $\mathcal{G}_{\text{out}}$ that, henceforth, are denoted by $\mathbf{v}_{\mathsf{x}_{m\ell}}$. The other side of this trade-off, however, does not offer a unique and natural choice. As a result, one can apply various meaningful expressions. In what follows, we apply two different sets of constraints, corresponding to the *parallel* and *successive* processing strategies at the RPU to recover all user signals. We then use *Variational Calculus* to derive the stationary solutions for all (local) quantizers and employ them as the heart of our devised algorithm to address both design problems. We further provide the convergence proofs of this iterative algorithm to a stationary point of the underlying objective functionals.

A. PARALLEL SCHEME: IGNORING SIDE-INFORMATION

A *parallel* processing scheme is considered as the first choice wherein, to retrieve the source signals, no side-information is used at the RPU. The design problem is then formulated as the constrained maximization (with $\mathbf{y}_m = \mathbf{Pa}^{\mathcal{G}_{\text{in}}}_{\mathsf{z}_m}$, i.e., the set of variables to be quantized by the m-th local compressor)

$$P^* = \{p^*(\mathsf{z}_1|\mathbf{y}_1), \ldots, p^*(\mathsf{z}_J|\mathbf{y}_J)\} = \underset{P:\, \forall m\; I(\mathbf{y}_m;\, \mathsf{z}_m) \le R_m}{\operatorname{argmax}} \sum_{m=1}^{J} \sum_{\ell=1}^{N_m} I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}), \tag{4}$$

where, $0 \le R_m \le \log_2 |\mathcal{Z}_m|$ bits, sets an upper-bound on the m-th compression rate, $I(\mathbf{y}_m; \mathsf{z}_m)$.[2] We can recast the above design problem into an unconstrained optimization (up to the validity of all compressor mappings) by applying the method of *Lagrange Multipliers* [64], namely,

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

in which, $\lambda_m \ge 0$, is associated with R_m, in (4).

[2] A multi-letter description is required for the asymptotic coding problems in Section III besides the considered *one-shot* formulations.

B. SUCCESSIVE SCHEME: USING SIDE-INFORMATION

A *successive* processing scheme is considered as the second choice in which, when recovering a certain source signal, we also take advantage of the side-information at the RPU. Compared to the parallel processing scheme, generally, this yields a superior "*informativity-compactness*" trade-off, but at the expense of higher processing complexity. In essence, this approach is aligned with the Wyner-Ziv setup for source coding [68] that utilizes some statistically correlated signals as the side-information at the decoder. The design problem is then formulated as the constrained maximization

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

with $0 \le R_m \le \log_2|\mathcal{Z}_m|$ bits, imposing an upper-bound on the m-th *conditional* compression rate, $I(\mathbf{y}_m;\mathsf{z}_m|\mathbf{t}_{1:m-1})$. It is important to note that, here, an extra degree of freedom exists, which is the processing order. Generally speaking, it affects the performance, and therefore, it should be optimized (e.g., via a brute-force search). Henceforth, we continue our technical discussion with a fixed choice of ordering. Like in the parallel scheme, the above design problem can be recast into an unconstrained optimization (up to the validity of all compressor mappings) by applying the *Lagrange Multipliers* method [64], namely,

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

in which, $\lambda_m \ge 0$, is associated with R_m, in (6).

Please note that, in the extreme case of *full-informativity*, which corresponds to letting $\lambda_m \to 0$ for all $m = 1$ to J, the objective functional of the successive processing coincides with that of the parallel processing. To clearly see this, it should be noted that the Lagrangians in (5) and (7) differ in their second terms that vanish, when letting $\lambda_m \to 0$.

IV. CHARACTERIZATION OF STATIONARY SOLUTIONS

In this section, by making use of the Variational Calculus, we derive the stationary solutions of the parallel and successive processing schemes. Later on, these solutions will become the central components of our devised iterative algorithm, the FAGEMIB, to efficiently address the pertinent design problems.

A. PARALLEL PROCESSING

Considering the design problem for parallel processing, the following theorem provides the form of stationary solutions for (local) quantizers:

Theorem 1 (PARALLEL SCHEME): Presume the joint distribution of input variables (i.e., all nodes in $\mathcal{G}_{\text{in}}$ except the leaves and their direct parents) as well as the forward transition probabilities, $p(\mathsf{t}_m|\mathsf{z}_m)$, and λ_m are given for all $m \in \{1,\ldots,J\}$. The set of local compressors, $\{p(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$, is a stationary point of the Lagrangian for IB-based parallel processing scheme

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

if and only if for each pair, $(\boldsymbol{y}_j, z_j) \in \boldsymbol{\mathcal{Y}}_j \times \mathcal{Z}_j$, it applies

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

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

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

with $\beta_j = \frac{1}{\lambda_j}$, $\mathbf{y}_j = \mathbf{Pa}^{\mathcal{G}_{\text{in}}}_{\mathsf{z}_j}$, $\mathbf{v}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$, $\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{v}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{t}_j\}$. Also, $\mathbf{w}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$ for the counterpart scenario of *error-free* forwarding, that is, replacing t_m with z_m in $\mathcal{G}_{\text{out}}$ for every $m \in \{1,\ldots,J\}$, and $\mathbf{w}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{w}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$.

The proof is presented in Appendix A.

B. SUCCESSIVE PROCESSING

Focusing on the design problem for successive processing, the next theorem provides the form of stationary solutions for (local) quantizers:

Theorem 2 (SUCCESSIVE SCHEME): Assume the joint distribution of input variables (i.e., all nodes in $\mathcal{G}_{\text{in}}$ except the leaves and their direct parents) as well as the forward transition probabilities, $p(\mathsf{t}_m|\mathsf{z}_m)$, and λ_m are given for $m \in \{1,\ldots,J\}$. The set of local compressors, $\{p(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$, is a stationary point of the Lagrangian for IB-based successive processing scheme

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

if and only if for each pair, $(\boldsymbol{y}_j, z_j) \in \boldsymbol{\mathcal{Y}}_j \times \mathcal{Z}_j$, it applies

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

wherein, $\psi^{\text{Suc.}}_{\mathsf{z}_j}(\boldsymbol{y}_j,\beta_j)$, is a partition function that ensures the validity of compressor mapping, and the relevant distortion, $d_{\text{Suc.}}(\boldsymbol{y}_j,z_j)$, is calculated through

$$d_{\text{Suc.}}\big(\boldsymbol{y}_j,z_j\big) = \beta_j \sum_{(m,\ell):\,\mathsf{t}_j\in\mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}}|\boldsymbol{y}_j\right)}\Bigg\{\sum_{\boldsymbol{v}_{\mathsf{x}_{m\ell}}} p\big(\boldsymbol{v}_{\mathsf{x}_{m\ell}}|\boldsymbol{w}_{\mathsf{x}_{m\ell}}\big) \times D_{\text{KL}}\Big(p\Big(\mathsf{x}_{m\ell}|\boldsymbol{y}_j,\boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}}\Big)\|p\big(\mathsf{x}_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big)\Big)\Bigg\}$$

$$-\sum_{\boldsymbol{t}_{1:j-1}} p(\boldsymbol{t}_{1:j-1}|\mathbf{y}_j)\log p(\boldsymbol{t}_{1:j-1}|z_j)$$
$$-\beta_j\sum_{k=j+1}^{J}\beta_k^{-1}\sum_{\boldsymbol{t}_{1:k-1},z_k} p(t_j|z_j)p\Big(\boldsymbol{t}_{1:k-1}^{-j},z_k|\mathbf{y}_j\Big)\log p(z_k|\boldsymbol{t}_{1:k-1}), \tag{13}$$

with $\beta_j=\frac{1}{\lambda_j}$, $\mathbf{y}_j=\mathbf{Pa}_{\mathsf{z}_j}^{\mathcal{G}_{\text{in}}}$, $\mathbf{v}_{\mathsf{x}_{m\ell}}=\mathbf{Pa}_{\mathsf{x}_{m\ell}}^{\mathcal{G}_{\text{out}}}$, $\mathbf{v}_{\mathsf{x}_{m\ell}}^{-j}=\mathbf{v}_{\mathsf{x}_{m\ell}}\setminus\{\mathsf{t}_j\}$. Also, $\mathbf{w}_{\mathsf{x}_{m\ell}}=\mathbf{Pa}_{\mathsf{x}_{m\ell}}^{\mathcal{G}_{\text{out}}}$ for the counterpart scenario of *error-free* forwarding, that is, replacing t_m with z_m in $\mathcal{G}_{\text{out}}$ for every $m\in\{1,\ldots,J\}$, and $\mathbf{w}_{\mathsf{x}_{m\ell}}^{-j}=\mathbf{w}_{\mathsf{x}_{m\ell}}\setminus\{\mathsf{z}_j\}$.

The proof is presented in Appendix B.

