2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON) | 978-1-6654-4108-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/SECON52354.2021.9491585

Bandwidth-Efficient Precoding in Cell-Free Massive MIMO Networks with Rician Fading Channels

Li Sun, Jing Hou, and Tao Shu
Department of Computer Science and Software Engineering
Auburn University
Auburn, AL 36849, USA
{lzs0070, jzh0141, tshu}@auburn.edu

***Abstract*—Global precoding is an effective way to suppress interference in cell-free massive MIMO systems. However, it requires all access points (APs) to upload their local instantaneous channel state information (CSI) to a central processor via capacity-constrained fronthaul links, consuming significant bandwidth resources. Such overhead may become unaffordable in an ultra-dense network (UDN) in future 5G systems, due to the large number of APs and the frequent CSI uploads required to combat the fast-changing state of the high-frequency channels. In order to address this issue, we propose a novel bandwidth-efficient global zero-forcing precoding strategy for downlink transmission in cell-free massive MIMO systems. By exploiting the physical structure of Rician fading channels, we propose a novel model-based CSI compression mechanism, which decomposes a channel matrix into a line-of-sight (LoS) and a non-line-of-sight (NLoS) components, and then compresses them using a model-based method and a singular-value-decomposition (SVD)-based method, respectively. We also present two optimization-based algorithms to obtain the phase information of the LoS component of the channel, which is then used by the proposed channel matrix decomposition. The simulation results demonstrate the efficiency of the proposed precoding strategy on reducing the upload overhead and improving the bandwidth efficiency.**

***Index Terms*—cell-free massive MIMO, Rician fading channel, zero-forcing precoding, overhead reduction, SVD**

I. Introduction

A cell-free massive MIMO consists of a large number of distributed access points (APs) that jointly serve user equipments (UEs) within a certain area. It has grabbed growing attention recently due to its ability to avoid handovers since there is no "cell-edge" users and to achieve high energy efficiency. However, the benefits of cell-free massive MIMO systems are being hindered by severe interference among the APs. To address this performance degradation, several precoding techniques have been proposed to mitigate the interference, but at the cost of high computational complexity and transmission overhead.

In downlink interference management of massive MIMO systems, local zero-forcing (LZF) precoding and its variants [1], [2] are generally implemented to eliminate the *local* inter-user interference and partially reduce the interference from non-serving APs. In further efforts to suppress the interference from all APs, global ZF (GZF) precoding [3], [4] has been proposed, in which all APs upload their local channel state information (CSI) to a centralized processing unit (CPU) which then conducts global precoding for them. It is expected that the CPU can make a better precoding decision based on the full knowledge of CSI of the whole network. However, a key challenge of the GZF is the huge overhead caused by uploading the CSI to the CPU for precoding. In particular, the GZF requires the CPU to collect instantaneous CSI from all APs through the fronthaul links with limited capacity. The transmission of CSI is carried out in every coherent block, thus consumes considerable bandwidth resources and aggravates communication delay. This overhead issue becomes even more severe with higher carrier frequencies, because the channel's coherence time becomes shorter with the increase of the carrier frequency. This has become a bottleneck that prohibits the GZF to be implemented in an ultra-dense network (UDN) with large number of APs and UEs operating on the high-frequency 4G and 5G bands [2].

Hence, for practical implementation of interference management in cell-free massive MIMO systems, the CSI overhead issue has to be addressed. Although many solutions have been developed to compress the CSI feedback in massive MIMO, they are not tailored for cell-free massive MIMO systems. On one hand, most of them are built for conventional massive MIMO systems in which the channels are modeled as Rayleigh fading due to a large amount of reflection and scattering. On the other hand, these methods are statistic-based and focus on exploring the statistical characteristics of channel matrix. However, in a cell-free massive MIMO system, channels are more suitable to be modeled as Rician fading because the APs are close to the UEs [5]. The short distance between AP and UE makes the channels in cell-free massive MIMO have remarkably different physical structure from the channels in conventional massive MIMO. Therefore, a model-based CSI compression method that considers and fully exploits this unique physical structure of channels is able to reduce the overhead in cell-free massive MIMO more efficiently than the statistic-based methods, but is overlooked in the literature.

Unlike Rayleigh fading channels, a Rician fading channel consists of two components: the line-of-sight (LoS) component and the non-line-of-sight (NLoS) component. Compared with the NLoS component that is caused by multi-path reflection and scattering, the LoS component is more deterministic because it relies on the direct propagation path between the

978-1-6654-4108-7/21/$31.00 ©2021 IEEE

transmitter and the receiver, and contributes the majority of signal power since it suffers less energy loss. This characteristic enables us to decompose a Rician fading channel matrix into two components: a main component and a residual component, corresponding to the LoS component and the NLoS component, respectively. This decomposition motivates us to conduct different overhead-reduction operations on these two parts separately, rather than directly compressing the original channel matrix: the main component can be calculated based on the information of direct propagation path, and the residual component can be compressed using a matrix compression method. By doing so, we are able to reduce the information loss in CSI compression and thus achieve comparable performance to the GZF without CSI compression at a significantly reduced overhead.

In this paper, we propose a novel bandwidth-efficient GZF precoding strategy in order to suppress interference in downlink transmission for cell-free massive MIMO systems, associated with a model-based CSI overhead-reduction mechanism customized for Rician fading channels. Specifically, the AP exploits side information of direct path, i.e., the signal phase and the angle of arrival (AoA), to decompose a Rician fading channel into a LoS comonent and a NLoS component, and then conducts model-based compression and singular-value-decomposition (SVD) on them, respectively. The contributions of this paper are summarized as follows:

1) Exploiting the physical structure of Rician fading channels, we propose a model-based CSI overhead-reduction mechanism for the GZF precoding in cell-free massive MIMO. We decompose the channel matrix into a main matrix and a residual matrix, and then conduct a model-based and an SVD-based compression methods on them, respectively.
2) One of the key challenges in the proposed channel matrix decomposition is that this decomposition requires the phase information of the LoS component of the channel. While the phase information of the channel (i.e., the phase of the superposition of the LoS and the NLoS components) is provided in CSI, the phase of the LoS component is not available. To tackle this problem, we propose two optimization problems to obtain an optimal estimation of the LoS phase for APs with large and small antenna arrays respectively.
3) To verify the efficacy of our strategy, we evaluate its performances in various cell-free massive MIMO scenarios through simulations. It is demonstrated that our strategy can save 83.4% of the overhead while achieving up to 96.5% of the performance compared with the conventional GZF precoding without CSI compression. The appropriate circumstances under which our strategy can perform to its full advantage are also explored.

The rest of this paper is organized as follows. In Sec. II, the related work is briefly reviewed. We introduce our system model in Sec. III and then present our bandwidth-efficient GZF precoding strategy in Sec. IV. In Sec. V, the simulation results are analyzed to compare our strategy with comparable counterparts. Finally, we conclude our work in Sec. VI.

