Relevance-Based Multi-User Data Compression for Fronthaul Rate Reduction in Cell-Free Massive 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

***Abstract*—In the Cell-Free massive Multiple-Input Multiple-Output (CF-mMIMO) systems, a large number of distributed users are simultaneously served by multiple Radio Access Points (RAPs). In the uplink, each RAP receives *noisy* observations from several users and must *locally* compress these signals before forwarding them to the corresponding Central Processing Unit (CPU) via multiple fronthaul channels, each subject to a rate limitation. The challenge is to design the compressed signals at the RAPs such that the received signals at the CPU retain as much information as possible about the users (to be retrieved). To address this, we adopt compression techniques based on the *Information Bottleneck* (IB) principle to design local quantizers at the RAPs by ensuring an efficient balance between the *informativity* and *compactness* of the compressed signals. We discuss here two different compression schemes: one that processes the signals *independently* across fronthaul links and another that leverages the side information from previously retrieved signals at the CPU. By using side information, the latter generally provides a better trade-off between the compression efficiency and performance, albeit with an increased complexity. Through numerical simulations, we demonstrate the effectiveness of both IB-based schemes compared to the conventional compression methods, showing their potential for improving fronthaul rate efficiency and overall system performance in typical digital transmission scenarios.**

***Index Terms*—6G, Cell-Free massive MIMO, distributed data compression, information bottleneck method, remote source coding**

I. Introduction

Massive Multiple-Input Multiple-Output (MIMO) technology has become a cornerstone for addressing the ever-growing demand for higher data rates and improved spectral efficiency in modern wireless communication systems [1]. Massive MIMO exploits spatial multiplexing and beamforming techniques by utilizing a large array of antennas at base stations, enabling the simultaneous service of multiple users over the same time-frequency resources. This approach significantly enhances both spectral and energy efficiency, positioning the massive MIMO as a pivotal technology for next-generation wireless networks [2]. Building on this, Cell-Free massive MIMO (CF-mMIMO) has recently become a key focus as an innovative network architecture. Its ability to eliminate cell boundaries and provide ubiquitous coverage aligns well with the vision for 6G networks [3], which aim to offer seamless connectivity, Ultra-Reliable Low-Latency Communication (URLLC), and massive Machine-Type Communication (mMTC) across diverse environments. As 6G evolves, CF-mMIMO is expected to play a key role in enabling the high-density, low-latency, and energy-efficient communication demands of future wireless systems. Unlike traditional cellular systems, CF-mMIMO eliminates the concept of cells and cell boundaries, allowing the users to be served by all available Radio Access Points (RAPs) simultaneously. These RAPs are connected to a Central Processing Unit (CPU) through a fronthaul network [4]. However, the conventional CF-mMIMO model, where every RAP processes and transmits signals for all users, faces scalability issues due to the linear (or faster) growth in computational complexity and fronthaul rates as the number of users increases [5]. The User-Centric CF-mMIMO has introduced a more scalable solution than the conventional CF-mMIMO model [6] where only a cluster of RAPs, specifically those that provide the most benefit to a given user, are responsible for serving that user. This approach offers enhanced coverage, improved fairness, and increased capacity [5], [7], thereby serving as the basis for our study. Furthermore, distributed processing and cooperation among RAPs in CF-mMIMO enables more efficient resource allocation and interference management, boosting the overall network performance [8].

Despite these advantages, the practical deployment of CF-mMIMO presents challenges, particularly concerning fronthaul capacity and signal processing overhead. The fronthaul network, which links distributed RAPs to the CPU, becomes a critical bottleneck, especially due to the large number of antennas and high-dimensional signal processing in the uplink. These factors impose strict fronthaul capacity requirements. Thus, to mitigate this issue, signal compression techniques are essential for reducing the volume of data transmitted over the fronthaul network, while maintaining the necessary information for accurate signal recovery at the CPU. Therefore, the development of efficient compression algorithms tailored to the specific characteristics of massive MIMO signals is crucial to unleashing the full potential of CF-mMIMO in future networks. In response to the performance bottleneck outlined, several fronthaul compression methods have been proposed, such as those in [9]–[11], which primarily utilize uniform quantization. In this paper, we adopt the *Information Bottleneck* (IB) principle [12], [13] to design compression schemes aimed at reducing fronthaul rates in the uplink of CF-mMIMO systems.