C. FURTHER DISCUSSION

Theorems 1 & 2 extend the presented results in [48, Thms 1 & 2]. As opposed to there, here, apart from the (potentially) common source signals, different (local) quantizers compress various noisy observations from uncommon sources as well. The major difference of the successive processing scheme compared to the parallel processing, is the fact that the *side-information* is taken into account during the compression to further take advantage of the potentially present correlations among output signals of (local) compressors. Compared to the calculated relevant distortion (10) for parallel processing, the derived relevant distortion (13) for successive processing contains two additional terms that appear due to *conditioning* the compression rates.

Moreover, it is noteworthy that Theorems 1 & 2 generalize the presented results in [49, Thms 1 & 2] from noisy / remote source coding to the context of joint source-channel coding. The main contrasting point between the derived solutions here and the ones in [49] is the appearance of the forward statistics, $p(\boldsymbol{v}_{\mathsf{x}_{m\ell}}|\boldsymbol{w}_{\mathsf{x}_{m\ell}})$, in the relevant distortion terms that is the very way by which the impacts of the imperfect signal forwarding are taken into account. The "*backward compatibility*" of the derived solutions here for both parallel and successive processing to the obtained results in [49] for the case of *error-free* forwarding can be easily verified by setting $p(\boldsymbol{v}_{\mathsf{x}_{m\ell}}|\boldsymbol{w}_{\mathsf{x}_{m\ell}})=\delta_{\boldsymbol{v}_{\mathsf{x}_{m\ell}},\boldsymbol{w}_{\mathsf{x}_{m\ell}}}$ (with the *Kronecker delta* denotation).

V. A GENERIC DESIGN ALGORITHM

The purpose of this part of the article is to present an iterative algorithm that can be applied to both parallel and successive scheme's design problems. To that end, we take advantage of the common structure in the calculated stationary solutions for both schemes. We further present the convergence proofs to a stationary point of the objective functionals, along with an in-depth analysis of the behavior of this algorithm over the whole range of its (main) input parameters, namely, β_j for $j=1$ to J, which are associated with the rate-limits, R_j, of the pertinent forward channels at the considered system model in Fig. 2a.

A. THE DEVISED ALGORITHM

The derived stationary solution for each compressor's input-output pair, $(\mathbf{y}_j,z_j)\in\mathcal{Y}_j\times\mathcal{Z}_j$, has an *implicit* form (for both parallel and successive schemes), namely,

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

in which $r\in\{\text{Par.},\text{Suc.}\}$. The relevant distortion, $d_r(\mathbf{y}_j,z_j)$, depends on all quantizer mappings. So, the right side of (14) can be seen as a functional of all (local) quantizer mappings. When we sweep through all quantizers, a non-linear system is achieved which directly extends the form of *Multivariate Fixed-Point Systems* [65]. Explicitly, multivariate functions are replaced by multi-mapping functionals:

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

wherein, $\mathcal{F}_j$, denotes a specific functional. Therefore, we can resort to the conventional techniques for solving this system. Here, an iterative algorithm is proposed with the *synchronous* updating rule, fully aligned with the *Jacobi* method for linear systems [65].

Specifically, a set of random mappings, $\{p^{(0)}(\mathsf{z}_j|\mathbf{y}_j)\ \lfloor j\}$ is taken for the sake of initialization. Then, those mappings are iteratively updated (till convergence by $\varepsilon\ll 1$) via

$$p^{(i+1)}(z_j|\mathbf{y}_j)=\frac{p^{(i)}(z_j)}{\psi_{\mathsf{z}_j}^{r\,(i+1)}(\mathbf{y}_j,\beta_j)}\exp\big(-d_r^{(i)}(\mathbf{y}_j,z_j)\big), \tag{16}$$

with i, denoting the iteration counter. Through a synchronous updating, at each iteration, we recalculate (local) quantizers, $\{p^{(i+1)}(\mathsf{z}_j|\mathbf{y}_j)\ \lfloor j\}$, based upon the previous configuration of the same set, $\{p^{(i)}(\mathsf{z}_j|\mathbf{y}_j)\ \lfloor j\}$. We present the pseudo-code of this algorithm, the FAGEMIB, in Algorithm 1.

In the following part, the convergence of FAGEMIB to a stationary point of the corresponding functionals is proven for both parallel and successive schemes. Hence, as a popular workaround for avoiding poor local optima, FAGEMIB can be repeated with different starting points (to retain the best outcome, since the quality of the outcome directly depends on the choice of initialization).

B. CONVERGENCE PROOFS

Here, we will pursue the following roadmap to provide the convergence proofs for both parallel and successive processing schemes: the design problems are first reformulated as an alternating minimization w.r.t. the set of (local) compressor mappings, P, and another set of *auxiliary* distributions, Q. To that end, a (tight) *variational upper-bound*, $\bar{\mathcal{F}}_r(P,Q)$, is introduced on the corresponding objective functional, $\mathcal{F}_r(P)$. Thereupon, through an unfolding trick, it is shown that the central update step in FAGEMIB is reached by merging the corresponding updates for P and Q. Therefore, it is directly deduced that our proposed algorithm lies within the class of *Successive Upper-bound Minimization (SUM)* [69], ensuring its convergence to a stationary point (by satisfying certain

Algorithm 1 Forward-Aware GEneralized Multivariate IB (FAGEMIB)

Input: Joint input distribution, forwarding statistics $p(\mathsf{t}_j|\mathsf{z}_j)$, $\beta_j = \frac{1}{\lambda_j} > 0$, $\varepsilon > 0$, $r \in \{\text{Par.}, \text{Suc.}\}$

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

  end for
  if $\forall j, \forall \boldsymbol{y}_j$:$D_{\text{JS}}^{\{\frac{1}{2},\frac{1}{2}\}}\big(p^{(i+1)}(\mathsf{z}_j|\boldsymbol{y}_j)\|p^{(i)}(\mathsf{z}_j|\boldsymbol{y}_j)\big) \le \varepsilon$ **then**
    Break
  else
    $i \leftarrow i + 1$
  end if
end while

mild assumptions). It is noteworthy that, in [42], [43], [48], a similar roadmap has been followed to prove the convergence of their devised iterative algorithms.

1) PARALLEL PROCESSING

The design problem for parallel scheme can be recast into minimizing the following functional (over P)

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

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

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

the next four Lemmas apply:

Lemma 1: The following equivalence holds

$$\min_P \mathcal{F}_{\text{Par.}}(P) = \min_P \min_Q \bar{\mathcal{F}}_{\text{Par.}}(P, Q)\ . \quad (19)$$

Proof: For the difference of $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$ and $\mathcal{F}_{\text{Par.}}(P)$, it applies

$$\bar{\mathcal{F}}_{\text{Par.}}(P, Q) - \mathcal{F}_{\text{Par.}}(P) = \sum_{m=1}^{J}\sum_{\ell=1}^{N_m}\sum_{\boldsymbol{v}_{\mathsf{x}_{m\ell}}} p\big(\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big) D_{\text{KL}}\big(p\big(\mathsf{x}_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big)\|q\big(\mathsf{x}_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big)\big) \ge 0\ . \quad (20)$$

The equality holds true iff $q(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}}) = p(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}})$ for all $m \in \{1, \ldots, J\}$ and $\ell \in \{1, \ldots, N_m\}$. ■

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

Proof: This is straightly deduced by applying the *log-sum* inequality [62]. ■

Lemma 3: For a given P, there is a unique Q minimizing $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$, namely,

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

wherein, $p(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}})$, is calculated from P.

Proof: This is straightforward to realize from the proof of Lemma 1. ■

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

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

for every pair $(\boldsymbol{y}_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$, where $\bar{\psi}^{\text{Par.}}_{\mathsf{z}_j}(\boldsymbol{y}_j, \beta_j)$, acts as a partition function to ensure the compressor mapping's validity. For the relevant distortion, $\bar{d}_{\text{Par.}}(\boldsymbol{y}_j, z_j)$, it applies

$$\bar{d}_{\text{Par.}}\big(\boldsymbol{y}_j, z_j\big) = \beta_j \sum_{(m,\ell):\,\mathsf{t}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}}|\boldsymbol{y}_j\right)}\Bigg\{\sum_{\boldsymbol{v}_{\mathsf{x}_{m\ell}}} p\big(\boldsymbol{v}_{\mathsf{x}_{m\ell}}|\boldsymbol{w}_{\mathsf{x}_{m\ell}}\big) \times D_{\text{KL}}\Big(p\Big(\mathsf{x}_{m\ell}|\boldsymbol{y}_j, \boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}}\Big)\|q\Big(\mathsf{x}_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\Big)\Big)\Bigg\}, \quad (23)$$

wherein, $\mathbf{w}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$ for the counterpart case of *error-free* forwarding, that is, replacing t_m with z_m in $\mathcal{G}_{\text{out}}$ for every $m \in \{1, \ldots, J\}$, and $\mathbf{w}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{w}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$.

