This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

IEEE Open Journal of the Communications Society

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/OJCOMS.2025.XXXXXXX

A Deep Variational Approach to Multiterminal Joint Source-Channel Coding Based on Information Bottleneck Principle

SHAYAN HASSANPOUR (Member, IEEE), MATTHIAS HUMMERT (Student 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 partly funded by the German ministry of education and research (BMBF) under grants 16KISK016 (Open6GHub), 16KISK068 (6G-TakeOff), and 16KISK109 (6G-ANNA).

ABSTRACT In this article, we concentrate on a generic multiterminal joint source-channel coding scenario, appearing in a wide variety of real-world applications. Specifically, several *noisy* observations from a source signal must be compressed at some intermediate nodes, before getting forwarded over multiple *error-prone* and *rate-limited* channels towards a (remote) processing unit. The imperfections of the forward channels should be integrated into the design of (local) compressor units. By following the *Information Bottleneck* principle, the *Mutual Information* is selected here as the fidelity criterion, and a novel (data-driven) design approach is presented for two distinct types of processing flow / strategy at the remote unit. To that end, tractable objective functions are developed, together with the pertinent learning architectures, generalizing the concepts of *Variational Auto-Encoders* and *(Distributed) Deep Variational Information Bottleneck* for (remote) source coding to the context of distributed joint source-channel coding. Unlike the conventional approaches, the proposed schemes here work based upon a finite sample set, thereby obviating the call for full prior knowledge of the joint statistics of input signals. The effectiveness of these novel sample-based compression schemes is substantiated as well by a couple of simulations over typical transmission setups.

INDEX TERMS 6G, auto-encoders, deep learning, information bottleneck, joint source-channel coding

I. INTRODUCTION

THE original formulation of the *Information Bottleneck (IB)* method [1] was based upon the Shannon's seminal work on *lossy* source coding [2]. To quantify the fundamental limits of the inevitable "*complexity-precision*" trade-off, the concept of *Rate-Distortion (RD) function* has been defined. The central optimization in the IB formulation then applies an intuitive twist on the *single-letter* characterization of this RD function by lower-bounding a *Mutual Information* term w.r.t. a *target / relevant variable* rather than upper-bounding an expected distortion term. The cogent reason behind this modification comes from the fact that, in many real-world data compression applications, pinpointing a target variable whose information must be retained becomes a considerably simpler task compared to the (often challenging) problem of determining the suitable distortion function. With the certain choice of *Logarithmic Loss* distortion [3], it was shown later on (see, e.g., [4]) that solving the IB constrained optimization problem delivers the boundary of achievable rate-distortion region for a *remote / indirect* source coding problem [5]–[9]. To obtain a clearer picture on this variational principle from the vantage points of both Information and Learning Theory, interested readers are referred to [10]–[12]. Important to note are the connections to some other (classic) problems, ranging from the Wyner-Ahlswede-Körner problem [13], [14], to the efficiency of investment information [15], the privacy funnel [16]–[18], and last but not least, the (distributed) functional compression [19]–[23].

From a practical standpoint and apart from purely theoretical investigations, it is further noteworthy that advanced data transmission systems have already adopted the IB principle. The pertinent applications range from efficient construction of the Polar Codes [24], [25], to discrete (channel) decoding schemes [26]–[29], Analog-to-Digital converters for receiver front ends [30], [31], and also in the semantic / task-oriented communication schemes [32]–[35].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

The multiterminal / distributed *model-based* extensions of the IB principle have also been considered in the pertinent literature (see, e.g., [36]–[43]). These quantization schemes mostly consider a certain scenario in which several *noisy* observations from a source signal must be compressed (by various strategies and at potentially different rates) ahead of a forward transmission to a (remote) processing unit via multiple rate-limited channels. Depending on whether these forward links are presumed to be *error-free* or *error-prone*, the underlying design problems become instances of either (remote) source or joint source-channel coding, respectively. The interested readers are referred to [44] for further recent studies on both the theory and applications of the IB method.

The recent rise of machine learning-based approaches has brought a new and fresh perspective to the design of IB-based compression schemes. Through leveraging the capabilities of neural networks and deep learning, innovative architectures have been explored to (efficiently) optimize the IB trade-off (i.a., [45]–[47]). These data-driven approaches which work based upon a finite sample set, obviate the call for prior knowledge of the joint statistics of input signals. Moreover, they can efficiently handle high-dimensional and potentially continuous data. These points make them a more attractive choice compared to the conventional model-based algorithms in many applications of interest. Motivated by this, herein, we present a direct generalization of these approaches, which have been built upon generative latent variable models, to the context of distributed joint source-channel coding.

A. CONTRIBUTIONS

Within the scope of this article, we develop deep variational approaches to address the design problems for a (generic) distributed / multiterminal joint source-channel coding setup in which several *noisy* observations from a *common* source signal shall be (locally) compressed at multiple intermediate nodes, before getting forwarded over a couple of *error-prone* and *rate-limited* channels towards a (remote) processing unit. Explicitly, by following the *Information Bottleneck* method, we choose the *Mutual Information* as the fidelity criterion to formulate the pertinent design problems for two different types of processing strategies. Thereupon, we derive tractable objective functions and introduce the corresponding learning architectures to tackle the given design problems by standard training of several Deep Neural Networks (DNNs), e.g., via Stochastic Gradient Descent (SGD) with back-propagation. The devised schemes in this article generalize (well-known) concepts of *Variational Auto-Encoder (VAE)* [48], [49], *Deep Variational Information Bottleneck* [45], and its distributed extension for (remote / indirect) source coding [39], [46], to the context of multiterminal joint source-channel coding.

To clearly realize the generality of the considered setup here, note that it appears in a broad variety of real-world applications regarding the fifth (5G) and sixth (6G) generations of wireless networking technologies, from cooperative relaying schemes with the *Quantize-and-Forward* approach [50], [51], to Cloud-based Radio Access Networks (Cloud-RANs) [37], [52], [53], as well as Cell-Free massive Multiple-Input Multiple-Output systems (CF-mMIMO) with the *error-prone* and *rate-limited* fronthaul links [54]–[59], and in distributed inference sensor networks with *imperfect* links to the fusion center [60], [61].

B. OUTLINE

We start our technical discussion with the point-to-point IB-based joint source-channel coding scheme in Section II as a prelude towards distributed extensions. In Section III, we present both the system model and the corresponding design problems for the *parallel* and *successive* processing schemes. Section IV has been fully dedicated to the presentation of our deep variational approaches to address the introduced design problems. Subsequently, in Section V, a couple of numerical results are presented to corroborate the effectiveness of these (data-driven) compression schemes. Finally, a brief wrap-up in Section VI concludes this article.

C. NOTATIONS

The realizations, $a \in \mathcal{A}$, from the (discrete) random variable, a, occur according to the distributions, $p/q/r/s$. The same holds true for the random vector, $\mathbf{a}_{1:J} = \{\mathsf{a}_1, \cdots, \mathsf{a}_J\}$, with the boldface counterparts. Further, $\mathbf{a}_{1:J}^{\bar{j}} = \mathbf{a}_{1:J} \setminus \{\mathsf{a}_j\}$, and $\mathbb{E}_{\bullet}\{\cdot\}$, denotes the expectation operator. Finally, $D_{\mathrm{KL}}(\cdot\|\cdot)$, $H(\cdot)$, and $I(\cdot;\cdot)$, stand for *Kullback-Leibler (KL)* divergence, Shannon's entropy, and Mutual Information [62], and $\{\cdot\}_{j=1}^{J}$ denotes a set of J elements.

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

We start our discussion by introducing the point-to-point IB-based joint source-channel coding setup and design problem, together with a roadmap for recasting the original design formulation into a certain form which can be addressed by a deep variational approach. A similar roadmap is followed later on for distributed / multiterminal extensions.

A. SYSTEM MODEL AND PROBLEM FORMULATION

The illustrated system model in Fig. 1 is considered. Explicitly, an intermediate node must compress a noisy observation, y, from the source signal, x, ahead of a forward transmission over an *error-prone* and *rate-limited* channel to a (remote) processing unit. Denoting by, z, the compressed signal at the input of the forward channel, and by, t, the noisy counterpart at its output, it is presumed that both the statistics, $p(\mathsf{t}|\mathsf{z})$, of the forward channel and its capacity, R, are known. The source statistics, $p(\mathsf{x})$, and the transition probabilities, $p(\mathsf{y}|\mathsf{x})$, of the access channel are also presumed to be known. The goal is to design the compressor such that the imperfections of the forward channel are taken into account. Fully aligned with the main idea behind the *Information Bottleneck (IB)* framework [1], the design problem has then been formulated in [63] by establishing a basic trade-off among two *Mutual Information* terms.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

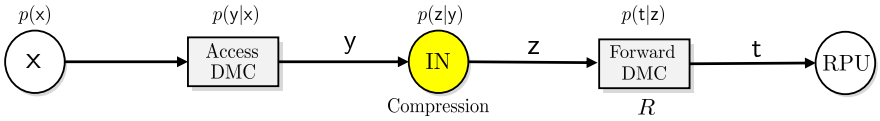


FIGURE 1. The system model for point-to-point joint source-channel coding. DMC, IN, and RPU stand for Discrete Memoryless Channel, Intermediate Node, and Remote Processing Unit, respectively. The Markov chain, $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z} \leftrightarrow \mathsf{t}$,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

On the one hand, the *relevant information*, $I(\mathsf{x};\mathsf{t})$, is the term which quantifies the *informativity* of outcome. On the other hand, the *compression rate*, $I(\mathsf{y};\mathsf{z})$, is the term that quantifies its *compactness*. The design problem is then mathematically formulated as a constrained optimization task wherein, the relevant information is maximized such that the compression rate does not exceed the capacity / rate-limit of the forward link. Explicitly, it applies

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

in which, $0 \le R \le \log_2 |\mathcal{Z}|$ bits, sets an upper-bound on the compression rate, $I(\mathsf{y};\mathsf{z})$. Applying the method of *Lagrange Multipliers* [64], we can then recast this problem into an unconstrained maximization (up to the validity of quantizer mapping), namely,

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

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., via applying a bisection search on a finite range. Utilizing *Variational Calculus*, the stationary solution of (2) has been derived in [63] as

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

for each $(y,z) \in \mathcal{Y}\times\mathcal{Z}$, with $\beta = \frac{1}{\lambda}$, and $\omega(y,\beta)$, being a partition function to ensure a valid mapping. Specifically, for every realization $y \in \mathcal{Y}$, the sum of calculated terms in (3) (ignoring ω) over all output bins / clusters, $z \in \mathcal{Z}$, acts as the partition function. Furthermore, an iterative algorithm, called Forward-Aware Vector Information Bottleneck (FAVIB), has been presented in [63] performing the *Fixed-Point Iterations* [65] on the stationary solution (3). Its convergence proof to a stationary point of the pertinent objective functional in (2) has also been provided.

In the following part, we build up a deep learning approach to (approximately) address the design problem (2), when instead of the full joint statistics, $p(\mathsf{x},\mathsf{y})$, solely a finite sample set, $\{(x^{(i)},y^{(i)})\}_{i=1}^{M}$, is available. This approach which directly generalizes the *Deep Variational Information Bottleneck (DVIB)* [45] to the context of joint source-channel coding by considering an *error-prone* forward link, is based upon latent variable models, specifically, the well-known concept of *Variational Auto-Encoders (VAEs)* [48], [49].