The IB method offers a variational principle for compressing a Random Variable (RV) such that the quantized signal preserves as much information as possible about another statistically correlated, relevant variable. This preservation of information is flexible, allowing control of the trade-off between the *compactness* (i.e., size reduction) and the *informativity* (i.e.,

979-8-3315-2289-6/25/$31.00 ©2025 IEEE

retention of meaningful content) of the compressed output signal. The IB method formalizes this trade-off using the Mutual Information (MI) [14] as a symmetric measure to balance these competing objectives, enabling it to be highly adaptable in various scenarios. A general setup happens where several noisy observations of several RVs should be compressed at multiple terminals such that the information of all RVs is preserved in the compressed signals. This setup was investigated in [15] and an iterative algorithm, the GEneralized Multivariate IB (GEMIB) was proposed to design the local compressors.

In wireless communication systems, the IB method has been applied in numerous ways, ranging from its usage in analog-to-digital (A/D) converters for receiver front ends [16], to discrete channel decoding techniques [17], [18], and more recently, in task or goal-oriented communications [19], [20]. Beyond communication systems, IB has also been leveraged as a powerful clustering approach in machine learning [21]. It facilitates clustering by grouping data based on their relevance to a target task which can be seen as an unsupervised learning method that focuses on maximizing the relevant information in the compressed representation. This connection between IB and machine learning highlights its importance in designing efficient, data-driven solutions for wireless networks.

In the next section, we start with the system model for IB-based distributed data compression and propose the framework to adapt the CF-mMIMO setup to this compression scheme. Then we present two different approaches to design the IB-based (local) compressors. In Section III, several numerical investigations are presented to confirm the effectiveness of the proposed approach and a brief wrap-up in Section IV concludes this paper. Note that 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 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 the Kullback-Leibler (KL) divergence [14], respectively.

II. Distributed Information Bottleneck Compression for Cell-Free Massive MIMO Systems

In this section, first, we detail the system model for distributed data compression based on the Information Bottleneck (IB) method and propose a framework to tailor this general system model to the CF-mMIMO systems. Then, we present two different schemes to design the local IB-based compressors in the same context.

A. System Model

We consider the system model illustrated in Figure 1(a) for the uplink transmission in the CF-mMIMO system with N distributed users and J RAPs connected to a CPU by J Ideal (error-free) Rate-limited Channels (IRCs) in the fronthaul. Each RAP j, $j \in \{1, \ldots, J\}$, receives a set of different (non-interfering) *noisy* observations, $\{\mathsf{y}_{m\ell}^{(j)}\}$ from the set of source signals, $\{\mathsf{x}_{m\ell}\}$ of users served by it. For a clear enumeration and simplicity of formulation in what follows, we allocate N_m