II. Related Work

Interference management is rather challenging due to the distributed nature of the cell-free massive MIMO system with numerous APs. In an effort to eliminate the inter-user interference in downlink transmission, LZF precoding and its variants [1], [2] are widely used, in which each AP conducts ZF precoding independently for its served UEs. However, these precoding techniques can only partially mitigate the interference since the precoding is conducted locally at each AP without any CSI sharing and the precoding decision is "blindly" in some way. In addition, it is required that the number of antennas at each AP is larger than the number of mutually orthogonal pilots, which limits their application in the systems where the APs are equipped with few antennas. Although the GZF is able to further mitigate the interference theoretically, its critical issue is the huge overhead on fronthaul links. Existing research on communication overhead-reduction mainly focuses on the CSI feedback from remote readio heads (RRHs) to CPU in C-RANs, and from mobile station (MS) to base station (BS) in frequency-division duplexing (FDD) massive MIMO systems. Due to the existence of a huge number of antennas, the CSI feedback is too large to be transmitted without compression. How to efficiently reduce the CSI feedback is of great importance to improving the practicality and performance of wireless networks [6], [7].

As an early CSI compression method, the quantization-based compression has been widely utilized in C-RANs and MIMO networks [8]–[10]. Because of its low-complexity, it is suitable for RRHs with limited computation ability. But it is too simple to be adequate to more complex application scenarios. In recent years, compressing sensing (CS) has been applied to reduce the CSI feedback in FDD massive MIMO systems [11]–[13]. It exploits the spatial correlation of CSI and represents the sparse signal by a few elements through random projections [13]. The two major issues of the CS-related methods are their complex recovery computation at BS, and the limitations and requirements of the sparsifying-basis. In order to address these issues, the PCA (principle component analysis)-based schemes are proposed [14]–[16]. The PCA-based compression matrix, being signal-dependent and a statistical basis, is a trade-off between the DCT basis and the KLT basis [14], [15]. It only requires the MS and the BS to have the same spatial correlation matrix in a long-term period. Another advantage of PCA-based compression schemes is the simplified decoding computation at BS, which only needs to conduct matrix operation on the received low-dimension CSI, instead of solving the underdetermined linear system as in CS.

Machine learning (ML) is another powerful tool to solve the CSI feedback-reduction problem [17]. The advantage of ML over CS lies in that, it requires no sparsifying-basis or random projection, and has a simpler CSI recovery procedure [18]. This motivates a growing research effort in this field [18]–[23]. Specifically, convolutional neural networks (CNNs)

have been introduced as the encoder and the decoder, which conducts random projection and inverse transformation from codewords to original channels, respectively [18], [19], [22], [23]. Moreover, the recurrent neural network (RNN) is utilized to extract interframe correlation [20], and redesign the feature compression and decompression modules [21]. However, the ML-based overhead-reduction method requires massive training data sets and nontrivial hyper-parameter tuning.

Our study differs from the existing research in the following three aspects:

1) In most of the related work, the wireless channels are modeled to be Rayleigh fading channels which do not fit the characteristics of UDNs. By contrast, we consider Rician fading channels in UDNs based on which to propose a customized GZF precoding strategy.
2) Most of the current CSI feedback compression methods only consider the statistical characteristics of the channel matrix. However, our strategy fully exploits the physical structure of Rician fading channels to efficiently reduce the information loss in CSI compression and therefore is more reliable.
3) Unlike many existing methods that rely on sparsefying-basis to make the channel vector sparse and need time-consuming iterative algorithms to recover CSI, our strategy is built on simple SVD and only involves basic matrix operation. Moreover, our method does not need off-line training or hyper-parameter tuning that is necessary for ML-based methods, and hence is easy to implement.

III. System Model

We consider a cell-free massive MIMO system consisting of one CPU, L APs and K UEs. Denote $\mathcal{L}$ and $\mathcal{K}$ as the set of APs and that of UEs, respectively. Each AP is equipped with N antennas, and each UE is equipped with a single antenna. The APs jointly serve the UEs and communicate with the CPU through the fronthaul links with limited capacity. Throughout this paper, we use upper and lower case boldfaced letter to describe matrix and vector, respectively. Moreover, we denote the transpose of matrix by $(\cdot)^{\mathrm{T}}$, the Hermitian transpose of matrix by $(\cdot)^{\mathrm{H}}$, and the l_2-norm of vector by $\|\cdot\|$.

A. Propagation Model

The channel matrix of the AP $l \in \mathcal{L}$ is denoted by $\mathbf{G}_l = [\mathbf{g}_{l1}, \mathbf{g}_{l2}, \ldots, \mathbf{g}_{lK}] \in \mathbb{C}^{N\times K}$. The column $\mathbf{g}_{lk} \in \mathbb{C}^{N\times 1}$ is the channel vector between the AP l and the UE $k \in \mathcal{K}$. Each element of $\mathbf{g}_{lk}$ is modeled as a Rician fading channel, which is a combination of a dominant line-of-sight (LoS) component and a non-line-of-sight (NLoS) component following Rayleigh distribution. Thus, the channel vector $\mathbf{g}_{lk}$ is represented as

$$\mathbf{g}_{lk} = \sqrt{\beta_{lk}}\left(\sqrt{\frac{K_{lk}}{K_{lk}+1}}\bar{\mathbf{h}}_{lk} + \sqrt{\frac{1}{K_{lk}+1}}\check{\mathbf{h}}_{lk}\right), \quad (1)$$

where β_{lk} is the large-scale fading coefficient and K_{lk} is the K-factor that represents the ratio of signal power in dominant component over the scattered power on the corresponding channel. $\bar{\mathbf{h}}_{lk}$ and $\check{\mathbf{h}}_{lk}$ denote the LoS and the NLoS components, respectively.

The value of β_{lk} depends on the distance d_{lk} between the AP l and the UE k and is modeled as [24]

$$\beta_{lk} = \alpha + 10\rho\log_{10}(d_{lk}) + \xi[dB], \xi \sim \mathcal{N}(0, \sigma_{\mathrm{S}}^2), \quad (2)$$

where α and ρ are the least square fittings of floating intercept and slope respectively over the measured distance, and ξ represents a lognormal shadowing with variance σ_{S}^2. Since β_{lk} does not change frequently, we assume that it is known at both CPU and AP. The large-scale fading coefficients for the LoS and the NLoS components are computed as $\bar{\beta}_{lk} = \frac{K_{lk}}{K_{lk}+1}\beta_{lk}$ and $\check{\beta}_{lk} = \frac{1}{K_{lk}+1}\beta_{lk}$, respectively.

The LoS components $\bar{\mathbf{h}}_{lk}$ are modeled as

$$\bar{\mathbf{h}}_{lk} = \left[e^{j\varphi_{lk}^{(1)}}, e^{j\varphi_{lk}^{(2)}}, \ldots, e^{j\varphi_{lk}^{(N)}}\right]^{\mathrm{T}}, \quad (3)$$

where $\varphi_{lk}^{(n)}, 1 \le n \le N$ is the instantaneous phase of the signal received by the nth antanna at the AP l from the UE k. The elements in the NLoS components $\check{\mathbf{h}}_{lk}$ are modeled as i.i.d. random variables drawn from $\mathcal{CN}(\mathbf{0}, \mathbf{R}_{lk})$, where $\mathbf{R}_{lk}$ describes the spatial correlation of the NLoS components.