B. THE ROADMAP TO DEEP VARIATIONAL APPROACH

The starting point towards developing a data-driven design method is to introduce a *Variational Lower-Bound (VLB)* on the IB-based point-to-point / centralized joint source-channel coding Lagrangian

$$\mathcal{L}_{\text{Cent.}}\big(p(\mathsf{z}|\mathsf{y})\big) = I(\mathsf{x};\mathsf{t}) - \lambda I(\mathsf{y};\mathsf{z})\,. \tag{4}$$

To that end, by introducing two *proxy* distributions, namely, $q(\mathsf{x}|\mathsf{t})$, as the decoder (to retrieve the source, x), and, $r(\mathsf{z})$, as the latent prior, the following relations hold true

$$I(\mathsf{x};\mathsf{t}) = \underbrace{H(\mathsf{x})}_{\ge 0} - H(\mathsf{x}|\mathsf{t}) \tag{5a}$$

$$\ge \underbrace{\sum_{t\in\mathcal{T}} p(t) D_{\mathrm{KL}}\big(p(\mathsf{x}|t)\|q(\mathsf{x}|t)\big)}_{\ge 0} + \sum_{x\in\mathcal{X},\, t\in\mathcal{T}} p(x,t)\log q(x|t) \tag{5b}$$

$$\ge \mathbb{E}_{\mathsf{x},\mathsf{t}}\{\log q(\mathsf{x}|\mathsf{t})\}, \tag{5c}$$

wherein, from (5a) to (5b), the non-negativity of entropy (for the discrete source signal, x), and, from (5b) to (5c), the non-negativity of KL divergence, also known as the *information inequality*, has been applied [62]. Moreover, it holds true that

$$I(\mathsf{y};\mathsf{z}) = \sum_{y\in\mathcal{Y},\, z\in\mathcal{Z}} p(y,z)\log\frac{p(z|y)}{r(z)} - \underbrace{D_{\mathrm{KL}}\big(p(\mathsf{z})\|r(\mathsf{z})\big)}_{\ge 0} \tag{6a}$$

$$\le \mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{\log\frac{p(\mathsf{z}|\mathsf{y})}{r(\mathsf{z})}\right\} = \sum_{y\in\mathcal{Y}} p(y) D_{\mathrm{KL}}\big(p(\mathsf{z}|y)\|r(\mathsf{z})\big)\,. \tag{6b}$$

From (6a) to (6b), the non-negativity of KL divergence has been applied. Now, we can define our VLB as

$$\mathcal{L}^{\text{VLB}}_{\text{Cent.}}\big(p(\mathsf{z}|\mathsf{y}), q(\mathsf{x}|\mathsf{t}), r(\mathsf{z})\big) = \mathbb{E}_{\mathsf{x},\mathsf{t}}\{\log q(\mathsf{x}|\mathsf{t})\} - \lambda \mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{\log\frac{p(\mathsf{z}|\mathsf{y})}{r(\mathsf{z})}\right\}, \tag{7}$$

since from (5) and (6), it is immediately inferred that[1]

$$\mathcal{L}_{\text{Cent.}}\big(p(\mathsf{z}|\mathsf{y})\big) \ge \max_{q,r}\; \mathcal{L}^{\text{VLB}}_{\text{Cent.}}\big(p(\mathsf{z}|\mathsf{y}), q(\mathsf{x}|\mathsf{t}), r(\mathsf{z})\big), \tag{8}$$

and as a direct consequence,

$$\max_{p}\; \mathcal{L}_{\text{Cent.}}\big(p(\mathsf{z}|\mathsf{y})\big) \ge \max_{p}\max_{q,r}\; \mathcal{L}^{\text{VLB}}_{\text{Cent.}}\big(p(\mathsf{z}|\mathsf{y}), q(\mathsf{x}|\mathsf{t}), r(\mathsf{z})\big)\,. \tag{9}$$

The next step is then to consider a parameterized family for all input distributions of the introduced VLB. Specifically, denoting by $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$, the parameter sets for families of distributions regarding the encoder, p, the decoder, q, and the latent prior, r, and by $\mathcal{L}^{\text{DNN}}_{\text{Cent.}}$, the defined VLB with the parameterized input distributions, i.e.,

$$\mathcal{L}^{\text{DNN}}_{\text{Cent.}}\big(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}\big) = \mathcal{L}^{\text{VLB}}_{\text{Cent.}}\big(p_{\boldsymbol{\theta}}(\mathsf{z}|\mathsf{y}), q_{\boldsymbol{\phi}}(\mathsf{x}|\mathsf{t}), r_{\boldsymbol{\psi}}(\mathsf{z})\big), \tag{10}$$

[1] For clarity: $p = p(\mathsf{z}|\mathsf{y})$, $q = q(\mathsf{x}|\mathsf{t})$, $r = r(\mathsf{z})$.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

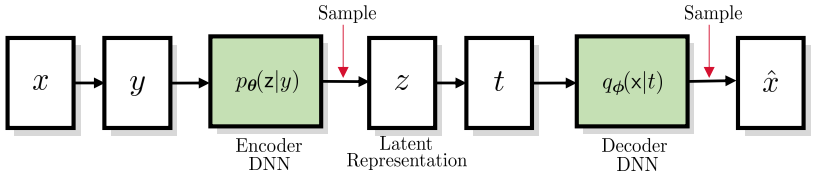


FIGURE 2. The considered learning architecture for point-to-point joint source-channel coding scheme, extending the ones for DVIB [45] and VAE [48].

it applies

$$\max_{p}\max_{q,r}\mathcal{L}^{\text{VLB}}_{\text{Cent.}}\big(p(\mathsf{z}|\mathsf{y}),q(\mathsf{x}|\mathsf{t}),r(\mathsf{z})\big)\geq\max_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}}\mathcal{L}^{\text{DNN}}_{\text{Cent.}}\big(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}\big)\,. \tag{11}$$

The inequality in (11) holds true since the search space over (valid) distributions, p, q, r, is restricted to the hypothesis space of pertinent parameterized families.

By considering two Deep Neural Networks (DNNs) for the encoder, $p_{\boldsymbol{\theta}}$, and the decoder, $q_{\boldsymbol{\phi}}$, the learning architecture illustrated in Fig. 2 is utilized to address the optimization of $\mathcal{L}^{\text{DNN}}_{\text{Cent.}}$, based upon a finite sample set, $\{(x^{(i)},y^{(i)})\}_{i=1}^{M}$, and the prior knowledge of the forward statistics, $p(\mathsf{t}|\mathsf{z})$. For estimating the gradients of $\mathcal{L}^{\text{DNN}}_{\text{Cent.}}$, the conventional approach of utilizing the *reparameterization trick* [48] to enable the *Monte-Carlo Sampling*, and subsequently, substituting the expectation terms with their (empirical) estimates can be exploited here as well. Since the principal focus of this work is on discrete latent spaces, the *Gumbel-Softmax / Concrete Distribution* [66], [67] can be utilized to do the trick for us, i.e., reparameterizing the underlying categorical distribution. This data-driven compression scheme was first introduced in [68], and is called Deep Forward-Aware Vector Information Bottleneck (Deep FAVIB).

It must be mentioned that the parameterized lower-bound, $\mathcal{L}^{\text{DNN}}_{\text{Cent.}}$, for the IB-based point-to-point joint source-channel coding generalizes the well-known *Evidence Lower-BOund (ELBO)* that is used to train VAEs [48], when the *Evidence* itself (i.e., the direct *Maximum Likelihood* objective function) becomes intractable [69].

III. MULTITERMINAL EXTENSIONS: SYSTEM MODEL AND PROBLEM FORMULATION

For distributed compression, we focus on the depicted system model in Fig. 3. Explicitly, a number, J, of intermediate nodes receive noisy observations, y_j for $j=1$ to J, from the source signal, x. Each node, j, should then compress y_j into another signal, namely, z_j, ahead of a forward transmission over an *error-prone* and *rate-limited* channel to the (remote) processing unit. The goal is then to *jointly* design the (local) compressor units, $p(\mathsf{z}_j|\mathsf{y}_j)$ for $j=1$ to J, such that the imperfections of the forward channels are taken into account. Similar to the point-to-point case, it is presumed that the source statistics, $p(\mathsf{x})$, as well as both access, $p(\mathsf{y}_j|\mathsf{x})$, and forward transition probabilities, $p(\mathsf{t}_j|\mathsf{z}_j)$, and pertinent capacities, R_j, are known for $j=1$ to J. Furthermore, it is presumed that the Markovian relation $\mathsf{x}\leftrightarrow\mathsf{y}_j\leftrightarrow\mathsf{z}_j\leftrightarrow\mathsf{t}_j$ applies per branch, j, and the counterpart signals of every two distinct branches are (conditionally) independent, given the source signal, x.

Pursuing the IB philosophy, one shall formulate the design problem(s) as a basic trade-off among the *informativity* and *compactness* of the resultant outcomes. The informativity is naturally quantified by the Mutual Information, $I(\mathsf{x};\mathbf{t}_{1:J})$, between the source, x, and the set of all forward channels outputs, $\mathbf{t}_{1:J}$, that are fed into the processing unit to retrieve the source. However, there is no natural unique choice for the other side of trade-off. So, different meaningful expressions can be applied. In the next part, based upon the capacities of forward channels and the chosen processing strategy / flow at the remote processing unit, two different sets of constraints are introduced to stipulate the compactness of outcomes.