Proof: The derivation goes through the exact same track as in the proof of Theorem 1, noting that

$$\frac{\delta\Big(\sum_{m=1}^{J}\sum_{\ell=1}^{N_m}\mathbb{E}_{\mathsf{x}_{m\ell},\mathbf{v}_{\mathsf{x}_{m\ell}}}\big\{\log q\big(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}}\big)\big\}\Big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = p\big(\boldsymbol{y}_j\big) \sum_{(m,\ell):\,\mathsf{t}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}}|\boldsymbol{y}_j\right)}\Bigg\{\sum_{\boldsymbol{v}_{\mathsf{x}_{m\ell}}} p\big(\boldsymbol{v}_{\mathsf{x}_{m\ell}}|\boldsymbol{w}_{\mathsf{x}_{m\ell}}\big) \times \sum_{x_{m\ell}} p\Big(x_{m\ell}|\boldsymbol{y}_j, \boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}}\Big) \log q\big(x_{m\ell}|\boldsymbol{v}_{\mathsf{x}_{m\ell}}\big)\Bigg\}\ . \quad (24)$$

By merging the solutions for P and Q from the last two Lemmas, the FAGEMIB's central update step is achieved, indicating that it follows the SUM framework. Furthermore, [69, Thm 1] ensures the convergence to a stationary point as $\bar{\mathcal{F}}_{\text{Par.}}(P, Q)$ and $\mathcal{F}_{\text{Par.}}(P)$ satisfy [69, Proposition 1]. ■

2) SUCCESSIVE PROCESSING

Like the parallel scheme, the design problem for successive processing can be reformulated as minimizing the following functional (over P)

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

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

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

the next four Lemmas apply:

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: For the difference of $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$ and $\mathcal{F}_{\text{Suc.}}(P)$, it applies

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

wherein equality applies iff $q(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}}) = p(\mathsf{x}_{m\ell}|\mathbf{v}_{\mathsf{x}_{m\ell}})$ and also $q(\mathsf{z}_m|\mathbf{t}_{1:m-1}) = p(\mathsf{z}_m|\mathbf{t}_{1:m-1})$ 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 Q and P.

Proof: It is immediately deduced by applying the *log-sum* inequality [62]. ■

Lemma 7: For a given P, there is a unique Q minimizing $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$, namely,

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

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

Proof: This is straightforward to realize from the proof of Lemma 5. ■

Lemma 8: For a given Q, there exists a P that minimizes $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$, namely,

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

for every $(\boldsymbol{y}_j, z_j) \in \boldsymbol{\mathcal{Y}}_j \times \mathcal{Z}_j$, with $\bar{\psi}^{\text{Suc.}}_{\mathsf{z}_j}(\boldsymbol{y}_j, \beta_j)$, acting as a normalization function to ensure the compressor mapping's validity. For the relevant distortion, $\bar{d}_{\text{Suc.}}(\boldsymbol{y}_j, z_j)$, it applies

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

wherein, $\mathbf{w}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$ for the counterpart case of *error-free* forwarding, that is, replacing t_j with z_j in $\mathcal{G}_{\text{out}}$ for every $j \in \{1, \ldots, J\}$, and $\mathbf{w}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{w}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$.

Proof: The derivation goes through the exact same track as in the proof of Theorem 2, noting that

$$\frac{\delta\Big(\mathbb{E}_{\mathbf{t}_{1:j-1},\mathsf{z}_j}\Big\{\log\frac{q(\mathsf{z}_j|\mathbf{t}_{1:j-1})}{p(\mathsf{z}_j)}\Big\}\Big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = p\big(\boldsymbol{y}_j\big) \times\sum_{\boldsymbol{t}_{1:j-1}} p\big(\boldsymbol{t}_{1:j-1}|\boldsymbol{y}_j\big)\log\frac{q\big(z_j|\boldsymbol{t}_{1:j-1}\big)}{p\big(z_j\big)}, \tag{32}$$

and for $j < m \leq J$

$$\frac{\delta\Big(\mathbb{E}_{\mathbf{t}_{1:m-1},\mathsf{z}_m}\Big\{\log\frac{q(\mathsf{z}_m|\mathbf{t}_{1:m-1})}{p(\mathsf{z}_m)}\Big\}\Big)}{\delta p\big(\mathsf{z}_j|\mathbf{y}_j\big)} = p\big(\boldsymbol{y}_j\big) \times\sum_{\boldsymbol{t}_{1:m-1},z_m} p\big(t_j|z_j\big)p\Big(\boldsymbol{t}^{-j}_{1:m-1}, z_m|\boldsymbol{y}_j\Big)\log\frac{q(z_m|\boldsymbol{t}_{1:m-1})}{p(z_m)} . \tag{33}$$

By merging the solutions for P and Q from the last two Lemmas, the FAGEMIB's central update step is achieved. Like before, this indicates that FAGEMIB follows the SUM framework. Moreover, [69, Thm 1] ensures the convergence to a stationary point as $\bar{\mathcal{F}}_{\text{Suc.}}(P, Q)$ and $\mathcal{F}_{\text{Suc.}}(P)$ satisfy [69, Proposition 1]. ■

C. FURTHER DISCUSSION

In this part, based upon an in-depth analysis, further insights are provided into the way the FAGEMIB algorithm behaves over the entire range of its main input parameters. By this, we intend to answer two important questions: what to expect from FAGEMIB when working on various regimes of its input parameters, $\beta_j > 0$ for $j = 1$ to J, and more importantly, how to justify the observed behaviors. A brief discussion about an important extension of FAGEMIB will also be presented to the case of *uneven* user retrieval preferences.

1) Letting $\beta_j \to 0$ and assuming fixed $p(\mathsf{z}_m|\mathbf{y}_m)$ and finite λ_m for all $m = 1$ to J and $m \neq j$, the parallel processing's design problem will boil down to minimization of the j-th compression rate, $I(\mathbf{y}_j; \mathsf{z}_j)$, w.r.t. the j-th (local) compressor mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$. This corresponds to one extreme state, that is, the *full diffusion*, wherein, every compressor input realization, $\boldsymbol{y}_j \in \boldsymbol{\mathcal{Y}}_j$, will be allocated equiprobably to every output cluster, $z_j \in \mathcal{Z}_j$. In this manner, the input and output of the j-th (local) compressor become statistically independent, and, as a direct result, for the respective compression rate, it applies $I(\mathbf{y}_j; \mathsf{z}_j) = 0$, yielding its global minimum.

For finite (and non-zero) values of β_j, FAGEMIB typically generates *soft/stochastic* compressor mappings, $p(\mathsf{z}_j|\mathbf{y}_j)$. Contrarily, when we let $\beta_j \to \infty$, it yields *hard/deterministic* compressor mappings, $p(\mathsf{z}_j|\mathbf{y}_j)$, that corresponds to the other extreme state, namely, the *full concentration*, wherein each input realization, $\boldsymbol{y}_j \in \boldsymbol{\mathcal{Y}}_j$, is allocated to one (and only one) output cluster, $z_j^*(\boldsymbol{y}_j) \in \mathcal{Z}_j$. How to justify this? By letting $\beta_j \to \infty$ and assuming fixed $p(\mathsf{z}_m|\mathbf{y}_m)$ and (finite) λ_m for all $m = 1$ to J and $m \neq j$, the parallel scheme's design problem boils down to the following optimization

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

This turns out to be a *convex maximization* task. To perceive that, first note 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 [62]. Furthermore, the relation among $p(\mathsf{z}_j|\mathbf{y}_j)$ and $p(\mathbf{v}_{\mathsf{x}_{m\ell}}|\mathsf{x}_{m\ell})$ is established by

$$p\left(\mathbf{v}_{\mathsf{x}_{m\ell}}|\mathsf{x}_{m\ell}\right) = \frac{p\left(\mathsf{x}_{m\ell}, \mathbf{v}_{\mathsf{x}_{m\ell}}^{-j}, \mathsf{t}_j\right)}{p(\mathsf{x}_{m\ell})} = \frac{\sum_{\boldsymbol{y}_j, z_j} p\left(\mathsf{x}_{m\ell}, \mathbf{v}_{\mathsf{x}_{m\ell}}^{-j}, \boldsymbol{y}_j\right) p\left(z_j|\boldsymbol{y}_j\right) p\left(\mathsf{t}_j|z_j\right)}{p(\mathsf{x}_{m\ell})}. \quad (35)$$

This is an *affine* transform which preserves convexity [70, Sec. 3.2]. Therefore, it is directly inferred that $I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}})$ is also convex w.r.t. $p(\mathsf{z}_j|\mathbf{y}_j)$. Since the sum of several convex functions yields a convex function as well, we immediately realize that the objective in (34) is a convex function of the j-th local quantizer mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$. It is well known that a convex function which is defined over a closed and convex set reaches its global maximum at an *extreme point* of that set [71, Ch. 4]. This directly implies that the focus can be restricted solely to the *hard/deterministic* mappings. To see this, it should be noted that the space of valid probability mappings, $p(\mathsf{z}_j|\mathbf{y}_j)$, is a closed and convex polytope formed by the Cartesian product of $|\boldsymbol{\mathcal{Y}}_j|$ probability simplices [72]. The extreme points of this polytope are at its corners, and each corner corresponds to the Cartesian product of some corners of the constituent probability simplices. This simply means that every corner of this polytope corresponds to a *hard/deterministic* mapping.