B. User-Centric Association

In this paper, we consider user-centric downlink transmission, i.e., each UE is only served by the APs that are not far away. The motivation to consider the user-centric service comes from the intention to increase the energy efficiency of downlink transmission [25]. For illustration purpose, we assume that the AP l serves the UE k when their propagation distance, is smaller than a predefined threshold $\bar{d}$. We denote the set of APs that serve the UE k as $\mathcal{L}(k)$ and the set of UEs that are served by the AP l as $\mathcal{K}(l)$. Note that other criteria such as the signal strength can also be used to determine the serving range of the APs, and this would not affect the design and performance of our precoding strategy proposed below.

C. Communication Process

The transmission between APs and UEs is conducted in TDD mode. Specifically, each coherence block is divided into three stages: uplink training, CSI uploading, and downlink transmission.

1) Uplink Training: In the uplink training stage, all UEs simultaneously send their assigned pilot signals to all APs. Upon receiving the copies of these pilots, each AP utilizes a well designed channel estimation method, such as MMSE estimation, to estimate the channels between itself and the UEs. In this study, we assume that the AP $l \in \mathcal{L}$ has already obtained its perfect CSI, i.e., the channel matrix $\mathbf{G}_l$, and focus on the GZF precoding strategy involved in the subsequent stages.

2) CSI Uploading: After channel estimation, the AP l transforms its local channel matrix $\mathbf{G}_l$ into compressed CSI and uploads it to the CPU through the fronthaul link. Then the CPU computes the $\tilde{\mathbf{G}}_l$ as a recovery of the original channel matrix $\mathbf{G}_l$ based on the compressed CSI. The details will be discussed in Sec. IV.

3) Downlink Transmission: In this stage, the CPU computes the GZF precoding matrix for downlink transmission based on the recovered channel matrix $\tilde{\mathbf{G}} = \left[\tilde{\mathbf{G}}_1, \tilde{\mathbf{G}}_2, \ldots, \tilde{\mathbf{G}}_L\right]^{\mathrm{T}} \in \mathbb{C}^{LN\times K}$. In global precoding, the whole cell-free massive MIMO system consisting of L APs each with N antennas can be treated as a virtual AP with LN antennas. Therefore, for the UE $k \in \mathcal{K}$, the normalized precoding vector $\mathbf{w}_k \in \mathbb{C}^{LN\times 1}$ is calculated as

$$\mathbf{w}_k = \frac{\mathbf{f}_k}{\|\mathbf{f}_k\|_F^2}, \tag{4}$$

where $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_K] = \tilde{\mathbf{G}} / \left(\tilde{\mathbf{G}}^{\mathrm{H}}\tilde{\mathbf{G}}\right)$. The collection of the precoding vectors for all UEs is $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K] = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L]^{\mathrm{T}}$, where $\mathbf{W}_l \in \mathbb{C}^{N\times K}$ is the precoding matrix for the AP $l \in \mathcal{L}$. The CPU sends back the precoding matrix $\mathbf{W}_l$ to the AP l that then calculates the transmitted signal as

$$\mathbf{x}_l = \sum_{k\in\mathcal{K}(l)} \sqrt{\eta_k}\mathbf{w}_{lk}q_k, \tag{5}$$

where $\mathbf{w}_{lk} \in \mathbb{C}^{N\times 1}$ is the kth column of $\mathbf{W}_l$, η_k denotes the power allocated to the UE k, and q_k is the data intended to the UE k, $\mathbb{E}\{|q_k|^2\} = 1$.

The received signal at the UE k is modeled as

$$y_k = \sum_{l\in\mathcal{L}(k)} \mathbf{g}_{lk}^{\mathrm{H}}\sqrt{\eta_k}\mathbf{w}_{lk}q_k + \sum_{\substack{t\in\mathcal{K}\\ t\neq k}}\sum_{l\in\mathcal{L}(t)} \mathbf{g}_{lk}^{\mathrm{H}}\sqrt{\eta_t}\mathbf{w}_{lt}q_t + n_k, \tag{6}$$

where the first item is the desired signal, the second stands for the interference and $n_k \sim \mathcal{CN}(0, \sigma_{\mathrm{N}}^2)$ is the additive white Gaussian noise at the UE k.

D. Performance Metrics

We will evaluate the GZF precoding strategy based on two key performance metrics: the spectral efficiency and the upload overhead.

1) Spectral Efficiency: Based on Eq. (6), the SINR at the UE k can be calculated as

$$SINR_k^{\mathrm{G}} = \frac{\left|\sum_{l\in\mathcal{L}(k)} \mathbf{g}_{lk}^{\mathrm{H}}\sqrt{\eta_k}\mathbf{w}_{lk}\right|^2}{\sum_{\substack{t\in\mathcal{K}\\ t\neq k}}\left|\sum_{l\in\mathcal{L}(t)} \mathbf{g}_{lk}^{\mathrm{H}}\sqrt{\eta_t}\mathbf{w}_{lt}\right|^2 + \sigma_{\mathrm{N}}^2}. \tag{7}$$

The achievable data rate for the UE k is calculated as

$$r_k^{\mathrm{G}} = \log_2\left(1 + SINR_k^{\mathrm{G}}\right). \tag{8}$$

2) Upload Overhead: In the GZF precoding, all APs upload their local channel matrices to the CPU through the fronthaul links for global precoding. If each AP uploads its full channel matrix, the total overhead is $2L \times N \times K$ (consider each element in a channel matrix as a complex number containing a real part and an imaginary part). A large overhead from a huge number of APs in UDNs with limited bandwidth would cause unaffordable transmission delay. Therefore, the upload overhead should be considered as an important metric to evaluate the performance of the precoding strategy.

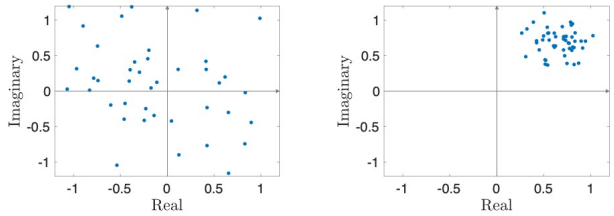
IV. Bandwidth-Efficient Global ZF Precoding Strategy

In order to improve the efficiency of GZF precoding in cell-free massive MIMO systems, we propose a bandwidth-efficient precoding strategy by exploiting the physical structure of Rician fading channels to address the upload overhead challenge. This strategy consists of two parts: the AP-side operation and the CPU-side operation, where the CSI is compressed and recovered, respectively. The details are discussed as follows.

A. AP-Side Operation

In the AP-side operation, the AP compresses its channel matrix before uploading it to the CPU. In contrast to the conventional CSI compression methods that conduct compression directly on the original channel matrix, in our proposed strategy, the AP first decomposes the original channel matrix into a main matrix and a residual matrix, and then compresses them respectively.