A. PARALLEL SCHEME: IGNORING SIDE-INFORMATION

Presuming a fully *parallel* processing flow at the remote unit (to retrieve the source signal, x), the first set of constraints comprises individual compression rates of local compressors. In this manner, no side-information is leveraged at the remote unit. Mathematically, the design problem is formulated as the following constrained maximization

$$P^{*}=\{p^{*}(\mathsf{z}_1|\mathsf{y}_1),\cdots,p^{*}(\mathsf{z}_J|\mathsf{y}_J)\}=\underset{P:\,\forall j\;I(\mathsf{y}_j;\mathsf{z}_j)\leq R_j}{\operatorname{argmax}}\;I(\mathsf{x};\mathbf{t}_{1:J}), \tag{12}$$

in which, $0\leq R_j\leq\log_2|\mathcal{Z}_j|$ bits, sets an upper-bound on the j-th compression rate, $I(\mathsf{y}_j;\mathsf{z}_j)$. By exploiting the method of *Lagrange Multipliers* [64], we can recast this design problem into an unconstrained maximization (up to the validity of all compressor mappings), namely,

$$P^{*}=\underset{P}{\operatorname{argmax}}\;I(\mathsf{x};\mathbf{t}_{1:J})-\sum_{j=1}^{J}\lambda_j I(\mathsf{y}_j;\mathsf{z}_j), \tag{13}$$

wherein, $\lambda_j\geq 0$, denotes the counterpart of the upper-bound, R_j, in (12). By utilizing *Variational Calculus*, the stationary solution of (13) for each local compressor has been derived in [42] (for every pair $(y_j,z_j)\in\mathcal{Y}_j\times\mathcal{Z}_j$) as

$$p(z_j|y_j)=\frac{p(z_j)}{\omega^{\text{Par.}}_{\mathsf{z}_j}(y_j,\beta_j)}\exp\Big(-d_{\text{Par.}}(y_j,z_j)\Big), \tag{14}$$

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

$$\begin{aligned}d_{\text{Par.}}(y_j,z_j)=\beta_j\sum_{\boldsymbol{z}^{\neg j}_{1:J}}p(\boldsymbol{z}^{\neg j}_{1:J}|y_j)\times\\ \sum_{\boldsymbol{t}_{1:J}}p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J})D_{\text{KL}}\big(p(\mathsf{x}|y_j,\boldsymbol{z}^{\neg j}_{1:J})\|p(\mathsf{x}|\boldsymbol{t}_{1:J})\big),\end{aligned} \tag{15}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

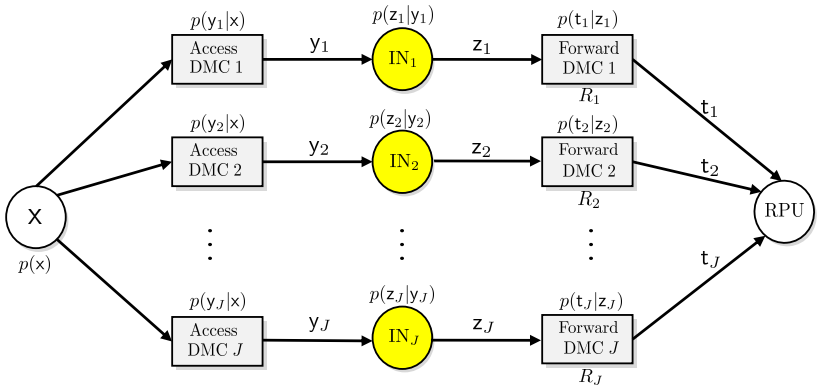


FIGURE 3. The considered system model for distributed joint source-channel coding. All access and forward channels are presumed to be Discrete and Memoryless with known statistics. Given the source, x, the counterpart signals of different branches are presumed to be independent.

with $\beta_j = \frac{1}{\lambda_j}$, and $p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J}) = \prod_{j=1}^{J} p(t_j|z_j)$. Note that the statistics of forward links directly appear in the calculation of relevant distortion in (15). This, indeed, is the very way by which the imperfections of the *error-prone* forward links are taken into account by the compression scheme. Moreover, an iterative algorithm, called Multivariate Forward-Aware Vector Information Bottleneck (MFAVIB-Parallel), has been devised in [42] which performs the *Multivariate Fixed-Point Iterations* [65] on the stationary solution (14), together with its convergence proof to a stationary point of the pertinent objective functional.

B. SUCCESSIVE SCHEME: USING SIDE-INFORMATION

Considering a *successive* processing flow at the remote unit, in which, the side-information from previous branches is used when treating the signal of current branch, the second set of constraints consists of *conditional* compression rates of (local) compressors. This scheme, essentially, follows the Wyner-Ziv approach for source coding [70], in which, some statistically correlated signals are utilized as side-information at the decoder. The design problem is then mathematically formulated as

$$P^* = \underset{P:\,\forall j\ I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1})\le R_j}{\operatorname{argmax}} I(\mathsf{x};\mathbf{t}_{1:J}), \tag{16}$$

with $0 \le R_j \le \log_2|\mathcal{Z}_j|$ bits, setting an upper-bound on the j-th *conditional* compression rate, $I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1})$. It should be noted that here, generally, the processing order affects the resultant performance. Hence, it must be optimized (e.g., via brute-force search). Henceforth, we continue our discussion with a fixed choice of ordering. Similar to the previous case, by exploiting the method of *Lagrange Multipliers* [64], we can recast this problem into an unconstrained maximization (up to the validity of all compressor mappings), namely[2],

$$P^* = \underset{P}{\operatorname{argmax}}\ I(\mathsf{x};\mathbf{t}_{1:J}) - \sum_{j=1}^{J}\lambda_j I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1}), \tag{17}$$

[2]Please note that, for the special case of *full-informativity*, corresponding to letting $\lambda_j \to 0$ for all $j=1$ to J, the objective functions of parallel and successive processing schemes coincide, as the difference in the Lagrangians in (13) and (17) is in their second terms that vanishes, when letting $\lambda_j \to 0$.

with $\lambda_j \ge 0$, being associated with the upper-bound, R_j, in (16). Utilizing *Variational Calculus,* the stationary solution of (17) for each local compressor has been derived in [42] (for every pair $(y_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$) as

$$p(z_j|y_j) = \frac{p(z_j)}{\omega^{\text{Suc.}}_{\mathsf{z}_j}(y_j,\beta_j)} \exp\Big(-d_{\text{Suc.}}(y_j,z_j)\Big), \tag{18}$$

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

$$\begin{aligned} d_{\text{Suc.}}(y_j,z_j) = \beta_j \sum_{\boldsymbol{z}^{\bar{j}}_{1:J}} p(\boldsymbol{z}^{\bar{j}}_{1:J}|y_j)\times \\ \sum_{\boldsymbol{t}_{1:J}} p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J}) D_{\text{KL}}\big(p(\mathsf{x}|y_j,\boldsymbol{z}^{\bar{j}}_{1:J})\|p(\mathsf{x}|\boldsymbol{t}_{1:J})\big) \\ - \sum_{\boldsymbol{t}_{1:j-1}} p(\boldsymbol{t}_{1:j-1}|y_j)\log p(\boldsymbol{t}_{1:j-1}|z_j) \\ -\beta_j \sum_{k=j+1}^{J} \beta_k^{-1} \sum_{\boldsymbol{t}_{1:k-1},\,z_k} p(t_j|z_j)p(\boldsymbol{t}^{\bar{j}}_{1:k-1},z_k|y_j)\log p(z_k|\boldsymbol{t}_{1:k-1}), \end{aligned} \tag{19}$$

with $\beta_j = \frac{1}{\lambda_j}$, and $p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J}) = \prod_{j=1}^{J} p(t_j|z_j)$. Like in the parallel processing scheme, an iterative algorithm, called Multivariate Forward-Aware Vector Information Bottleneck (MFAVIB-Successive), has been presented in [42] which performs the *Multivariate Fixed-Point Iterations* [65] on the stationary solution (18), along with its proof of convergence to a stationary point of the pertinent objective functional.

The major difference between the successive and parallel processing schemes shows itself in the consideration of side-information at the compression rates to further leverage the correlations in the output signals of local compressors. This extra level of complexity in the design formulation reflects itself also in the derived stationary solutions. Comparing the obtained relevant distortion (19) for successive processing with its counterpart (15) for parallel processing reveals that it extends it by two extra terms appearing due to conditioning the compression rates that translates into the consideration of side-information in the respective design problem.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

IV. DEEP VARIATIONAL CASE: OBJECTIVE FUNCTIONALS AND LEARNING ARCHITECTURES

In this part of the article, by following the same roadmap as the one given for the point-to-point scenario, we present the corresponding deep variational approaches to approximately address the design problems for both parallel and successive processing schemes. Specifically, first, we introduce tractable lower-bounds on the pertinent objective functionals. Then, we parameterize their input arguments. Finally, we train the DNNs, which represent the parameterized input arguments.

A. PARALLEL PROCESSING

Analogous to the point-to-point scenario, we commence our discussion towards developing a data-driven design approach by first introducing a *Variational Lower-Bound (VLB)* on the IB-based *parallel* distributed joint source-channel coding Lagrangian, i.e.,

$$\mathcal{L}_{\text{PDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}\big) = I(\mathsf{x};\mathbf{t}_{1:J}) - \sum_{j=1}^{J}\lambda_j I(\mathsf{y}_j;\mathsf{z}_j)\ . \quad (20)$$

To that end, by considering some *proxy* distributions, namely, $q(\mathsf{x}|\mathbf{t}_{1:J})$, as the joint decoder (to retrieve the source, x), and, $\{r(\mathsf{z}_j)\}_{j=1}^{J}$, as the latent priors, the following relations hold true

$$\begin{aligned} I(\mathsf{x};\mathbf{t}_{1:J}) &= \underbrace{H(\mathsf{x})}_{\geq 0} - H(\mathsf{x}|\mathbf{t}_{1:J}) && (21a)\\ &\geq \underbrace{\sum_{\boldsymbol{t}_{1:J}} p(\boldsymbol{t}_{1:J}) D_{\text{KL}}\big(p(\mathsf{x}|\boldsymbol{t}_{1:J})\|q(\mathsf{x}|\boldsymbol{t}_{1:J})\big)}_{\geq 0} \\ &\quad + \sum_{x,\,\boldsymbol{t}_{1:J}} p(x,\boldsymbol{t}_{1:J})\log q(x|\boldsymbol{t}_{1:J}) && (21b)\\ &\geq \mathbb{E}_{\mathsf{x},\mathbf{t}_{1:J}}\{\log q(\mathsf{x}|\mathbf{t}_{1:J})\}, && (21c) \end{aligned}$$

wherein, from (21a) to (21b), the non-negativity of entropy (for the discrete source signal, x), and, from (21b) to (21c), the non-negativity of KL divergence has been applied [62]. Further, it holds true that

$$\begin{aligned} I(\mathsf{y}_j;\mathsf{z}_j) &= \sum_{y_j\in\mathcal{Y}_j,\,z_j\in\mathcal{Z}_j} p(y_j,z_j)\log\frac{p(z_j|y_j)}{r(z_j)} - \underbrace{D_{\text{KL}}\big(p(\mathsf{z}_j)\|r(\mathsf{z}_j)\big)}_{\geq 0} && (22a)\\ &\leq \mathbb{E}_{\mathsf{y}_j,\mathsf{z}_j}\left\{\log\frac{p(\mathsf{z}_j|\mathsf{y}_j)}{r(\mathsf{z}_j)}\right\} = \sum_{y_j\in\mathcal{Y}_j} p(y_j) D_{\text{KL}}\big(p(\mathsf{z}_j|y_j)\|r(\mathsf{z}_j)\big)\ . && (22b) \end{aligned}$$

From (22a) to (22b), the non-negativity of KL divergence has been applied. Now, we can define the VLB for parallel distributed processing scheme as

$$\begin{aligned} &\mathcal{L}^{\text{VLB}}_{\text{PDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}, q(\mathsf{x}|\mathbf{t}_{1:J}), \{r(\mathsf{z}_j)\}_{j=1}^{J}\big) = \\ &\qquad \mathbb{E}_{\mathsf{x},\mathbf{t}_{1:J}}\{\log q(\mathsf{x}|\mathbf{t}_{1:J})\} - \sum_{j=1}^{J}\lambda_j \mathbb{E}_{\mathsf{y}_j,\mathsf{z}_j}\left\{\log\frac{p(\mathsf{z}_j|\mathsf{y}_j)}{r(\mathsf{z}_j)}\right\}, \end{aligned} \quad (23)$$

since from (21) and (22), it is immediately inferred that[3]

$$\begin{aligned} &\mathcal{L}_{\text{PDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}\big) \geq \\ &\qquad \max_{q,\{r\}} \mathcal{L}^{\text{VLB}}_{\text{PDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}, q(\mathsf{x}|\mathbf{t}_{1:J}), \{r(\mathsf{z}_j)\}_{j=1}^{J}\big), \end{aligned} \quad (24)$$

and as a direct consequence,

$$\begin{aligned} &\max_{P=\{p\}} \mathcal{L}_{\text{PDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}\big) \geq \\ &\max_{P=\{p\}} \max_{q,\{r\}} \mathcal{L}^{\text{VLB}}_{\text{PDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}, q(\mathsf{x}|\mathbf{t}_{1:J}), \{r(\mathsf{z}_j)\}_{j=1}^{J}\big)\ . \end{aligned} \quad (25)$$

