Relevance-Based Information Processing for Fronthaul Rate Reduction in Cell-Free MIMO Systems

Alireza Danaee Shayan Hassanpour Dirk Wübben Armin Dekorsy
Department of Communications Engineering, University of Bremen, Germany
{danaee; hassanpour; wuebben; dekorsy}@ant.uni-bremen.de

2024 19th International Symposium on Wireless Communication Systems (ISWCS) | 979-8-3503-6251-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISWCS61526.2024.10639087

***Abstract*—Consider a user equipment in a Cell-Free massive Multiple-Input Multiple-Output (CF-mMIMO) system that is served by several Radio Access Points (RAPs). In the uplink of this setup, these RAPs receive *noisy* observations of the user/source signal and must locally compress their signals before forwarding them to the Central Processing Unit (CPU) through multiple rate-limited fronthaul channels. To retrieve the source signal at CPU, we are interested in maximizing the Mutual Information (MI) between the received signals at CPU and the user/source signal, and purposefully choose the Information Bottleneck (IB)-based compression techniques to design the quantizers at RAPs. We consider both separate and joint designs of the local compressors by establishing basic trade-offs between the *informativity* and *compactness* of the outcomes. For the joint design, two different schemes are presented, based on whether to leverage the side-information at CPU. Finally, the effectiveness of both compression schemes will be shown as well by means of numerical investigations over typical digital data transmission scenarios.**

***Index Terms*—6G, Cell-Free massive MIMO, distributed joint source-channel coding, information bottleneck method, mutual information**

I. Introduction

Massive Multiple-Input Multiple-Output (MIMO) technology has emerged as a promising solution to meet the ever-increasing demand for high data rates and spectral efficiency in wireless communication systems [1]. By employing a large number of antennas at the base stations, massive MIMO leverages the spatial multiplexing and beamforming techniques to serve multiple User Equipments (UEs) simultaneously in the same time-frequency resource. This leads to significant gains in both spectral efficiency and energy efficiency, making it a key technology for next-generation of wireless networks [2]. Building upon the foundation of massive MIMO, the concept of Cell-Free massive MIMO (CF-mMIMO) has gained attraction in recent years as a novel architecture for wireless networks. Unlike traditional cellular technology, cell-free networks eliminate the concept of individual cells and cell boundaries, allowing users to be served by all Radio Access Points (RAPs) simultaneously where RAPs are connected to a Central Processing Unit (CPU) through a fronthaul network [3], [4]. The simplistic approach of CF-mMIMO, where every RAP is tasked with processing and transmitting data signals for all UEs, lacks the scalability since there is a linear (or even faster) growth in computational complexity and fronthaul rates associated with these tasks as the number of UEs increases [5]. The User-Centric CF-mMIMO (UC-CF-mMIMO) approach as the more scalable version of CF-mMIMO offers several advantages, including enhanced coverage, improved user fairness, and increased network capacity [5], [6]. Moreover, by leveraging distributed processing and cooperation among RAPs, CF-mMIMO enables efficient resource allocation and interference management, further enhancing the overall network performance [7].

Despite the promising benefits of CF-mMIMO, its practical deployment poses challenges, particularly in terms of fronthaul capacity and signal processing overhead. The fronthaul network, responsible for transporting signals from distributed RAPs to CPU, is a critical bottleneck in these systems. The massive number of antennas and the high-dimensional signal processing tasks in the uplink impose stringent requirements on the fronthaul capacity. To alleviate this bottleneck, signal compression techniques play a crucial role in reducing the amount of data transmitted over the fronthaul while preserving the essential information for accurate signal recovery at CPU. Thus, efficient compression algorithms tailored to the characteristics of massive MIMO signals are essential for realizing the full potential of CF-mMIMO systems in future wireless networks. To address the performance bottleneck highlighted previously, several fronthaul compression schemes have been developed, see, e.g., [8]–[10]. These schemes usually focus on the uniform quantization. In this work, we consider the Information Bottleneck (IB) method [11], [12] to design the compression schemes for fronthaul rate reduction at the uplink of CF-mMIMO systems.

The central concept of IB lies in compressing a Random Variable (RV) such that its information content with respect to a statistically correlated (relevant) variable is preserved to a significant extent. This ability to retain information is highly adaptable and can be adjusted by manipulating a parameter that governs a fundamental trade-off between the *compactness* and *informativity* of the resultant outcome. The IB method formulates this basic trade-off in a symmetric manner, utilizing the Mutual Information (MI) [13] to quantify both aspects. The applications of the IB method in various parts of modern communication systems range from A/D converters for receiver front ends [14], to discrete channel decoding schemes [15], [16], task/goal-oriented communications [17], [18], and more.