2) An analogous behavior is observed by the successive scheme and a similar justification is made for that as well. Explicitly, by letting $\beta_j \to 0$ and assuming fixed $p(\mathsf{z}_m|\mathbf{y}_m)$ and (finite) λ_m for all $m = 1$ to J and $m \neq j$, the successive processing's design problem boils down to minimizing the j-th conditional compression rate, $I(\mathbf{y}_j; \mathsf{z}_j|\mathbf{t}_{1:j-1})$, w.r.t. the j-th local compressor mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$. Like before, this corresponds to one extreme state, that is, the *full diffusion*, where each quantizer input realization, $\boldsymbol{y}_j \in \boldsymbol{\mathcal{Y}}_j$, is allocated equiprobably to every output cluster, $z_j \in \mathcal{Z}_j$. In this manner, the input and output of the j-th quantizer become statistically independent, and consequently, for the respective conditional compression rate, it holds $I(\mathbf{y}_j; \mathsf{z}_j|\mathbf{t}_{1:j-1}) = 0$.

Similar to the parallel scheme, for finite (non-zero) values of β_j, FAGEMIB typically yields *soft/stochastic* compressor mappings, $p(\mathsf{z}_j|\mathbf{y}_j)$. But when we let $\beta_j \to \infty$, it generates *hard/deterministic* mappings, $p(\mathsf{z}_j|\mathbf{y}_j)$, corresponding to the other extreme state, namely, the *full concentration*, wherein each input realization, $\boldsymbol{y}_j \in \boldsymbol{\mathcal{Y}}_j$, will be allocated to one (and only one) output cluster/bin, $z_j^*(\boldsymbol{y}_j) \in \mathcal{Z}_j$. To clearly justify this behavior, we present an analogous line of argumentation as the one already provided for the parallel processing. By letting $\beta_j \to \infty$ and presuming fixed $p(\mathsf{z}_m|\mathbf{y}_m)$ and (finite) λ_m for all $m = 1$ to J and $m \neq j$, the successive scheme's design problem boils down to the following optimization[3]

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

This, again, turns out to be a *convex maximization* task. To discern that, it is fully sufficient to show that $I(\mathsf{z}_m; \mathbf{t}_{1:m-1})$ is convex w.r.t. the j-th quantizer mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$. Why? Since if so, the second summation on the right side of (36) is also a convex function of $p(\mathsf{z}_j|\mathbf{y}_j)$, as λ_m is non-negative and the sum of several convex functions also yields a convex function. It should be noted that $I(\mathsf{z}_m; \mathbf{t}_{1:m-1})$ is a convex function of $p(\mathbf{t}_{1:m-1}|\mathsf{z}_m)$, when $p(\mathsf{z}_m)$ is given [62]. The relation among $p(\mathsf{z}_j|\mathbf{y}_j)$ and $p(\mathbf{t}_{1:m-1}|\mathsf{z}_m)$ is established by

$$p(\mathbf{t}_{1:m-1}|\mathsf{z}_m) = \frac{\sum_{\boldsymbol{y}_{1:m}} p(\boldsymbol{y}_{1:m}) p(\mathsf{z}_m|\boldsymbol{y}_m) p\left(\mathbf{t}_{1:m-1}^{-j}|\boldsymbol{y}_{1:m-1}^{-j}\right) \sum_{z_j} p(\mathsf{t}_j|z_j) p(z_j|\boldsymbol{y}_j)}{p(\mathsf{z}_m)}, \quad (37)$$

This is an *affine* transform which preserves convexity. Thus, we directly deduce that $I(\mathsf{z}_m; \mathbf{t}_{1:m-1})$ is also convex w.r.t. $p(\mathsf{z}_j|\mathbf{y}_j)$, and this concludes the claimed proposition's proof.

3) Finally, it is noteworthy that FAGEMIB can be readily adapted to the case wherein, via a "normalized preference" set, $\{0 \leq \alpha_{m\ell} \leq 1 \;|m \in \{1, \ldots, J\}, \ell \in \{1, \ldots, N_m\}\}$, the signal recoveries are prioritized based upon their importance.

[3]From the input BN, $\mathcal{G}_{\text{in}}$, it is inferred that z_m is independent of $\mathbf{t}_{1:m-1}$ given $\mathbf{y}_m$, leading to $I(\mathbf{y}_m; \mathsf{z}_m|\mathbf{t}_{1:m-1}) = I(\mathbf{y}_m; \mathsf{z}_m) - I(\mathsf{z}_m; \mathbf{t}_{1:m-1})$.

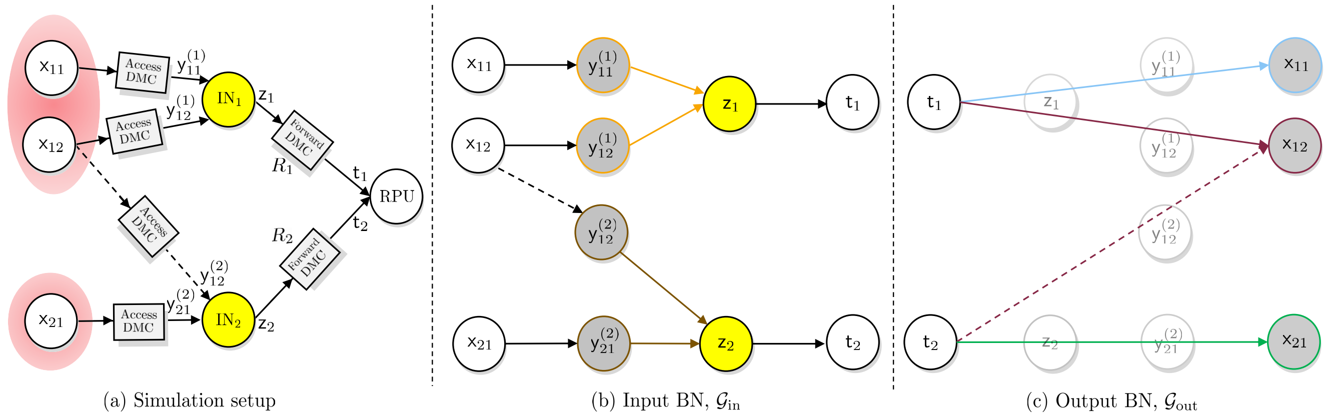


FIGURE 3. a) The considered setup for numerical simulations regarding distributed joint source-channel coding with two uncommon (x_{11} and x_{21}) and one common (x_{12}) source, together with b) input BN and c) output BN.

In such cases, the *informativity* is measured by the *weighted* sum of relevant information terms, i.e.,

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

By this, the form of derived stationary solutions for parallel and successive processing schemes given in (10) and (13) does not change, up to a multiplicative prefactor, $\alpha_{m\ell}$, in front of the expected values in the first term of both solutions. Hence, for *uneven* recovery preferences, the pertinent set of non-negative entries is fed as another input to the FAGEMIB.

VI. NUMERICAL RESULTS

In this part, we substantiate the effectiveness of FAGEMIB through a couple of numerical investigations. To that end, we consider a particular setup that has been depicted in Fig. 3a. Explicitly, three users are connected to the RPU via two INs. Two of these users, x_{11} and x_{21}, are *uncommon* since they get served by a single IN, namely, IN_1 and IN_2, respectively. The third user, x_{12}, is a *common* user as it gets served by both IN_1 and IN_2. Nevertheless, this user has been (arbitrarily) assigned to IN_1. The pertinent input and output BNs have been illustrated in Fig. 3b and Fig. 3c, respectively. They formalize both the *informativity* and *compression* sides of the design problem(s). Principally, they indicate that every IN should *jointly* compress its incoming *noisy* observations such that the pertinent received signal after (*imperfect*) forwarding preserves information about the original source signals (being served by it).