The next step is to consider a parameterized family for all input distributions of the derived VLB. Specifically, denoting by $\{\boldsymbol{\theta}_j\}_{j=1}^{J}$, $\boldsymbol{\phi}$, and $\{\boldsymbol{\psi}_j\}_{j=1}^{J}$, the parameter sets for families of distributions regarding the encoders, $\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}$, the decoder, $q(\mathsf{x}|\mathbf{t}_{1:J})$, and the latent priors, $\{r(\mathsf{z}_j)\}_{j=1}^{J}$, and by $\mathcal{L}^{\text{DNN}}_{\text{PDist.}}$, the introduced VLB with the parameterized input distributions, i.e.,

$$\begin{aligned} &\mathcal{L}^{\text{DNN}}_{\text{PDist.}}\big(\{\boldsymbol{\theta}_j\}_{j=1}^{J}, \boldsymbol{\phi}, \{\boldsymbol{\psi}_j\}_{j=1}^{J}\big) = \\ &\qquad \mathcal{L}^{\text{VLB}}_{\text{PDist.}}\big(\{p_{\boldsymbol{\theta}_j}(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}, q_{\boldsymbol{\phi}}(\mathsf{x}|\mathbf{t}_{1:J}), \{r_{\boldsymbol{\psi}_j}(\mathsf{z}_j)\}_{j=1}^{J}\big), \end{aligned} \quad (26)$$

the following applies

$$\begin{aligned} &\max_{\{p\}}\max_{q,\{r\}} \mathcal{L}^{\text{VLB}}_{\text{PDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}, q(\mathsf{x}|\mathbf{t}_{1:J}), \{r(\mathsf{z}_j)\}_{j=1}^{J}\big) \geq \\ &\qquad \max_{\{\boldsymbol{\theta}_j\}_{j=1}^{J},\boldsymbol{\phi},\{\boldsymbol{\psi}_j\}_{j=1}^{J}} \mathcal{L}^{\text{DNN}}_{\text{PDist.}}\big(\{\boldsymbol{\theta}_j\}_{j=1}^{J}, \boldsymbol{\phi}, \{\boldsymbol{\psi}_j\}_{j=1}^{J}\big)\ . \end{aligned} \quad (27)$$

The inequality in (27) holds true since the search space over valid distributions, $\{p\}$, q, $\{r\}$, is restricted to the hypothesis space of pertinent parameterized families.

By considering J DNNs for local encoders, $\{p_{\boldsymbol{\theta}_j}\}_{j=1}^{J}$, and one DNN for the joint decoder, $q_{\boldsymbol{\phi}}$, the learning architecture illustrated in Fig. 4 is exploited to address the optimization of $\mathcal{L}^{\text{DNN}}_{\text{PDist.}}$, based upon a finite sample set, $\{(x^{(i)}, \boldsymbol{y}^{(i)}_{1:J})\}_{i=1}^{M}$, and the prior knowledge of forward channels statistics, $p(\mathsf{t}_j|\mathsf{z}_j)$ for $j=1$ to J. Analogous to the point-to-point scenario, for estimating the gradients of $\mathcal{L}^{\text{DNN}}_{\text{PDist.}}$, the conventional approach of exploiting the *reparameterization trick* [48] to enable the *Monte-Carlo Sampling*, and subsequently, substituting the expectation terms with their (empirical) estimates can be applied here as well. Focusing on discrete latent spaces, once again, the *Gumbel-Softmax / Concrete Distribution* [66], [67] can be employed to do the trick., i.e., reparameterizing the underlying categorical distributions. Henceforth, we call this data-driven scheme Deep Multivariate Forward-Aware Vector Information Bottleneck (Deep MFAVIB-Parallel).

For the special case of *error-free* forward transmissions, corresponding to the *remote / noisy* source coding counterpart (over the considered scenario), the learning architecture will remain the same as before except for the case that the (joint) decoder, $q_{\boldsymbol{\phi}}(\mathsf{x}|\boldsymbol{z}_{1:J})$, is directly fed by the samples, $\boldsymbol{z}_{1:J}$, from all latent representations of individual branches. Obviously, the first expectation term in the VLB (23) will then change to $\mathbb{E}_{\mathsf{x},\mathbf{z}_{1:J}}\{\log q(\mathsf{x}|\mathbf{z}_{1:J})\}$.

[3]For clarity: $\{p\}=\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^{J}$, $q=q(\mathsf{x}|\mathbf{t}_{1:J})$, $\{r\}=\{r(\mathsf{z}_j)\}_{j=1}^{J}$.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

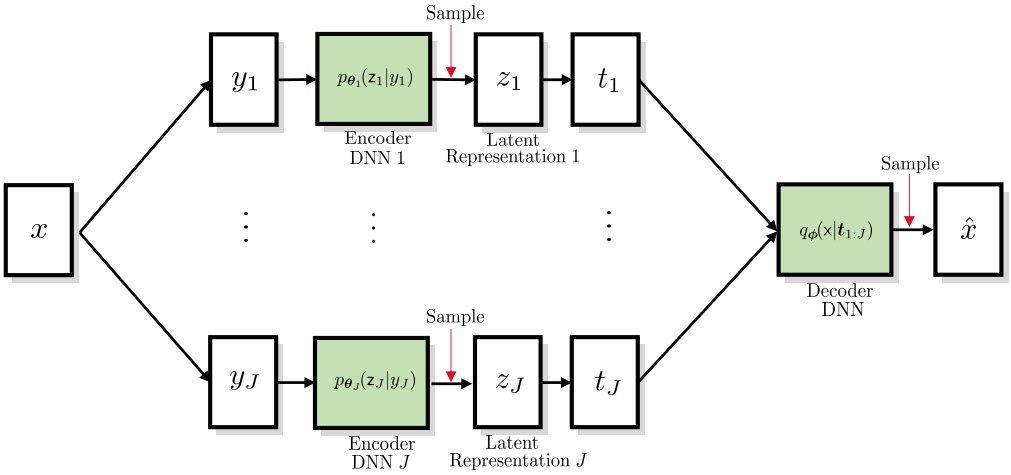


FIGURE 4. **The introduced learning architecture for parallel distributed joint source-channel coding, extending the point-to-point setup by J encoders.**

B. SUCCESSIVE PROCESSING

By following the same roadmap as in the previous scenario, first we introduce a tractable VLB on the IB-based *successive* distributed joint source-channel coding Lagrangian, i.e.,

$$\mathcal{L}_{\text{SDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J\big) = I(\mathsf{x};\mathbf{t}_{1:J}) - \sum_{j=1}^J \lambda_j I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1})\ . \tag{28}$$

To that end, by introducing some *proxy* distributions, namely, $q(\mathsf{x}|\mathbf{t}_{1:J})$, as the joint decoder (to retrieve the source, x), $\{r(\mathsf{z}_j)\}_{j=1}^J$ as the latent priors, and, $\{s\}=\{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})\}_{j=2}^J$, as the side-information components, first recall from (21) that the following holds true

$$I(\mathsf{x};\mathbf{t}_{1:J}) \geq \mathbb{E}_{\mathsf{x},\mathbf{t}_{1:J}}\{\log q(\mathsf{x}|\mathbf{t}_{1:J})\}\ . \tag{29}$$

Moreover, it applies

$$I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1}) = I(\mathsf{y}_j;\mathsf{z}_j) - I(\mathsf{z}_j;\mathbf{t}_{1:j-1})\ . \tag{30}$$

For the second term at the right side of (30), the following holds true

$$-I(\mathsf{z}_j;\mathbf{t}_{1:j-1}) = \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log \frac{p(z_j)}{p(z_j|\boldsymbol{t}_{1:j-1})} \tag{31a}$$

$$= \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log\left(\frac{p(z_j)}{s(z_j|\boldsymbol{t}_{1:j-1})} \times \frac{s(z_j|\boldsymbol{t}_{1:j-1})}{p(z_j|\boldsymbol{t}_{1:j-1})}\right) \tag{31b}$$

$$= \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log \frac{p(z_j)}{s(z_j|\boldsymbol{t}_{1:j-1})} - \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log \frac{p(z_j|\boldsymbol{t}_{1:j-1})}{s(z_j|\boldsymbol{t}_{1:j-1})} \tag{31c}$$

$$= \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log \frac{p(z_j)}{s(z_j|\boldsymbol{t}_{1:j-1})} \underbrace{- \sum_{\boldsymbol{t}_{1:j-1}} p(\boldsymbol{t}_{1:j-1}) D_{\text{KL}}\big(p(\mathsf{z}_j|\mathbf{t}_{1:j-1})\|s(\mathsf{z}_j|\mathbf{t}_{1:j-1})\big)}_{\geq 0} \tag{31d}$$

$$\leq \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log \frac{p(z_j)}{s(z_j|\boldsymbol{t}_{1:j-1})} \tag{31e}$$

$$= \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log\left(\frac{r(z_j)}{s(z_j|\boldsymbol{t}_{1:j-1})} \times \frac{p(z_j)}{r(z_j)}\right) \tag{31f}$$

$$= \sum_{z_j,\,\boldsymbol{t}_{1:j-1}} p(z_j,\boldsymbol{t}_{1:j-1}) \log \frac{r(z_j)}{s(z_j|\boldsymbol{t}_{1:j-1})} + D_{\text{KL}}\big(p(\mathsf{z}_j)\|r(\mathsf{z}_j)\big). \tag{31g}$$

From (31d) to (31e), the non-negativity of KL divergence has been applied. Recalling from (22) that, for the first term at the right side of (30), it applies

$$I(\mathsf{y}_j;\mathsf{z}_j) = \mathbb{E}_{\mathsf{y}_j,\mathsf{z}_j}\left\{\log \frac{p(\mathsf{z}_j|\mathsf{y}_j)}{r(\mathsf{z}_j)}\right\} - D_{\text{KL}}\big(p(\mathsf{z}_j)\|r(\mathsf{z}_j)\big), \tag{32}$$

from (32), (31), and (30), it is immediately deduced that the following holds true

$$I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1}) \leq \mathbb{E}_{\mathsf{y}_j,\mathsf{z}_j}\left\{\log \frac{p(\mathsf{z}_j|\mathsf{y}_j)}{r(\mathsf{z}_j)}\right\} - \mathbb{E}_{\mathsf{z}_j,\mathbf{t}_{1:j-1}}\left\{\log \frac{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})}{r(\mathsf{z}_j)}\right\}. \tag{33}$$