In the following, starting with the point-to-point setup of the original IB method for (remote) source coding, we gradually build up a comprehensive discussion on the generalizations to the uniterminal joint source-channel coding and multiterminal techniques for both source and joint source-channel coding that are applicable to the architecture of CF-mMIMO systems.

979-8-3503-6251-0/24/$31.00 ©2024 IEEE

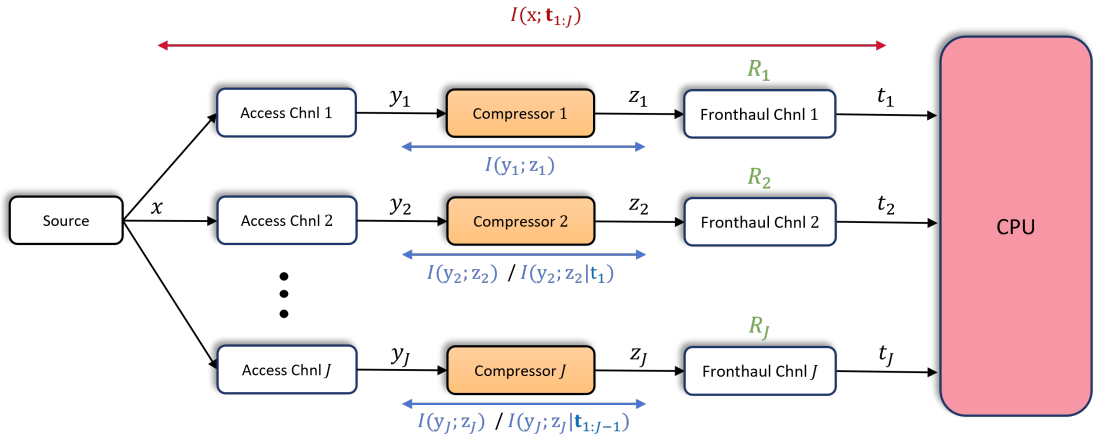

Fig. 1. Considered system model for a given UE in the CF-mMIMO network. Multiple RAPs compress noisy observations $y_1, \ldots, y_J$ from a common source signal x to $z_1, \ldots, z_J$, and send to CPU via fronthaul links. The CPU should retrieve the source signal from the received signals $t_1, \ldots, t_J$.

Notation: According to the distribution, $p(\mathsf{a})$, the realizations, $a \in \mathcal{A}$, of the (discrete) random variable, a, happen. With boldface counterparts, the same holds true for the (discrete) random vector, $\mathbf{a}_{1:J} = \{\mathsf{a}_1, \cdots, \mathsf{a}_J\}$ and $\mathbf{a}_{1:J}^{-j} = \mathbf{a}_{1:J} \setminus \{\mathsf{a}_j\}$. Moreover, $I(\cdot\,;\cdot)$ and $D_{\mathrm{KL}}(\cdot\|\cdot)$ stand for the Mutual Information and Kullback-Leibler (KL) divergence [13], respectively.

II. Information Bottleneck Compression for Cell-Free massive MIMO Systems

Let us consider the uplink transmission in a UC-CF-mMIMO system with M UEs and K RAPs connected to a CPU by K (error-free/error-prone) rate-limited fronthaul links. We presume a linear equalization [19] at each RAP to cancel the spatial interference of different UEs (which get served by it) and separate their signals. We consider a given UE served by J RAPs where $J \leq K$. The interrelation between the UE signal, x, and the corresponding output of the linear equalizer, i.e., y_j, at each RAP j where $j = 1, 2, \ldots, J$, is termed the access channel j in Figure 1 which is modeled by a Discrete Memoryless Channel (DMC). Each RAP j compresses y_j to z_j to transmit to CPU through a fronthaul channel with a limited rate of R_j. The CPU should retrieve the source signal from the received signals $\mathsf{t}_1, \ldots, \mathsf{t}_J$. To design the IB-based compression, the source distribution, $p(\mathsf{x})$ as well as both the access and fronthaul transition probabilities, $p(\mathsf{y}_j|\mathsf{x})$ and $p(\mathsf{t}_j|\mathsf{z}_j)$, respectively, are supposed to be known. For the case of the error-free fronthaul channels, all fronthaul transition probability matrices $p(\mathsf{t}_j|\mathsf{z}_j)$ are evidently identity matrices. It is further presumed that the counterpart signals of different branches are conditionally independent, given the source signal, x. Then the design problem can be formulated as a basic trade-off between two MI terms. The first term is the MI of the source signal and the corresponding received signal(s) at CPU which is called the relevant information. The second term is the MI of the noisy observation(s) of the source signal and the outcome of the compressor(s) which is called the compression rate. The goal of IB-based compression is to maximize the relevant information such that the compression rate does not exceed the capacity of the corresponding fronthaul channel(s). To do so, we consider two scenarios; designing a compressor at each RAP separately from other RAPs and designing all local compressors at RAPs jointly.