1) Rician Fading Channel Decomposition: Different from a Rayleigh fading channel, a Rician fading channel exhibits higher directivity as a result of the LoS component that contributes the most of the signal power. For a Rayleigh fading channel and a Rician fading channel between two arbitrary pairs of transmitting and receiving antennas, we collect 50 samples of the channel gain respectively and display them in Fig. 1. As illustrated in Fig. 1(a), all sample points of the Rayleigh fading channel are uniformly distributed in the whole signal space; while for the Rician fading channel, the points are constricted within a limited area, shown in Fig. 2(b). This



(a) Sample points of Rayleigh fading channel (b) Sample points of Rician fading channel

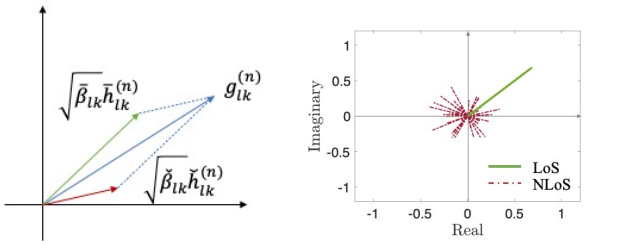
Fig. 1. Spacial correlations of different fading types

characteristic allows us to decompose a Rician fading channel matrix into a main and a residual matrices, which correspond to the LoS and the NLoS components, respectively. According to Eq. (1), the channel between the nth antenna of the AP l and the UE k can be decomposed as follows:

$$g_{lk}^{(n)} = \sqrt{\bar{\beta}_{lk}}\bar{h}_{lk}^{(n)} + \sqrt{\check{\beta}_{lk}}\check{h}_{lk}^{(n)}, \tag{9}$$

where $\bar{h}_{lk}^{(n)}$ is the deterministic LoS component that relies on the phase of the received signal at the receiving antenna, and $\check{h}_{lk}^{(n)}$ denotes the NLoS component that is modeled as an i.i.d. random variable. The relationship between these two components is described in Fig. 2(a). Given a Rician channel gain $g_{lk}^{(n)}$, if the LoS component $\bar{h}_{lk}^{(n)}$ is available, the NLoS

component $\check{h}_{lk}^{(n)}$ can be easily calculated. In this way, the sample points of the Rician fading channel shown in Fig. 1(b) can be decomposed and the corresponding LoS component and NLoS components are shown in Fig. 2(b). Because the NLoS component contributes only the minority of the signal power, i.e., its magnitude is rather small, the information loss caused by compressing the residual matrix brings smaller impact on CSI recovery than that coming from directly compressing the original channel matrix $\mathbf{G}_l$. The prerequisite for this implementation is the ability to extract the LoS component from a Rician fading channel, as will be discussed below.



(a) Decomposition of Rician fading channel (b) Decomposition of the sample points in Fig. 1(b)

Fig. 2. Rician fading channel decomposition

2) LoS Components Extraction: We consider a horizontal uniform linear array (ULA) with antenna spacing $d_A = 1/2\lambda$, where λ is the wavelength. Since the UEs are located in the far-field of the AP, the signal that reaches the antenna array at the AP from a UE can be treated as a plane wave from a generic azimuth angle [26]. Denote the distance between the closest antenna (suppose to be the 1st antenna) of the AP l and the UE k as d, the LoS components $\bar{\mathbf{h}}_{lk}$ described in Eq. (3) can be further modeled as

$$\begin{aligned}\bar{\mathbf{h}}_{lk} &= \left[e^{j\varphi_{lk}^{(1)}}, \ldots, e^{j\varphi_{lk}^{(N)}}\right]^{\mathrm{T}} \\ &= \left[e^{j\left(2\pi\frac{d}{\lambda}+\varphi_{lk}^{(0)}\right)}, e^{j\left(2\pi\frac{d+d_A\sin(\theta_{lk})}{\lambda}+\varphi_{lk}^{(0)}\right)}, \ldots,\right. \\ &\quad \left. e^{j\left(2\pi\frac{d+(N-1)d_A\sin(\theta_{lk})}{\lambda}+\varphi_{lk}^{(0)}\right)}\right]^{\mathrm{T}},\end{aligned} \tag{10}$$

where θ_{lk} is the angle of arrival (AoA) from the UE k to the AP l and $\varphi_{lk}^{(0)}$ is the initial phase. If the AoA θ_{lk} and the signal phase $\varphi_{lk}^{(1)} = 2\pi\frac{d}{\lambda} + \varphi_{lk}^{(0)}$ at the 1st antenna are known, the phases at all antennas can be aligned and Eq. (10) can be simplified as [7]:

$$\bar{\mathbf{h}}_{lk} = e^{j\varphi_{lk}^{(1)}}\left[1, e^{j\pi\sin(\theta_{lk})}, \ldots, e^{j(N-1)\pi\sin(\theta_{lk})}\right]^{\mathrm{T}}. \tag{11}$$

In the rest of the paper, we use φ_{lk} instead by omitting the superscript of $\varphi_{lk}^{(1)}$. As illustrated in Fig. 3, the phase shifts between any two adjacent antennas are identical to be $\pi\sin(\theta_{lk})$. This property enables us to infer each element in $\bar{\mathbf{h}}_{lk}$ based on the given φ_{lk} and θ_{lk}. In this way, we can extract the LoS components $\bar{\mathbf{h}}_{lk}$ from the observed channel vector $\mathbf{g}_{lk}$ and then obtain the NLoS components $\check{\mathbf{h}}_{lk}$ as

$$\check{\mathbf{h}}_{lk} = \frac{1}{\sqrt{\check{\beta}_{lk}}}\left(\mathbf{g}_{lk} - \sqrt{\bar{\beta}_{lk}}\bar{\mathbf{h}}_{lk}\right) \tag{12}$$

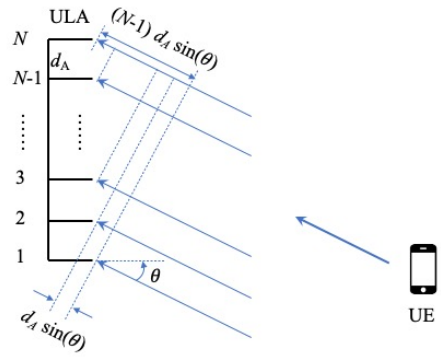


Fig. 3. LoS propagation

3) Side Information Calculation: It is obvious that the signal phase and the AoA play important roles in the LoS components extraction. Their values keep changing with the movement of UEs, which may cause significant change in the LoS components. Fortunately, these changes are identical for all antennas at the AP [5], and can be tracked as follow.

Firstly, we estimate the AoA θ_{lk}, for $l \in \mathcal{L}$, $k \in \mathcal{K}$. There have been many efficient algorithms to estimate the AoA on ULA with omni-directional antenna elements, such as MUSIC and ESPRIT. In this study, we use MUSIC to estimate θ_{lk}. The details are omitted.

