Cell-Free Massive MIMO: Joint Maximum-Ratio and Zero-Forcing Precoder with Power Control

Liutong Du, *Student Member, IEEE*, Lihua Li, *Member, IEEE*, Hien Quoc Ngo, *Senior Member, IEEE*, Trang C. Mai, *Member, IEEE* and Michail Matthaiou, *Senior Member, IEEE*

***Abstract*—Cell-free massive multiple-input multiple-output (MIMO) system is a promising architecture for next generation wireless systems by deploying a very large number of distributed access points (APs), which simultaneously serve a smaller number of user equipments (UEs) over the same time-frequency resources. It guarantees uniformly good service at high spectral efficiency with simple linear precoding techniques and max-min power control. In this paper, we propose a new joint maximum-ratio and zero-forcing (JMRZF) precoding scheme, where part of APs are combined to perform centralized zero-forcing (ZF), while other APs apply simple maximum-ratio transmission (MRT). Our proposed precoder offers an adaptable trade-off between the spectral efficiency and front-haul signalling overhead. A corresponding AP subset selection scheme is also proposed which is based on large-scale fading coefficients. A closed-form expression for the achievable spectral efficiency of our proposed scheme is derived, which represents a generalized result including both fully distributed MRT and fully centralized ZF cases. Based on this closed-form expression, max-min power control is formulated and solved via the second order cone and first order methods. The former can obtain the global optimal solution, but its computational complexity is very high. On the other hand, the latter technique is sub-optimal, yet, it has very low computational complexity. Hence, it is suitable for large-scale cell-free massive MIMO systems with hundreds or thousands of APs and users. Numerical results show that our proposed JMRZF scheme can substantially outperform the local precoding schemes, even when a small part of APs are combined to deploy ZF and is implementable even when each AP has very few antennas. In addition, it is shown that our max-min power controls improves the spectral efficiency significantly, compared to the uniform power control scheme.**

***Index Terms*—Cell-free massive MIMO, maximum-ratio transmission, max-min fairness power control, zero-forcing.**

I. Introduction

MASSIVE multiple-input multiple-output (MIMO) is a system, where a large number of antennas are deployed at the base station (BS) to simultaneously serve many user equipments (UEs) in the same time-frequency resource. Since massive MIMO can offer high throughput, reliability, and energy efficiency with simple signal processing [1], it has been included as a core technology in 5G new radio (NR) standard. However, the performance of the current mobile networks is still limited by the inherent inter-cell interference due to the cellular structure design, especially for UEs close to cell boundaries.

Cell-free massive MIMO, which can be seen as a scalable implementation of network MIMO and distributed antenna systems (DAS) concepts [2]–[4], is a promising physical layer technology for next generation wireless systems [5], which is expected to harness massive MIMO and mitigate inter-cell interference at the same time [6]. In cell-free massive MIMO, a large number of access points (APs) are distributed over a wide geographic area to simultaneously serve all the UEs via time-division duplex (TDD) operation. Each AP is connected to one or several central processing units (CPUs) through a front-haul network, while the CPUs are connected via a back-haul network. Within such configuration, UEs can get closer to APs, and hence, avail of a high degree of macro-diversity and low path losses. As a result, many UEs can be served simultaneously with uniformly good quality-of-service (QoS), which means cell-free massive MIMO can offer a higher coverage probability compared with co-located massive MIMO [6], [7]. In canonical cell-free massive MIMO, the key feature that makes it so attractive, is that channel estimation and precoding are performed locally at each AP by leveraging the channel reciprocity in TDD. Hence, front-hauling is greatly reduced since there is no instantaneous channel state information (CSI) sharing [8].

A. Motivation

There has been a broad amount of work on deriving closed-form expressions for the ergodic spectral efficiency (SE) and corresponding optimal power allocation scheme [9] for both uplink (UL) and downlink (DL) cell-free massive MIMO systems, when maximum-ratio combining/transmission (MRC/MRT) are considered to sustain system scalability [6]. The work of [10] further analyzed the performance of cell-free massive MIMO when centralized zero-forcing (CZF) is applied. However, all these works considered single antenna APs. In [11], the authors investigated the energy efficiency (EE) of cell-free massive MIMO systems when conjugate beamforming is applied on multiple-antenna APs. The more sophisticated local full-pilot ZF (FZF) precoding schemes were further considered in [12] to reduce the performance

L. Du and L. Li are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, P. R. China, (email: {liutongdu, lilihua}@bupt.edu.cn.

H. Q. Ngo, T. C. Mai and M. Matthaiou are with the Institute of Electronics, Communications and Information Technology (ECIT), Queen's University Belfast, BT3 9DT, Belfast, U.K. (e-mail: {hien.ngo, trang.mai, m.matthaiou}@qub.ac.uk).

The work of L.Li and L.Du was supported by China National Key Research and Development Plan under grant 2018YFE0205501. The work of L.Du was partly supported by China Scholarship Council under grant 201906470068. The work of H. Q. Ngo and T. C. Mai was supported by the U.K. Research and Innovation Future Leaders Fellowships under Grant MR/S017666/1. The work of M. Matthaiou was supported by the EPSRC, U.K., under Grant EP/P000673/1.

gap between MRT and ZF. The performance of local ZF depends strongly on the number of AP antennas, M, and the number of mutually orthogonal pilots, τ_{P}, since the available spatial degrees of freedom (DoFs) to mitigate interference is determined by $M - \tau_{\mathrm{P}}$.

For cell-free massive MIMO, where the APs have a small number of antenna elements, local *protective partial zero-forcing* (PPZF) was proposed in [12], in which each AP suppresses only the interference it causes to part of UEs, and the interference between the other UEs is tolerated. However, the performance of local ZF is mainly limited by two aspects: 1) since local ZF only suppress the self-interference of each AP, its performance is limited by the inter-AP interference; 2) as shown in our simulation results, even when the most advanced local ZF in [12] is applied, the DL SE is still limited by the number of antennas on each AP M, i.e., at most M UEs can avoid pilot contamination and inter-AP interference at the same time.

In this paper, we propose a joint maximum-ratio and zero-forcing (JMRZF) scheme which can fully avoid such limitations and provides an adaptable trade-off between interference cancellation and front-haul signalling overhead by performing MRT and ZF partially among APs.

B. Contributions

The main technical contributions of the paper can be summarized as follows

- We propose a joint maximum-ratio and zero-forcing-based precoding scheme, referred to as JMRZF, where a subset of APs uses the MR technique while the remaining APs employ the ZF technique to precode the symbols sent to all users. An AP selection scheme is proposed to select these subsets. Our proposed scheme can achieve a SE close to that offered by centralized ZF with considerable reduction in the front-haul overhead.
- We derive an achievable DL SE, under the assumption of independent Rayleigh fading, considering channel estimation errors and the pilot contamination effect.
- We propose an algorithm to solve the max-min fairness power control optimization problem, subject to per-AP power constraints, based on the second order cone (SOC) program.
- We propose to use the first order method to further boost the speed of solving the power control optimization problem.
- We compare the performance of the proposed scheme against that of the state-of-art precoding schemes, including MRT [6], PPZF [12] and CZF [10], with max-min fairness power control.

The rest of paper is organized as follows Section II describes the system model of cell-free massive MIMO. Section III introduces the proposed JMRZF precoder with a new AP selection scheme and the derivation of the corresponding SE. Section IV and V describe the max-min power control policy with SOC and first order method, respectively. The front-haul requirements are discussed in Section VI. Finally, numerical results are presented in Section VII, while Section VIII concludes the paper.

TABLE I: Notations and variables

$\beta_{l,k}$	large-scale fading coefficients for AP l and UE k
$\boldsymbol{\phi}_{\mathrm{i}_k}$	pilot assigned to UE k
$\tau_{\mathrm{P}}, \tau_{\mathrm{U}}$ and τ_{D}	number of channel uses for pilots, UL data and DL data for each coherence interval
$\gamma_{l,k}$	variance of channel estimation for AP l and UE k
p_k	UL normalized transmit power
$\rho_l^{\max}$	DL per-AP power limit
$\lambda_{l,k}$	normalization factor for precoding vector between AP l and UE k
$\rho_{l,k}$	power coefficient at AP l assigned for UE k

Notation: Boldface lower and upper case letters denote vectors and matrices, respectively. $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$ stand for conjugate, transpose and conjugate-transpose, respectively; $\mathbf{I}_M$ stands for the identity matrix of size $M \times M$, whilst $\otimes$, $\circ$ and $\oslash$ are the notations for the Kronecker product, Hadamard (element-wise) product and Hadamard division, respectively. The Euclidean norm and the expectation operators are denoted by $\|\cdot\|$ and $\mathsf{E}\{\cdot\}$, respectively. Notation $\equiv$ means "is identically equal to" and $\cong$ stands for "is congruent to". We use $[\cdot]_+$ to denote the projection onto the positive orthant. Finally, $z \sim \mathcal{CN}(\mathbf{0}, \sigma^2)$ denotes a circularly symmetric complex Gaussian random variable (RV) z with zero mean and variance σ^2, $\mathcal{O}(\cdot)$ represents the big-O notation.

II. System model

We consider a cell-free massive MIMO system, where K single-antenna UEs are jointly served by L randomly deployed APs with M antennas each, such that $LM > K$. The APs are connected to the CPU via a front-haul link for exchanging the network information, i.e., channel estimates, precoding vectors, and power coefficients. As in [13], we adopt the standard block fading channel model, where the channels are invariant and frequency-flat within a coherence interval.

Suppose that $\mathbf{h}_{l,k}$ is the channel between AP l and UE k. We consider the following standard channel model as in [6]: $\mathbf{h}_{l,k} \sim \mathcal{CN}(\mathbf{0}, \beta_{l,k}\mathbf{I}_M)$, where $\beta_{l,k}$ represents the large-scale fading parameter, which changes slowly with time [14] and is assumed known a-priori at each AP. We assume the channel is reciprocal by perfect calibration of the hardware chains as in [15].

Let τ_{C} denote the length of the TDD frame, which is determined by the shortest coherence interval of all UEs in the network. Each TDD frame is divided into three phases: UL pilot transmission (or UL training), UL data transmission, and DL data transmission. We use τ_{P} to denote the channel use for UL pilots, τ_{U} for UL data, and τ_{D} for DL data. Infinite front-haul network capacity is considered in this paper, as the performance with front-haul capacity constraints was well investigated in [16] and [17]. The main parameters throughout this paper are summarized in Table I on the top of this page.