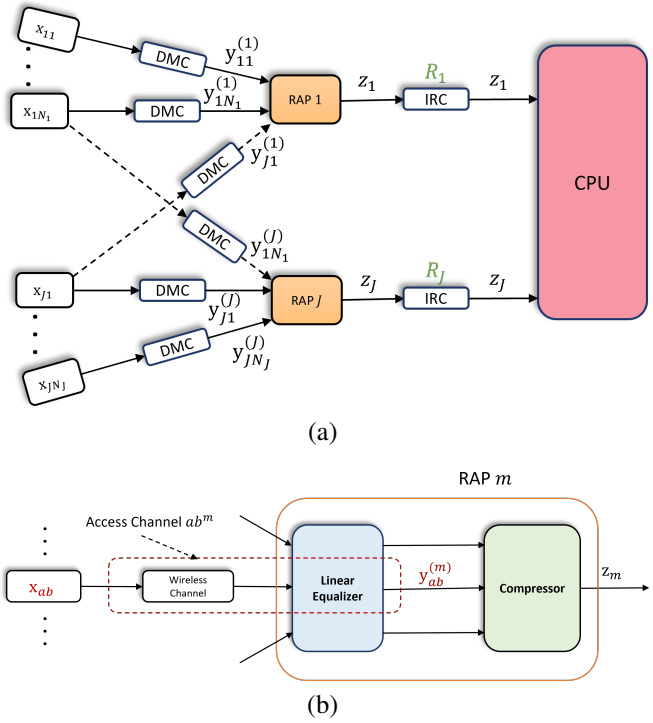


Fig. 1. **(a)** The system model for IB-based distributed data compression. Each Radio Access Point (RAP) has several noisy observations received via Discrete Memoryless Channels (DMCs) from users in its service area and must compress these signals before a forward transmission to the Central Processing Unit (CPU) through an Ideal Rate-limited Channel (IRC) in the fronthaul. **(b)** Applying a linear equalization at each RAP cancels the spatial interference of different user signals (which get served by it) and separates their signals.

users to each RAP m such that $N = \sum_{m=1}^{J} N_m$. Here, $\mathsf{x}_{m\ell}$ is an arbitrary source signal where $m = \{1, \ldots, J\}$ is the index of RAP to which the user is allocated and $\ell = \{1, \ldots, N_m\}$ is the index of the user within the allocated user set to m-th RAP in the network. Therefore, users that are allocated to RAP m are $\mathsf{x}_{m1}, \ldots, \mathsf{x}_{mN_m}$, however, some users in this group are served by a cluster of RAPs as intended for CF-mMIMO. We refer to the users served by only one RAP as uncommon (unshared) users and those served by a cluster of RAPs as common (shared) users, e.g., in Figure 1(a), the user with source signal x_{11} served by only RAP 1 is an uncommon user and the user with source signal x_{J1} served by RAPs 1 and J is a common user. Moreover, $\mathsf{y}_{m\ell}^{(j)}$ denotes a noisy observation of $\mathsf{x}_{m\ell}$ at j-th RAP. The interrelation between $\mathsf{x}_{m\ell}$ and $\mathsf{y}_{m\ell}^{(j)}$ is modeled through a Discrete Memoryless Channel (DMC) whose transition probabilities, $p(\mathsf{y}_{m\ell}^{(j)}|\mathsf{x}_{m\ell})$, and input distribution, $p(\mathsf{x}_{m\ell})$, are presumed to be known.

To fulfill these requirements in the CF-mMIMO system, we presume a linear equalization [22] at each RAP to cancel the spatial interference of different users (which get served by it) and separate their signals. Therefore, the interrelation between any arbitrary source signal, $\mathsf{x}_{m\ell}$, and the corresponding output of the linear equalizer, $\mathsf{y}_{m\ell}^{(j)}$, at RAP j is termed the access channel $m\ell^{(j)}$ in Figure 1(b) which is modeled by a DMC. Let $\mathbf{y}_j$ denote all noisy observations at RAP j. Each RAP j

compresses $\mathbf{y}_j$ to the signal z_j to transmit to the CPU through an error-free fronthaul channel with a limited rate of R_j. Since different noisy observations of the source signal $\mathsf{x}_{m\ell}$ can be received at different RAPs, the information of $\mathsf{x}_{m\ell}$ should be preserved in all corresponding compressed signals. For this purpose, let $\mathbf{v}_{\mathsf{x}_{m\ell}}$ denote the set of all z_j such that $\mathsf{y}_{m\ell}^{(j)} \in \mathbf{y}_j$, i.e., every compressed signal that a noisy observation of $\mathsf{x}_{m\ell}$ is included in its corresponding compressor input.