Next, we calculate the signal phase φ_{lk}. For different scenarios where the antenna array at AP is large or small, we design the following two algorithms.

Algorithm I: When the antenna array is large, we calculate the mean value of channel vector $\mathbf{g}_{lk}$ as follow

$$\frac{1}{N}\sum_{1\le n\le N} g_{lk}^{(n)} = \frac{1}{N}\sqrt{\bar{\beta}_{lk}}\sum_{1\le n\le N}\bar{h}_{lk}^{(n)} + \frac{1}{N}\sqrt{\check{\beta}_{lk}}\sum_{1\le n\le N}\check{h}_{lk}^{(n)} \tag{13}$$

We assume that the number of the antennas is large enough, then $\frac{1}{N}\sum\check{h}_{lk}^{(n)} \to 0$, since $\check{\mathbf{h}}_{lk} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{lk})$. Hence,

$$\frac{1}{N}\sum_{1\le n\le N} g_{lk}^{(n)} \approx \frac{1}{N}\sqrt{\bar{\beta}_{lk}}\sum_{1\le n\le N}\bar{h}_{lk}^{(n)}. \tag{14}$$

According to Eq. (11),

$$\sum_{1\le n\le N}\bar{h}_{lk}^{(n)} = e^{j\varphi_{lk}}\left(1 + e^{j\pi\sin(\theta_{lk})} + \cdots + e^{j(N-1)\pi\sin(\theta_{lk})}\right). \tag{15}$$

Substituting Eq. (15) into Eq. (14), we get

$$\frac{1}{N}\sum_{1\le n\le N} g_{lk}^{(n)} \approx \frac{1}{N}\sqrt{\bar{\beta}_{lk}}e^{j\varphi_{lk}}\left(1 + \cdots + e^{j(N-1)\pi\sin(\theta_{lk})}\right). \tag{16}$$

As $\mathbf{g}_{lk}$ is known and θ_{lk} has already been calculated, there is only one variable in Eq. (16), which then can be rewritten as

$$e^{j\varphi_{lk}} = a + bi, \tag{17}$$

where a and b are the calculation results. According to Euler's formula and Taylor's series for trigonometric functions,

$$\begin{aligned} e^{j\varphi_{lk}} &= \cos(\varphi_{lk}) + i\sin(\varphi_{lk}) \\ &= 1 - \frac{\varphi_{lk}^2}{2} + o(\varphi_{lk}^2) + i\left(\varphi_{lk} - \frac{\varphi_{lk}^3}{3!} + o(\varphi_{lk}^3)\right) \\ &\approx 1 - \frac{\varphi_{lk}^2}{2} + i\left(\varphi_{lk} - \frac{\varphi_{lk}^3}{3!}\right). \end{aligned} \tag{18}$$

Combining Eq. (17) and Eq. (18), the signal phase φ_{lk} is calculated by solving the following optimization problem

$$\text{minimize}_{\varphi_{lk}} \left(1 - \varphi_{lk}^2/2 - a\right)^2 + \left(\varphi_{lk} - \varphi_{lk}^3/6 - b\right)^2. \tag{19}$$

Algorithm II: When the antenna array is small, we treat the LoS component as a function of φ_{lk} according to Eq. (11), denoted by $\bar{\mathbf{h}}_{lk}(\varphi_{lk})$. Then the optimal φ_{lk} can be estimated by solving the following MMSE problem

$$\text{minimize}_{\varphi_{lk}} \ \|\sqrt{\bar{\beta}_{lk}}\bar{\mathbf{h}}_{lk}(\varphi_{lk}) - \mathbf{g}_{lk}\|^2. \tag{20}$$

The above two optimization problems can be easily solved. The collection of the calculated side information, i.e., $\boldsymbol{\varphi}_l = [\varphi_{l1}, \ldots, \varphi_{l,K}]$ and $\boldsymbol{\theta}_l = [\theta_{l1}, \ldots, \theta_{l,K}]$, is the compressed CSI of the LoS components $\bar{\mathbf{H}}_l = \left[\bar{\mathbf{h}}_{l1}, \ldots, \bar{\mathbf{h}}_{l,K}\right]$.

4) Compression of NLoS Components: After obtaining the NLoS components, denoted as $\check{\mathbf{H}}_l = \left[\check{\mathbf{h}}_1, \check{\mathbf{h}}_2, \ldots, \check{\mathbf{h}}_K\right]$, according to Eq. (12), the AP l compresses $\check{\mathbf{H}}_l$ using SVD technique. The basic idea is to compress a matrix by extracting its dominant part and discarding the insignificant parts.

Specifically, for the given channel matrix $\check{\mathbf{H}}_l$ of the AP l, its SVD is calculated as

$$\check{\mathbf{H}}_l = \mathbf{U}_l\mathbf{S}_l\mathbf{V}_l^{\mathrm{H}}, \tag{21}$$

where $\mathbf{U}_l = [\mathbf{u}_{l1}, \ldots, \mathbf{u}_{l,R}] \in \mathbb{C}^{N\times R}$ and $\mathbf{V}_l = [\mathbf{v}_{l1}, \ldots, \mathbf{v}_{l,R}] \in \mathbb{C}^{K\times R}$ are semi-unitary matrices that contain the left-singular vectors and the right-singular vectors of $\check{\mathbf{H}}_l$, respectively, with $R = \text{rank}(\check{\mathbf{H}}_l)$, and $\mathbf{S}_l = \text{diag}(s_{l1}, \ldots, s_{l,R}) \in \mathbb{C}^{R\times R}$ is a diagonal matrix whose diagonal elements $s_{li} > 0$ $(1 \le i \le R)$ are the singular values of $\check{\mathbf{H}}_l$ and sorted in descending order. Since the top largest singular values contain the most information, the rest can be discarded for the purpose of compression. That being said, the AP l is able to keep the most significant information of $\check{\mathbf{H}}_l$ by choosing the first $M(M \le R)$ singular values $\tilde{\mathbf{S}}_l = \text{diag}(s_{l1}, \ldots, s_{l,M})$ and the corresponding $\tilde{\mathbf{U}}_l = [\mathbf{u}_{l1}, \ldots, \mathbf{u}_{l,M}]$ and $\tilde{\mathbf{V}}_l = [\mathbf{v}_{l1}, \ldots, \mathbf{v}_{l,M}]$ to be the compressed CSI of the NLoS component. Together with the compressed CSI of $\bar{\mathbf{H}}_l$, the total volume of the uploaded data is $2M(N+K+1)+2K$. Compared with uploading $\mathbf{G}_l$ with the volume of $2\times N\times K$, our method is able to save the bandwidth of the fronthaul links when $M \le \lfloor\frac{(N-1)K}{N+K+1}\rfloor$, with a compression ratio of $\frac{M(N+K+1)+K}{NK}$.

B. CPU-Side Operation

In the CPU-side operation, the CPU recovers the original local channel matrix based on the received CSI from each AP. After that, the CPU implements global ZF precoding and sends the result back to all APs.