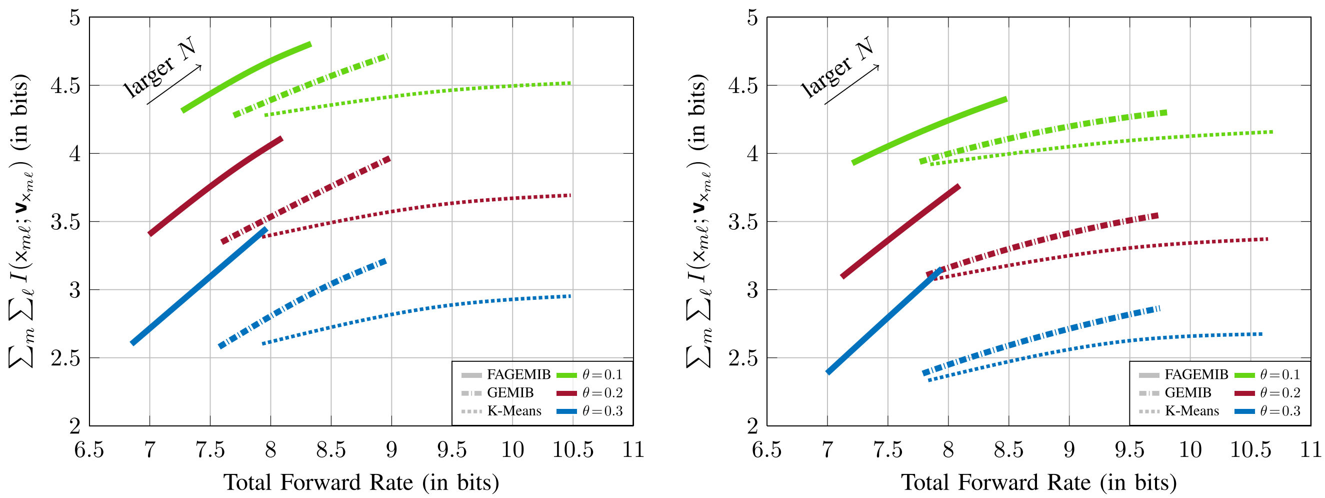
In what follows, first, we confirm the fact that bringing the impacts of *imperfect* forwarding into the design problem yields substantial performance gains (on the overall system dynamics) compared to the case of full forward-unawareness. Subsequently, we present and discuss separate performance curves of all three users. We consider the parallel processing scheme for these investigations. Finally, by comparing the obtained performance curves from the successive and parallel processing schemes, we corroborate the fact that leveraging the potential correlations in signals of different INs (through side-information) by successive processing scheme leads to a superior "*information-compression*" trade-off.

A. FORWARD-AWARENESS VS. -UNAWARENESS

A standard bipolar 4-ASK (Amplitude Shift Keying) source signaling is considered for all three users. Regarding every access channel from any user to the serving IN(s), a DMC is considered to approximate a discrete-time, discrete-input, and continuous-output AWGNC (Additive White Gaussian Noise Channel) that gets characterized by the noise variance, σ_{n}^2. Following a purely "Monte Carlo" approach, instead of prequantizing the output signals, we generate 40 samples per access link to get the transition probability matrices.

Regarding every forward link from each IN to the RPU, a symmetric $N \times N$ channel model is considered, where N denotes the allowed number of output clusters. This channel model is characterized by the reliability parameter, θ, in the following fashion: With the probability $1-\theta$, every input symbol is received correctly, and with the probability $\frac{\theta}{N-1}$, it is received erroneously (i.e., as every other output symbols). Hence, the lower the θ value, the more reliable the forward transmission and vice versa. Here, a fully symmetric setup is considered with the same noise variance, σ_{n}^2, and reliability, θ, for all access and forward links, respectively. Moreover, the trade-off parameters, λ_m for $m = 1, 2$, are set to 0.01. This indicates that, for this investigation, the main focus will be on the preservation of relevant information.

The required vector quantization at each IN is performed by three algorithms: the well-known K-Means [73], the GEMIB (Parallel) from [49] (that totally ignores the statistics of forward links) and its *forward-aware* generalization here, i.e., the FAGEMIB (Parallel). The main purpose of this numerical investigation is to confirm the fact that, integrating the impacts of the *error-prone* forward transmission into the design of *local* quantizers at INs yields some performance enhancement compared to the case of *forward-unawareness*, wherein the statistics of forward channels do not play any role in the design of *local* quantizers. All algorithms were randomly initialized 100 times, and then the best



(a) Access noise variance $\sigma_{\mathsf{n}}^2 = 0.3$

(b) Access noise variance $\sigma_{\mathsf{n}}^2 = 0.4$

FIGURE 4. Total relevant information of FAGEMIB (Parallel), GEMIB (Parallel), and K-Means vs. total forward rate for different access noise variances a) $\sigma_{\mathsf{n}}^2 = 0.3$ and b) $\sigma_{\mathsf{n}}^2 = 0.4$. Bipolar 4-ASK source signaling ($\sigma^2_{\mathsf{x}_{m\ell}} = 5$), with $\lambda_1 = \lambda_2 = 0.01$, the convergence parameter $\epsilon = 10^{-3}$, and $16 \le N \le 48$.

outcomes were retained. This procedure was repeated for 100 regenerations of access links, and the outcomes were averaged. We present the obtained results in Fig. 4.

In Fig. 4a, the relevant information, $\sum_m \sum_\ell I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}})$, has been depicted vs. the total forward rate, $\sum_m I(\mathbf{y}_m; \mathsf{z}_m)$, for a fixed value of access noise variance, namely, $\sigma_{\mathsf{n}}^2 = 0.3$, and three different values of the forward reliability, namely, $\theta = 0.1, 0.2, 0.3$. To obtain these results, the number of output clusters (per IN) was varied over a particular range, namely, $N = 16$ to 48. Furthermore, in Fig. 4b, a similar investigation has been performed by increasing the access noise variance to $\sigma_{\mathsf{n}}^2 = 0.4$. In this fashion, the impacts of the qualities of both access and forward channels on the obtained overall system performance are investigated.

Focusing on the presented results in Fig. 4a, it is clearly observed that, for a fixed access statistics, the obtained total relevant information increases via increasing the quality of forward transmissions (through decreasing the value of θ). This, indeed, is expected, since in the case of more reliable forward transmissions, less information loss will occur over second hops (INs to the RPU). Thus, more information about user signals is contained in the received signals at the RPU. Moreover, it is observed that by loosening the compression bottleneck through increasing the size, N, of each quantizer output alphabet (per IN) larger total relevant information is obtained. Naturally, by increasing N, more information can be flown into the system and get delivered to the RPU.

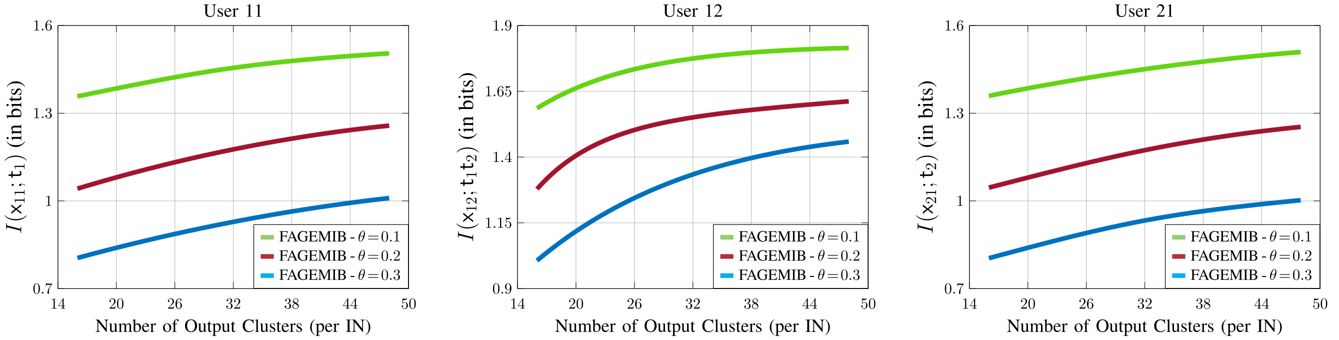
Concentrating on the results in Fig. 4b, similar patterns are observed as before. The main difference is that the obtained total relevant information drops for the corresponding curves with the same forward reliability. This is due to the fact that, by increasing the access noise variance, σ_{n}^2, the capacities of access links decrease. Consequently, less information about the user signals gets flown into the system.

Considering both results together, as the main takeaway, it is realized that FAGEMIB yields superior performance (in establishing better information-compression trade-off points) compared to GEMIB. This clearly corroborates the fact that integrating the impacts of *error-prone* forward transmissions into the design problem of distributed compression schemes is indeed beneficial. This result seems expected when noting that, by deploying the FAGEMIB algorithm, the compressed signals are designed such that, in addition to capturing well the information of all user / source signals, they further take the errors (which occur over *imperfect* forward channels) into account. In contrast, when applying the GEMIB algorithm, by ignoring the impacts of forward links, the goal is solely to preserve information about the source signals. Naturally, the less reliable the forwarding becomes, the wider becomes the performance gap between these algorithms. This is due to the fact that, by worse forwarding conditions, the overall system becomes prone to higher information loss over second hops. Consequently, the gained benefits from "*forward-awareness*" increase accordingly.