A. UL Training

Let $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{\tau_\mathrm{P}}$, where $\|\boldsymbol{\phi}_k\|^2 = \tau_\mathrm{P}$, be the set of orthogonal pilot sequences used for the UL training. We denote the index of the pilot assigned to UE k as $\mathrm{i}_k \in \{1, \ldots, \tau_\mathrm{P}\}$ and use the notation $\mathcal{P}_k \subset \{1, \ldots, K\}$ to denote the subset of UEs that use the same pilot as UE k, including k. The pilot signal sent by UE k is $\sqrt{p_k}\boldsymbol{\phi}_{\mathrm{i}_k}$, where p_k is the UL normalized transmit power. We assume all UEs transmit with full power. The pilot signal received at AP l is

$$\mathbf{Y}_l = \sum\nolimits_{k=1}^{K} \mathbf{h}_{l,k}\sqrt{p_k}\boldsymbol{\phi}_{\mathrm{i}_k}^{\mathrm{H}} + \mathbf{N}_l \in \mathbb{C}^{M\times\tau_\mathrm{P}}, \tag{1}$$

where $\mathbf{N}_l \in \mathbb{C}^{M\times\tau_\mathrm{P}}$ is a Gaussian noise matrix whose elements are i.i.d. $\mathcal{CN}(0,1)$. Considering the standard minimum mean square error (MMSE) estimation [18], the channel estimate of $\mathbf{h}_{l,k}$ can be derived as

$$\hat{\mathbf{h}}_{l,k} \triangleq c_{l,k}\mathbf{Y}_l\boldsymbol{\phi}_{\mathrm{i}_k}, \tag{2}$$

where

$$c_{l,k} \triangleq \frac{\sqrt{\tau_\mathrm{P}p_k}\beta_{l,k}}{\tau_\mathrm{P}\sum_{t\in\mathcal{P}_k} p_t\beta_{l,t}+1}. \tag{3}$$

The estimation error is given by $\tilde{\mathbf{h}}_{l,k} = \mathbf{h}_{l,k} - \hat{\mathbf{h}}_{l,k}$. The estimate and estimation error are independent and distributed as $\hat{\mathbf{h}}_{l,k} \sim \mathcal{CN}(\mathbf{0}, \gamma_{l,k}\mathbf{I}_M)$ and $\tilde{\mathbf{h}}_{l,k} \sim \mathcal{CN}(\mathbf{0}, (\beta_{l,k}-\gamma_{l,k})\mathbf{I}_M)$, respectively, where

$$\gamma_{l,k} = \frac{\tau_\mathrm{P}p_k\beta_{l,k}^2}{\tau_\mathrm{P}\sum_{t\in\mathcal{P}_k} p_t\beta_{l,t}+1}. \tag{4}$$

Owing to the limited length of the coherence interval, in general, *pilot contamination*[1] will appear when $\tau_\mathrm{P} < K$, for any pair of UEs k and t assigned the same pilot. The respective channel estimates to AP l are linearly dependent as

$$\hat{\mathbf{h}}_{l,k} = \frac{\sqrt{p_k}\beta_{l,k}}{\sqrt{p_t}\beta_{l,t}}\hat{\mathbf{h}}_{l,t}. \tag{5}$$

B. DL data transmission

The channel estimates from UL training are used from the APs to generate precoding vectors for DL transmission. The data signal transmitted from AP l to all the UEs is given by

$$\mathbf{x}_l = \sum_{k=1}^{K}\sqrt{\rho_{l,k}}\mathbf{w}_{l,k}q_k, \tag{6}$$

where $\mathbf{w}_{l,k} \in \mathbb{C}^{M\times 1}$ is the precoding vector used to form the signal from AP l towards UE k, with $\mathsf{E}\left\{\|\mathbf{w}_{l,k}\|^2\right\} = 1$, and $\rho_{l,k}$ is the normalized transmit power, satisfying a per-AP power constraint. The data symbol q_k has unit power as $\mathsf{E}\left\{|q_k|^2\right\} = 1$, and zero mean. In addition, we assume that the data symbols of different UEs are independent, i.e., $\mathsf{E}\{q_k q_t^*\} = 0$ for any $t \neq k$.

[1]Pilot contamination has received a lot of research interest in the literature [19], [20] since it can substantially compromise the system performance of cell-free massive MIMO. In this work, we do not aim to design a robust resource allocation scheme against this pilot contamination. Such design requires a comprehensive study, and hence, is left for future work.

At UE k, the received data signal can be denoted as

$$y_k = \sum_{l=1}^{L}\sqrt{\rho_{l,k}}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{w}_{l,k}q_k + \sum_{l=1}^{L}\sum_{t\neq k}^{K}\sqrt{\rho_{l,t}}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{w}_{l,t}q_t + n_k, \tag{7}$$

where the first term is the desired signal, the second term describes the multi-user interference, and the third term is the Gaussian noise at the receiver, $n_k \sim \mathcal{CN}(0,1)$.

C. Distributed precoding vs centralized precoding

Coherent joint transmission has higher front-haul requirements than non-coherent transmission as the former requires not only phase-synchronization between the APs, but also information exchange between the APs and the CPUs. To reduce the amount of overhead exchanged over the front-haul network, precoders can be designed with local CSI in a distributed manner by exploiting the channel reciprocity in TDD mode. Distributed signal processing also enables the scalability of the system. When $\tau_\mathrm{P} < K$, the channel estimates of UEs assigned with the same pilot are parallel. Thus, for notation simplicity, we introduce the corresponding full-rank matrix of channel estimates at AP l, that only contains channel direction information (CDI), as follows

$$\bar{\mathbf{H}}_l = \mathbf{Y}_l\boldsymbol{\phi} \in \mathbb{C}^{M\times\tau_\mathrm{P}}, \tag{8}$$

where $\boldsymbol{\phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{\tau_\mathrm{P}}]$ is the pilot codebook matrix. Local precoding vectors (e.g., local MRT [6], local ZF [12]) can be constructed using the full-rank channel estimates (8) from each AP (see Section III-A). To crystallize that the precoding vector for each UE is determined only by the assigned pilot, we denote the precoding vector from AP l towards UEs in $\mathcal{P}_k$ as $\mathbf{w}_{l,\mathrm{i}_k}$.

Remark 1: For APs whose number of antennas, $M < \tau_\mathrm{P}$, the local ZF schemes in [12] can only mitigate the interference among the channels estimated with at most M pilots. Under these conditions, the system has to choose between less efficient interference cancellation (less UEs with ZF) and more pilot contamination (more UEs share the same pilot), thereby experiencing a performance loss.

For centralized precoding, CPU can obtain the estimated channel matrix from each AP as

$$\hat{\mathbf{H}}_C = \left[\hat{\mathbf{H}}_1^{\mathrm{H}}, \ldots, \hat{\mathbf{H}}_L^{\mathrm{H}}\right]^{\mathrm{H}} \in \mathbb{C}^{LM\times K}, \tag{9}$$

and more specifically, $\hat{\mathbf{H}}_C^{\mathrm{H}} \in \mathbb{C}^{K\times LM}$ is assumed to have a right inverse to enable centralized ZF, which means that for the general case of cell-free massive MIMO $LM > K$, such that the matrix $\hat{\mathbf{H}}_C^{\mathrm{H}}$ is full row rank to make sure $\hat{\mathbf{H}}_C\left(\hat{\mathbf{H}}_C^{\mathrm{H}}\hat{\mathbf{H}}_C\right)^{-1}$ exists. However, when pilot reuse is considered, i.e., $\tau_\mathrm{P} < K$, the rank of the channel matrix estimated at AP l, denoted as $\hat{\mathbf{H}}_l^{\mathrm{H}} \in \mathbb{C}^{K\times M}$, is determined by the smaller value of K and τ_P, i.e., $r = \min\{K, \tau_\mathrm{P}\}$. As the APs are not assumed co-located, the number of APs L must fulfill $L \geq r/M$ for centralized zero-forcing.

Remark 2: For centralized zero-forcing precoding, it is essential that $L \geq r/M$, which can always be satisfied with cell-free massive MIMO. It can be seen that centralized ZF is

more suitable for APs with few antennas, as it can effectively avoid pilot contamination compared with local ZF, at the expense of additional front-haul overhead.

III. Joint Maximum-Ratio and Zero-Forcing Precoding and SE

In this section, we first propose a simple JMRZF precoding technique together with an efficient AP selection scheme. Then, we derive the SE of the proposed JMRZF in closed-form.

A. Proposed joint maximum-ratio and zero-forcing precoding

This section describes our proposed precoding scheme named JMRZF. The principle of JMRZF is that only a subset of APs are connected to the CPU via the front-haul link to implement the centralized ZF precoding, while the remaining APs perform MRT with local CSI to reduce the signalling overhead.

Let $\mathcal{S} \subset \{1,\ldots,L\}$ be the set of indices of APs which use ZF, and $\mathcal{W} \subset \{1,\ldots,L\}$ be the set of indices of APs which use MRT. Note that $\mathcal{S} \cap \mathcal{W} = \emptyset$ and $|\mathcal{S}| + |\mathcal{W}| = L$; we use S_i to denote the i-th AP in set $\mathcal{S}$ and further define $C_l \triangleq |\mathcal{S}|$ and $W_l \triangleq |\mathcal{W}|$ to represent the number of APs in the corresponding set, while C_l satisfies $L \geq C_l \geq r/M$ to make sure the estimated channel matrix of set $\mathcal{S}$ is full-rank.

As noted in Remark 2, with $C_l \geq r/M$, the CPU can get the full-rank channel estimates $\hat{\mathbf{H}}_S$ by concatenating the corresponding channel estimates of corresponding APs in set $\mathcal{S}$ as

$$\hat{\mathbf{H}}_S = \left[\hat{\mathbf{H}}_{S_1}^{\mathrm{H}},\ldots,\hat{\mathbf{H}}_{S_{C_l}}^{\mathrm{H}}\right]^{\mathrm{H}} \in \mathbb{C}^{C_l M \times K}, \tag{10}$$

where $\hat{\mathbf{H}}_{S_i}$ is the estimated channel of the i-th AP in set $\mathcal{S}$.

Let $R_{C_l} = \{r_{l,1},\ldots,r_{l,M}\}$ be the corresponding indices of AP l in set $\mathcal{S}$; then, the channel estimate vector between the AP $l \in \mathcal{S}$ and UE k can be expressed in terms of $\hat{\mathbf{H}}_S$ as

$$\hat{\mathbf{h}}_{l,k} = \mathbf{E}_{R_{C_l}} \hat{\mathbf{H}}_S \mathbf{e}_k = \hat{\mathbf{H}}_l \mathbf{e}_k, \tag{11}$$

where $\mathbf{E}_{R_{C_l}} = \left[\mathbf{e}_{r_{l,1}},\ldots,\mathbf{e}_{r_{l,M}}\right]^{\mathrm{H}} \in \mathbb{C}^{M \times C_l M}$, and $\mathbf{e}_{r_{l,i}}$ is the $r_{l,i}$ -th column of $\mathbf{I}_{C_l M}$, while $\mathbf{e}_k$ denotes the k-th column of $\mathbf{I}_K$. The transmit signals at APs in set $\mathcal{S}$ and $\mathcal{W}$ can be denoted as

$$\mathbf{x}_l = \sum_{k=1}^{K} \sqrt{\rho_{l,k}} \mathbf{w}_{l,k}^{\mathrm{ZF}} q_k, l \in \mathcal{S}, \tag{12}$$

$$\mathbf{x}_p = \sum_{k=1}^{K} \sqrt{\rho_{p,k}} \mathbf{w}_{p,\mathrm{i}_k}^{\mathrm{MRT}} q_k, p \in \mathcal{W}. \tag{13}$$