In the design of IB-based compressors, the fundamental goal is to optimize the trade-off between the information retention and compression rate. In the context of Cell-Free mMIMO, this involves maximizing the MI between the user signals and the pertinent compressed representatives at the CPU while minimizing the required fronthaul rate for transmission. Each RAP compresses its received *noisy* observations before forwarding them to the CPU. The first approach applies the IB principle independently at each RAP [12], [23], treating the observations for each user as separate entities. This method is straightforward but can be suboptimal, as it does not fully exploit the potential correlations across RAPs that are jointly serving a certain user. The second approach, more sophisticated but complex, leverages *joint* compression strategies. Through the joint design of all compressors at different RAPs that serve the same user, this approach can reduce the redundancy in the compressed data [24], thereby improving fronthaul efficiency while maintaining high reconstruction accuracy at the CPU. It was shown in [25] that the IB-based joint design of compressors in the CF-mMIMO systems outperforms the separate IB-based design as well as the conventional quantizers such as the Lloyd-Max scheme [26]. However, these schemes require a large number of compressors at each RAP when the number of users that it serves increases. Therefore, they have limited practical efficiency. Here, we take advantage of the multi-source IB-based data compression method [15] to design only one compressor at each RAP to compress all *noisy* observations from users served by that RAP. Multi-source IB-based data compression is a novel distributed noisy source coding scheme. It focuses on a generic setup wherein, several terminals receive different sets of *noisy* observations from the users and compress their signals before transmitting them over multiple ideal (error-free) rate-limited channels to a remote processing unit.

Then the design problem is formulated as a basic trade-off between two MI terms. The first one is the sum of MI terms between each source signal and its sets of the corresponding received signals at the CPU which is called the relevant information. The second one called total compression rate is the sum of MI terms between the noisy observations (of the source signals) which are the inputs of the local compressors and their output signals. The goal of IB-based compression is to maximize the relevant information such that the total compression rate does not exceed the capacity of the corresponding fronthaul network which is similar to the concept of data clustering, where the goal is to group observations in a way that retains the most relevant features while reducing the redundancy.

The relevant information is naturally quantified by the sum of MI terms among each source signal, $\mathsf{x}_{m\ell}$, and the corresponding set of $\mathbf{v}_{\mathsf{x}_{m\ell}}$, i.e., all the compressed signals that must preserve information about $\mathsf{x}_{m\ell}$. Therefore, for all users, the relevant information is given by

$$\text{Relevant Information} = \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}). \tag{1}$$

However, there is no single, definitive measure for compactness, and one can apply various terms depending on the context. We consider two sets of constraints to determine the compactness, corresponding to the parallel [27] and successive [28] (retrieval) processing strategies at the CPU which are used in the generic setup of the IB-based multi-source data compression [15].

B. Parallel Processing

As the first choice of the imposed constraint set, we consider a scenario where individual fronthaul links experience varying rate limitations, and from a compression standpoint, no side information is utilized when processing each observation $\mathbf{y}_j$, enabling independent parallel processing across the branches. Let $P^* = \{p^*(\mathsf{z}_1|\mathbf{y}_1), \cdots, p^*(\mathsf{z}_J|\mathbf{y}_J)\}$ denote the optimal set of compressors at all RAPs. The design problem is formulated as follows:

$$P^* = \underset{P:\, \forall m\ I(\mathbf{y}_m;\mathsf{z}_m)\le R_m}{\operatorname{argmax}} \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}), \tag{2}$$

wherein, $0 \le R_m \le \log_2 |\mathcal{Z}_m|$ bits, sets an upper-bound on the m-th compression rate, $I(\mathbf{y}_m; \mathsf{z}_m)$. Utilizing the method of Lagrange multipliers [29], the design problem (2) can be reformulated as the following unconstrained optimization problem, assuming the validity of all compressor mappings:

$$P^* = \underset{P}{\operatorname{argmax}} \sum_{m=1}^{J}\sum_{\ell=1}^{N_m} I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}) - \sum_{m=1}^{J} \lambda_m I(\mathbf{y}_m; \mathsf{z}_m), \tag{3}$$

where $\lambda_m \ge 0$ is associated with the rate R_m, in (2). The form of stationary solution for the (non-convex) design problem (3) is obtained in [15] for each local compressor $\{p(\mathsf{z}_j|\mathbf{y}_j)\ | j \in \{1, \cdots, J\}\}$ as follows:

$$p(z_j|\mathbf{y}_j) = \frac{p(z_j)}{\psi^{\text{Par.}}_{\mathsf{z}_j}(\mathbf{y}_j, \beta_j)} \exp\Big(-d_{\text{Par.}}(\mathbf{y}_j, z_j)\Big), \tag{4}$$

where, $\psi^{\text{Par.}}_{\mathsf{z}_j}(\mathbf{y}_j, \beta_j)$, serves as a normalization function that guarantees the validity of the corresponding quantizer mapping, and the relevant distortion, $d_{\text{Par.}}(\mathbf{y}_j, z_j)$ is given by