1) CSI Recovery: Upon receiving the compressed CSI uploaded from all APs, the CPU recovers the whole channel matrix $\mathbf{G}$ in the following steps.

Step 1: $\check{\mathbf{H}}_l$ recovery. Based on $\tilde{\mathbf{S}}_l$, $\tilde{\mathbf{U}}_l$ and $\tilde{\mathbf{V}}_l$, $\check{\mathbf{H}}_l$ can be recovered as

$$\tilde{\mathbf{H}}_l = \tilde{\mathbf{U}}_l\tilde{\mathbf{S}}_l\tilde{\mathbf{V}}_l^{\mathrm{H}}. \tag{22}$$

Step 2: $\bar{\mathbf{H}}_l$ recovery. Based on φ_{lk} and θ_{lk}, $\bar{\mathbf{h}}_l$ can be calculated according to Eq. (11). Then the collection of $\bar{\mathbf{h}}_{lk}$, $k \in \mathcal{K}$, is $\bar{\mathbf{H}}_l = \left[\bar{\mathbf{h}}_{l1}, \ldots, \bar{\mathbf{h}}_{l,K}\right]$.

Step 3: $\mathbf{G}_l$ recovery. Upon obtaining $\bar{\mathbf{h}}_{lk}$ and $\tilde{\mathbf{h}}_{lk}$ from $\bar{\mathbf{H}}_l$ and $\tilde{\mathbf{H}}_l$, respectively, the channel vector $\mathbf{g}_{lk}$ can be recovered as

$$\tilde{\mathbf{g}}_{lk} = \sqrt{\bar{\beta}_{lk}}\bar{\mathbf{h}}_{lk} + \sqrt{\check{\beta}_{lk}}\tilde{\mathbf{h}}_{lk}. \tag{23}$$

And $\tilde{\mathbf{G}}_l = [\tilde{\mathbf{g}}_{l1}, \ldots, \tilde{\mathbf{g}}_{l,K}]$ is used as the recovery of $\mathbf{G}_l$.

Step 4: $\mathbf{G}$ recovery. Collecting all $\mathbf{G}_l$, $l \in \mathcal{L}$, we get the recovery of $\mathbf{G}$ as $\tilde{\mathbf{G}} = \left[\tilde{\mathbf{G}}_1, \ldots, \tilde{\mathbf{G}}_L\right]^{\mathrm{T}}$.

2) GZF Precoding: After obtaining $\tilde{\mathbf{G}}$, the CPU computes the precoding matrix $\mathbf{W}$ according to Eq. (4).

3) Power Allocation: With the estimated channel matrix $\tilde{\mathbf{G}}$ and the precoding matrix $\mathbf{W}$, the CPU sets the transmit power for each UE. The optimal transmit power allocation can be determined by solving the following optimization problem

$$\begin{aligned} \text{maximize}_{\boldsymbol{\eta}} \ & \sum_{k\in\mathcal{K}} \log_2\left(1 + SINR_k^{\mathrm{G}}\right) \\ \text{s.t. } & \text{Eq.}(7), \forall k \in \mathcal{K}, \\ & \sum_{k\in\mathcal{K}(l)} \eta_k \le P_l, \forall l \in \mathcal{L}, \\ & \eta_k \ge 0, \forall k \in \mathcal{K}, \end{aligned} \tag{24}$$

where $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_K]$ and P_l is the maximum power that the AP l can provide. The power constraints require that the total transmitted power from an AP to its served UEs can not exceed its maximum power. The above power allocation problem is modeled as a non-linear programming. We can solve it by using the interior-point algorithm offered by the Matlab toolbox.

V. Simulations and Analysis

In this section, the performances of the proposed GZF precoding strategy in various scenarios are evaluated. We compare our strategy with three counterparts: GZF precoding with fully uploading (GZF-FU), GZF precoding with preliminary SVD (GZF-PSVD) and LZF precoding.

The communication process introduced in Sec. III-C also applies to both the GZF-FU and the GZF-PSVD strategies. However, different from our strategy, in the GZF-FU strategy, each AP uploads its original local channel matrix; and in the GZF-PSVD strategy, the SVD is conducted directly on the original local channel matrix without channel decomposition as in our strategy.

In the LZF precoding strategy, all APs implement downlink precoding independently basd on their own local channel matrices without any information exchanging with the CPU or

other APs. Specifically, in the LZF precoding, the AP $l \in \mathcal{L}$ computes its normalized precoding vector as

$$\mathbf{w}_{lk} = \frac{\mathbf{f}_{lk}}{\|\mathbf{f}_{lk}\|_F^2}, \tag{25}$$

where $\mathbf{F}_l = [\mathbf{f}_{l1}, \ldots, \mathbf{f}_{lK}] = \mathbf{G}_l / \left(\mathbf{G}_l^{\mathrm{H}} \mathbf{G}_l\right)$. The received signal at the UE k is

$$y_k = \sum_{l \in \mathcal{L}(k)} \mathbf{g}_{lk}^{\mathrm{H}} \sqrt{\eta_{lt}} \mathbf{w}_{lk} q_k + \sum_{\substack{t \in \mathcal{K} \\ t \neq k}} \sum_{l \in \mathcal{L}(t)} \mathbf{g}_{lk}^{\mathrm{H}} \sqrt{\eta_{lt}} \mathbf{w}_{lt} q_t + n_k, \tag{26}$$

where $\mathbf{g}_{lk} \in \mathbb{C}^{N \times 1}$ indicates the channel vector between the AP l and the UE k, η_{lt} denotes the power allocated to the UE k by the AP l, and $\mathbf{w}_{lk} \in \mathbb{C}^{N \times 1}$ is the corresponding precoding vector. The SINR at the UE k is calculated as

$$SINR_k^{\mathrm{L}} = \frac{\left| \sum_{l \in \mathcal{L}(k)} \mathbf{g}_{lk}^{\mathrm{H}} \sqrt{\eta_{lt}} \mathbf{w}_{lk} \right|^2}{\sum_{\substack{t \in \mathcal{K} \\ t \neq k}} \left| \sum_{l \in \mathcal{L}(t)} \mathbf{g}_{lk}^{\mathrm{H}} \sqrt{\eta_{lt}} \mathbf{w}_{lt} \right|^2 + \sigma_{\mathrm{N}}^2}, \tag{27}$$

and the achievable data rate is

$$r_k^{\mathrm{L}} = \log_2 \left(1 + SINR_k^{\mathrm{L}}\right). \tag{28}$$

Moreover, the AP l allocates its power to its served UEs by solving the following optimization problem

$$\begin{aligned} \text{maximize}_{\boldsymbol{\eta}_l} \quad & \sum_{k \in \mathcal{K}(l)} \log_2 \left(1 + SINR_k^{\mathrm{L}}\right) \\ \text{s.t. Eq.}(27), & \forall k \in \mathcal{K}(l), \\ & \sum_{k \in \mathcal{K}(l)} \eta_{lk} \leq P_l, \\ & \eta_{lk} \geq 0, \forall k \in \mathcal{K}(l), \end{aligned} \tag{29}$$

where $\boldsymbol{\eta}_l = [\eta_{l1}, \ldots, \eta_{lK}]$.