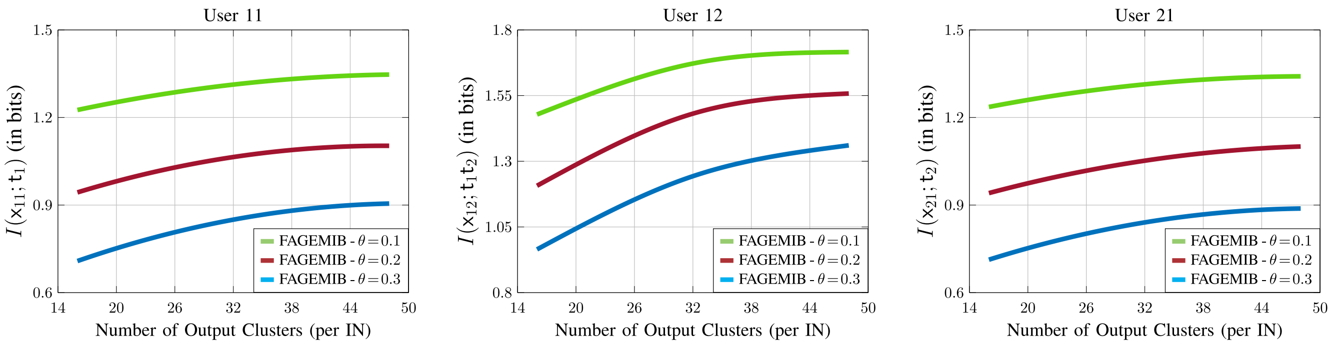
Finally, it should further be mentioned that an analogous line of argumentation (FAGEMIB vs. GEMIB) applies when comparing FAGEMIB with K-Means. The main difference is that, with the K-Means algorithm, the goal is to minimize the Mean Squared Error (MSE) between the noisy observations and the compressed representatives at each IN by following a *separate* design approach, leading to an inferior performance compared to the GEMIB algorithm, which follows the *joint* design of local compressors at different INs.

B. INDIVIDUAL PERFORMANCE RESULTS

The individual dynamics of all user signals are investigated in this part. The same specifications are considered as before, i.e., the ones already described in the previous part. In Fig. 5,



(a) Access noise variance $\sigma_{\mathsf{n}}^2=0.3$



(b) Access noise variance $\sigma_{\mathsf{n}}^2=0.4$

FIGURE 5. Individual relevant information of FAGEMIB (Parallel) vs. number of output clusters for different access noise variances a) $\sigma_{\mathsf{n}}^2=0.3$ and b) $\sigma_{\mathsf{n}}^2=0.4$. Bipolar 4-ASK source signaling ($\sigma^2_{\mathsf{x}_{m\ell}}=5$) with $\lambda_1=\lambda_2=0.01$, and the convergence parameter $\epsilon=10^{-3}$.

the per-user/individual relevant information, $I(\mathsf{x}_{m\ell};\mathbf{v}_{\mathsf{x}_{m\ell}})$, has been illustrated vs. the cardinality, N, of the quantizer's output alphabet (per IN).

Analogous to the previous investigations, in Fig. 5a, the access noise variance has been fixed to $\sigma_{\mathsf{n}}^2=0.3$, and the value of forward reliability has been varied ($\theta=0.1, 0.2, 0.3$). Furthermore, in Fig. 5b, similar investigations have been made by increasing the access noise variance to $\sigma_{\mathsf{n}}^2=0.4$. As before, the intention behind this is to investigate the effects of both access and forward channels on the obtained relevant information per user/source signal.

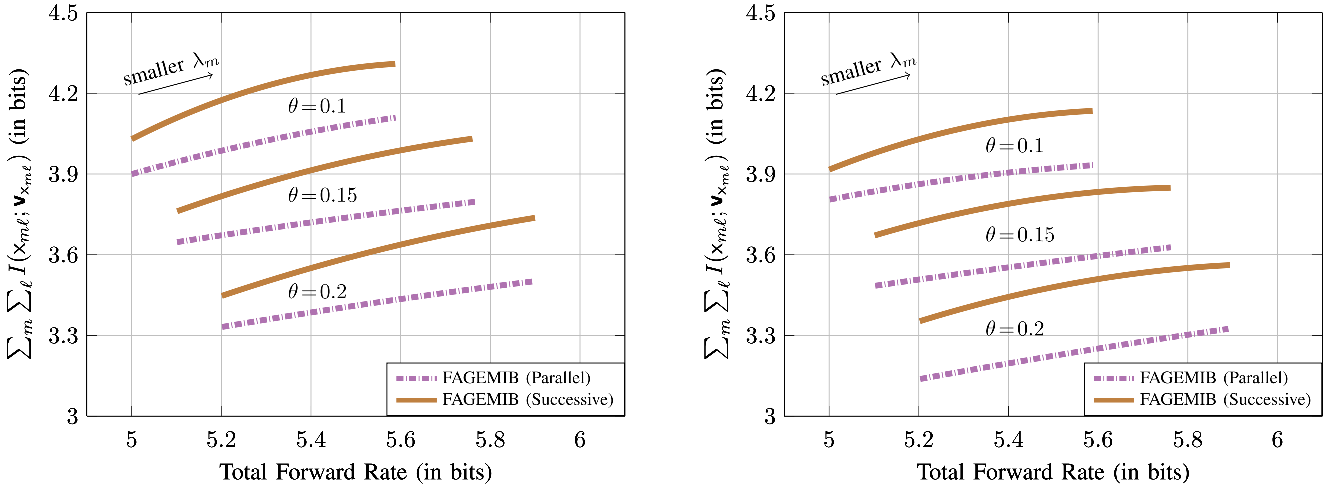
Considering all results together, we can easily distinguish the *uncommon* users, namely, x_{11}, x_{21}, from the *common* one, namely, x_{12}. More specifically, the common user, with the results in the middle, which gets served by both INs enjoys a performance boost compared to the uncommon users, with the results on both ends (i.e., left and right), which get served by a single IN. This seems expected by noticing the fact that both received signals, t_1 and t_2, at the RPU are informative about the common user signal, x_{12}, while solely one of them is informative about the other two uncommon users.

It is further observed that the performance curves of the uncommon users are quite similar. Principally, this is due to the present symmetry in the considered simulation scenario. Moreover, the same trends as the ones already observed and justified for the previous investigations are present here as well. Basically, either the looser the compression bottleneck (larger N) or the better the qualities of access and forward channels (smaller σ_{n}^2 and θ), the higher the obtained relevant information per user.

C. PARALLEL VS. SUCCESSIVE PROCESSING

The obtained overall performances by parallel and successive processing schemes are compared here as the final part of the numerical investigations. Specifically, the BPSK (Binary Phase Shift Keying) signaling is considered for uncommon sources, x_{11}, x_{21}, and the 8-ASK signaling is considered for the common source, x_{12}. To get the transition probabilities of the access links, like before, we follow a "Monte Carlo" approach and generate 20 samples per link for uncommon sources and 80 samples per link for the common source. We further set the number of output clusters (per IN) to $N=32$. The trade-off parameters λ_m (for $m=1,2$) are then varied over a particular range, and the total relevant information and the required forward rate for supporting it are calculated. The obtained results have been depicted in Fig. 6.

In Fig. 6a, the noise variance of all access links has been set to $\sigma_{\mathsf{n}}^2=0.3$ and the value of forward reliability has been varied ($\theta=0.1, 0.15, 0.2$). Moreover, similar investigations have been made in Fig. 6b for a larger access noise variance, namely, $\sigma_{\mathsf{n}}^2=0.4$. In the presented results, larger values of λ_m correspond to the solutions with the higher focus on the *compactness*, and smaller values of λ_m correspond to the solutions with the higher focus on the *informativity*.



(a) Access noise variance $\sigma_{\mathsf{n}}^2=0.3$ (b) Access noise variance $\sigma_{\mathsf{n}}^2=0.4$

FIGURE 6. Total relevant information of FAGEMIB (Parallel and Successive) vs. total forward rate for different access noise variances a) $\sigma_{\mathsf{n}}^2=0.3$ and b) $\sigma_{\mathsf{n}}^2=0.4$. BPSK signaling for uncommon sources ($\sigma_{\mathsf{x}_{11}}^2=\sigma_{\mathsf{x}_{21}}^2=1$) and 8-ASK signaling for the common source ($\sigma_{\mathsf{x}_{12}}^2=21$), with $N=32$ output clusters per IN, the convergence parameter $\epsilon=10^{-3}$, and $0.1\le\lambda_m\le 0.4$ for $m=1,2$.

Considering both results together, as the main takeaway, we can clearly observe that the successive processing scheme yields superior *information-compression* trade-offs compared to the parallel processing. Expectedly, this outperformance is achieved by better leveraging the present correlations in the signals of INs (since both commonly serve x_{12}). Specifically, by recalling the following relation (that is deduced based on the presumed Markovian relations)

$$I(\mathbf{y}_m; \mathsf{z}_m|\mathbf{t}_{1:m-1}) = I(\mathbf{y}_m; \mathsf{z}_m) - \underbrace{I(\mathsf{z}_m; \mathbf{t}_{1:m-1})}_{\geq 0}, \tag{39}$$

it is straightly realized that conditioning on previous signals can lead to a forward rate reduction. Finally, it is noteworthy that a similar trend as in previous investigations is observed here as well regarding the impacts of the qualities of access and forward links on the obtained overall performance.