$$\begin{aligned} d_{\text{Par.}}(\mathbf{y}_j, z_j) = \beta_j \sum_{(m,\ell):\ \mathsf{z}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} & \mathbb{E}_{p(\mathsf{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathbf{y}_j)} \\ & \Big\{ D_{\text{KL}}\big(p(\mathsf{x}_{m\ell}|\mathbf{y}_j, \mathsf{v}^{-j}_{\mathsf{x}_{m\ell}}) \| p(\mathsf{x}_{m\ell}|\mathsf{v}_{\mathsf{x}_{m\ell}})\big)\Big\}, \end{aligned} \tag{5}$$

with $\beta_j = \frac{1}{\lambda_j}$, and $\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{v}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$. In (5), the summation occurs over all pairs of (m, ℓ) at RAP j that have a noisy observation $\mathsf{y}_{m\ell}^{(j)}$ of source signal $\mathsf{x}_{m\ell}$ to compute $d_{\text{Par.}}(\mathbf{y}_j, z_j)$.

C. Successive Processing

As the second option for the processing flow, we consider a successive scheme where the side information from previously retrieved signals is utilized at the CPU during the recovery of a specific source signal. In general, this approach offers a better

"informativity-compactness" trade-off compared to parallel processing, albeit with increased processing complexity. This scheme aligns well with the Wyner-Ziv framework for source coding [30], where statistically correlated signals are used as side information at the decoder. The design problem for the optimal set of compressors $P^* = \{p^*(\mathsf{z}_1|\mathbf{y}_1), \cdots, p^*(\mathsf{z}_J|\mathbf{y}_J)\}$ is formulated as follows:

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

where, $0 \le R_m \le \log_2 |\mathcal{Z}_m|$ bits is the rate of the m-th fronthaul link and sets an upper-bound on the m-th conditional compression rate, $I(\mathbf{y}_m; \mathsf{z}_m|\mathbf{z}_{1:m-1})$. In this case, there is an additional degree of freedom: the processing order. This order generally influences the performance and should ideally be optimized. Henceforth, the discussion is continued with a fixed choice of ordering. Similar to the parallel processing, by applying the method of *Lagrange Multipliers*, the design problem (6) can be reformulated as an unconstrained optimization problem, assuming the validity of all compressor mappings:

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

wherein, $\lambda_m \ge 0$, is associated with the upper-bound, R_m, in (6). It is important to note that in the special case of *full-informativity*, achieved by setting $\lambda_m \to 0$ for $m = 1$ to J, the objective functionals for both the parallel and successive processing schemes become identical since the difference between the objective functionals in (3) and (7) lies in their second term, which vanishes when $\lambda_m \to 0$.

The stationary solution for the non-convex design problem (3) is derived in [15] for each local compressor $\{p(\mathsf{z}_j|\mathbf{y}_j)|j \in 1, \cdots, J\}$, and is given as follows:

$$p(z_j|\mathbf{y}_j) = \frac{p(z_j)}{\psi^{\text{Suc.}}_{\mathsf{z}_j}(\mathbf{y}_j, \beta_j)} \exp\Big(- d_{\text{Suc.}}(\mathbf{y}_j, z_j)\Big), \quad (8)$$

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

$$\begin{aligned} d_{\text{Suc.}}(\mathbf{y}_j, z_j) = & \beta_j \sum_{(m,\ell):\, \mathsf{z}_j \in \mathbf{v}_{\mathsf{x}_{m\ell}}} \mathbb{E}_{p(\mathsf{v}^{-j}_{\mathsf{x}_{m\ell}}|\mathbf{y}_j)} \\ & \Big\{ D_{\text{KL}}\big(p(\mathsf{x}_{m\ell}|\mathbf{y}_j, \mathsf{v}^{-j}_{\mathsf{x}_{m\ell}}) \| p(\mathsf{x}_{m\ell}|\mathsf{v}_{\mathsf{x}_{m\ell}})\big)\Big\} \\ & - \sum_{\boldsymbol{z}_{1:j-1}} p(\boldsymbol{z}_{1:j-1}|\mathbf{y}_j) \log p(\boldsymbol{z}_{1:j-1}|z_j) \\ & - \beta_j \sum_{k=j+1}^{J} \frac{1}{\beta_k} \sum_{\boldsymbol{z}^{-j}_{1:k}} p(\boldsymbol{z}^{-j}_{1:k}|\mathbf{y}_j) \log p(z_k|\boldsymbol{z}_{1:k-1}), \end{aligned} \quad (9)$$