A. Simulation Setup

In our simulations, we consider a cell-free massive MIMO system consisting of 25 APs (i.e., $L = 25$), each of which is equipped with 20 antennas (i.e., $N = 20$), and 16 single-antenna UEs (i.e., $K = 16$). The APs and the UEs are randomly distributed in a 100(m) $\times$ 100(m) square area. The UEs are moving in the area with random directions at the speed of 1(m) per iteration. In all of our simulations, we consider 10 iterations. The AP's serving range is set to be 50(m) (i.e., $\bar{d} = 50$) and the maximum transmit power is set to be 1. The noise power is $\sigma_{\mathrm{N}}^2 = -92$ dBm [27]. For Eq. (2), we consider the conventional 1900 MHz frequency band and set $\alpha = -30.18$, $\rho = -2.6$ and $\sigma_{\mathrm{S}} = 4$ [5]. The K-factor in Eq. (1) is modeled as [5]

$$K_{lk} = 10^{1.3 - 0.003 d_{lk}} [dB], \tag{30}$$

where d_{lk} is the distance between the AP l and the UE k. The spatial correlation of the NLoS components $\mathbf{R}_{lk}$ is calculated according to [5], and the details are omitted in this paper. Given the φ_{lk} and θ_{lk}, $l \in \mathcal{L}$ and $k \in \mathcal{K}$ that could be estimated by Sec. IV-A3, we implement the simulations as follows.

B. Compression Ratio

Since the number of chosen singular values determines the compression ratio and impacts the CSI recovery accuracy, here we evaluate these strategies under different values of M. The results of the spectral efficiencies and the overhead are shown in Fig. 4 and Fig. 5, respectively.

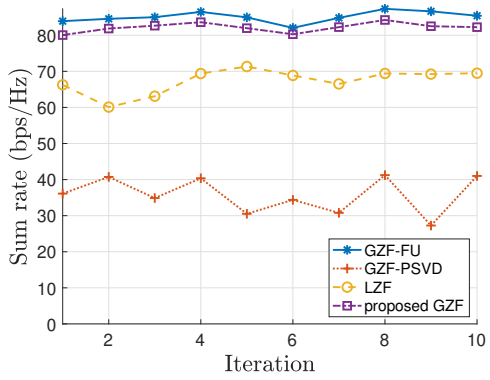
In Fig. 4, the spectral efficiencies of the considered strategies are compared under $M = 1$, 2 and 3. It is illustrated that, the GZF-FU precoding outperforms the LZF precoding since the CPU has more comprehensive CSI knowledge and is able to make a better precoding decision to suppress interference. Although the GZF-PSVD is also a global precoding strategy, its performance is even worse than the LZF's performance when $M = 1$, as shown in Fig. 4(a). The reason is that when only one singular value is used, very limited information is contained in the compressed CSI, which leads to low CSI recovery accuracy. Compared with the GZF-PSVD, the proposed GZF precoding strategy significantly increases the spectral efficiency by 134.4% on average. Besides, our strategy is slightly suboptimal, with an average gap of only 3.5% relative to the GZF-FU strategy. It is also observed in Fig. 5 that our strategy drastically cuts the overhead by 83.4% compared with the GZF-FU. That being said, our proposed precoding strategy can provide slightly suboptimal performance with much less overhead and hence can reach outstanding bandwidth efficiency.

Moreover, the spectral efficiencies of both the GZF-PSVD strategy and our strategy increase with the number of chosen singular values, with the increasing rate of the former being much larger than that of the later. The reason is that, in the GZF-PSVD, the more singular values are chosen, the more information is kept in the compressed CSI. Therefore, the CPU can recover the channel matrix with higher accuracy and consequently make a better precoding decision. However, in our strategy, since the compressed CSI only contains the NLoS components that contribute little signal power, choosing more singular values does not bring obvious performance improvement as in the GZF-PSVD.

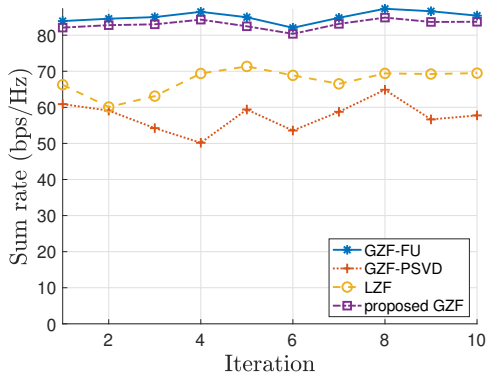
But there is no free lunch. For the GZF-PSVD, its performance is improved at the cost of significantly increased overhead. As displayed in Fig. 5, the overhead when $M = 2$ and $M = 3$ is two-fold and three-fold of that when $M = 1$, respectively. Moreover, this overhead increase could not improve the performance proportionally. By contrast, the performance of our strategy when $M = 1$ is even higher than that of the GZF-PSVD when $M = 3$, but with much less overhead. This indicates that our strategy is able to achieve a suboptimal performance with substantially low overhead.

C. K-Factor

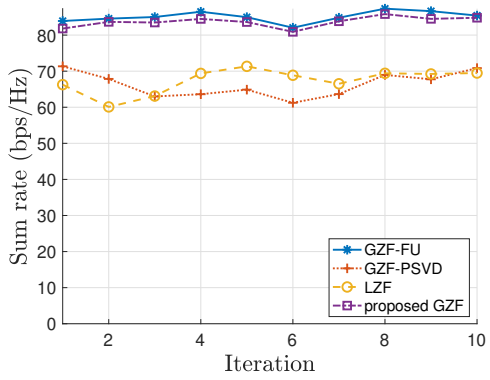
The K-factor indicates the degree of dominance of the LoS component over the NLoS component. A larger K-factor means that the LoS component contributes more power. If the K-factor equals to 0, there is no direct path between the transmitter and the receiver, meaning that the Rician fading channel decays to a Rayleigh fading channel. We also evaluate



(a) $M = 1$



(b) $M = 2$



(c) $M = 3$

Fig. 4. Comparison of spectral efficiency with different values of M

the candidate strategies in the scenario of Rayleigh fading channels. The simulation results when $M = 1$ are shown in Fig. 6. It is demonstrated that the performance of the proposed precoding strategy is identical to that of the GZF-PSVD strategy. The reason is that there is no LoS component in this case. That means, in Rayleigh fading channels, our strategy degrades to the GZF-PSVD whose performance is the worst. Therefore, our strategy is not suitable for in Rayleigh fading channels.