A. Separate Design

1) Error-Free Fronthauling: Consider a single RAP that receives a noisy observation y_j from a single source x with distribution, $p(\mathsf{x})$, through a DMC described by transition probabilities, $p(\mathsf{y}_j|\mathsf{x})$. It must then compress y_j to the signal z_j and forward it to the CPU through an *error-free* and *rate-limited* channel with the capacity, R_j. The IB framework [11] formulates the design problem as a trade-off between the relevant information, $I(\mathsf{x};\mathsf{z}_j)$ and the compression rate, $I(\mathsf{y}_j;\mathsf{z}_j)$. The objective is to design a compressor $p(\mathsf{z}_j|\mathsf{y}_j)$ that maximizes the relevant information while the compression rate does not exceed the capacity of the fronthaul channel. The design problem is written as follows:

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

where $0 \leq R_j \leq \log_2|\mathcal{Z}_j|$ bits, sets an upper-bound on the compression rate, $I(\mathsf{y}_j;\mathsf{z}_j)$. Using the method of Lagrange Multipliers (LM) [20], one can formulate (1) as the following unconstrained optimization (up to the validity of the quantizer mapping)

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

where $\lambda_j \geq 0$ is associated with R_j in the original formulation. By exploiting the Variational Calculus, the form of stationary solution for the (non-convex) design problem (2) is obtained (for each pair $(y_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$) in [11] as follows:

$$p(z_j|y_j) = \frac{p(z_j)}{\psi_{\mathsf{z}_j}(y_j,\beta_j)} \exp\Big(-\beta_j D_{\mathrm{KL}}\big(p(\mathsf{x}|y_j)\|p(\mathsf{x}|z_j)\big)\Big), \tag{3}$$

where $\beta_j = \lambda_j^{-1}$, and $\psi_{\mathsf{z}_j}(y_j,\beta_j)$, is a partition function to ensure the validity of the compressor mapping. Additionally,

an iterative algorithm, the Iterative IB (IIB), has been presented in [11] to address the design problem, which performs the *fixed-point iterations* [21] on the derived *implicit* solution (3).

2) Error-Prone Fronthauling: In practice, it happens quite often that, the fronthaul channels (to the CPU) are *error-prone*. Therefore, the impacts of the noisy fronthaul channels should also be considered within the design formulation of the compressors. Here, the fronthaul transition probabilities, $p(\mathsf{t}_j|\mathsf{z}_j)$ are supposed to be known in addition to $p(\mathsf{x})$ and $p(\mathsf{y}_j|\mathsf{x})$. By following the IB philosophy, the design problem has been formulated in [22] as a trade-off between two MI terms, i.e., the relevant information, $I(\mathsf{x};\mathsf{t}_j)$, and the compression rate, $I(\mathsf{y}_j;\mathsf{z}_j)$. The objective here is to maximize the informativity while the compression rate does not exceed the fronthaul channel capacity, R_j, namely,

$$p^*(\mathsf{z}_j|\mathsf{y}_j) = \underset{p(\mathsf{z}_j|\mathsf{y}_j):I(\mathsf{y}_j;\mathsf{z}_j)\le R_j}{\operatorname{argmax}} I(\mathsf{x};\mathsf{t}_j). \tag{4}$$

This problem can be reformulated as an unconstrained optimization using LM as follows:

$$p^*(\mathsf{z}_j|\mathsf{y}_j) = \underset{p(\mathsf{z}_j|\mathsf{y}_j)}{\operatorname{argmax}} I(\mathsf{x};\mathsf{t}_j) - \lambda_j I(\mathsf{y}_j;\mathsf{z}_j). \tag{5}$$

The stationary solution has been derived in [22] (for each pair $(y_j,z_j)\in\mathcal{Y}_j\times\mathcal{Z}_j$) as follows:

$$p(z_j|y_j) = \frac{p(z_j)}{\psi_{\mathsf{z}_j}(y_j,\beta_j)}\exp\Big(-\beta_j\sum_{t_j\in\mathcal{T}_j} p(t_j|z_j)\times D_{\mathrm{KL}}\big(p(\mathsf{x}|y_j)\|p(\mathsf{x}|t_j)\big)\Big), \tag{6}$$

where $\beta_j=\lambda_j^{-1}$, and $\psi_{\mathsf{z}_j}(y_j,\beta_j)$, is a partition function to ensure the validity of the compressor mapping. An iterative algorithm, the Forward-Aware Vector IB (FAVIB), has also been presented in [22] for addressing this design problem together with its convergence proof to a stationary point of the pertinent objective functional. By this, [22] fully extends the original IB framework for point-to-point Noisy Source Coding (NSC) [23] to the Joint Source-Channel Coding (JSCC) [24].