Now, we define the pertinent VLB for successive distributed processing scheme as

$$\mathcal{L}^{\text{VLB}}_{\text{SDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J, q(\mathsf{x}|\mathbf{t}_{1:J}), \{r(\mathsf{z}_j)\}_{j=1}^J, \{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})\}_{j=2}^J\big) = \mathbb{E}_{\mathsf{x},\mathbf{t}_{1:J}}\{\log q(\mathsf{x}|\mathbf{t}_{1:J})\} - \sum_{j=1}^J \lambda_j \left(\mathbb{E}_{\mathsf{y}_j,\mathsf{z}_j}\left\{\log \frac{p(\mathsf{z}_j|\mathsf{y}_j)}{r(\mathsf{z}_j)}\right\} - \mathbb{E}_{\mathsf{z}_j,\mathbf{t}_{1:j-1}}\left\{\log \frac{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})}{r(\mathsf{z}_j)}\right\}\right), \tag{34}$$

since from (32), (31), and (21), it is inferred that

$$\mathcal{L}_{\text{SDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J\big) \geq \max_{q,\{r\},\{s\}} \mathcal{L}^{\text{VLB}}_{\text{SDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J, q(\mathsf{x}|\mathbf{t}_{1:J}), \{r(\mathsf{z}_j)\}_{j=1}^J, \{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})\}_{j=2}^J\big), \tag{35}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

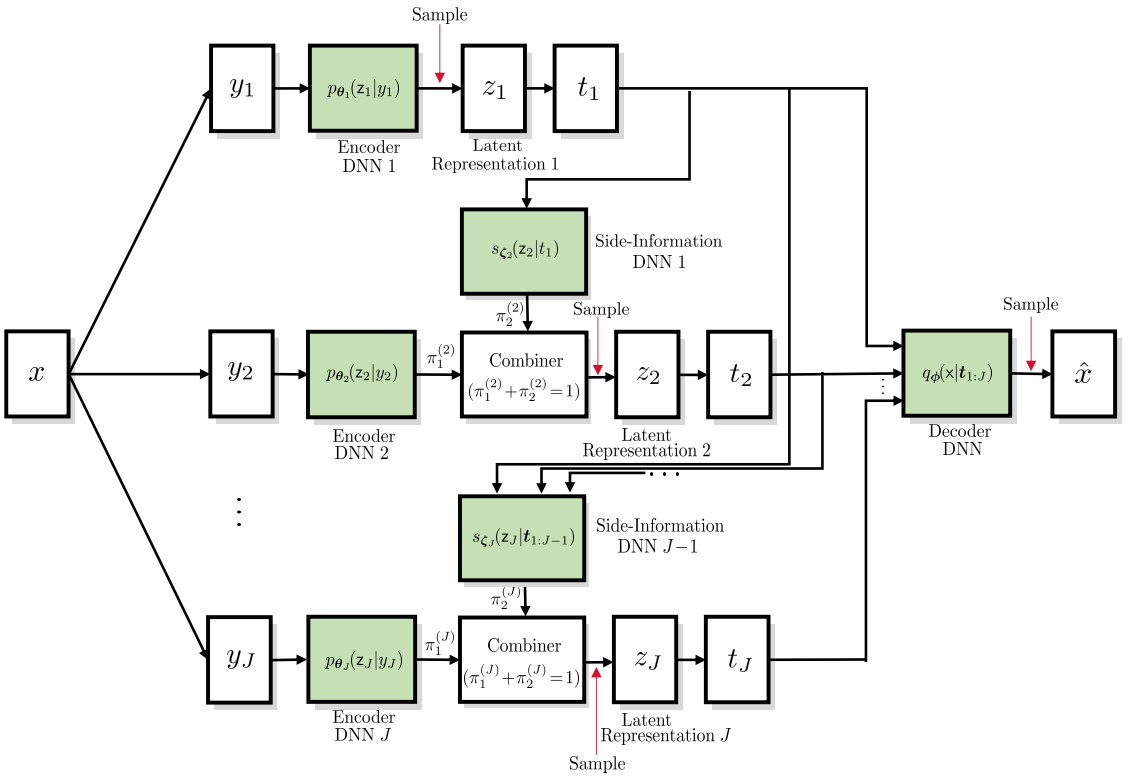


FIGURE 5. The introduced learning architecture for successive distributed joint source-channel coding scheme, featuring J local encoders and $J-1$ side-information DNNs, together with $J-1$ distribution combiner units for the flow of side-information across branches, and a joint decoder.

and as a direct consequence,

$$\max_{\{p\}} \mathcal{L}_{\text{SDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J\big) \geq \max_{\{p\}} \max_{q,\{r\},\{s\}} \mathcal{L}^{\text{VLB}}_{\text{SDist.}}\big(\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J, q(\mathsf{x}|\mathbf{t}_{1:J}), \{r(\mathsf{z}_j)\}_{j=1}^J, \{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})\}_{j=2}^J\big) \,. \tag{36}$$

The next step is then to consider a parameterized family for all input distributions of the introduced VLB. Specifically, representing by $\{\boldsymbol{\theta}_j\}_{j=1}^J$, $\boldsymbol{\phi}$, $\{\boldsymbol{\psi}_j\}_{j=1}^J$, and $\{\boldsymbol{\zeta}_j\}_{j=2}^J$, the sets of parameters for families of distributions regarding the encoders, $\{p(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J$, the decoder, $q(\mathsf{x}|\mathbf{t}_{1:J})$, the latent priors, $\{r(\mathsf{z}_j)\}_{j=1}^J$, and the side-information components, $\{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})\}_{j=2}^J$, and by $\mathcal{L}^{\text{DNN}}_{\text{SDist.}}$, the defined VLB with the parameterized distributions, i.e.,

$$\mathcal{L}^{\text{DNN}}_{\text{SDist.}}\big(\{\boldsymbol{\theta}_j\}_{j=1}^J, \boldsymbol{\phi}, \{\boldsymbol{\psi}_j\}_{j=1}^J, \{\boldsymbol{\zeta}_j\}_{j=2}^J\big) = \mathcal{L}^{\text{VLB}}_{\text{SDist.}}\big(\{p_{\boldsymbol{\theta}_j}(\mathsf{z}_j|\mathsf{y}_j)\}_{j=1}^J, q_{\boldsymbol{\phi}}(\mathsf{x}|\mathbf{t}_{1:J}), \{r_{\boldsymbol{\psi}_j}(\mathsf{z}_j)\}_{j=1}^J, \{s_{\boldsymbol{\zeta}_j}(\mathsf{z}_j|\mathbf{t}_{1:j-1})\}_{j=2}^J\big), \tag{37}$$

it applies

$$\max_{\{p\}} \max_{q,\{r\},\{s\}} \mathcal{L}^{\text{VLB}}_{\text{SDist.}}\big(\{p\}, q, \{r\}, \{s\}\big) \geq \max_{\{\boldsymbol{\theta}_j\}_{j=1}^J, \boldsymbol{\phi}, \{\boldsymbol{\psi}_j\}_{j=1}^J, \{\boldsymbol{\zeta}_j\}_{j=2}^J} \mathcal{L}^{\text{DNN}}_{\text{SDist.}}\big(\{\boldsymbol{\theta}_j\}_{j=1}^J, \boldsymbol{\phi}, \{\boldsymbol{\psi}_j\}_{j=1}^J, \{\boldsymbol{\zeta}_j\}_{j=2}^J\big) \,. \tag{38}$$

The inequality in (38) holds true since the search space over (valid) probability distributions for the local encoders, $\{p\}$, the decoder, q, the latent priors, $\{r\}$, and the side-information components, $\{s\}$, gets restricted to the hypothesis space of pertinent parameterized families.

By considering J DNNs for the encoders, $\{p_{\boldsymbol{\theta}_j}\}_{j=1}^J$, $J-1$ DNNs for the side-information components, $\{s_{\boldsymbol{\zeta}_j}\}_{j=2}^J$, $J-1$ (2 distributions') combiners for the flow of side-information across branches, and one DNN for the (joint) decoder, $q_{\boldsymbol{\phi}}$, the learning architecture illustrated in Fig. 5 is exploited to address the optimization of $\mathcal{L}^{\text{DNN}}_{\text{SDist.}}$, based upon a finite sample set, $\{(x^{(i)}, \boldsymbol{y}_{1:J}^{(i)})\}_{i=1}^M$, and the prior knowledge of the forward channels statistics, $p(\mathsf{t}_j|\mathsf{z}_j)$ for $j=1$ to J. Similar to the parallel scheme, to estimate the gradients of $\mathcal{L}^{\text{DNN}}_{\text{SDist.}}$, the conventional approach of utilizing the *reparameterization trick* [48] to enable the *Monte-Carlo Sampling* and replacing the expectation terms with their (empirical) estimates can be employed here as well. Focusing on discrete latent spaces, once again, the *Gumbel-Softmax / Concrete Distribution* [66], [67] can be employed to do the trick, i.e., reparameterizing the underlying categorical distributions. Henceforth, we call this data-driven scheme Deep Multivariate Forward-Aware Vector Information Bottleneck (Deep MFAVIB-Successive).

For the special case of *error-free* forward transmissions, corresponding to the *remote / noisy* source coding counterpart (over the considered scenario), the learning structure remains the same except for the case that every side-information DNN, $s_{\boldsymbol{\zeta}_j}(\mathsf{z}_j|\boldsymbol{z}_{1:j-1})$, is directly fed by the samples, $\boldsymbol{z}_{1:j-1}$, from the latent variables of previous branches. Moreover, the (joint) decoder, $q_{\boldsymbol{\phi}}(\mathsf{x}|\boldsymbol{z}_{1:J})$, is directly fed by the outputs of combiner units. Obviously, the first and the last expectation terms in the VLB (34) will change to $\mathbb{E}_{\mathsf{x},\mathbf{z}_{1:J}}\{\log q(\mathsf{x}|\mathbf{z}_{1:J})\}$ and $\mathbb{E}_{\mathbf{z}_{1:j}}\big\{\log \frac{s(\mathsf{z}_j|\mathbf{z}_{1:j-1})}{r(\mathsf{z}_j)}\big\}$, respectively.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

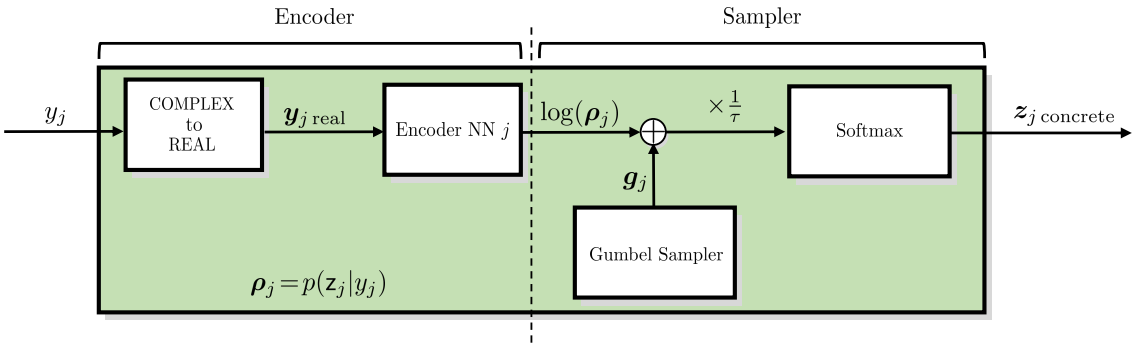


FIGURE 6. **The detailed schematic of the encoder and concrete distribution sampler in j-th branch of the learning architecture for parallel processing. The same individual units are applied in the learning architecture for successive processing as well.**

It should be reminded that the side-information DNNs as well as the joint decoder DNN are located at the (remote) processing unit. Consequently, the interconnections between different branches only happen at the receiver side, and still the (local) compressor units do not have to exchange any information in this successive compression scheme, similar to the previous case of parallel processing.