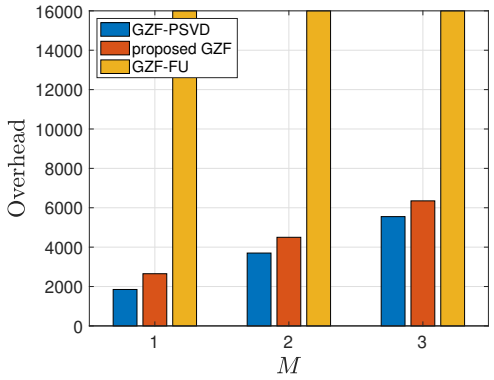


Fig. 5. Comparison of overhead with different values of M

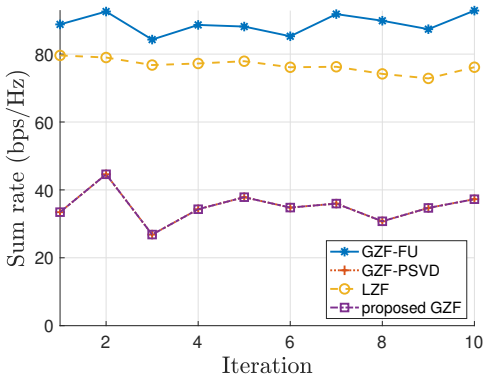


Fig. 6. Comparison of spectral efficiency when the K-factor = 0

D. Frequency Band

In this section, we evaluate the impact of the frequency band on the performance of the proposed precoding strategy. We choose a different set of path loss parameters, i.e., $\alpha = -61.4$, $\rho = -2$ and $\sigma_{\mathrm{S}} = 5.8$ for mmWave 28 GHz frequency band [24], and set $M = 1$. The simulation results are shown in Fig. 7. Comparing Fig. 4(a) with Fig. 7, we can find that, in the conventional band, our strategy achieves 96.5% of the spectral efficiency of the GZF-FU strategy on average; and in the mmWave band, this proportion is 96.7%, showing that the proposed strategy has similar performances in these two frequency bands and performs well in both scenarios. This indicates a great flexibility and reliability of our strategy when being applied to various application scenarios with different frequency bands.

VI. Conclusions

In this paper, we propose a bandwidth-efficient GZF precoding strategy for cell-free massive MIMO systems considering Rician fading. In order to reduce the CSI overhead on fronthaul links, we propose to decompose the channel matrix into two components and design two corresponding efficient compression methods. In addition, two effective algorithms have been developed to calculate the necessary information

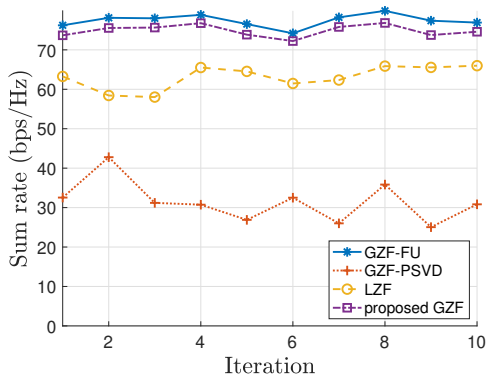


Fig. 7. Comparison of spectral efficiency in 28 GHz frequency band

that is required for the proposed CSI compression mechanism. The simulation results have demonstrated that our strategy can significantly reduce the upload overhead while achieving comparable performance to the GZF without CSI compression.

VII. Acknowledgement

This work is supported in part by NSF under grants CNS-2006998, CNS-1837034, and CNS-1745254. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of NSF.

References

[1] E. Björnson, E. G. Larsson, and M. Debbah, "Massive mimo for maximal spectral efficiency: How many users and pilots should be allocated?," *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1293–1308, 2015.

[2] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive mimo," *IEEE Transactions on Wireless Communications*, 2020.

[3] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445–4459, 2017.

[4] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy efficiency in cell-free massive mimo with zero-forcing precoding design," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1871–1874, 2017.

[5] Ö. Özdogan, E. Björnson, and E. G. Larsson, "Massive mimo with spatially correlated rician fading channels," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3234–3250, 2019.

[6] J. C. Roh and B. D. Rao, "Efficient feedback methods for mimo channels based on parameterization," *IEEE Transactions on Wireless Communications*, vol. 6, no. 1, pp. 282–292, 2007.

[7] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive mimo systems with arbitrary-rank channel means," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 966–981, 2014.

[8] J. Kang, O. Simeone, J. Kang, and S. S. Shitz, "Joint signal and channel state information compression for the backhaul of uplink network mimo systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1555–1567, 2014.

[9] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for c-rans over ergodic fading channels," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5022–5032, 2015.

[10] T. X. Vu, H. D. Nguyen, and T. Q. Quek, "Adaptive compression and joint detection for fronthaul uplinks in cloud radio access networks," *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 4565–4575, 2015.

[11] P.-H. Kuo, H. Kung, and P.-A. Ting, "Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 492–497, IEEE, 2012.

[12] M. E. Eltayeb, T. Y. Al-Naffouri, and H. R. Bahrami, "Compressive sensing for feedback reduction in mimo broadcast channels," *IEEE Transactions on communications*, vol. 62, no. 9, pp. 3209–3222, 2014.

[13] M. S. Sim, J. Park, C.-B. Chae, and R. W. Heath, "Compressed channel feedback for correlated massive mimo systems," *Journal of Communications and Networks*, vol. 18, no. 1, pp. 95–104, 2016.

[14] A. Ge, T. Zhang, Z. Hu, and Z. Zeng, "Principal component analysis based limited feedback scheme for massive mimo systems," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 326–331, IEEE, 2015.

[15] T. Zhang, A. Ge, N. C. Beaulieu, Z. Hu, and J. Loo, "A limited feedback scheme for massive mimo systems based on principal component analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–12, 2016.

[16] J. Joung, E. Kurniawan, and S. Sun, "Channel correlation modeling and its application to massive mimo channel feedback reduction," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 3787–3797, 2016.

[17] D. Gündüz, P. de Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. van der Schaar, "Machine learning in the air," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2184–2199, 2019.

[18] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive mimo csi feedback," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 748–751, 2018.

[19] Y. Liao, H. Yao, Y. Hua, and C. Li, "Csi feedback based on deep learning for massive mimo systems," *IEEE Access*, vol. 7, pp. 86810–86820, 2019.

[20] T. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based csi feedback approach for time-varying massive mimo channels," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 416–419, 2018.

[21] C. Lu, W. Xu, H. Shen, J. Zhu, and K. Wang, "Mimo channel information feedback using deep recurrent network," *IEEE Communications Letters*, vol. 23, no. 1, pp. 188–191, 2018.

[22] Q. Yang, M. B. Mashhadi, and D. Gündüz, "Deep convolutional compression for massive mimo csi feedback," in *2019 IEEE 29th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6, IEEE, 2019.

[23] Z. Liu, L. Zhang, and Z. Ding, "Exploiting bi-directional channel reciprocity in deep learning for low rate massive mimo csi feedback," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 889–892, 2019.

[24] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE journal on selected areas in communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[25] 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, 2019.

[26] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive mimo networks: Spectral, energy, and hardware efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[27] E. Björnson and L. Sanguinetti, "Making cell-free massive mimo competitive with mmse processing and centralized implementation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77–90, 2019.