with $\beta_j = \frac{1}{\lambda_j}$, and $\mathbf{v}^{-j}_{\mathsf{x}_{m\ell}} = \mathbf{v}_{\mathsf{x}_{m\ell}} \setminus \{\mathsf{z}_j\}$. To address the design problems in parallel (4) and successive (8) processing schemes, an iterative algorithm, the *GEneralized Multivariate IB* (GEMIB), was presented in [15] that we use here to design the IB-based compressors in the proposed framework for CF-mMIMO systems.

It is worth mentioning that the above distributed IB-based data compression techniques extend the results presented in [25]. In that work, as in [24], [28], the input signal to each local compressor was a *noisy* observation of a *single* common source signal. In contrast, here, in addition to the common source signals, different local compressors also quantize distinct sets of *noisy* observations from different uncommon source signals. Moreover, the key distinction of successive processing from the parallel scheme lies in the inclusion of side information at the compression rates, which further exploits the potential correlations between the output signals of the local compressors. This added layer of complexity in the design is reflected in the resulting stationary solutions. When comparing the derived relevant distortion in (9) for successive processing to its counterpart in (5) for the parallel scheme, we observe that two additional terms emerge. These terms arise from conditioning the compression rates, corresponding to the incorporation of side information in the design problem. To address the design problems in (3) and (7), the GEneralized Multivariate IB (GEMIB) was presented in [15]. In the next section, we investigate the performance of GEMIB for parallel (P-GEMIB) and successive (S-GEMIB) processing schemes.

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. Let us assume $N = 3$ users that are served by $J = 2$ RAPs as depicted in Figure 2(a). After the linear equalization at RAPs, RAP 1 has the noisy observations $\mathsf{y}^{(1)}_{11}$ and $\mathsf{y}^{(1)}_{12}$ of the source signals x_{11} and x_{12}, respectively and RAP 2 has the noisy observations $\mathsf{y}^{(2)}_{12}$ and $\mathsf{y}^{(2)}_{21}$ of the source signals x_{12} and x_{21}, respectively. Note that although the user with the source signal x_{12} is (arbitrarily) assigned to RAP 1 following the enumeration notation of the users in Section II.A, it gets served by both RAPs, therefore both compressors in RAPs should preserve information of this source signal in their outputs. Based on this, $\mathbf{v}_{\mathsf{x}_{12}} = \{\mathsf{z}_1, \mathsf{z}_2\}$ while $\mathbf{v}_{\mathsf{x}_{11}} = \{\mathsf{z}_1\}$ and $\mathbf{v}_{\mathsf{x}_{21}} = \{\mathsf{z}_2\}$. 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^2_{n}, for all access channels from the source signals to the corresponding outputs of equalizers at the RAPs.

To evaluate the compression schemes, we use the relevant information that is basically the overall transmission rate, i.e., the sum of MI terms between the source signals and the corresponding received signals at the CPU

$$\sum_{j=1}^{J} \sum_{\ell=1}^{N_m} I(\mathsf{x}_{m\ell}; \mathbf{v}_{\mathsf{x}_{m\ell}}) = I(\mathsf{x}_{11}; \mathsf{z}_1) + I(\mathsf{x}_{12}; \mathsf{z}_1\mathsf{z}_2) + I(\mathsf{x}_{21}; \mathsf{z}_2), \quad (10)$$

as the performance indicator to compare the GEMIB algorithm in this setup with a popular vector quantization method, the