VII. SUMMARY

In this article, based on the *Information Bottleneck* method, we devised novel distributed multi-user joint source-channel coding schemes for a generic scenario appearing in a broad variety of real-world applications. The setup we focused on, with the highest flexibility w.r.t. the assignment of users to the serving nodes, goes beyond the State-of-the-Art methods (which work exclusively for a single common user signal) by considering uncommon user / source signals for various local compressing unints, alongside the potentially common ones. We selected the *Mutual Information* as the fidelity criterion and characterized the stationary solutions for the introduced design problems with the help of *Variational Calculus*. Then, we proposed an (iterative) algorithm, the FAGEMIB, which leverages the derived stationary solutions to efficiently tackle both design problems. We further provided the convergence proofs (to a stationary point of the objective functionals), along with a detailed analysis on the behavior of FAGEMIB over the entire range of its main parameters. To corroborate the effectiveness of FAGEMIB, we further presented several numerical results over typical digital transmission scenarios.

APPENDIX

In the following, we will provide the detailed proofs of two main theorems regarding the form of stationary solutions for both parallel and successive processing schemes. In these derivations, the *functional derivative* in *Variational Calculus* appears as the central player, since it generalizes the concept of gradient to the cases of dealing with the optimization of functionals (i.e., functions of functions) w.r.t. input functions.

A. PROOF OF THEOREM 1 (PARALLEL PROCESSING)

The parallel processing's Lagrangian, $\mathcal{L}_{\text{Par.}}$, depends on all quantizer mappings, $\{p(\mathsf{z}_m|\mathbf{y}_m)\ |m\}$. Thus, by definition of a stationary solution, all (functional) derivatives w.r.t. these mappings should vanish. We associate a Lagrange multiplier, $\lambda_{\boldsymbol{y}_m}$, to every input realization, $\boldsymbol{y}_m \in \boldsymbol{\mathcal{Y}}_m$, of the m-th local compressor mapping's input set of variables, $\mathbf{y}_m = \mathbf{Pa}^{\mathcal{G}_{\text{in}}}_{\mathsf{z}_m}$, to further include the validity conditions into the analysis. Subsequently, the overall Lagrangian for parallel processing scheme, $\mathcal{L}^{\text{Ov.}}_{\text{Par.}}$, is introduced as

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

in which, $\mathbf{v}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\text{out}}}_{\mathsf{x}_{m\ell}}$. Now, we calculate the (functional) derivative w.r.t. the j-th compressor mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$. First, note that $(\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{v}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{t}_j\})$

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

wherein, $\mathbf{w}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}_{\mathsf{x}_{m\ell}}^{\mathcal{G}_{\text{out}}}$ for the counterpart case of *error-free* forwarding, that is, replacing t_j with z_j in $\mathcal{G}_{\text{out}}$ for every $j \in \{1,\ldots,J\}$, and $\mathbf{w}_{\mathsf{x}_{m\ell}}^{-j} = \mathbf{w}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$. Moreover, it applies

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

and

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

Equating the functional derivative to zero, from (41), (42), (43), and by assuming (without loss of generality) $p(\boldsymbol{y}_j) > 0$, the following is immediately inferred

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

wherein, it applies

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

Thereupon, we perform these steps: we bring the second term in (44) to the other side of equality, then we multiply both sides by $\beta_j = \frac{1}{\lambda_j}$, and exponentiate them afterwards. Finally, we again multiply both sides by $p(z_j)$. Subsequently, the following holds true

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

Enforcing the validity condition for the compressor mapping, that is, $\sum_{z_j} p(z_j|\boldsymbol{y}_j) = 1$, the term $\exp(-\beta_j\tilde{\lambda}_{\boldsymbol{y}_j}^{\text{Par.}})$ can then be regarded as the normalization function, $\psi_{\mathsf{z}_j}^{\text{Par.}}$, to achieve the stated solution in Theorem 1. ■

B. PROOF OF THEOREM 2 (SUCCESSIVE PROCESSING)

The successive scheme's Lagrangian, $\mathcal{L}_{\text{Suc.}}$, depends on all compressor mappings, $\{p(\mathsf{z}_m|\mathbf{y}_m)\ |m\}$. Thus, by definition of a stationary solution, all functional derivatives w.r.t. these mappings must vanish. Associating a Lagrange multiplier to each realization, $\boldsymbol{y}_m \in \boldsymbol{\mathcal{Y}}_m$, of the m-th (local) compressor mapping's input set of variables, $\mathbf{y}_m = \mathbf{Pa}_{\mathsf{z}_m}^{\mathcal{G}_{\text{in}}}$, we introduce the overall Lagrangian for successive scheme, $\mathcal{L}_{\text{Suc.}}^{\text{Ov.}}$, as

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

in which $\mathbf{v}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}_{\mathsf{x}_{m\ell}}^{\mathcal{G}_{\text{out}}}$. Now, we calculate the (functional) derivative of the above overall Lagrangian w.r.t. the j-th compressor mapping, $p(\mathsf{z}_j|\mathbf{y}_j)$. For that, we need to compute only the derivative of the second term on the right side of (47), since the other derivatives (i.e., for the first and third term) have been already calculated in (41) and (43). To this end, first note that it applies

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

simply since the first $j-1$ terms in the summation do not depend on $p(\mathsf{z}_j|\mathbf{y}_j)$. Further, it holds true that

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

and for $j < m \le J$

$$\frac{\delta I\left(\mathbf{y}_m;\mathsf{z}_m|\mathbf{t}_{1:m-1}\right)}{\delta p\left(\mathsf{z}_j|\mathbf{y}_j\right)} = \frac{\delta H\left(\mathsf{z}_m|\mathbf{t}_{1:m-1}\right)}{\delta p\left(\mathsf{z}_j|\mathbf{y}_j\right)} - \frac{\delta H\left(\mathsf{z}_m|\mathbf{y}_m,\mathbf{t}_{1:m-1}\right)}{\delta p\left(\mathsf{z}_j|\mathbf{y}_j\right)}$$
$$= \frac{\delta H\left(\mathsf{z}_m|\mathbf{t}_{1:m-1}\right)}{\delta p\left(\mathsf{z}_j|\mathbf{y}_j\right)} - \underbrace{\frac{\delta H\left(\mathsf{z}_m|\mathbf{y}_m\right)}{\delta p\left(\mathsf{z}_j|\mathbf{y}_j\right)}}_{0}$$
$$= \frac{\delta H\left(\mathsf{z}_m|\mathbf{t}_{1:m-1}\right)}{\delta p\left(\mathsf{z}_j|\mathbf{y}_j\right)}$$
$$= p\left(\boldsymbol{y}_j\right)\sum_{\boldsymbol{t}_{1:m-1},z_m} p\left(t_j|z_j\right)p\left(\boldsymbol{t}_{1:m-1}^{-j},z_m|\boldsymbol{y}_j\right)$$
$$\times\log\frac{1}{p\left(z_m|\boldsymbol{t}_{1:m-1}\right)}. \tag{50}$$

Analogous to the previous case of parallel processing, when setting the functional derivative to zero, from (41), (43), (49), (50), and by assuming (without loss of generality) $p(\boldsymbol{y}_j) > 0$, the following is directly deduced

$$-\sum_{(m,\ell):\, \mathsf{t}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p\left(\boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}} | \boldsymbol{y}_j\right)} \left\{ \sum_{\boldsymbol{v}_{\mathsf{x}_{m\ell}}} p\left(\boldsymbol{v}_{\mathsf{x}_{m\ell}} | \boldsymbol{w}_{\mathsf{x}_{m\ell}}\right) \times D_{\mathrm{KL}}\left(p\left(\mathsf{x}_{m\ell} | \boldsymbol{y}_j, \boldsymbol{w}^{-j}_{\mathsf{x}_{m\ell}}\right) \| p\left(\mathsf{x}_{m\ell} | \boldsymbol{v}_{\mathsf{x}_{m\ell}}\right)\right) \right\} - \lambda_j \log \frac{p(z_j|\boldsymbol{y}_j)}{p(z_j)} + \lambda_j \sum_{\boldsymbol{t}_{1:j-1}} p(\boldsymbol{t}_{1:j-1}|\boldsymbol{y}_j) \log p(\boldsymbol{t}_{1:j-1}|z_j) + \sum_{k=j+1}^{J} \lambda_k \sum_{\boldsymbol{t}_{1:k-1}, z_k} p(t_j|z_j) p\left(\boldsymbol{t}^{-j}_{1:k-1}, z_k | \boldsymbol{y}_j\right) \log p(z_k|\boldsymbol{t}_{1:k-1}) + \tilde{\lambda}^{\mathrm{Suc.}}_{\boldsymbol{y}_j} = 0, \tag{51}$$

wherein, $\mathbf{w}_{\mathsf{x}_{m\ell}} = \mathbf{Pa}^{\mathcal{G}_{\mathrm{out}}}_{\mathsf{x}_{m\ell}}$ for the counterpart case of *error-free* forwarding, that is, replacing t_m with z_m in $\mathcal{G}_{\mathrm{out}}$ for every $m \in \{1, \ldots, J\}$, $\mathbf{w}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{w}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$, and ($\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{v}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{t}_j\}$)