The precoding vectors constructed for AP $l \in \mathcal{S}$ and $p \in \mathcal{W}$ to UE k are defined as

$$\mathbf{w}_{l,k}^{\mathrm{ZF}} \triangleq \sqrt{\lambda_{l,k}} \mathbf{E}_{R_{C_l}} \hat{\mathbf{H}}_S \left(\hat{\mathbf{H}}_S^{\mathrm{H}} \hat{\mathbf{H}}_S\right)^{-1} \mathbf{e}_k, \tag{14}$$

$$\mathbf{w}_{p,\mathrm{i}_k}^{\mathrm{MRT}} \triangleq \frac{\bar{\mathbf{H}}_p \mathbf{e}_{\mathrm{i}_k}}{\sqrt{\mathsf{E}\left\{\left\|\bar{\mathbf{H}}_p \mathbf{e}_{\mathrm{i}_k}\right\|^2\right\}}}, \tag{15}$$

where $\lambda_{l,k}$ is the normalization factor to satisfy

$$\mathsf{E}\left\{\left\|\mathbf{w}_{l,k}^{\mathrm{ZF}}\right\|^2\right\} \leq 1, \forall l \in \mathcal{S}, \tag{16}$$

and $\mathbf{e}_{\mathrm{i}_k}$ is the i_k-th column of $\mathbf{I}_{\tau_{\mathrm{p}}}$. In [21], it was pointed out that ZF is, in general, sub-optimal for power control schemes subject to per-antenna power constraints, whilst finding an optimal precoder involves numerical algorithms. The precoding matrix for APs in set $\mathcal{S}$ can be rewritten as $\mathbf{W}_S^{\mathrm{ZF}} = \Lambda \circ \mathbf{W}_S$, where $[\Lambda]_{l,k} = \sqrt{\lambda_{l,k}}$. The elements of $\mathbf{W}_S^{\mathrm{ZF}}$ are given as

$$\left[\mathbf{W}_S^{\mathrm{ZF}}\right]_{l,k} = \mathbf{w}_{l,k}^{\mathrm{ZF}} = \sqrt{\lambda_{l,k}} \left[\mathbf{W}_S\right]_{l,k}, \forall l \in \mathcal{S}. \tag{17}$$

In order for $\hat{\mathbf{H}}_S^{\mathrm{H}} \mathbf{W}_S^{\mathrm{ZF}}$ to be diagonal, similar to CZF [10], it is necessary to have $\lambda_{1,k} = \ldots = \lambda_{C_l,k}$ for any UE k, which means that the normalization factors should only be functions of k, i.e., $\lambda_{l,k} = \lambda_k, \forall l \in \mathcal{S}$. For the same reason, we should also have the power coefficients for set $\mathcal{S}$ as $\rho_{l,k} = \rho_{S_k}, \forall l \in \mathcal{S}$. Therefore, the precoding matrix can be further expressed as

$$\mathbf{W}_S^{\mathrm{ZF}} = \mathbf{W}_S \mathbf{\Lambda}_S \in \mathbb{C}^{C_l M \times K}, \tag{18}$$

where $\mathbf{\Lambda}_S$ is a diagonal matrix with $\sqrt{\lambda_1},\ldots,\sqrt{\lambda_K}$ on its diagonal. For a given UE k, the corresponding precoding vector of AP set $\mathcal{S}$ can be given by

$$\mathbf{W}_{S,k}^{\mathrm{ZF}} = \sqrt{\lambda_k} \hat{\mathbf{H}}_S \left(\hat{\mathbf{H}}_S^{\mathrm{H}} \hat{\mathbf{H}}_S\right)^{-1} \mathbf{e}_k \in \mathbb{C}^{C_l M \times 1}, \tag{19}$$

The received signal at UE k is, thus, given by

$$y_k = \left(\sum_{l \in \mathcal{S}} \sqrt{\rho_{l,k}} \mathbf{h}_{l,k}^{\mathrm{H}} \mathbf{w}_{l,k}^{\mathrm{ZF}} + \sum_{p \in \mathcal{W}} \sqrt{\rho_{p,k}} \mathbf{h}_{p,k}^{\mathrm{H}} \mathbf{w}_{p,\mathrm{i}_k}^{\mathrm{MRT}}\right) q_k + \sum_{\substack{t=1 \\ t \neq k}}^{K} \left(\sum_{l \in \mathcal{S}} \sqrt{\rho_{l,t}} \mathbf{h}_{l,k}^{\mathrm{H}} \mathbf{w}_{l,t}^{\mathrm{ZF}} + \sum_{p \in \mathcal{W}} \sqrt{\rho_{p,t}} \mathbf{h}_{p,k}^{\mathrm{H}} \mathbf{w}_{p,\mathrm{i}_t}^{\mathrm{MRT}}\right) q_t + n_k. \tag{20}$$

B. DL SE

A lower bound on the ergodic capacity, i.e. an achievable SE, can be obtained by using the use-and-then-forget bounding technique, where the detection at the users relies only on the channel statistics as in [1], [13]. This technique is widely used in the massive MIMO literature because: (i) it yields a simple and insightful closed-form achievable rate expression which enables us to characterize and optimize the system performance; (ii) this bound is inherently tight due to the channel hardening property of massive MIMO systems, i.e., the effective channel gain fluctuates only slightly around its mean. This tightness has been discussed in Remark 4 and Figure 2 of [6] among others; and (iii) with the detection using only the channel statistics, the need for DL training is avoided which significantly reduces the DL channel estimation overhead. The received signal y_k in (7) can be rewritten as

$$y_k = \mathrm{CP}_k \cdot q_k + \mathrm{PU}_k \cdot q_k + \sum_{t \neq k}^{K} \mathrm{UI}_{k,t} \cdot q_t + n_k, \tag{21}$$

where CP_k, PU_k, and $\mathrm{UI}_{k,t}$ represent the coherent precoding gain, precoding gain uncertainty, and multi-user interference, respectively, defined as

$$\mathrm{CP}_k \triangleq \sum_{l=1}^{L} \sqrt{\rho_{l,k}} \mathsf{E}\left\{\mathbf{h}_{l,k}^{\mathrm{H}} \mathbf{w}_{l,k}\right\}, \tag{22}$$

$$\mathrm{PU}_k \triangleq \sum_{l=1}^{L} \left(\sqrt{\rho_{l,k}} \mathbf{h}_{l,k}^{\mathrm{H}} \mathbf{w}_{l,k} - \sqrt{\rho_{l,k}} \mathsf{E}\left\{\mathbf{h}_{l,k}^{\mathrm{H}} \mathbf{w}_{l,k}\right\}\right), \tag{23}$$

$$\mathrm{UI}_{k,t} \triangleq \sum_{l=1}^{L} \sqrt{\rho_{l,t}} \mathbf{h}_{l,k}^{\mathrm{H}} \mathbf{w}_{l,t}. \tag{24}$$

As described in (21), UE k effectively sees a deterministic channel (CP_k) with some unknown noise. Since q_k and q_t are independent for any $k \neq t$, the first term in (21) is uncorrelated with the third term. Furthermore, q_k is independent of PU_k, thus, the first and the second terms are also uncorrelated. By assumption, the fourth term (noise) is independent of the first term. Therefore, the sum of the second, third and fourth term in (21) can be treated as an uncorrelated effective noise. By invoking the arguments in [1], an achievable DL SE for UE k is given by

$$\mathsf{SE}_k = \frac{\tau_{\mathrm{D}}}{\tau_{\mathrm{U}} + \tau_{\mathrm{D}}} \left(1 - \frac{\tau_{\mathrm{P}}}{\tau_{\mathrm{C}}}\right) \log_2\left(1 + \mathsf{SINR}_k\right) \quad [\mathrm{bit/s/Hz}], \tag{25}$$

where the effective SINR of UE k is given as (26) on the top of next page, and $T_k \triangleq \left| \sum_{l \in \mathcal{S}} \sqrt{\rho_{l,k}} \mathsf{E}\left\{\mathbf{h}_{l,k}^{\mathrm{H}} \mathbf{w}_{l,k}^{\mathrm{ZF}}\right\} + \sum_{p \in \mathcal{W}} \sqrt{\rho_{p,k}} \mathsf{E}\left\{\mathbf{h}_{p,k}^{\mathrm{H}} \mathbf{w}_{p,i_k}^{\mathrm{MRT}}\right\} \right|^2$.

Theorem 1: The closed-form expression for the achievable SE of UE k is given by (25), where the SINR_k is given as (27) on the top of next page, where $\lambda_k \triangleq \min_l \frac{1}{\upsilon_{l,k}}$, and $\upsilon_{l,t} \triangleq \mathsf{E}\left\{\left\|\mathbf{w}_{l,t}^{\mathrm{ZF}}\right\|^2\right\}, \forall l \in \mathcal{S}$.

Proof: See Appendix. □

Remark 3: Note that, if $\mathcal{S} = \emptyset, \mathcal{W} = \{1, \ldots, L\}$, then all APs would generate the precoding vector locally by MRT, and (27) will reduce to the scheme in [6]. By contrast, if $\mathcal{W} = \emptyset$, then our scheme becomes the centralized ZF scheme [10].

C. Access point subset selection

The aim of JMRZF is to combine a set of APs to mitigate part of the inter-AP interference. It is computationally challenging to determine the subset of C_l APs that generates the most inter-AP interference. In this section, we propose a large-scale fading-based AP selection method with a sub-optimal performance, where the AP selection changes only when the large-scale fading parameters change..

The selection scheme is mainly based on the observation in [6], that for a given UE, there are many APs which are located very far away. These APs will not contribute much to the overall spatial diversity gains. From a user-centric point of view, for a given UE, its neighboring APs with the largest channel gains will probably contribute most of the interference. Therefore, the APs with the largest channel gains will generate more inter-AP interference, thus, should be considered to perform centralized ZF. Motivated by the above observation, we propose to choose the set $\mathcal{S}$ for a given $C_l \in [K/r, L]$ as follows

1) Let $\mathcal{T}$ be the set of $\{\beta_1, \ldots, \beta_L\}$, where $\beta_l = \sum_{k=1}^{K} \beta_{l,k}$ corresponds to AP l.
2) Sort set $\mathcal{T}$ in descending order, and choose the first C_l APs corresponding to the largest C_l elements in $\mathcal{T}$.

IV. Max-min Fairness Power Control

In this section, we consider the problem of maximizing the minimum SE of all UEs (which is also known as max-min fairness) subject to per-AP power constraints.[2] With (12) and (17), we can further denote the transmit signal at AP $l \in \mathcal{S}$ as

$$\mathbf{x}_l = \mathbf{W}_l \mathbf{P}_S \mathbf{q}_l \in \mathbb{C}^{M \times 1}, \forall l \in \mathcal{S}, \tag{28}$$

where $\mathbf{W}_l = \mathbf{E}_{R_{C_l}} \mathbf{W}_S$ is the un-normalized precoding matrix for APs in $\mathcal{S}$, and $\mathbf{q}_l = [q_1, \ldots, q_K]^{\mathrm{H}}$ is the corresponding data symbol vector, and $\mathbf{P}_S$ is a $K \times K$ power-related diagonal matrix with the $k-$th element $[\mathbf{P}_S]_{k,k} = \sqrt{\lambda_k \rho_{S_k}}$. We now define the vector

$$\boldsymbol{\mu}_l \triangleq \mathrm{diag}\left\{\mathsf{E}\{\mathbf{W}_l^{\mathrm{H}} \mathbf{W}_l\}\right\} \in \mathbb{C}^{K \times 1}, \tag{29}$$

then, the transmit signal power from AP $l \in \mathcal{S}$ can be given as

$$\mathsf{E}\{\mathbf{x}_l^{\mathrm{H}} \mathbf{x}_l\} = \sum_{k=1}^{K} \lambda_k \rho_{S_k} \mu_{l,k}, \tag{30}$$

where $\mu_{l,k}$ is the k-th element of $\boldsymbol{\mu}_l$.