B. Joint Design

Compared to the separate design of local compressors, the joint design targets a more efficient usage of the correlations among signals of different RAPs, thereby leading to a better overall performance as we will see in the numerical results. Consider the system model in Figure 1 where J RAPs want to jointly compress their noisy observations, $\mathsf{y}_1,\ldots,\mathsf{y}_J$, of the source signal x to $\mathsf{z}_1,\ldots,\mathsf{z}_J$ and forward them to the CPU through the *error-free* and *rate-limited* fronthaul links. Similar to the separate IB method, to formulate the design problem for the joint compression, one should specify the responsible terms regarding both the *informativity* and *compactness* of the outcomes. The end-to-end transmission rate, $I(\mathsf{x};\mathbf{z}_{1:J})$, (i.e., the amount of information that the compressed signals collectively preserve about the data source) is naturally chosen as the term measuring the informativity. In contrast, there is no natural and unique choice for compactness and different meaningful expressions can be applied. Herein, two separate sets of constraints are considered to determine the compactness, corresponding to the parallel [25] and successive [26] (retrieval) processing strategies at CPU.

1) Error-Free Fronthauling: In the *parallel* processing, no side-information is used at CPU to retrieve the source signal, x. Explicitly, let $P^*=\{p^*(\mathsf{z}_1|\mathsf{y}_1),\ldots,p^*(\mathsf{z}_J|\mathsf{y}_J)\}$ denote the optimal set of compressors at RAPs. The design problem is formulated in [25] as follows:

$$P^* = \underset{P:\,\forall j\ I(\mathsf{y}_j;\mathsf{z}_j)\le R_j}{\operatorname{argmax}} I(\mathsf{x};\mathbf{z}_{1:J}), \tag{7}$$

where $0\le R_j\le\log_2|\mathcal{Z}_j|$ bits, sets an upper-bound on the j-th compression rate, $I(\mathsf{y}_j;\mathsf{z}_j)$. Using the method of LM, the design problem (7) can be stated as the following unconstrained optimization (up to the validity of the corresponding mappings):

$$P^* = \underset{P}{\operatorname{argmax}}\, I(\mathsf{x};\mathbf{z}_{1:J}) - \sum_{j=1}^{J}\lambda_j I(\mathsf{y}_j;\mathsf{z}_j), \tag{8}$$

where $\lambda_j\ge 0$ is associated with the rate, R_j in (7). The form of stationary solution for the (non-convex) design problem (8) is obtained in [25] for each pair $(y_j,z_j)\in\mathcal{Y}_j\times\mathcal{Z}_j$ as follows:

$$p(z_j|y_j) = \frac{p(z_j)}{\psi_{\mathsf{z}_j}(y_j,\beta_j)}\exp\big(-\beta_j d^{\text{NSC}}_{\text{Par}}(y_j,z_j)\big), \tag{9}$$

where $\beta_j=\lambda_j^{-1}$, and $\psi_{\mathsf{z}_j}(y_j,\beta_j)$, is a normalization function that ensures the validity of pertinent quantizer mapping, and the relevant distortion, $d^{\text{NSC}}_{\text{Par}}(y_j,z_j)$, is given by

$$d^{\text{NSC}}_{\text{Par}}(y_j,z_j) = \sum_{\boldsymbol{z}^{-j}_{1:J}} p(\boldsymbol{z}^{-j}_{1:J}|y_j)D_{\mathrm{KL}}\big(p(\mathsf{x}|y_j,\boldsymbol{z}^{-j}_{1:J})\|p(\mathsf{x}|\boldsymbol{z}_{1:J})\big). \tag{10}$$

An iterative algorithm, the MultiIB, has also been presented in [25] for addressing this design problem. Principally, it applies the *multivariate fixed-point iterations* on the derived *implicit* solutions (9).

In the *successive* processing scheme, as the second choice regarding the set of imposed constraints, unlike the parallel scheme, from a compression perspective, side-information is used when handling the signal, y_j. The main idea behind this scheme is fully aligned with the well-known Wyner-Ziv [27], [28] setup for source coding, where a statistically correlated signal is used as side-information at the decoder. The design problem is formulated in [26] for the optimal set of compressors, $P^*=\{p^*(\mathsf{z}_1|\mathsf{y}_1),\ldots,p^*(\mathsf{z}_J|\mathsf{y}_J)\}$ as follows:

$$P^* = \underset{P:\,\forall j\ I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{z}_{1:j-1})\le R_j}{\operatorname{argmax}} I(\mathsf{x};\mathbf{z}_{1:J}), \tag{11}$$