REMARKS: First, it should be noted that the coefficients of the combiner units are the extra (hyper)parameters to be twiddled in the successive architecture. Second, by recalling from (30) that it applies

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

it is immediately inferred that the derived upper-bound for the unconditional compression rate, $I(\mathsf{y}_j;\mathsf{z}_j)$, in (22) also applies to the conditional counterpart, $I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1})$. If so, why looking for another upper-bound in case of successive processing? To answer this, note that for the second term at the right side of (33), it applies

$$\begin{aligned}&\mathbb{E}_{\mathsf{z}_j,\mathbf{t}_{1:j-1}}\left\{\log\frac{s(\mathsf{z}_j|\mathbf{t}_{1:j-1})}{r(\mathsf{z}_j)}\right\} = \\ &\sum_{\boldsymbol{t}_{1:j-1}} p(\boldsymbol{t}_{1:j-1}) \sum_{z_j} p(z_j|\boldsymbol{t}_{1:j-1}) \log\frac{s(z_j|\boldsymbol{t}_{1:j-1})}{r(z_j)} = \\ &\sum_{\boldsymbol{t}_{1:j-1}} p(\boldsymbol{t}_{1:j-1}) \Big(D_{\text{KL}}\big(p(\mathsf{z}_j|\boldsymbol{t}_{1:j-1})\|r(\mathsf{z}_j)\big) \\ &\qquad - D_{\text{KL}}\big(p(\mathsf{z}_j|\boldsymbol{t}_{1:j-1})\|s(\mathsf{z}_j|\boldsymbol{t}_{1:j-1})\big)\Big)\,.\end{aligned} \tag{40}$$

Consequently, if for every $\boldsymbol{t}_{1:j-1}$, the second KL divergence in (40) becomes less than its first KL divergence, a *tighter* upper-bound compared to the parallel processing is obtained (since the expectation term above becomes positive). This should, naturally, be the case as $s(\mathsf{z}_j|\boldsymbol{t}_{1:j-1})$ is a proxy for $p(\mathsf{z}_j|\boldsymbol{t}_{1:j-1})$ (for every $\boldsymbol{t}_{1:j-1}$) and shall better approximate it compared to $r(\mathsf{z}_j)$ that does not take any side-information into account. This, indeed, is the exact point which yields the benefits of bringing the side-information into play, when formulating the respective design problem.

TABLE 1. **The configuration of encoder/side-information/decoder DNNs.**

Denotation	# Hidden Layers	Widths	# Weights
$p_{\boldsymbol{\theta}_j}(\mathsf{z}_j\|\mathsf{y}_j)$	3	300, 200, 100	81200+101×N
$s_{\boldsymbol{\zeta}_j}(\mathsf{z}_j\|\mathbf{t}_{1:j-1})$	3	300, 200, 100	80600+(300j−199)×N
$q_{\boldsymbol{\phi}}(\mathsf{x}\|\mathbf{t}_{1:3})$	3	300, 200, 100	82216+900×N

V. NUMERICAL RESULTS

A. SPECIFICATIONS & IMPLEMENTATION DETAILS

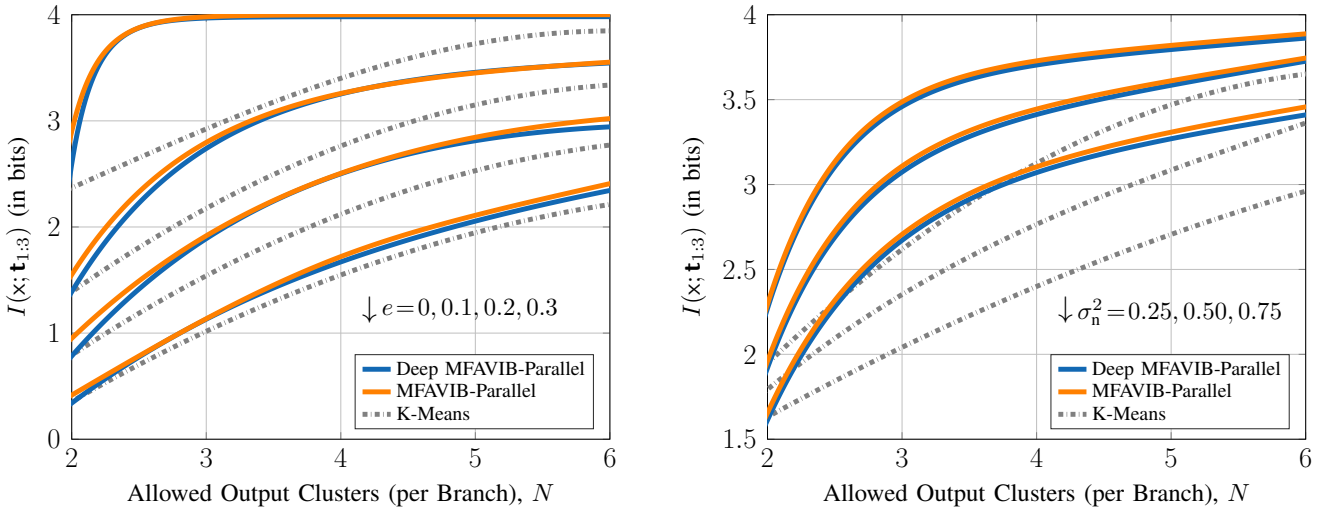
In what follows, a standard 16-QAM (Quadrature Amplitude Modulation) source signaling ($\sigma_{\mathsf{x}}^2 = 10$) is considered over $J=3$ branches. Each access link is modeled by a DMC that approximates a discrete-time, discrete-input, and continuous-output AWGN (Additive White Gaussian Noise) channel with the noise variance, $\sigma_{\mathsf{n}_j}^2$ for $j=1,2,3$. Every forward link is modeled by a symmetric $N\times N$ DMC (N denoting the number of output clusters per branch) that is specified by the error probability, e_j for $j=1,2,3$. Explicitly, every input symbol is received correctly with probability $1-e_j$, and erroneously (to any other symbol) with probability $\frac{e_j}{N-1}$. This indicates that higher values of e_j correspond to less reliable forward transmissions and vice versa.

To conduct the training, 10^6 samples are generated via a Monte Carlo approach. These samples are then utilized with the batch size of 10^4 and a maximum of 10^4 epochs. Further, to regularize, the *Early Stopping* is applied. The learning rate is set to 10^{-5} and *Adam* [71] is used as the chosen learning algorithm to update the weights. The detailed configurations of the encoder DNNs, the side-information DNNs, and the (joint) decoder DNN have been presented in Table 1. All 3 types of DNNs (i.e., encoder / side-information / decoder) are *Multi-Layer Perceptrons (MLPs)* with 3 hidden layers, each featuring a *Rectified Linear Unit (ReLU)* activation function. The training is performed once per parameter set, and the weights are saved and used for inference without retraining.

The detailed schematic of the encoder and the sampler in j-th branch of the learning architecture for parallel scheme has been depicted in Fig. 6. Generally, the *noisy* observation, $y_j\in\mathcal{Y}_j$ for $j=1,2,3$, is complex-valued. Thus, to be fed into the j-th local encoder DNN, the signal, $y_j\in\mathcal{Y}_j$, is stacked into a two-dimensional vector, $\boldsymbol{y}_{j\,\text{real}}\in\mathbb{R}^2$, that contains the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446



(a) Fixed access noise variance, $\sigma_{\mathrm{n}}^2=0.2$

(b) Fixed forward error probability, $e=0.03$

FIGURE 7. The relevant information, $I(\mathrm{x};\mathbf{t}_{1:3})$, vs. the number of allowed output clusters, N, for a) fixed access noise variance, and b) fixed forward error probability. 16-QAM source signaling ($\sigma_{\mathrm{x}}^2=10$) over AWGN access links, with $\lambda_j=0.01$ for $j=1,2,3$, and temperature $\tau=2$.

inphase and quadrature parts, separately. We apply this since DNNs cannot handle the complex numbers straightforwardly in the training phase (as complex derivatives are not always straightforward to calculate). The j-th local encoder that is fed by $\boldsymbol{y}_{j\,\mathrm{real}}$ outputs a categorical distribution, $\boldsymbol{\rho}_j\in(0,1)^N$, where N denotes the number of categories / clusters for the discrete latent variable, z_j. To generate samples from the pertinent *concrete variable*, first we draw N Independent and Identically Distributed (IID) samples from the Gumbel $(0,1)$ distribution and stack them into the vector, $\boldsymbol{g}_j$ for $j=1,2,3$. The sum signal $\log(\boldsymbol{\rho}_j)+\boldsymbol{g}_j$ is then multiplied by the inverse of a (positive) hyperparameter, τ, which is known as the *temperature* in the literature. This scaled signal is then fed into a *Softmax* unit. Consequently, the k-th entry of the j-th sample vector, $\boldsymbol{z}_{j\,\mathrm{concrete}}$, is calculated as ($k=1$ to N)

$$z_{j\,\mathrm{concrete}}^{[k]}=\frac{\exp\Big(\big(\log(\rho_j^{[k]})+g_j^{[k]}\big)/\tau\Big)}{\sum_{\ell=1}^{N}\exp\Big(\big(\log(\rho_j^{[\ell]})+g_j^{[\ell]}\big)/\tau\Big)}\in[0,1],\quad(41)$$

wherein $\rho_j^{[k]}$ and $g_j^{[k]}$ denote the k-th entries of the vectors, $\boldsymbol{\rho}_j$ and $\boldsymbol{g}_j$, respectively. The lower the temperature value, τ, becomes, the closer behavior to an *Argmax* is achieved. For $\tau=1$, the unmodified *Softmax* is achieved. By letting $\tau\to\infty$, a uniform distribution is obtained over N categories. Hence, the temperature, τ, should be chosen carefully, as either too small or too large values of it may lead to a poor performance (due to either experiencing very rapid changes in gradients or an overly smoothed behavior). It should be mentioned that, for the successive processing scheme, the same individual units are used in the respective learning architecture as well. Finally, it should also be noted that, the latent priors, $r_{\boldsymbol{\psi}_j}(\mathrm{z}_j)$ for $j=1,2,3$, although having their own sets of parameters (i.e., the probabilities of different categories / clusters), are not implemented by DNNs.

B. MODEL-BASED VS. DATA-DRIVEN DESIGN

In the first part, we would like to compare the performance of the State-of-the-Art (SotA) model-based MFAVIB with the devised data-driven Deep MFAVIB algorithm. For that, by focusing on the case of parallel processing, we consider a symmetric setup in which all access noise variances are set to the same value, σ_{n}^2. Moreover, all forward error probabilities are set to the same value, e, as well. We further set $\lambda_j=0.01$ for $j=1,2,3$, indicating that the main focus will be on the preservation of the relevant information, i.e., $I(\mathrm{x};\mathbf{t}_{1:3})$. For the described setup, we further present an extra performance comparison with the case in which, per IN, the well-known K-Means algorithm [72] is used. To avoid poor local optima, for both MFAVIB and K-Means algorithms, the best outcome has been chosen out of 100 runs with different initialization.