For the APs in set $\mathcal{S}$, the (normalized) transmitted power from AP $l \in \mathcal{S}$ given by (30) is constrained by the per-AP power limit ρ_l^{max} as

$$\rho_l = \sum_{k=1}^{K} \nu_k \mu_{l,k} \leq \rho_l^{\mathsf{max}}, \forall l \in \mathcal{S}, \tag{31}$$

where ν_k is defined as $\nu_k \triangleq \lambda_k \rho_{S_k}, \forall l \in \mathcal{S}$, ρ_{S_k} denotes the power coefficient allocated for APs in set $\mathcal{S}$ to UE k. For the set $\mathcal{W}$, we also have the per-AP power limit for each AP as

$$\rho_p = \sum_{t=1}^{K} \rho_{p,t} \leq \rho_l^{\mathsf{max}}, \forall p \in \mathcal{W}. \tag{32}$$

Then, a general power limit for JMRZF can be expressed as

$$\rho_i = \sum_{k=1}^{K} \rho_{i,k} \lambda_{i,k} \mu_{i,k} \leq \rho_l^{\mathsf{max}}, i = 1, \ldots, L, \tag{33}$$

[2]Theoretically, it is possible to design the beamforming vectors based on the obtained CSI. However, to do this, the optimization problem has to be solved for every small-scale fading realization (i.e. has to be done over the small-scale fading time scale) which changes very quickly with time. This has a huge computational complexity, especially in massive MIMO where the beamforming matrices have very high dimension due to the use of large numbers of antennas and users. With our scheme, the power control is only implemented on a large-scale fading time scale which changes very slowly with time (e.g. about some 40 times slower than the small-scale fading coefficient does [14]). In addition, in massive MIMO, due to the favorable propagation property, the linear beamformers work very well [1]. Therefore, all the designs of massive MIMO in literature are similar to the design in our work, i.e. the linear beamformers are first deployed, and then optimal power control is designed (see, for example, [1] and references therein).

$$\mathsf{SINR}_k = \frac{\left|\sum\limits_{l\in\mathcal{S}}\sqrt{\rho_{l,k}}\mathsf{E}\left\{\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{w}_{l,k}^{\mathrm{ZF}}\right\}+\sum\limits_{p\in\mathcal{W}}\sqrt{\rho_{p,k}}\mathsf{E}\left\{\mathbf{h}_{p,k}^{\mathrm{H}}\mathbf{w}_{p,\mathrm{i}_k}^{\mathrm{MRT}}\right\}\right|^2}{\sum\limits_{t=1}^{K}\mathsf{E}\left\{\left|\sum\limits_{l\in\mathcal{S}}\sqrt{\rho_{l,t}}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{w}_{l,t}^{\mathrm{ZF}}+\sum\limits_{p\in\mathcal{W}}\sqrt{\rho_{p,t}}\mathbf{h}_{p,k}^{\mathrm{H}}\mathbf{w}_{p,\mathrm{i}_t}^{\mathrm{MRT}}\right|^2\right\}-T_k+1}, \tag{26}$$

$$\mathsf{SINR}_k = \frac{\left(\sqrt{\rho_{S_k}\lambda_k}+\sum\limits_{p\in\mathcal{W}}\sqrt{M\rho_{p,k}\gamma_{p,k}}\right)^2}{\sum\limits_{t=1}^{K}\sum\limits_{p\in\mathcal{W}}\rho_{p,t}\beta_{p,k}+\sum\limits_{t=1}^{K}\left(\rho_{S_t}\sum\limits_{l\in\mathcal{S}}\upsilon_{l,t}\left(\beta_{l,k}-\gamma_{l,k}\right)\right)+\sum\limits_{t\in\mathcal{P}_k\setminus\{k\}}\left(\sum\limits_{p\in\mathcal{W}}\sqrt{M\rho_{p,t}\gamma_{p,k}}\right)^2+1}. \tag{27}$$

where $\lambda_{l,k}=\lambda_k,\rho_{l,k}=\rho_{S_k},\forall l\in\mathcal{S}$ and $\lambda_{p,k}=\mu_{p,k}=1,\forall p\in\mathcal{W}$. For simplicity, in the following, the expression in (33) can be further rewritten as

$$\rho_l=\sum_{k=1}^{K}\nu_{l,k}\mu_{l,k}\leq\rho_l^{\mathsf{max}},\forall l, \tag{34}$$

where $\nu_{l,k}=\nu_k,\forall l\in\mathcal{S}$, and $\nu_{p,k}\triangleq\lambda_{p,k}\rho_{p,k},\forall p\in\mathcal{W}$.

We can further rewrite the SINR given in (27) as (35) on the top of next page, Then, the max-min power allocation scheme with per-AP power constraint can be rewritten as

$$\max_{\nu_{l,k}\geq 0}\ \min_k\ \mathsf{SINR}_k, \tag{36a}$$

$$\text{s.t.}\quad\sum_{k=1}^{K}\nu_{l,k}\mu_{l,k}\leq\rho_l^{\mathsf{max}},\forall l, \tag{36b}$$

which is equivalent to

$$\max_{\nu_{l,k}\geq 0}\ \omega, \tag{37a}$$

$$\text{s.t.}\quad\mathsf{SINR}_k\geq\omega,\forall k, \tag{37b}$$

$$\sum_{k=1}^{K}\nu_{l,k}\mu_{l,k}\leq\rho_l^{\mathsf{max}},\forall l, \tag{37c}$$

To simplify the optimization problem, we introduce the following notations:

$$\begin{aligned}
\mathbf{u}_S&=\left[\sqrt{\nu_1},\ldots,\sqrt{\nu_K}\right],\\
\mathbf{u}_k&=\left[\sqrt{\nu_{1,k}},\ldots,\sqrt{\nu_{W_{l,k}}}\right]^{\mathrm{T}},\\
\boldsymbol{\nu}_k&=\left[\sqrt{\nu_k},\sqrt{\nu_{1,k}},\ldots,\sqrt{\nu_{W_{l,k}}}\right]^{\mathrm{T}},\\
\mathbf{u}_l^{'}&=\begin{cases}\mathbf{u}_S, & l\in\mathcal{S}\\ \left[\sqrt{\nu_{l,1}},\ldots,\sqrt{\nu_{l,K}}\right], & l\in\mathcal{W}\end{cases},\\
\mathbf{G}_k&=\left[1,\sqrt{M\gamma_{1,k}},\ldots,\sqrt{M\gamma_{W_{l,k}}}\right]^{\mathrm{T}},\\
\mathbf{g}_k&=\left[\sqrt{M\gamma_{1,k}},\ldots,\sqrt{M\gamma_{W_{l,k}}}\right]^{\mathrm{T}},\\
\mathbf{d}_k&=\left[\sqrt{\mu_{1,k}},\ldots,\sqrt{\mu_{C_{l,k}}}\right]^{\mathrm{T}},\\
\mathbf{D}_l^{'}&=\begin{cases}\left[\sqrt{\mu_{l,1}},\ldots,\sqrt{\mu_{l,K}}\right], & l\in\mathcal{S}\\ \mathbf{1}^K, & l\in\mathcal{W}\end{cases},\\
\mathbf{c}_k&=\left[\sqrt{(\beta_{1,k}-\gamma_{1,k})},\ldots,\sqrt{\left(\beta_{C_{l,k}}-\gamma_{C_{l,k}}\right)}\right]^{\mathrm{T}},\\
\mathbf{b}_k&=\left[\sqrt{\beta_{1,k}},\ldots,\sqrt{\beta_{W_{l,k}}}\right]^{\mathrm{T}},
\end{aligned}$$

where $t_1,\ldots,t_{|\mathcal{P}_k\setminus\{k\}|}$ are the UE indices $\in\mathcal{P}_k\setminus\{k\}$ and s_k is given as (38) on the top of next page . Finally, the max-min power allocation problem can be reformulated as

$$\max_{\nu_{l,k}\geq 0}\ \omega, \tag{39a}$$

$$\text{s.t.}\ \|\mathbf{s}_k\|-\mathbf{G}_k^{\mathrm{T}}\boldsymbol{\nu}_k\leq 0,\forall k, \tag{39b}$$

$$\left\|\mathbf{u}_l^{'}\circ\mathbf{D}_l^{'}\right\|\leq\sqrt{\rho_l^{\mathsf{max}}},\forall l. \tag{39c}$$

Similarly to [12], since (39b) is an increasing function of ω, the solution to (39) can be obtained by solving the corresponding feasibility problem, through the *bisection method* [22] together with SOC programs.

V. First order method to reduce the computation complexity

In Section IV, the max-min optimization problem was solved with the bisection method, where the interior-point

$$\mathsf{SINR}_k = \frac{\left(\sqrt{\nu_k} + \sum\limits_{p\in\mathcal{W}} \sqrt{M\nu_{p,k}\gamma_{p,k}}\right)^2}{\sum\limits_{t=1}^{K}\sum\limits_{p\in\mathcal{W}} \nu_{p,t}\beta_{p,k} + \sum\limits_{t=1}^{K}\left(\nu_t \sum\limits_{l\in\mathcal{S}} \mu_{l,t}\left(\beta_{l,k}-\gamma_{l,k}\right)\right) + \sum\limits_{t\in\mathcal{P}_k\setminus\{k\}}\left(\sum\limits_{p\in\mathcal{W}}\sqrt{M\nu_{p,t}\gamma_{p,k}}\right)^2 + 1}. \tag{35}$$

$$\mathbf{s}_k = \left[\sqrt{\omega}\left(\mathbf{g}_k^{\mathrm{T}}\mathbf{u}_{t_1},\ldots,\mathbf{g}_k^{\mathrm{T}}\mathbf{u}_{t_{|\mathcal{P}_k\setminus\{k\}|}},\left\|\mathbf{u}_1\circ\mathbf{b}_k\right\|,\ldots,\left\|\mathbf{u}_K\circ\mathbf{b}_k\right\|,\left\|\sqrt{\nu_1}\mathbf{d}_1\circ\mathbf{c}_k\right\|,\ldots,\left\|\sqrt{\nu_k}\mathbf{d}_k\circ\mathbf{c}_k\right\|,1\right)\right]^{\mathrm{T}}, \tag{38}$$

method is involved and requires the computation of the Hessian matrix, thus, is called second-order cone method. This method can provide global optimal solution. However, the SOC methods entails substantial execution time and memory requirements, especially when the number of APs and UEs is large. Therefore, in this section, we propose to use a first order method to solve (39) which is sub-optimal but its computational complexity is low.