where $0\le R_j\le\log_2|\mathcal{Z}_j|$ bits, sets an upper-bound on the j-th *conditional* compression rate, $I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{z}_{1:j-1})$. By using the LM method, the design problem (11) can be stated as the following unconstrained optimization (up to the validity of the corresponding quantizer mappings):

$$P^* = \underset{P}{\operatorname{argmax}}\, I(\mathsf{x};\mathbf{z}_{1:J}) - \sum_{j=1}^{J}\lambda_j I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{z}_{1:j-1}), \tag{12}$$

where $\lambda_j \geq 0$ is associated with the rate, R_j in (11). The form of stationary solution for the (non-convex) design problem (12) is obtained in [26] for each pair $(y_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$ as follows:

$$p(z_j|y_j) = \frac{p(z_j)}{\psi_{\mathsf{z}_j}(y_j, \beta_j)} \exp\big(-\beta_j d^{\text{NSC}}_{\text{Suc}}(y_j, z_j)\big), \tag{13}$$

where $\beta_j = \lambda_j^{-1}$, and $\psi_{\mathsf{z}_j}(y_j, \beta_j)$, is a normalization function that ensures the validity of pertinent quantizer mapping, and the relevant distortion, $d^{\text{NSC}}_{\text{Suc}}(y_j, z_j)$, is given by

$$\begin{aligned} d^{\text{NSC}}_{\text{Suc}}(y_j, z_j) = &\sum_{\boldsymbol{z}^{-j}_{1:J}} p(\boldsymbol{z}^{-j}_{1:J}|y_j) D_{\text{KL}}\big(p(\mathsf{x}|y_j, \boldsymbol{z}^{-j}_{1:J}) \| p(\mathsf{x}|\boldsymbol{z}_{1:J})\big) \\ &- \lambda_j \sum_{\boldsymbol{z}_{1:j-1}} p(\boldsymbol{z}_{1:j-1}|y_j) \log p(\boldsymbol{z}_{1:j-1}|z_j) \\ &- \sum_{k=j+1}^{J} \lambda_k \sum_{\boldsymbol{z}^{-j}_{1:J}} p(\boldsymbol{z}^{-j}_{1:k}|y_j) \log p(z_k|\boldsymbol{z}_{1:k-1}). \end{aligned} \tag{14}$$

It can be observed that, by considering the side-information in the design problem (11), two extra terms appear in the relevant distortion (14) compared to the derived distortion for the parallel scheme in (10). An iterative algorithm, the GDIB, has also been presented in [26] for addressing this design problem that, analogous to the parallel processing, applies the *multivariate fixed-point iterations* on the derived *implicit* solutions (13).

2) Error-Prone Fronthauling: The *error-prone* fronthaul channels shall be considered for practical implementation of CF-mMIMO systems. Therefore, the impacts of the noisy fronthaul channels should be considered within the design formulation for the IB-based compression. The above parallel and successive processing schemes have been extended for the case of *error-prone* fronthaul links in [29].

For the *parallel* processing, the design problem is formulated as finding the optimal set, $P^* = \{p^*(\mathsf{z}_1|\mathsf{y}_1), \ldots, p^*(\mathsf{z}_J|\mathsf{y}_J)\}$ given by

$$P^* = \underset{P:\, \forall j\ I(\mathsf{y}_j;\mathsf{z}_j) \leq R_j}{\operatorname{argmax}} I(\mathsf{x}; \mathbf{t}_{1:J}), \tag{15}$$

where $0 \leq R_j \leq \log_2 |\mathcal{Z}_j|$ bits, sets an upper-bound on the j-th compression rate, $I(\mathsf{y}_j; \mathsf{z}_j)$. Using the method of LM, the design problem (15) can be restated as the following unconstrained optimization (up to the validity of the corresponding mappings):

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

where $\lambda_j \geq 0$ is associated with the rate, R_j in (15). The form of stationary solution for the (non-convex) design problem (16) is obtained in [29] for each pair $(y_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$ as follows:

$$p(z_j|y_j) = \frac{p(z_j)}{\psi_{\mathsf{z}_j}(y_j, \beta_j)} \exp\big(-\beta_j d^{\text{JSCC}}_{\text{Par}}(y_j, z_j)\big), \tag{17}$$

where $\beta_j = \lambda_j^{-1}$, and $\psi_{\mathsf{z}_j}(y_j, \beta_j)$, is a normalization function that ensures the validity of pertinent quantizer mapping, and the relevant distortion, $d^{\text{JSCC}}_{\text{Par}}(y_j, z_j)$, is given by

$$\begin{aligned} d^{\text{JSCC}}_{\text{Par}}(y_j, z_j) = &\sum_{\boldsymbol{z}^{-j}_{1:J}} p(\boldsymbol{z}^{-j}_{1:J}|y_j) \sum_{\boldsymbol{t}_{1:J}} p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J}) \times \\ &D_{\text{KL}}\big(p(\mathsf{x}|y_j, \boldsymbol{z}^{-j}_{1:J}) \| p(\mathsf{x}|\boldsymbol{t}_{1:J})\big), \end{aligned} \tag{18}$$

wherein, from the presumed conditional independence relations, it applies $p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J}) = \prod_{j=1}^{J} p(t_j|z_j)$. Note that, by bringing the *error-prone* fronthaul channels into the design problem (15), their statistics, $p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J})$, directly appear in the derived relevant distortion (18). An iterative algorithm, the M-FAVIB (Parallel), has also been proposed in [29] for addressing this design problem alongside its convergence proof to a stationary point of the pertinent objective functional.