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

Thereupon, we perform these steps: we bring the second term in (44) to the other side of equality, then we multiply both sides by $\beta_j = \frac{1}{\lambda_j}$, and exponentiate them afterwards. Finally, we again multiply both sides by $p(z_j)$. Subsequently, the following holds true

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

Enforcing the validity condition for the compressor mapping, that is, $\sum_{z_j} p(z_j|\boldsymbol{y}_j) = 1$, the term $\exp(-\beta_j \tilde{\lambda}^{\mathrm{Suc.}}_{\boldsymbol{y}_j})$ can then be regarded as the normalization function, $\psi^{\mathrm{Suc.}}_{\mathsf{z}_j}$, to achieve the stated solution in 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.*, 1999, pp. 1–16.
[2] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE Int. Conv. Rec.*, vol. 7, pp. 142–163, Mar. 1959.
[3] R. Dobrushin and B. Tsybakov, “Information transmission with additional noise,” *IRE Trans. Inf. Theory*, vol. 8, no. 5, pp. 293–304, Sep. 1962.
[4] D. Sakrison, “Source encoding in the presence of random disturbance,” *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 165–167, Jan. 1968.
[5] J. Wolf and J. Ziv, “Transmission of noisy information to a noisy receiver with minimum distortion,” *IEEE Trans. Inf. Theory*, vol. 16, no. 4, pp. 406–411, Jul. 1970.
[6] H. Witsenhausen, “Indirect rate distortion problems,” *IEEE Trans. Inf. Theory*, vol. 26, no. 5, pp. 518–521, Sep. 1980.
[7] E. Ayanoglu, “On optimal quantization of noisy sources,” *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1450–1452, Nov. 1990.
[8] 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.
[9] 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.
[10] 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.
[11] S. Hu, Z. Lou, X. Yan, and Y. Ye, “A survey on information bottleneck,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5325–5344, Aug. 2024.
[12] A. Wyner, “On source coding with side information at the decoder,” *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.
[13] 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.
[14] E. Erkip and T. M. Cover, “The efficiency of investment information,” *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.
[15] 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.
[16] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, “Information extraction under privacy constraints,” *Information*, vol. 7, no. 1, p. 15, Mar. 2016.
[17] B. Razeghi, P. Rahimi, and S. Marcel, “Deep variational privacy funnel: General modeling with applications in face recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 4920–4924.
[18] V. Doshi, D. Shah, M. Médard, and M. Effros, “Functional compression through graph coloring,” *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3901–3917, Aug. 2010.
[19] S. Feizi and M. Médard, “On network functional compression,” *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5387–5401, Sep. 2014.
[20] Y. M. Saidutta, A. Abdi, and F. Fekri, “Analog joint source-channel coding for distributed functional compression using deep neural networks,” in *Proc. IEEE Int. Symp. Inf. Theory*, 2021, pp. 2429–2434.
[21] D. Malak and M. Médard, “A distributed computationally aware quantizer design via hyper binning,” *IEEE Trans. Signal Process.*, vol. 71, pp. 76–91, Jan. 2023.
[22] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in *Proc. IEEE Inf. Theory Workshop*, 2015, pp. 1–5.
[23] R. Shwartz-Ziv and N. Tishby, “Opening the black box of deep neural networks via information,” 2017, *arXiv:1703.00810*.
[24] A. M. Saxe et al., “On the information bottleneck theory of deep learning,” *J. Stat. Mech., Theory Exp.*, vol. 2019, no. 12, Dec. 2019, Art. no. 124020.
[25] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–19.
[26] H. Hafez-Kolahi and S. Kasaei, “Information bottleneck and its applications in deep learning,” 2019, *arXiv:1904.03743*.
[27] B. C. Geiger and G. Kubin, “Information bottleneck: Theory and applications in deep learning,” *Entropy*, vol. 22, no. 12, p. 1408, 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] 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.
[32] J. Lewandowsky and G. Bauch, “Information-optimum LDPC decoders based on the information bottleneck method,” *IEEE Access*, vol. 6, pp. 4054–4071, 2018.
[33] 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.

[34] 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.
[35] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6562–6582, Oct. 2013.
[36] 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.
[37] 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.
[38] T. Monsees, D. Wübben, and A. Dekorsy, "Optimum quantization of memoryless channels with N-ary input," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2022, pp. 158–162.
[39] J. Lewandowsky and G. Bauch, "Theory and application of the information bottleneck method," *Entropy*, vol. 26, no. 3, p. 187, Feb. 2024.
[40] I. Estella-Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources," in *Proc. Int. Zurich Semin. Inf. Commun.*, 2018, pp. 1–5.
[41] I. Estella-Aguerri, A. Zaidi, G. Caire, and S. Shamai, "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4575–4596, Jul. 2019.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] S. Hassanpour, M. Hummert, D. Wübben, and A. Dekorsy, "Deep learning-based distributed remote source coding via information bottleneck method: The parallel processing scheme," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2025, pp. 1–6.
[47] S. Hassanpour, M. Hummert, D. Wübben, and A. Dekorsy, "A deep variational approach to multiterminal joint source-channel coding based on information bottleneck principle," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 4462–4475, 2025.
[48] 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.
[49] S. Hassanpour, A. Danaee, D. Wübben, and A. Dekorsy, "Multi-source distributed data compression based on information bottleneck principle," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 4171–4185, 2024.
[50] 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.
[51] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.
[52] 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.
[53] A. Danaee, S. Hassanpour, D. Wübben, and A. Dekorsy, "Relevance-based information processing for fronthaul rate reduction in cell-free MIMO systems," in *Proc. 19th Int. Symp. Wireless Commun. Syst.*, 2024, pp. 1–6.
[54] A. Danaee, S. Hassanpour, D. Wübben, and A. Dekorsy, "Relevance-based multi-user data compression for fronthaul rate reduction in cell-free massive MIMO systems," in *Proc. Int. ITG Conf. Syst., Commun. Coding*, 2025, pp. 1–6.
[55] D. Wubben 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.
[56] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "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.
[57] 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.
[58] 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.
[59] 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.
[60] P. Popovski et al., "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.
[61] C. Roth, C. Benkeser, C. Studer, G. Karakonstantis, and A. Burg, "Data mapping for unreliable memories," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput.*, 2012, pp. 679–685.
[62] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[63] 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.
[64] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Cambridge, MA, USA: Academic, 1982.
[65] J. H. Mathews and K. D. Fink, *Numerical Methods Using MATLAB*, 4th ed. Hoboken, NJ, USA: Pearson Prentice Hall, 2004.
[66] M. Hummert, S. Hassanpour, D. Wübben, and A. Dekorsy, "Deep FAVIB: Deep learning-based forward-aware quantization via information bottleneck method," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 1885–1890.
[67] N. Slonim, N. Friedman, and N. Tishby, "Multivariate information bottleneck," *Neural Comput.*, vol. 18, no. 8, pp. 1739–1789, Aug. 2006.
[68] 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.
[69] 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.
[70] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[71] R. Horst, P. M. Pardalos, and N. Van Thoai, *Introduction to Global Optimization*, 2nd ed. New York, NY, USA: Springer, 2000.
[72] 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.
[73] 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 in communications engineering (program's award winner) from the Ulm University, Germany, in 2014, and the Dr.-Ing. degree (with summa cum laude) in electrical engineering (communications) from the University of Bremen, Germany, in 2022. He is a Postdoctoral Researcher with the Department of Communications Engineering, University of Bremen, Germany. Over the last couple of years, he has been prolifically contributing to the top-tier international journals and IEEE flagship conferences on his Ph.D. topic, that is, the information bottleneck method. His other research interests include information theory, MU/MIMO systems, wireless communications, statistical signal processing, and the applications of modern deep learning in the design of communication systems. He received the 2021's VDE/ITG Best Paper Award and the 2023's OHB Best Doctoral Dissertation Award.

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. He is currently a Senior Research Group Leader and a Lecturer with the Department of Communications Engineering, University of Bremen. He has published more than 160 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 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." He has been an Editor of IEEE WIRELESS COMMUNICATIONS LETTERS. He received the VDE/ITG Best Paper Award 2021 and the VDE/ITG Certificate of Honor in 2024.

ARMIN DEKORSY (Senior Member, IEEE) is a Professor with the University of Bremen, where he is the Director of the Gauss-Olbers Space Technology Transfer Center and Heads the Department of Communications Engineering. With over 11 years of industry experience, including distinguished research positions, such as a DMTS with Bell Labs and a Research Coordinator Europe with Qualcomm, he has actively participated in more than 65 international research projects, with leadership roles in 17 of them. He co-authored the textbook *Nachrichtenübertragung* (Release 6, Springer Vieweg), which is a bestseller in the field of communication technologies in German-speaking countries. His research focuses on signal processing and wireless communications for 5G/6G, industrial radio, and 3-D networks. He is a Senior Member of the IEEE COMMUNICATIONS AND SIGNAL PROCESSING SOCIETY and a member of the VDE/ITG Expert Committee on Information and System Theory.