A. Problem reformulation

To simplify the optimization problem with the first order method, we introduce the following notations: $\boldsymbol{\varrho} \in \mathbb{C}^{(W_l+1)K}$, $\mathbf{A}_k \in \mathbb{C}^{(W_l+1)\times(W_l+1)K}$, $\mathbf{B}_{E_k} \in \mathbb{C}^{(W_l+1)}$, $\mathbf{G}_{E_k} \in \mathbb{C}^{(W_l+1)}$, $\mathbf{Q}_S \in \mathbb{C}^{(W_l+1)\times(W_l+1)K}$. These vectors and matrices are defined as

$$\begin{aligned}
\boldsymbol{\varrho} &= \left[\mathbf{u}_S^{\mathrm{T}}; \mathbf{u}_1^{'\mathrm{T}}; \ldots; \mathbf{u}_{W_l}^{'\mathrm{T}}\right],\\
\mathbf{A}_k &= \mathbf{I}_{W_l+1}\otimes\mathbf{e}_k^{\mathrm{T}},\\
\mathbf{B}_{E_k} &= \left[0, \sqrt{\beta_{1,k}}, \ldots, \sqrt{\beta_{W_l,k}}\right]^{\mathrm{T}},\\
\mathbf{G}_{E_k} &= \left[0, \sqrt{M\gamma_{1,k}}, \ldots, \sqrt{M\gamma_{W_l,k}}\right]^{\mathrm{T}},\\
\mathbf{Q}_S &= \bar{\mathbf{e}}_s^{\mathrm{T}}\otimes\mathbf{I}_{W_l+1},\\
\mu_{S_{t,k}} &= \sum_{l\in\mathcal{S}}\mu_{l,t}\left(\beta_{l,k}-\gamma_{l,k}\right),
\end{aligned}$$

where $\mathbf{e}_k \in \mathbb{R}^K$ is the k-th unit vector. With these definitions, we can rewrite SINR_k as a function of $\boldsymbol{\varrho}$ as (40) on the top of next page, where $\bar{\mathbf{e}}_s$ is the first row of $\mathbf{I}_K$. Also, $\mathbf{J}_l \in \mathbb{C}^{K\times(W_l+1)K}$ is constructed by the rows of $\mathbf{I}_{(W_l+1)K}$, and fulfills

$$\left(\mathbf{J}_l\boldsymbol{\varrho}\right)^{\mathrm{T}} = \mathbf{u}_l^{'}, \forall l = 1,\ldots,L. \tag{41}$$

Now, the max-min fairness problem can be rewritten with the help of $\boldsymbol{\varrho}$ as

$$\max_{\boldsymbol{\varrho}\geq\mathbf{0}} \min_k \mathsf{SINR}_k\left(\boldsymbol{\varrho}\right), \tag{42a}$$

$$\text{s.t.} \quad \left\|\left(\mathbf{J}_l\boldsymbol{\varrho}\right)^{\mathrm{T}}\circ\mathbf{D}_l^{'}\right\| \leq \sqrt{\rho_l^{\max}}, \forall l. \tag{42b}$$

With the log-sum exponential approximation, we have

$$\min_k \mathsf{SINR}_k \cong \left(-\frac{1}{\chi}\right)\log\left(\sum_{k=1}^{K}\exp\left(-\chi\mathsf{SINR}_k\right)\right), \chi\to\infty. \tag{43}$$

Such approximation is tight when $\chi\to\infty$. In simulations, we use a very large number instead of infinity as a practical approach. Note that χ should be carefully selected depending on the software/hardware resources. With (43), the max-min problem (42) is equivalent to the following problem as

$$\min_{\boldsymbol{\varrho}\geq 0} \frac{1}{\chi}\log\left(\sum_{k=1}^{K}\exp\left(-\chi\mathsf{SINR}_k\left(\boldsymbol{\varrho}\right)\right)\right), \chi\to\infty, \tag{44a}$$

$$\text{s.t.} \quad \left\|\left(\mathbf{J}_l\boldsymbol{\varrho}\right)^{\mathrm{T}}\circ\mathbf{D}_l^{'}\right\| \leq \sqrt{\rho_l^{\max}}, \forall l. \tag{44b}$$

The first order method, which mainly focuses on the non-convex problem has the following general form:

$$\min_{\boldsymbol{\varrho}}\left\{F\left(\boldsymbol{\varrho}\right) \equiv f\left(\boldsymbol{\varrho}\right) + g\left(\boldsymbol{\varrho}\right)\right\} \tag{45}$$

where $f\left(\boldsymbol{\varrho}\right)$ is a differentiable function (but possibly non-convex) and $g\left(\boldsymbol{\varrho}\right)$ could be both non-convex and non-smooth. Further assumptions on $f\left(\boldsymbol{\varrho}\right)$ and $g\left(\boldsymbol{\varrho}\right)$ are listed as below:

1) $f\left(\boldsymbol{\varrho}\right)$ is a proper function with Lipschitz continuous gradients.
2) $g\left(\boldsymbol{\varrho}\right)$ is proper and lower semi-continuous.
3) $F\left(\boldsymbol{\varrho}\right)$ is coercive, which means $F\left(\boldsymbol{\varrho}\right)$ is bounded from below and $F\left(\boldsymbol{\varrho}\right)\to\infty$ when $\left\|\boldsymbol{\varrho}\right\|\to\infty$.

A special case of problem (45) is the constrained optimization. Let Ω be a closed convex set and let $\delta_\Omega\left(\boldsymbol{\varrho}\right)$ be its indicator function defined as

$$\delta_\Omega\left(\boldsymbol{\varrho}\right) \triangleq \begin{cases} 0, & \boldsymbol{\varrho}\in\Omega,\\ +\infty, & \boldsymbol{\varrho}\notin\Omega. \end{cases} \tag{46}$$

Then, the constrained minimization problem $\min\{f\left(\boldsymbol{\varrho}\right)|\boldsymbol{\varrho}\in\Omega\}$ can be equivalently rewritten in the form of (45) with $g\left(\boldsymbol{\varrho}\right) \equiv \delta_\Omega\left(\boldsymbol{\varrho}\right)$. The corresponding proximal operator function of $g\left(\boldsymbol{\varrho}\right) \equiv \delta_\Omega\left(\boldsymbol{\varrho}\right)$ is the projection onto Ω as

$$\begin{aligned}
\mathsf{prox}_{\varsigma g}\left(\boldsymbol{\varrho}\right) :=& \mathsf{arg}\min_{\mathbf{a}} \delta_\Omega\left(\boldsymbol{\varrho}\right) + \frac{1}{2\varsigma}\left\|\boldsymbol{\varrho}-\mathbf{a}\right\|^2\\
=& \mathsf{arg}\min_{\mathbf{a}\in\Omega}\left\|\boldsymbol{\varrho}-\mathbf{a}\right\|^2 \triangleq \mathrm{P}_\Omega\left(\boldsymbol{\varrho}\right).
\end{aligned} \tag{47}$$

The feasible set of problem (44) can be expressed as

$$\Omega = \left\{\boldsymbol{\varrho}\middle|\left\|\left(\mathbf{J}_l\boldsymbol{\varrho}\right)^{\mathrm{T}}\circ\mathbf{D}_l^{'}\right\| \leq \sqrt{\rho_l^{\max}}, l=1,\ldots,L; \boldsymbol{\varrho}\geq\mathbf{0}\right\}. \tag{48}$$

The corresponding first order problem of (44) is

$$\min_{\boldsymbol{\varrho}\in\mathbb{C}^{(W_l+1)K\times 1}}\left\{F\left(\boldsymbol{\varrho}\right) \equiv f\left(\boldsymbol{\varrho}\right) + g\left(\boldsymbol{\varrho}\right)\right\}, \tag{49}$$

where $f\left(\boldsymbol{\varrho}\right) = \frac{1}{\chi}\log\left(\sum_{k=1}^{K}\exp\left(-\chi\mathsf{SINR}_k\left(\boldsymbol{\varrho}\right)\right)\right)$, and $g\left(\boldsymbol{\varrho}\right) = \delta_\Omega\left(\boldsymbol{\varrho}\right)$. The authors in [23] proposed an accelerated

$$\mathsf{SINR}_k\left(\boldsymbol{\varrho}\right)=\frac{\left(\mathbf{G}_k^{\mathrm{T}}\mathbf{A}_k\boldsymbol{\varrho}\right)^2}{\sum\limits_{t=1}^{K}\left\|\mathbf{A}_t\boldsymbol{\varrho}\circ\mathbf{B}_{E_k}\right\|^2+\sum\limits_{t=1}^{K}\left(\mathbf{e}_t\mathbf{Q}_S\boldsymbol{\varrho}\right)^2\mu_{S_{t,k}}+\sum\limits_{t\in\mathcal{P}_k\setminus\{k\}}\left(\mathbf{G}_{E_k}^{\mathrm{T}}\mathbf{A}_t\boldsymbol{\varrho}\right)^2+1},\tag{40}$$

proximal gradient (APG) method for solving (45). The implementation of the APG algorithm heavily depends on the calculation of $\nabla f\left(\boldsymbol{\varrho}\right)$ and the projection of the feasible set, where the corresponding gradient function can be denoted as

$$\nabla f\left(\boldsymbol{\varrho}\right)=-\frac{\sum_{k=1}^{K}\exp\left(-\chi\mathsf{SINR}_k\left(\boldsymbol{\varrho}\right)\right)\nabla\mathsf{SINR}_k}{\sum_{k=1}^{K}\exp\left(-\chi\mathsf{SINR}_k\left(\boldsymbol{\varrho}\right)\right)}.\tag{50}$$

To get the gradient of SINR_k, we recall the following equations:

$$\nabla\left(\mathbf{G}_k^{\mathrm{T}}\mathbf{A}_k\boldsymbol{\varrho}\right)^2=2\mathbf{A}_k^{\mathrm{T}}\mathbf{G}_k\mathbf{G}_k^{\mathrm{T}}\mathbf{A}_k\boldsymbol{\varrho},\tag{51}$$

$$\nabla\left\|\mathbf{A}_t\boldsymbol{\varrho}\circ\mathbf{B}_{E_k}\right\|^2=2\mathbf{A}_t^{\mathrm{T}}\mathbf{B}_{E_k}\mathbf{B}_{E_k}^{\mathrm{T}}\mathbf{A}_t\boldsymbol{\varrho},\tag{52}$$

$$\nabla\left(\mathbf{e}_t\mathbf{Q}_S\boldsymbol{\varrho}\right)^2\mu_{S_{t,k}}=2\mu_{S_{t,k}}\mathbf{Q}_S^{\mathrm{T}}\mathbf{e}_t^{\mathrm{T}}\mathbf{e}_t\mathbf{Q}_S\boldsymbol{\varrho}.\tag{53}$$

Together with the composition rule of gradient, we can finally get $\nabla\mathsf{SINR}_k\left(\boldsymbol{\varrho}\right)$ as (54) on the top of next page, where $\vartheta_k=\left(\mathbf{G}_k^{\mathrm{T}}\mathbf{A}_k\boldsymbol{\varrho}\right)^2$ and $\varpi_k=\sum\limits_{t=1}^{K}\left\|\mathbf{A}_t\boldsymbol{\varrho}\circ\mathbf{B}_{E_k}\right\|^2+\sum\limits_{t=1}^{K}\left(\mathbf{e}_t\mathbf{Q}_S\boldsymbol{\varrho}\right)^2\mu_{S_{t,k}}+\sum\limits_{t\in\mathcal{P}_k\setminus\{k\}}\left(\mathbf{G}_{E_k}^{\mathrm{T}}\mathbf{A}_t\boldsymbol{\varrho}\right)^2+1$ are the numerator and the denominator of SINR_k, respectively.