In Fig. 7a, by fixing the access noise variance to $\sigma_{\mathrm{n}}^2=0.2$, we vary the number of allowed output bins / clusters (per branch), N, and illustrate the obtained relevant information, $I(\mathrm{x};\mathbf{t}_{1:3})$, for different forward error probabilities, namely, $e=0,0.1,0.2,0.3$. In Fig. 7b, by fixing the forward error probability to $e=0.03$, again we vary the number of allowed output bins / clusters (per branch), N, and depict the obtained relevant information, $I(\mathrm{x};\mathbf{t}_{1:3})$, for 3 different access noise variances, namely, $\sigma_{\mathrm{n}}^2=0.25,0.50,0.75$.

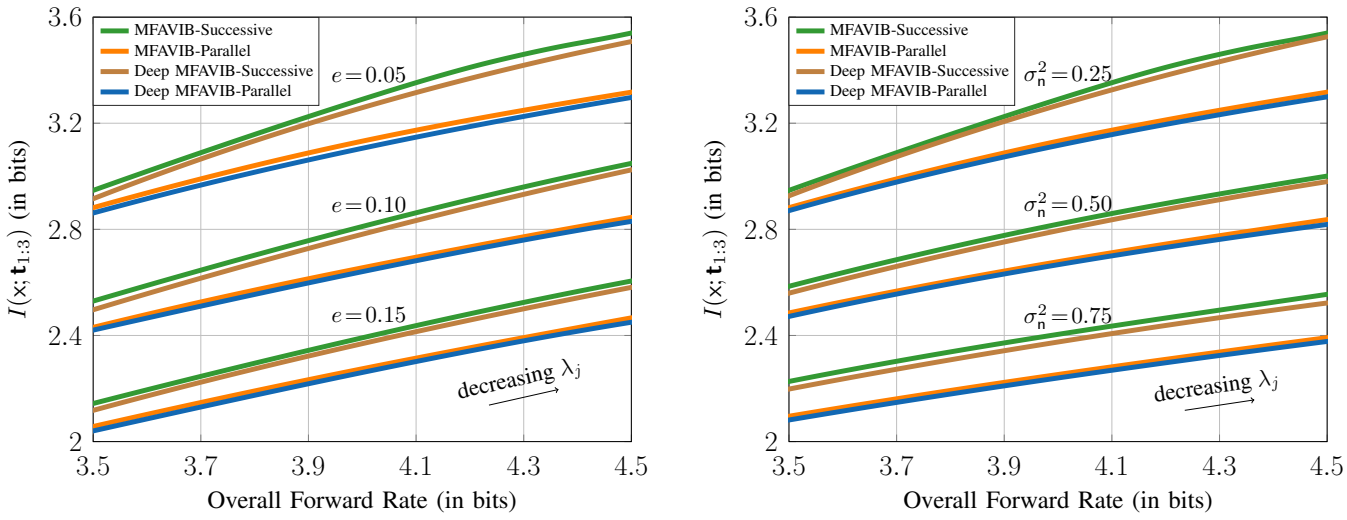
Focusing on the illustrated results in Fig. 7a, it is readily observed that, for fixed access statistics, the relevant information increases by decreasing the forward error probability. This can be justified by noting the fact that the capacity of forward channels calculated as [73]

$$R(N,e)=\log_2 N+(1-e)\log_2(1-e)+e\log_2\frac{e}{N-1},\quad(42)$$

increases by decreasing the error probability, e. Hence, by decreasing e more information can be flown into the forward links, and consequently, the relevant information increases as well. It is further observed that, expectedly, by increasing the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446



(a) Fixed access noise variance, $\sigma_{\mathsf{n}}^2=0.25$

(b) Fixed forward error probability, $e=0.05$

FIGURE 8. **The relevant information, $I(\mathsf{x};\mathbf{t}_{1:3})$, vs. the overall forward rate, $\sum_j I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1})$ successive / $\sum_j I(\mathsf{y}_j;\mathsf{z}_j)$ parallel for a) fixed access noise variance, and b) fixed forward error probability. 16-QAM source signaling ($\sigma_{\mathsf{x}}^2=10$) over AWGN access links, allowed output clusters, $N=4$, symmetric 4×4 forward channels, with $0.28\leq\lambda_j\leq0.66$ for $j=1,2,3$, $\tau=0.5$ (Par.), $\tau=0.225$ (Suc.), with $(\pi_1^{(j)},\pi_2^{(j)})=(0.9,0.1)$, for $j=2,3$.**

number of allowed output clusters, the relevant information increases since the compression bottleneck is loosened. It is further observed that, regardless of having either error-free or error-prone forward links, the K-Means algorithm yields an inferior performance in comparison with the proposed joint source-channel coding schemes in terms of preserving the relevant information, $I(\mathsf{x};\mathbf{t}_{1:3})$. The superior performance by (Deep) MFAVIB, for the case of error-prone forwarding, evidently corroborates the benefits of "*forward-awareness*" by incorporating the impact of imperfect forwarding into the design of local compressors.

Focusing on the depicted results in Fig. 7b, it is observed that, for a given forward statistics, the relevant information increases by decreasing the access noise variance. This can be justified through an analogous line of argumentation as the one already provided for the previous results, by noting the fact that the capacity of access links increases by decreasing the noise variance, σ_{n}^2, thereby allowing more information (about the user / source signal, x) to be flown into the system. Here as well, like in the previous investigation in Fig. 7a, the outperformance of the (Deep) MFAVIB scheme compared to the K-Means algorithm is observed, corroborating the gains obtained by the joint source-channel coding schemes which, for the compression of noisy observations at different INs, they take the imperfections of forward links into account.

Considering both results together, as the main takeaway, it is observed that regardless of the specific choice of model parameters, the devised Deep MFAVIB algorithm performs (almost) on par with the SotA model-based MFAVIB routine, without requiring the prior knowledge of the joint statistics of the source, x, and the noisy observations, $\mathbf{y}_{1:3}$, and solely based on a (finite) sample set. This becomes quite important, especially, in applications where the joint statistics are either unavailable or hard to estimate.

C. PARALLEL VS. SUCCESSIVE PROCESSING

In the second part of numerical investigations, we would like to compare the achievable performances of the parallel and successive processing schemes. For that, like in the previous case, we consider a symmetric scenario in which all access noise variances are set to the same value, σ_{n}^2. Furthermore, all forward error probabilities are set to the same value, e, as well. We fix the number of allowed output clusters (per branch) to $N=4$, and vary the trade-off parameters, λ_j for $j=1,2,3$, within a certain range of finite non-zero values to sweep various points in the *information-compression* plane.

In Fig. 8a, by fixing the access noise variance to $\sigma_{\mathsf{n}}^2=0.25$, we depict the obtained relevant information, $I(\mathsf{x};\mathbf{t}_{1:3})$, versus the overall forward rate, i.e., $\sum_j I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1})$ successive processing / $\sum_j I(\mathsf{y}_j;\mathsf{z}_j)$ parallel processing, for 3 different forward error probabilities, namely, $e=0.05,0.10,0.15$.

Conversely, in Fig. 8b, by presuming a fixed value for the forward error probability, i.e., $e=0.05$, we illustrate the obtained relevant information, $I(\mathsf{x};\mathbf{t}_{1:3})$, versus the overall forward rate, i.e., $\sum_j I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1})$ successive processing / $\sum_j I(\mathsf{y}_j;\mathsf{z}_j)$ parallel processing, for 3 different access noise variances, namely, $\sigma_{\mathsf{n}}^2=0.25,0.50,0.75$.

Focusing on the presented results in Fig. 8a, it is directly observed that, regardless of the chosen processing scheme, by decreasing the forward error probability, larger values of relevant information are achieved for a given overall forward rate. To justify this, an exact line of argumentation as the one already provided for Fig. 7a is applicable here as well. Also, focusing on the results depicted in Fig. 8b, it is readily seen that, regardless of the chosen processing scheme, by decreasing the access noise variance, larger values of relevant information are achieved for a given overall forward rate. Again, this can be justified by the same line of reasoning as the one already provided for Fig. 7b in the previous part.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

Focusing on both figures together, it is seen that various points in the *information-compression* plane are achieved by varying the value of trade-off parameters, λ_j for $j=1,2,3$. Explicitly, large values of λ_j correspond to the solutions with more focus on compactness, while the small values of λ_j correspond to solutions with more focus on informativity.

As the main takeaways, first, it should be noted that the successive processing scheme outperforms the parallel one by better leveraging the present correlations in the signals of different intermediate nodes (since all of them are noisy versions of the common user signal, x). This clearly indicates the gained benefits by using the side-information at the RPU. Second, it should also be noted that, regardless of the chosen specifications and the type of processing scheme, the introduced (data-driven) approaches here, perform (almost) on par with the SotA model-based algorithms without requiring the full prior knowledge of the joint statistics, $p(\mathsf{x}, \mathbf{y}_{1:3})$, of input signals, and purely based on a finite sample set, instead. Once again, this shows that the developed compression schemes in this article can be applied (as a promising alternative) in those applications where the joint input statistics are either unavailable, hard to estimate, or rapidly changing, as in cases of dealing with dynamic environments.

VI. SUMMARY

In this article, we developed new deep variational approaches to address the challenging design problems for a (generic) distributed / multiterminal joint source-channel coding setup which appears in a broad range of real-world applications. In the focused setup, a number of *noisy* observations from a *common* user / source signal should be compressed at several intermediate nodes before getting forwarded over multiple *error-prone* and *rate-limited* links to a (remote) processing unit. By following the *Information Bottleneck* principle, the *Mutual Information* was then chosen as the fidelity criterion, and subsequently, tractable objective functions were derived for two different types of retrieval strategies, together with the pertinent learning architectures. The underlying design problems were then addressed by the standard training of the Deep Neural Networks in the introduced architectures, e.g., by Stochastic Gradient Descent with back-propagation. The proposed (data-driven) latent variable-based approaches here, whose effectiveness was substantiated through a couple of numerical investigations over typical (real-world) digital transmission setups, generalize several well-known concepts, including *Variational Auto-Encoders* [48], *Deep Variational Information Bottleneck* [45], and its distributed extension for (indirect / remote) source coding [39], [46], to the context of multiterminal joint source-channel coding.

REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," in *37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 1999.

[2] C. E. Shannon, "Coding Theorems for a Discrete Source with a Fidelity Criterion," *IRE International Convention Record,* part 4, vol. 7, pp. 142–163, March 1959.

[3] T. A. Courtade and T. Weissman, "Multiterminal Source Coding Under Logarithmic Loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, January 2014.

[4] P. Harremoës and N. Tishby, "The Information Bottleneck Revisited or How to Choose a Good Distortion Measure," in *IEEE International Symposium on Information Theory*, Nice, France, June 2007.

[5] R. Dobrushin and B. Tsybakov, "Information Transmission with Additional Noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, September 1962.

[6] D. J. Sakrison, "Source Encoding in the Presence of Random Disturbance," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 165–167, January 1968.