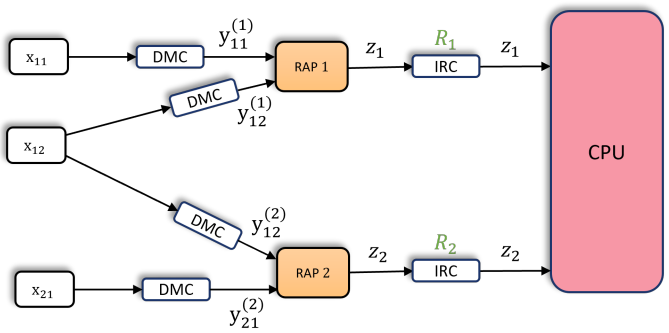
K-Means algorithm [31]. Since these methods are initialized randomly, the same starting points are used for all schemes to ensure fairness, and the best results from 100 trials are retained. We consider the uniformly distributed source signals using a bipolar 8-ASK (Amplitude Shift Keying) constellation with $\sigma_{\mathsf{x}}^2 = 24$ for three users. A total of 100 samples per access channel were generated based on a Monte Carlo simulation.

In the first experiment, we intend to compare the performance of P-GEMIB and K-Means methods in terms of the overall transmission rate (10) versus different numbers of output clusters (per RAP), and the results are illustrated in Figure 2(b). The allowed number of output clusters of each compressor is denoted by $C = |\mathcal{Z}_j|$. Note that, we choose the trade-off parameters, $\lambda_1 = \lambda_2 = 0.01$ for the P-GEMIB algorithm as we prioritize the information preservation to maximize the overall transmission rate. The key takeaway is the clear performance advantage of the P-GEMIB algorithm over the standard K-Means routine. This superiority is evident in both information preservation and compactness. For example, with $\sigma_{\mathsf{n}}^2 = 1$, the P-GEMIB algorithm achieves approximately 5 bits of relevant information while requiring only 12 clusters, compared to K-Means, which requires 16 clusters for the same information. Alternatively, if the number of output clusters is fixed at 8, P-GEMIB supports up to 4.5 bits of relevant information, outperforming K-Means, which can support up to 4 bits.

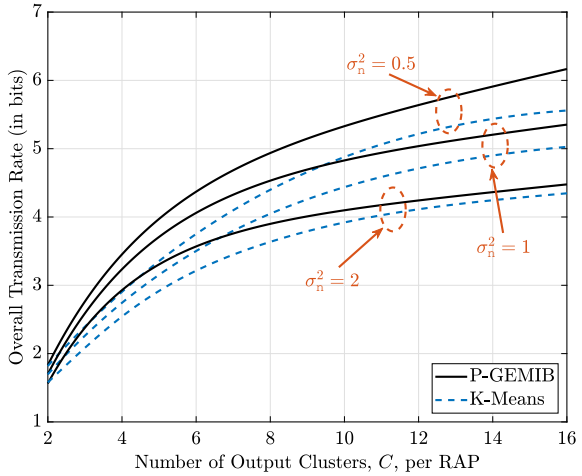
In the last part of our numerical investigations, we compare the performances of P-GEMIB and S-GEMIB schemes for the design of compressors in terms of overall transmission rate (10) and the required total fronthaul rate for supporting it. Note that the total fronthaul rates for P-GEMIB and S-GEMIB are $I(\mathbf{y}_1;\mathsf{z}_1) + I(\mathbf{y}_2;\mathsf{z}_2)$ and $I(\mathbf{y}_1;\mathsf{z}_1) + I(\mathbf{y}_2;\mathsf{z}_2|\mathbf{z}_1)$, respectively. The allowed output clusters of compressors are set to $C = 16$ and $\lambda_1 = \lambda_2 \in (0.33, 0.43)$. Figure 2(c) illustrates the obtained results for different noise powers in the access channels. The key point of this section is that the successive scheme outperforms the parallel one, providing better information-compression trade-offs. This is because the successive scheme leverages the available side information to reduce the total fronthaul rate required to achieve the desired overall transmission rate. In contrast, the parallel processing approach neglects the side information arising from the correlations between different fronthaul channels. To support this, observe that based on the assumed independence relations, the following holds

$$I(\mathbf{y}_j;\mathsf{z}_j|\mathbf{z}_{1:j-1}) = I(\mathbf{y}_j;\mathsf{z}_j) - \underbrace{I(\mathsf{z}_j;\mathbf{z}_{1:j-1})}_{\geq 0}. \qquad (11)$$