Finally, based on the APG method, we propose Algorithm 1 to solve problem (44).

Algorithm 1 Monotone accelerated proximal gradient method in non-convex case

Input: $\boldsymbol{\varrho}_0\in\mathbb{R}_+, 0<\alpha_x<\frac{1}{L_f}, 0<\alpha_y<\frac{1}{L_f}$
Output: $\boldsymbol{\varrho}$

Set $t_1=t_0=1, \boldsymbol{\varrho}_1=\mathbf{z}_1=\boldsymbol{\varrho}_0$
for $k=1,2,\ldots$ **do**
  $\mathbf{y}_k=\boldsymbol{\varrho}_k+\frac{t_{k-1}}{t_k}\left(\mathbf{z}_k-\boldsymbol{\varrho}_k\right)+\frac{t_{k-1}-1}{t_k}\left(\boldsymbol{\varrho}_k-\boldsymbol{\varrho}_{k-1}\right)$
  $\mathbf{z}_{k+1}=\mathrm{P}_\Omega\left(\mathbf{y}_k-\alpha_y\nabla f\left(\mathbf{y}_k\right)\right)$
  $\mathbf{v}_{k+1}=\mathrm{P}_\Omega\left(\boldsymbol{\varrho}_k-\alpha_x\nabla f\left(\boldsymbol{\varrho}_k\right)\right)$
  if $F\left(\mathbf{z}_{k+1}\right)\le F\left(\mathbf{v}_{k+1}\right)$, **then**
    $\boldsymbol{\varrho}_{k+1}=\mathbf{z}_{k+1}$
  else
    $\boldsymbol{\varrho}_{k+1}=\mathbf{v}_{k+1}$
  end if
  $t_{k+1}=\frac{\sqrt{4(t_k)^2+1}+1}{2}$
end for

B. Projection onto Ω

Recall that the projection function $\mathrm{P}_\Omega\left(\boldsymbol{\varrho}\right)$ in (47) is written as

$$\min_{\boldsymbol{\varrho}\in\mathbb{R}^{(W_l+1)K}}\quad\left\|\boldsymbol{\varrho}-\mathbf{a}\right\|^2\tag{55a}$$

$$\text{s.t.}\quad\left\|\left(\mathbf{J}_l\boldsymbol{\varrho}\right)^{\mathrm{T}}\circ\mathbf{D}_l^{'}\right\|\le\sqrt{\rho_l^{\max}}, l=1,\ldots,L,\tag{55b}$$

$$\boldsymbol{\varrho}\ge 0,\tag{55c}$$

where $\mathbf{a}$ is the vector to be projected in Algorithm 1 and has the same structure as $\boldsymbol{\varrho}$, while $\mathbf{a}_S\in\mathbb{C}^K$ contains the first K elements of $\mathbf{a}$. Note that the objective in (55) is separable with $\mathbf{u}_l^{'}$, thus, is equivalent to solving the sub-problem for each $l=1,\ldots,L$:

$$\min_{\mathbf{u}_l^{'}\in\mathbb{R}^K}\quad\left\|\mathbf{u}_l^{'}-\mathbf{a}_l\right\|^2\tag{56a}$$

$$\text{s.t.}\quad\left\|\mathbf{u}_l^{'}\circ\mathbf{D}_l^{'}\right\|\le\sqrt{\rho_l^{\max}},\tag{56b}$$

$$\mathbf{u}_l^{'}\ge 0.\tag{56c}$$

For APs belong to set $\mathcal{W}$, as $\mathbf{D}_l^{'}=\mathbf{1}^K, \forall p\in\mathcal{W}$, the sub-problems (56) of AP $l\in\mathcal{W}$ can be rewritten as

$$\min_{\mathbf{u}_l^{'}\in\mathbb{C}^K}\quad\left\|\mathbf{u}_l^{'}-\mathbf{a}_l\right\|^2\tag{57a}$$

$$\text{s.t.}\quad\left\|\mathbf{u}_l^{'}\right\|\le\sqrt{\rho_l^{\max}},\tag{57b}$$

$$\mathbf{u}_l^{'}\ge 0.\tag{57c}$$

Problem (57) admits the following analytical solution in [24]:

$$\mathbf{u}_l^{'}=\frac{\sqrt{\rho_l^{\max}}}{\max\left(\left\|[\mathbf{a}_i]_+\right\|,\sqrt{\rho_l^{\max}}\right)}[\mathbf{a}_l]_+, \forall l\in\mathcal{W}.\tag{58}$$

For the APs in set $\mathcal{S}$, the C_l sub-problems are equivalent to

$$\min_{\mathbf{u}_{D_l}\in\mathbb{R}^K}\quad\left\|\mathbf{u}_{D_l}-\mathbf{a}_{D_l}\right\|^2\tag{59a}$$

$$\text{s.t.}\quad\left\|\mathbf{u}_{D_l}\right\|\le\sqrt{\rho_l^{\max}},\tag{59b}$$

$$\mathbf{u}_{D_l}\ge 0,\tag{59c}$$

where $\mathbf{u}_{D_l}\triangleq\mathbf{u}_{S_l}\circ\mathbf{D}_l^{'}, \mathbf{a}_{D_l}\triangleq\mathbf{a}_s\circ\mathbf{D}_l^{'}$, and $\mathbf{u}_{S_l}=\mathbf{u}_{D_l}\oslash\mathbf{D}_l^{'}$ denotes the possible solution for $l\in\mathcal{S}$, we can now get the corresponding solution for each sub-problem as (58)

$$\mathbf{u}_{S_l}=\frac{\sqrt{\rho_l^{\max}}}{\max\left(\left\|[\mathbf{a}_{D_l}]_+\right\|,\sqrt{\rho_l^{\max}}\right)}[\mathbf{a}_s]_+, \forall l\in\mathcal{S}.\tag{60}$$

From (60), we can see that the solutions to all sub-problems are a scaled version of $[\mathbf{a}_s]_+$, which means $\mathbf{u}_{S_l}=\alpha_l[\mathbf{a}_s]_+, \forall l\in\mathcal{S}$, with $0<\alpha_l<1$. Among these C_l solutions of (60), only one fulfills $\left\|\mathbf{u}_{S_l}\circ\mathbf{D}_p^{'}\right\|\le\sqrt{\rho_l^{\max}}, \forall p\in\mathcal{S}, p\ne l$. To get this unique solution, we first define

$$\mathbf{D}_s^{'}\triangleq\max_{\mathbf{D}_l^{'}}\left\|[\mathbf{a}_s]_+\circ\mathbf{D}_l^{'}\right\|^2, \forall l\in\mathcal{S}.\tag{61}$$

Then, we can get the projection for set $\mathcal{S}$ as

$$\mathbf{u}_S=\frac{\sqrt{\rho_l^{\max}}}{\max\left(\left\|[\mathbf{u}_{D_s}]_+\right\|,\sqrt{\rho_l^{\max}}\right)}[\mathbf{a}_s]_+,\tag{62}$$

where $\mathbf{u}_{D_s}=\mathbf{a}_s\circ\mathbf{D}_s^{'}$ is an intermediate variable determined by $\mathbf{D}_s^{'}$.

$$\nabla \mathsf{SINR}_k = \left(\frac{2}{\varpi_k} \mathbf{A}_k^{\mathrm{T}} \mathbf{G}_k \mathbf{G}_k^{\mathrm{T}} \mathbf{A}_k - \frac{2\vartheta_k}{\varpi_k^2} \left(\sum_{t=1}^{K} \mathbf{A}_t^{\mathrm{T}} \mathbf{B}_{E_k} \mathbf{B}_{E_k}^{\mathrm{T}} \mathbf{A}_t + \sum_{t \in \mathcal{P}_k \setminus k} \mathbf{A}_t^{\mathrm{T}} \mathbf{G}_{E_k} \mathbf{G}_{E_k}^{\mathrm{T}} \mathbf{A}_t + \sum_{t=1}^{K} \mu_{S_{t,k}} \mathbf{Q}_S^{\mathrm{T}} \mathbf{e}_t^{\mathrm{T}} \mathbf{e}_t \mathbf{Q}_S \right) \right) \boldsymbol{\varrho}, \quad (54)$$

C. Backtracking linear search

From (50) and (54) we can see that both forms of $\nabla f(\boldsymbol{\varrho})$ are Lipschitz continuous, following the results in [25]. In fact, we can implement Algorithm 1 with the backtracking line search technique following the Barzilai-Borwein (BB) rule [26] instead of finding a Lipschitz constant L_f of $\nabla f(\boldsymbol{\varrho})$. For example, other than using a fixed step size as described in line 4 and line 5 of Algorithm 1, we can perform a line search as described in Algorithm 2 on top of next page. The backtracking line search algorithm starts with a large step size and decreases it until a better feasible solution is found. As $\nabla f(\mathbf{z}_k)$ is Lipschitz continuous for some Lipschitz constant, the line search procedure is proved to terminate after a finite number of iterations.

Algorithm 2 Backtracking Line Search Algorithm

Input: $\rho_{BB} < 1, \delta_{BB} > 0$
$\mathbf{s}_k = \mathbf{z}_k - \mathbf{y}_{k-1}, \mathbf{r}_k = \nabla f(\mathbf{z}_k) - \nabla f(\mathbf{y}_{k-1})$
Set $\alpha_y = \frac{\mathbf{s}_k^{\mathrm{T}} \mathbf{s}_k}{\mathbf{s}_k^{\mathrm{T}} \mathbf{r}_k}$ or $\alpha_y = \frac{\mathbf{s}_k^{\mathrm{T}} \mathbf{r}_k}{\mathbf{r}_k^{\mathrm{T}} \mathbf{r}_k}$
repeat
  $\mathbf{z}_{k+1} = \mathrm{P}_{\Omega}(\mathbf{y}_k - \alpha_y \nabla f(\mathbf{y}_k))$
  $\alpha_y = \rho_{BB} \alpha_y$
until $F(\mathbf{z}_{k+1}) \leq F(\mathbf{y}_k) - \delta_{BB} \|\mathbf{z}_{k+1} - \mathbf{y}_k\|^2$

VI. Front-haul overhead and computational complexity analysis

A. Front-haul overhead analysis

Front-haul overhead refers to the amount of information conveyed over the front-haul link to perform joint transmission/detection, which can be expressed as the number of exchanged complex scalars between the CPU and APs. The UL front-haul requirements of cell-free massive MIMO system have been investigated in [27] for different levels of cooperation between APs and CPU. For distributed precoding schemes (i.e. MRT), precoders are designed locally by the corresponding full-rank channel estimate $\bar{\mathbf{H}}_l$ of each AP, and from (8) we can see that only CDI is needed for distributed precoding schemes. Centralized ZF has the most advanced capability of interference cancellation with the help of channel magnitude information (CMI). The CPU can acquire CMI directly through the channel estimates from each AP, which can be represented by MK complex scalars. The CPU can also get CMI as in (3) with $M\tau_{\mathrm{P}}$ complex scalars for the pilot signals and the large-scale fading coefficient $\beta_{l,k}$ known a-priori. By defining $T \triangleq \min\{K, \tau_{\mathrm{P}}\}$, the UL training cost for CZF can be expressed as MT complex scalars.