For the *successive* processing scheme to gain benefit from the side-information to retrieve the source signal, x, at CPU, the design problem is presented as looking for the optimal set of compressors, $P^* = \{p^*(\mathsf{z}_1|\mathsf{y}_1), \ldots, p^*(\mathsf{z}_J|\mathsf{y}_J)\}$ as follows:

$$P^* = \underset{P:\, \forall j\ I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1}) \leq R_j}{\operatorname{argmax}} I(\mathsf{x}; \mathbf{t}_{1:J}), \tag{19}$$

where $0 \leq R_j \leq \log_2 |\mathcal{Z}_j|$ bits, sets an upper-bound on the j-th *conditional* compression rate, $I(\mathsf{y}_j; \mathsf{z}_j|\mathbf{t}_{1:j-1})$. By using the LM method, the design problem (19) can be stated as the following unconstrained optimization (up to the validity of the corresponding quantizer mappings):

$$P^* = \underset{P}{\operatorname{argmax}}\ I(\mathsf{x}; \mathbf{t}_{1:J}) - \sum_{j=1}^{J} \lambda_j I(\mathsf{y}_j; \mathsf{z}_j|\mathbf{t}_{1:j-1}), \tag{20}$$

where $\lambda_j \geq 0$ is associated with the rate, R_j in (19). The form of stationary solution for the (non-convex) design problem (20) is obtained in [29] for each pair $(y_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$ as follows:

$$p(z_j|y_j) = \frac{p(z_j)}{\psi_{\mathsf{z}_j}(y_j, \beta_j)} \exp\big(-\beta_j d^{\text{JSCC}}_{\text{Suc}}(y_j, z_j)\big), \tag{21}$$

where $\beta_j = \lambda_j^{-1}$, and $\psi_{\mathsf{z}_j}(y_j, \beta_j)$, is a normalization function that ensures the validity of pertinent quantizer mapping, and the relevant distortion, $d^{\text{JSCC}}_{\text{Suc}}(y_j, z_j)$, is given by

$$\begin{aligned} d^{\text{JSCC}}_{\text{Suc}}(y_j, z_j) = &\sum_{\boldsymbol{z}^{-j}_{1:J}} p(\boldsymbol{z}^{-j}_{1:J}|y_j) \sum_{\boldsymbol{t}_{1:J}} p(\boldsymbol{t}_{1:J}|\boldsymbol{z}_{1:J}) \times \\ &D_{\text{KL}}\big(p(\mathsf{x}|y_j, \boldsymbol{z}^{-j}_{1:J}) \| p(\mathsf{x}|\boldsymbol{t}_{1:J})\big) \\ &- \lambda_j \sum_{\boldsymbol{t}_{1:j-1}} p(\boldsymbol{t}_{1:j-1}|y_j) \log p(\boldsymbol{t}_{1:j-1}|z_j) \\ &- \sum_{k=j+1}^{J} \lambda_k \sum_{\boldsymbol{t}^{-j}_{1:J}, z_k} p(t_j|z_j) \times \\ &p(\boldsymbol{t}^{-j}_{1:k-1}, z_k|y_j) \log p(z_k|\boldsymbol{t}_{1:k-1}). \end{aligned} \tag{22}$$

An iterative algorithm, the M-FAVIB (Successive), has also been presented in [29] for addressing this design problem alongside its convergence proof to a stationary point of the pertinent objective functional.