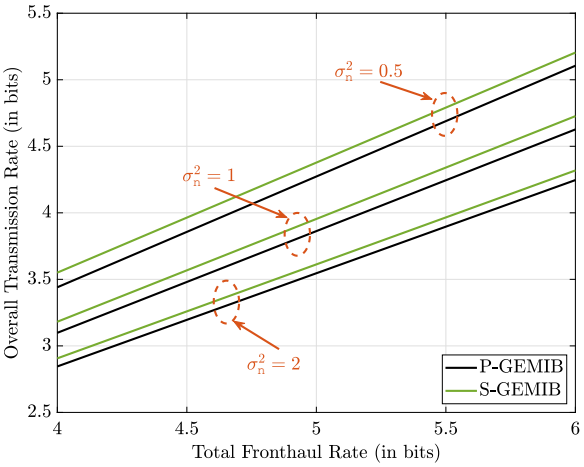
Hence, it can be directly concluded that conditioning on the prior fronthaul channel output signals can either reduce the current unconditional compression rate, $I(\mathbf{y}_j;\mathsf{z}_j)$, or leave it unchanged, as mutual information is non-negative. As the Signal-to-Noise Ratios (SNRs) of the access links increase, the correlations between the signals at two intermediate nodes become stronger. This, in turn, enhances the benefit of utilizing the side information, further widening the performance gap between P-GEMIB and S-GEMIB schemes (in favor of the latter) as can be observed in Figure 2(c).



(a)



(b)



(c)

Fig. 2. **(a)** System model for simulation setup; $N = 3$ users employing bipolar 8-ASK source signaling are served by $J = 2$ RAPs in a CF-mMIMO network. **(b)** A performance comparison between P-GEMIB (3) and K-Means [31] schemes to design the compressors at RAPs in terms of the overall transmission rate vs the number of output clusters (per RAP). **(c)** A performance comparison between P-GEMIB (parallel processing) (3) and S-GEMIB (successive processing) (7) schemes in terms of the overall transmission rate vs total required fronthaul rate.

IV. Summary and Outlook

In this work, we focused on the design of the distributed Information Bottleneck (IB)-based compression schemes for fronthaul rate reduction in the uplink transmission of Cell-Free massive MIMO systems where several Radio Access Points (RAPs) receive different noisy observations of distributed users and must compress their signals before a forward transmission through several *rate-limited* fronthaul channels to the Central Processing Unit (CPU). We considered the scenario of dealing with *error-free* fronthaul links, thereby addressing the distributed (remote) source coding problems. The stationary solutions were provided for different types of processing, i.e., the successive scheme that considers the side information from already retrieved signals at the CPU and the parallel scheme that ignores this side information. Through numerical simulations, we have explicitly demonstrated the superiority of the IB-based distributed compression schemes compared to the conventional vector quantization techniques in the preservation of the information of users. We further showed the outperformance of the successive processing scheme compared to the parallel approach, at the cost of an increase in the design complexity. Future research directions include exploring IB-based compression schemes that take the error-prone fronthaul links into account for the design, as well as investigating IB-based compressor designs for fully centralized systems without requiring equalization at the RAPs.

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 Journal on Selected Areas in Information Theory*, vol. 8, no. 5, pp. 742–758, April 2014.

[3] C. Wang et al., "On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 905–974, February 2023.

[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, C. D'Andrea, A. Zappone, and C. D'Elia, "User-Centric 5G Cellular Networks: Resource Allocation and Comparison with the Cell-Free Massive MIMO Approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1250–1264, November 2019.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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*, July 2024.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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, April 2020.

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

[23] 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.

[24] 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.

[25] A. Danaee, S. Hassanpour, D. Wübben, and A. Dekorsy, "Relevance-Based Information Processing for Fronthaul Rate Reduction in Cell-Free MIMO Systems," in *19th International Symposium on Wireless Communication Systems (ISWCS)*, Rio de Janeiro, Brazil, July 2024.

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

[27] 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.

[28] 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.

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

[30] 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.

[31] Anil K Jain, "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, June 2010.