In each coherence block of the DL data transmission, for centralized ZF precoding, CPU needs to send back MK complex scalars for precoding vectors and $\tau_{\mathrm{D}} K$ complex scalars for data payload (i.e., q_k) to each AP. Moreover, if the centralized power control schemes are applied, an extra number of K real-valued scalars, i.e., the power control coefficients $\rho_{l,k}$ need to be send back to each AP. For distributed precoding schemes, as the precoding vectors are generated locally at each AP, there exists only the overhead of data payload and power control coefficients of the front-haul link. However, as the power coefficients depend only on the large-scale fading coefficients, the corresponding overhead for $\beta_{l,k}$ and $\rho_{l,k}$ can be ignored if the channel changes slowly. Interestingly, the resolution of signal quantization can be optimized as in [16] and [17].

For JMRZF, C_l APs are combined for centralized zero-forcing, while the other W_l APs adopt MRT, naturally, the front-haul overhead of JMRZF can be seen as the sum of a fraction of CZF and MRT, respectively. The corresponding front-haul overhead for each precoding scheme is summarized in Table II.

TABLE II: Number of complex scalars sent via the front-haul per coherence block

Precoding Schemes	UL training	DL data
CZF [10]	LMT	$LKM + \tau_{\mathrm{D}} LK$
JMRZF (proposed)	$C_l MT$	$C_l KM + \tau_{\mathrm{D}} LK$
MRT [6]	-	$\tau_{\mathrm{D}} LK$

B. Complexity analysis

As the computation complexity of generalized inverse for a matrix $\mathbb{C}^{m \times n}$ is $\mathcal{O}(n^3)$ and the main computation complexity of precoding schemes is dominated by such inverse procedure, the corresponding complexity of the aforementioned precoding schemes can be summarized as in Table III.

TABLE III: Computation complexity comparison of precoding matrix design

Precoding Schemes	CZF [10]	PPZF [12]	proposed JMRZF
Computation complexity	$\mathcal{O}(L^3 M^3)$	$\mathcal{O}(M^3)$	$\mathcal{O}(C_l{}^3 M^3)$

Problem (39) has K SOC constraints of dimension $2K + P_k + 1$, where $P_k = |\mathcal{P}_k \setminus \{k\}|$, and L SOC constraints of dimension K, with a number of decision variables on the order of $d_v = \mathcal{O}((W_l + 1)L + 1)$. We call the set of $(\boldsymbol{\nu}^{\epsilon}, \omega^{\epsilon})$ an ϵ-solution to problem (39) if

$$\omega^{\epsilon} \leq \omega^* + \epsilon, \quad (63)$$

where ω^* is the globally optimum solution to (39). With the same methodology as [28], [29], the computational complexity to obtain the ϵ-solution to problem (39) is

$$\zeta(4K^3 + 4P_k K^2 + LK^2 + 4K^2 + P_k^2 K + 2P_k K + K + d_v^2) d_v, \quad (64)$$

where $\zeta = \ln(\epsilon^{-1})\sqrt{2L+2K}$ is the order of iterations required, while the remaining terms represent the per-iteration computation costs.

For the proposed FOM scheme, as noted in [30], it is clear that the computation complexity of Algorithm 1 is dominated by three parts: the objective function F, the gradient ∇f, and the projection P_{Ω}. It is easy to see that LK multiplications are required to compute SINR_k and, thus, the complexity of getting $F(\boldsymbol{\varrho})$ is $\mathcal{O}(LK^2)$. Similarly, we can find the complexity of ∇f which is $\mathcal{O}(LK^2)$. The projection of $\boldsymbol{\varrho}$ is given in Section V-B, where the computation of the l_2-norm entails a complexity of $\mathcal{O}(LK)$. We can conclude that the per-iteration computation complexity in big-$\mathcal{O}$ notation for FOM is $\mathcal{O}(LK^2)$. However, the worst case of computation complexity depends on the number of iterations of Algorithm 1, which is related to the convergence performance. The complexity of differentiable convex optimization has been reported in [31], however, the computation saving for non-convex problems has not been reported to the best of the authors' knowledge.

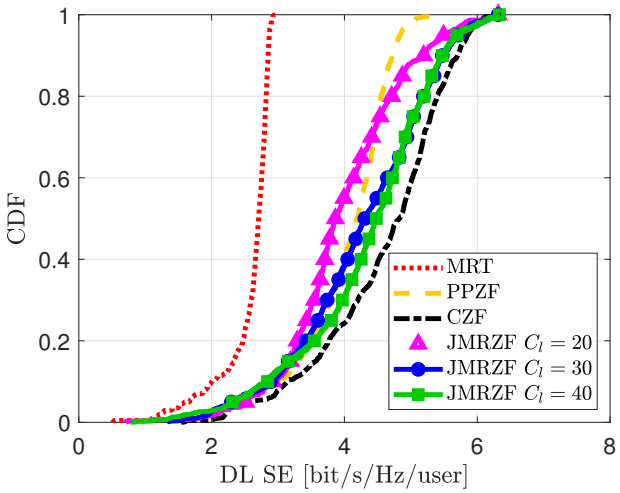


Fig. 1: CDFs of the per-user SE achieved by different precoding schemes under Scenario 1.

VII. Numerical Results and Discussion

The performance of the proposed precoding scheme and power allocation scheme are numerically evaluated, analyzed and discussed in this section. We firstly introduce the system setup as well as the parameters considered in simulations.

A. Simulation scenario

We assume K UEs and L APs are located in an area of size $D \times D$ square meters, APs and UEs are assumed uniformly and randomly distributed unless otherwise noted.

The large-scale fading coefficients $\{\beta_{l,k}\}$, incorporate pathloss and shadow fading, as follows

$$\beta_{l,k} = \mathsf{PL}_{l,k} \cdot 10^{\frac{\sigma_{sh} z_{l,k}}{10}}, \tag{65}$$

where $\mathsf{PL}_{l,k}$ represents the pathloss, and $10^{\frac{\sigma_{sh} z_{l,k}}{10}}$ models log-normal shadow fading with standard deviation σ_{sh} and $z_{l,k} \sim \mathcal{N}(0,1)$. The pathloss follows the 3GPP Urban Microcell Model [32], which assumes a 2GHz carrier frequency, and is given by

$$\mathsf{PL}_{l,k}[\mathrm{dB}] = -30.5 - 36.7\log_{10}\left(\frac{d_{l,k}}{1\mathrm{m}}\right), \tag{66}$$

where $d_{l,k}$ is the distance between AP l and UE k including AP and UE's heights. The shadow fading accounts for spatial correlations between APs and between UEs, and follows [33]

$$z_{l,k} = \sqrt{\kappa} a_l + \sqrt{1-\kappa} b_k, \tag{67}$$

where $a_l \sim \mathcal{N}(0,1)$ and $b_k \sim \mathcal{N}(0,1)$ are independent RVs modeling the shadow fading impact that caused by the obstructing objects in the vicinity of the lth AP and the kth UE, respectively, while the parameter κ provides the weighting between these phenomena. The shadowing terms are correlated as

$$\mathsf{E}\{a_l a_i\} = 2^{\frac{d_{l,i}^{\mathrm{AP}}}{9m}}, \quad \mathsf{E}\{b_k b_t\} = 2^{\frac{d_{k,t}^{\mathrm{UE}}}{9m}},$$

where $d_{l,i}^{\mathrm{AP}}$ is the distance between AP l and AP i, $d_{k,t}^{\mathrm{UE}}$ is the distance between UE k and UE t, and 9 meters is the decorrelation distance [32]. The standard deviation is set as $\sigma_{sh} = 4$ dB, APs and UEs are assumed with heights of 10 m and 1.5 m, respectively, while the channel bandwidth $B = 20$ MHz. The maximum transmit power is 200 mW for each AP, and 100 mW for each UE, with the noise power is $w_p^{(\mathrm{dBm})} = -92$ dBm. Hence,

$$\rho_l^{\max}\,[\mathrm{dBm}] = 10\log_{10}(200) - w_p^{(\mathrm{dBm})}, \forall l,$$
$$p_k\,[\mathrm{dBm}] = 10\log_{10}(100) - w_p^{(\mathrm{dBm})}, \forall k,$$

Finally, we assume the pilots are randomly assigned to UEs. The simulation parameters are summarized as in Table IV.

TABLE IV: Simulation parameters for different scenarios

Parameters	UE Distribution	L	K	D	M	τ_{P}	τ_{C}
Scenario 1	Uniform	100	10	500	8	7	200
Scenario 2	Uniform	40	20	200	5	15	200
Scenario 3	Non-Uniform	40	40	200	5	40	500

B. SE with power control

Fig. 1 compares the SE of the proposed scheme with some state-of-art precoding schemes: MRT in [6], CZF in [10], and PPZF in [12] under Scenario 1. For JMRZF, the number of total combined APs C_l increases from 20 to 40, and PPZF follows the setting in [12]. The max-min fairness power control is adopted together with SOC. As $K - \tau_{\mathrm{P}} = 3$ pilots are reused, all precoding schemes will experience pilot contamination. From Fig. 1, we first observe that all ZF based schemes significantly outperform the MRT, especially for the high percentiles, due to the fact that MRT cannot inherently suppress the interference. Among the ZF-based schemes, CZF has the best performance, as all APs jointly process to cancel all inter-user interference. It is not surprising that the performance of JMRZF lies in between MRT and CZF, in fact, its capability of interference suppression and available DoFs grow with C_l. We can also see that the lowest

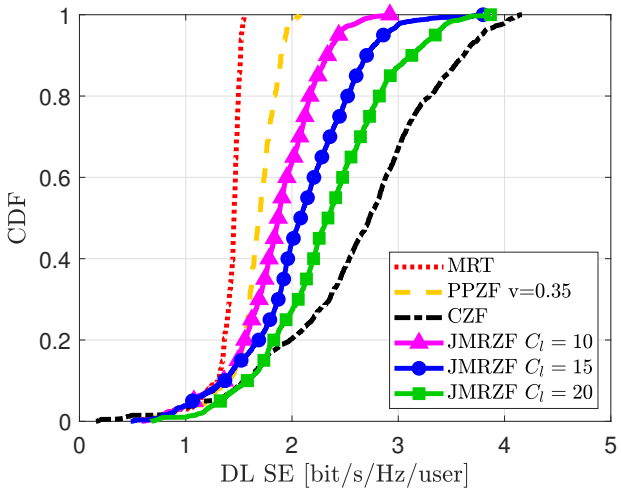


Fig. 2: CDFs of the per-user SE achieved by different precoding schemes under Scenario 2.