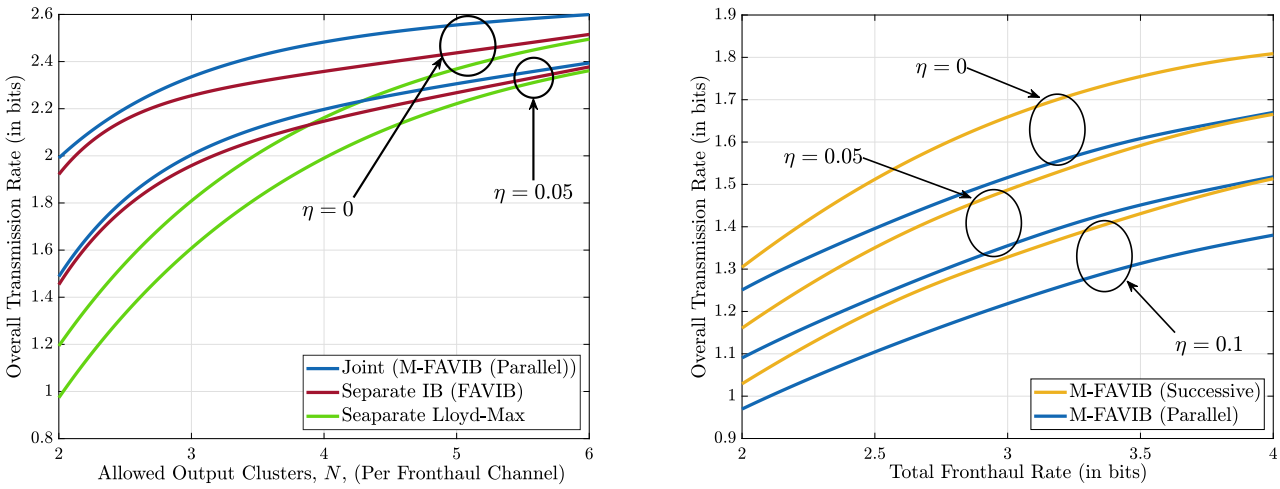

Fig. 2. **(a)** A comparison between IIB (3), FAVIB (6) and Lloyd-Max [30] as the methods for separate design of compressors and joint design using MultiIB (9) and M-FAVIB (Parallel) (17). **(b)** M-FAVIB (Parallel) (17) versus M-FAVIB (Successive) (21) as the methods for joint design of compressors.

III. Numerical Results

Here, we present some numerical results regarding typical transmission scenarios in the uplink of a CF-mMIMO system in which we apply different types of compression in the RAPs. To compare these schemes, we use the overall transmission rate, i.e., the MI between the source signal and the received signals at CPU, $I(\mathsf{x};\mathbf{t}_{1:J})$, as the performance indicator, and since these approaches are initialized randomly, for the sake of fairness, the same starting points are applied for the schemes and the best outcomes are retained out of 100 trials. We assume a DMC that approximates a discrete-time, discrete-input, and continuous-output AWGN (Additive White Gaussian Noise) channel with identical noise variance, σ_{n}^2, for all access connections from the UE to the RAPs where we consider $J=3$ RAPs in our simulation setup.

For the first experiment, we consider an equiprobable source signaling from a bipolar 8-ASK (Amplitude Shift Keying) constellation with $\sigma_{\mathsf{x}}^2=24$ to $J=3$ RAPs while $\sigma_{\mathsf{n}}^2=1$ in the AWGN access channels and 100 samples per access channel have been generated, following a Monte Carlo approach. We compare the performance of the separate and joint methods to design the compressors at RAPs. For the separate design, we used FAVIB (6) to compress each noisy observation y_j to signal z_j at each RAP as well as the Lloyd-Max quantization [30] to design the compressors at each RAP. For the joint compression, the M-FAVIB (Parallel) (17) has been applied to compress the noisy observations at RAPs. Note that, for FAVIB and M-FAVIB (Parallel), we choose $\beta_j \to \infty$ as we want to maximize the overall transmission rate, i.e., $I(\mathsf{x};\mathbf{t}_{1:3})$. With $N=|\mathcal{Z}_j|$ denoting the allowed number of output clusters of compressors, we consider an $N\times N$ symmetric channel model in each fronthaul link, characterized by the reliability parameter, η, as follows: each input symbol is correctly received with probability $1-\eta$ and incorrectly (to any other output symbol) with probability $\frac{\eta}{N-1}$. Therefore, higher η values indicate less reliable transmission and vice versa. We consider a symmetric setup, i.e., the same σ_{n}^2 for access links and the same η for fronthaul links. Figure 2(a) illustrates the obtained results.

Specifically, for two cases of *error-free* fronthaul ($\eta=0$) and *error-prone* fronthaul ($\eta=0.05$), the allowed number of output clusters (per link) has been varied from $N=2$ to 6, and the overall transmission rate has been calculated. It is observed that the joint design of local compressors performs better than the separate design. To clearly justify this, it should be noted that, from the presumed independence relations, it applies

$$I(\mathsf{x};\mathbf{t}_{1:3}) = \sum_{j=1}^{3} I(\mathsf{x};\mathsf{t}_j) - \Big(I(\mathsf{t}_1;\mathsf{t}_2) + I(\mathbf{t}_{1:2};\mathsf{t}_3)\Big). \quad (23)$$

While the separate design of local quantizers focuses solely on maximizing the first component at the right side of (23), the joint design strikes a good balance between maximizing the first component and simultaneously keeping the second component as low as possible. Moreover, it is seen that the conventional Lloyd-Max algorithm (applied separately on different RAPs) yields an inferior performance since it does not directly consider the source/user signal, x, by minimizing the Mean-Square-Error (MSE) between the input and output of the local compressors.