[7] J. Wolf and J. Ziv, "Transmission of Noisy Information to a Noisy Receiver with Minimum Distortion," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 406–411, July 1970.

[8] H. Witsenhausen, "Indirect Rate Distortion Problems," *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 518–521, September 1980.

[9] E. Ayanoglu, "On Optimal Quantization of Noisy Sources," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1450–1452, November 1990.

[10] A. Zaidi, I. Estella-Aguerri, and S. Shamai (Shitz), "On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views," *Entropy*, vol. 22, no. 2, Art. no. 151, January 2020.

[11] Z. Goldfeld and Y. Polyanskiy, "The Information Bottleneck Problem and Its Applications in Machine Learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[12] S. Hu, Z. Lou, X. Yan, and Y. Ye, "A Survey on Information Bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5325–5344, February 2024.

[13] A. Wyner, "On Source Coding with Side Information at the Decoder," *IEEE Transactions on Information Theory*, vol. 21, no. 3, pp. 294–300, May 1975.

[14] R. Ahlswede and J. Körner, "Source Coding with Side Information and a Converse for Degraded Broadcast Channels," *IEEE Transactions on Information Theory*, vol. 21, no. 6, pp. 629–637, November 1975.

[15] E. Erkip and T. M. Cover, "The Efficiency of Investment Information," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[16] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the Information Bottleneck to the Privacy Funnel," in *IEEE Information Theory Workshop*, Hobart, TAS, Australia, November 2014.

[17] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information Extraction Under Privacy Constraints," *Information*, vol. 7, no. 1, Art. no. 15, March 2016.

[18] B. Razeghi, P. Rahimi, and S. Marcel, "Deep Variational Privacy Funnel: General Modeling with Applications in Face Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, April 2024.

[19] V. Doshi, D. Shah, M. Médard, and M. Effros, "Functional Compression Through Graph Coloring," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3901–3917, August 2010.

[20] S. Feizi and M. Médard, "On Network Functional Compression," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5387–5401, September 2014.

[21] Y. M. Saidutta, A. Abdi, and F. Fekri, "Analog Joint Source-Channel Coding for Distributed Functional Compression using Deep Neural Networks," in *IEEE International Symposium on Information Theory*, Melbourne, Australia, July 2021.

[22] D. Malak and M. Médard, "A Distributed Computationally Aware Quantizer Design via Hyper Binning," *IEEE Transactions on Signal Processing*, vol. 71, pp. 76–91, January 2023.

[23] A. Alamoudi, Y. Saidutta, and F. Fekri, "Distributed Functional Compression for Independent Component Analysis in Wireless Networks," in *IEEE Global Communications Conference*, Cape Town, South Africa, December 2024.

[24] I. Tal and A. Vardy, "How to Construct Polar Codes," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6562–6582, October 2013.

[25] M. Stark, A. Shah, and G. Bauch, "Polar Code Construction Using the Information Bottleneck Method," in *IEEE Wireless Communications and Networking Conference Workshops*, Barcelona, Spain, April 2018.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

[26] F. J. C. Romero and B. M. Kurkoski, "LDPC Decoding Mappings That Maximize Mutual Information," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 9, pp. 2391–2401, September 2016.

[27] M. Stark, J. Lewandowsky, and G. Bauch, "Information-Optimum LDPC Decoders with Message Alignment for Irregular Codes," in *IEEE Global Communications Conference*, Abu Dhabi, United Arab Emirates, Dec. 2018.

[28] 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 Journal of the Communications Society*, vol. 1, pp. 646–660, May 2020.

[29] 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, Art. no. 19, October 2022.

[30] G. Zeitler, A. C. Singer, and G. Kramer, "Low-Precision A/D Conversion for Maximum Information Rate in Channels with Memory," *IEEE Transactions on Communications*, vol. 60, no. 9, pp. 2511–2521, September 2012.

[31] T. Monsees, D. Wübben, and A. Dekorsy, "Optimum Quantization of Memoryless Channels with N-ary Input," in *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, October 2022.

[32] J. Shao, Y. Mao, and J. Zhang, "Learning Task-Oriented Communication for Edge Inference: An Information Bottleneck Approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, January 2022.

[33] F. Pezone, S. Barbarossa, and P. Di Lorenzo, "Goal-Oriented Communication for Edge Learning Based on the Information Bottleneck," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, May 2022.

[34] D. Gündüz, Z. Qin, I. Estella-Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, January 2023.

[35] E. Beck, C. Bockelmann, and A. Dekorsy, "Semantic Information Recovery in Wireless Networks," *Sensors*, vol. 23, no. 14, Art. no. 6347, July 2023.

[36] I. Estella-Aguerri and A. Zaidi, "Distributed Information Bottleneck Method for Discrete and Gaussian Sources," in *International Zurich Seminar on Information and Communication*, Zurich, Switzerland, February 2018.

[37] I. Estella-Aguerri, A. Zaidi, G. Caire, and S. Shamai (Shitz), "On the Capacity of Cloud Radio Access Networks with Oblivious Relaying," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4575–4596, July 2019.

[38] Y. Uğur, I. Estella-Aguerri, and A. Zaidi, "Vector Gaussian CEO Problem Under Logarithmic Loss and Applications," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4183–4202, July 2020.

[39] I. Estella-Aguerri and A. Zaidi, "Distributed Variational Representation Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 120–138, January 2021.

[40] S. Hassanpour, D. Wübben, and A. Dekorsy, "A Novel Approach to Distributed Quantization via Multivariate Information Bottleneck Method," in *IEEE Global Communications Conference*, Waikoloa, HI, USA, December 2019.

[41] ——, "Generalized Distributed Information Bottleneck for Fronthaul Rate Reduction at the Cloud-RANs Uplink," in *IEEE Global Communications Conference*, Taipei, Taiwan, December 2020.

[42] S. Hassanpour, D. Wübben, and A. Dekorsy, "Forward-Aware Information Bottleneck-Based Vector Quantization: Multiterminal Extensions for Parallel and Successive Retrieval," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 6633–6646, October 2021.

[43] S. Hassanpour, A. Danaee, D. Wübben, and A. Dekorsy, "Multi-Source Distributed Data Compression Based on Information Bottleneck Principle," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4171–4185, July 2024.

[44] J. Lewandowsky and G. Bauch, "Theory and Application of the Information Bottleneck Method," *Entropy*, vol. 26, no. 3, Art. no. 187, February 2024.

[45] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep Variational Information Bottleneck," in *International Conference on Learning Representations*, Toulon, France, Apr. 2017.

[46] A. Zaidi and I. Estella-Aguerri, "Distributed Deep Variational Information Bottleneck," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications*, Atlanta, GA, USA, May 2020.

[47] Q. Wang, C. Boudreau, Q. Luo, P.-N. Tan, and J. Zhou, "Deep Multi-View Information Bottleneck," in *SIAM International Conference on Data Mining*, Calgary, Alberta, Canada, May 2019.

[48] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv 2013*, arXiv:1312.6114.

[49] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Factor Analysis, Probabilistic Principal Component Analysis, Variational Inference, and Variational Autoencoder: Tutorial and Survey," *arXiv 2021*, arXiv:2101.00734.

[50] G. Zeitler, G. Bauch, and J. Widmer, "Quantize-and-Forward Schemes for the Orthogonal Multiple-Access Relay Channel," *IEEE Transactions on Communications*, vol. 60, no. 4, pp. 1148–1158, April 2012.

[51] I. Avram, N. Aerts, H. Bruneel, and M. Moeneclaey, "Quantize and Forward Cooperative Communication: Channel Parameter Estimation," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1167–1179, March 2012.

[52] D. Wübben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible Centralization through Cloud-RAN," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, November 2014.

[53] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Fronthaul Compression for Cloud Radio Access Networks: Signal Processing Advances Inspired by Network Information Theory," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 69–79, November 2014.

[54] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, March 2017.

[55] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, July 2020.

[56] M. Bashar *et al.*, "Limited-Fronthaul Cell-Free Massive MIMO With Local MMSE Receiver Under Rician Fading and Phase Shifts," *IEEE Wireless Communications Letters*, vol. 10, no. 9, pp. 1934–1938, September 2021.

[57] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "User-Centric Cell-Free Massive MIMO Networks: A Survey of Opportunities, Challenges and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 611–652, First Quarter 2021.

[58] A. Danaee, S. Hassanpour, D. Wübben, and A. Dekorsy, "Relevance-Based Information Processing for Fronthaul Rate Reduction in Cell-Free MIMO Systems," in *International Symposium on Wireless Communication Systems*, Rio de Janeiro, Brazil, July 2024.

[59] ——, "Relevance-Based Multi-User Data Compression for Fronthaul Rate Reduction in Cell-Free Massive MIMO Systems," in *International ITG Conference on Systems, Communications and Coding*, Karlsruhe, Germany, March 2025.

[60] B. Chen, L. Tong, and P. K. Varshney, "Channel-Aware Distributed Detection in Wireless Sensor Networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 16–26, 2006.

[61] S. Movaghati and M. Ardakani, "Distributed Channel-Aware Quantization Based on Maximum Mutual Information," *Int. J. Distrib. Sensor Netw.*, vol. 12, no. 5, Art. no. 3595389, May 2016.

[62] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, 2006.

[63] S. Hassanpour, T. Monsees, D. Wübben, and A. Dekorsy, "Forward-Aware Information Bottleneck-Based Vector Quantization for Noisy Channels," *IEEE Transactions on Communications*, vol. 68, no. 12, pp. 7911–7926, December 2020.

[64] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

[65] J. H. Mathews and K. D. Fink, *Numerical Methods Using MATLAB*, 4th ed., Pearson Prentice Hall, 2004.

[66] E. Jang, S. Gu, and B. Poole, "Categorical Reparameterization with Gumbel-Softmax," *arXiv 2016*, arXiv:1611.01144.

[67] C. J. Maddison, A. Mnih, and Y. W. Teh, "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables," *arXiv 2016*, arXiv:1611.00712.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3570446

[68] M. Hummert, S. Hassanpour, D. Wübben, and A. Dekorsy, "Deep FAVIB: Deep Learning-Based Forward-Aware Quantization via Information Bottleneck Method," in *IEEE International Conference on Communications*, Denver, CO, USA, June 2024.
[69] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[70] A. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, January 1976.
[71] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv 2014*, arXiv:1412.6980.
[72] A. K. Jain, "Data Clustering: 50 Years beyond K-Means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, June 2010.
[73] D. Hankerson, G. A. Harris, and P. D. Johnson Jr., *Introduction to Information Theory and Data Compression*, 2nd ed., Boca Raton, FL, USA: CRC Press, 2003.

SHAYAN HASSANPOUR (Member, IEEE) works as a Post-Doctoral Researcher in the Department of Communications Engineering at the University of Bremen, Germany. He 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. Over the last couple of years, he has been prolifically contributing to the top-tier international journals and IEEE flagship conferences on his PhD 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.

MATTHIAS HUMMERT (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from University of Bremen, Germany, in 2015 and 2017, respectively, where currently he is pursuing the Ph.D. degree at the Department of Communications Engineering (ANT). His research interest centers around the applications of machine learning in the design of communication systems with a special focus on coding, MIMO and NOMA for wireless communications.

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.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/