For the second experiment, we compare the parallel and successive schemes for the joint design of compressors. To that end, we apply the same system setup described for the joint design compression in the previous experiment, except for the choice of source signaling, where we consider a standard QPSK (Quadrature Phase Shift Keying) constellation with $\sigma_{\mathsf{x}}^2=1$ and $\sigma_{\mathsf{n}}^2=0.3$ for the access channels. The allowed output clusters of compressors are set to $N=4$ and $\beta_j \in (0.1, 0.2)$ for $j=1,2,3$. Figure 2(b) illustrates the obtained results for different reliability values of the fronthaul links. It can be observed that the usage of the available side-information can decrease the required total fronthaul rate to have a desired overall transmission rate. In parallel processing, the side-

information that comes from the correlations between different fronthaul channels is failed to care for. To justify this, note that from the presumed independence relations, it applies

$$I(\mathsf{y}_j;\mathsf{z}_j|\mathbf{t}_{1:j-1}) = I(\mathsf{y}_j;\mathsf{z}_j) - \underbrace{I(\mathsf{z}_j;\mathbf{t}_{1:j-1})}_{\geq 0}. \tag{24}$$

Therefore, it is directly inferred that conditioning on previous fronthaul channel output signals can either deduct from the current unconditional compression rate, $I(\mathsf{y}_j;\mathsf{z}_j)$, or keep it unchanged as the MI is non-negative.

IV. Conclusions

In this work, we focused on the separate and joint design of Information Bottleneck (IB)-based compression schemes for fronthaul rate reduction at the uplink of Cell-Free massive MIMO systems where several Radio Access Points receive noisy observations of a common user/source and compress their signals prior to a forward transmission through several *rate-limited* fronthaul channels to the Central Processing Unit. We considered both scenarios of dealing with *error-free* and *error-prone* fronthaul links, thereby addressing the respective distributed (remote) source and joint source-channel coding problems. The stationary solutions were provided for both separate and joint design, with different types of processing for the latter. Through some numerical simulations, we clearly demonstrated the effectiveness of the IB-based compression schemes, alongside the fact that the *joint* design of local compressors is advantageous compared to the simpler approach of separately designing the local compressors. The IB-based design of compressors for a fully centralized scheme, without equalization at RAPs, is an interesting point for further research.

Acknowledgement

This was partly funded by the German ministry of education and research (BMBF) under grants 16KISK109 (6G-ANNA), 16KISK016 (Open6GHub), and 16KISK068 (6G-TakeOff).

References

[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for Next Generation Wireless Systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[2] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE J. Sel. Areas Inf. Theory*, vol. 8, no. 5, pp. 742–758, April 2014.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO: Uniformly Great Service for Everyone," in *IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Stockholm, Sweden, June 2015.

[4] 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, January 2017.

[5] O. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of User-Centric Cell-Free Massive MIMO," *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162–472, January 2021.

[6] S. Buzzi and C. D'Andrea, "Cell-Free Massive MIMO: User-Centric Approach," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 706–709, August 2017.

[7] 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, December 2021.

[8] D. Maryopi, M. Bashar, and A. Burr, "On the Uplink Throughput of Zero Forcing in Cell-Free Massive MIMO with Coarse Quantization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7220–7224, June 2019.

[9] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, "Max–Min Rate of Cell-Free Massive MIMO Uplink with Optimal Uniform Quantization," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 6796–6815, July 2019.

[10] M. Bashar, H. Q. Ngo, K. Cumanan, A. G. Burr, P. Xiao, E. Björnson, and E. G. Larsson, "Uplink Spectral and Energy Efficiency of Cell-Free Massive MIMO with Optimal Uniform Quantization," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 223–245, October 2020.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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, Jan. 2023.

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

[19] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

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

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

[22] 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, August 2020.

[23] 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.

[24] A. Kurtenbach and P. Wintz, "Quantizing for Noisy Channels," *IEEE Trans. on Commun. Tech.*, vol. 17, no. 2, pp. 291–302, April 1969.

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

[26] S. Hassanpour, D. Wübben, and A. Dekorsy, "Generalized Distributed Information Bottleneck for Fronthaul Rate Reduction at the Cloud-RANs Uplink," in *IEEE Global Communications Conference (GLOBECOM)*, Taipei, Taiwan, December 2020.

[27] 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.

[28] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed Source Coding for Sensor Networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, August 2004.

[29] 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, July 2021.

[30] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, March 1982.