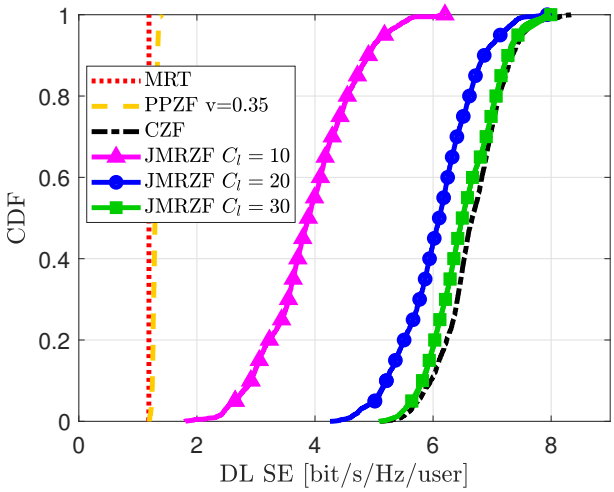


Fig. 3: CDFs of the per-user SE achieved by different precoding schemes under Scenario 3.

percentiles' performance is not limited by interference but pilot contamination. By improving the interference suppression capabilities, only the high percentiles' SE improves.

We will now evaluate the DL SE of the precoding schemes in a more practical manner. In Scenario 2, by setting $M = 5$, we assume that each AP deploys a small number of antennas, which makes fewer DoFs be available for interference suppression when local ZF is applied. Fig. 2 shows the cumulative distribution of the SE with the precoding schemes introduced in Fig. 1, while the setup follows Scenario 2. We keep the AP density constant by reducing the simulation area to 200×200 m^2. For JMRZF, the number of total combined APs C_l increases from 10 to 20. The max-min fairness power control is realized with the bisection method. We can first observe that, all schemes' performance deteriorates compared with Scenario 1, due to the partial loss of array gain. We can also see that the performance of the local ZF scheme becomes the worst across the ZF-based schemes. The reason is that, as noted in Remark 1, to offer enough DoF for interference suppression with local ZF, each AP selects less UEs to perform ZF by setting $v = 0.35$. As such, less interference can be canceled, yet, the achievable performance is still superior compared to MRT as noted in [12]. Together with Fig. 1, we can infer that the performance of JMRZF always lies in between MRT and CZF, and is able to work well even when each AP has few antennas.

Furthermore, in Scenario 3, we investigate the SE performance when a large number of UEs are clustered and not evenly distributed, which may occur in relatively open spaces such as a stadium, airport terminals or shopping malls. We assume that most UEs are clustered around some of the APs while others are evenly distributed in whole area. APs are assumed to be uniformly distributed following a Poisson point process (PPP), while the non-uniformly distributed UEs follow a Thomas cluster process (TCP) as in [34]. We assume $L = 40$ APs and $K_1 = 8$ UEs are uniformly distributed in an square area with $D = 200$ meters, while the other $K_2 = 32$ UEs are located in clusters following TCP around $L_N = 8$ APs with a maximum distance of 5 meters.

Fig. 3 shows the cumulative distribution of the SE with precoders introduced in Fig. 1. The setup difference with scenario 2 is $K = 40$. For JMRZF, the number of total combined APs C_l increases from 10 to 30. With MRT, most APs will suffer from path loss due to the non-uniform UE distribution, and the UEs can achieve a uniformly good service thanks to the max-min power control. We can see that the local ZF scheme PPZF has a closer performance to MRT compared with Scenario 2. The reason is that, as K/M increases from 4 to 8, a smaller part of UEs can avail of interference cancellation. We can see that JMRZF significantly outperforms local precoding schemes, and is comparable with CZF when $C_l = 30$. The reason is that, JMRZF can take advantage of all pilots and avoid inter-AP interference partially. We can see that CZF yields the best performance as it has the strongest capability of interference suppression.

C. SE and front-haul overhead evaluation

Fig. 4 shows the average SE and the front-haul overhead of different precoding schemes versus C_l, with max-min power allocation. The simulation set up is $L = 8, M = 100, K = 10, \tau_{\mathrm{P}} = 7, \tau_{\mathrm{D}} = 100$. As shown in Fig. 4, both the SE and front-haul overhead of JMRZF fall in between CZF and MRT and grows linearly with the increase of C_l. We can see that, JMRZF can offer about 20% more SE with a cost of 60% front-haul overhead when comparing with MRT at $C_l = 3$. With $C_l = 6$, JMRZF can achieve more than 93% SE with only 80% front-haul overhead compared with CZF.

D. SE without pilot reuse

As shown in Fig. 1, the performance of UEs with the lowest SE is limited by pilot contamination. To further demonstrate the advantage of the proposed scheme when the number of available pilots is greater or equal than the number of UEs, we examine the performance of JMRZF and other state-of-art precoding schemes in Scenario 1 by setting $K = \tau_{\mathrm{P}}$

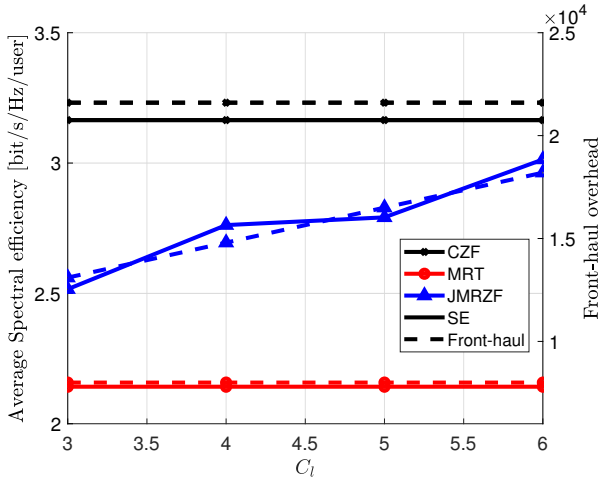


Fig. 4: Average SE and front-haul overhead per-coherence block of different precoding schemes versus C_l. Simulation setup: $L = 8, M = 100, K = 10, \tau_{\mathrm{P}} = 7, \tau_{\mathrm{D}} = 100$.

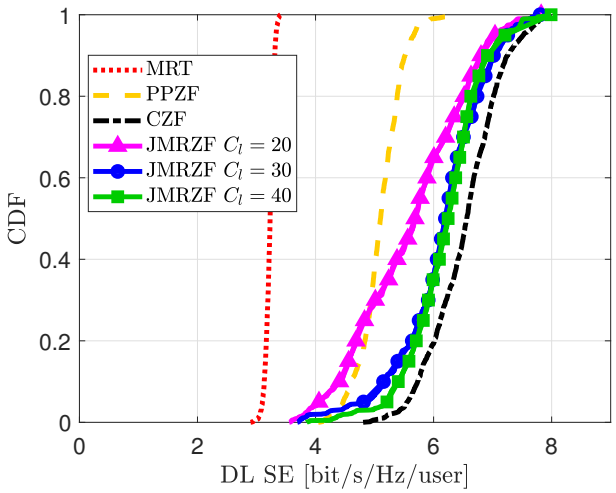


Fig. 5: CDFs of the per-user SE achieved by different precoding schemes without pilot reuse. Simulation setup same as Scenario 1, but $K = \tau_{\mathrm{P}}$.

and keep other parameters fixed. Fig. 5 shows the cumulative distribution of DL SE with no pilot reuse. Compared with Fig. 1, we can see that all schemes will have a better performance without pilot reuse. With MRT almost all UEs have a uniform performance, and JMRZF outperforms local ZF schemes when $C_l = 30$, since the gain of interference cancellation is small when C_l is greater than 30. What is different with Fig.1 is that, without pilot contamination, the system performance is mainly interference limited. Therefore, we can observe a big performance gap between PPZF and MRT, which showcases the performance improvement by cancelling each AP's interference. We can also see that there still exists a big gap between PPZF and CZF, which can be eliminated by increasing C_l, which implies that local ZF cannot fully reap the potentials of cell-free massive MIMO.

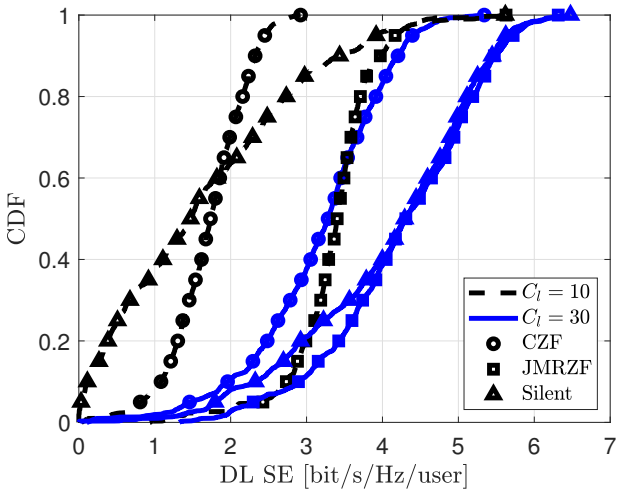


Fig. 6: CDFs of the per-user SE achieved by different precoding schemes under Scenario 1.

E. SE when partial APs activated

In this scenario, we want to find out the impact on the system performance if only the APs in set $\mathcal{S}$ participate in the transmission. In other words, the APs in set $\mathcal{W}$ are turned off in an energy saving approach. By setting the power of APs in set $\mathcal{W}$ to zero as $\rho_{p,k} = 0, \forall p \in \mathcal{W}, k = 1, \ldots, K$, with (27), we can have the corresponding SINR as

$$\mathsf{SINR}_k = \frac{\rho_{S_k} \lambda_k}{\sum\limits_{t=1}^{K} \left(\rho_{S_t} \sum\limits_{l \in \mathcal{S}} \upsilon_{l,t} \left(\beta_{l,k} - \gamma_{l,k} \right) \right) + 1}. \tag{68}$$

Note that (68) is quite similar to the SINR of CZF in [10, Eq. (13)] that has C_l APs. The only difference is that with JMRZF, the C_l activated APs are selected from L ones, hence with an additional AP clustering gain are expected to offer a better performance than CZF. The max-min power control can be performed easily following the procedure described in Section IV.

Fig. 6 simulates the corresponding cumulative distribution of SE, the simulation setup is Scenario 1. For JMRZF, the number of total combined APs C_l varies from 10 to 30. For CZF, a total number of C_l APs are located in a squared area with $D = 5C_l$ to keep the density of APs constant, such that the UEs will share a similar channel quality as JMRZF. We can see that, when C_l is small, e.g., $C_l = 10$, the contributions of APs in set $\mathcal{W}$ cannot be ignored, especially for the lower percentiles. However the performance gap between 'silent mode' and original JMRZF is quite small, when C_l becomes large, especially for higher percentiles, which means that APs with MRT contribute most in the lower percentiles. If we take CZF into account, we can see when C_l is small, CZF outperforms the 'silent' mode due to a smaller average AP-UE distance. However, when C_l becomes larger, JMRZF silent mode outperforms CZF due to the AP clustering gain.

F. SE with FOM power control

In this part, we evaluate the performance of first order method described in Section V together with the second order

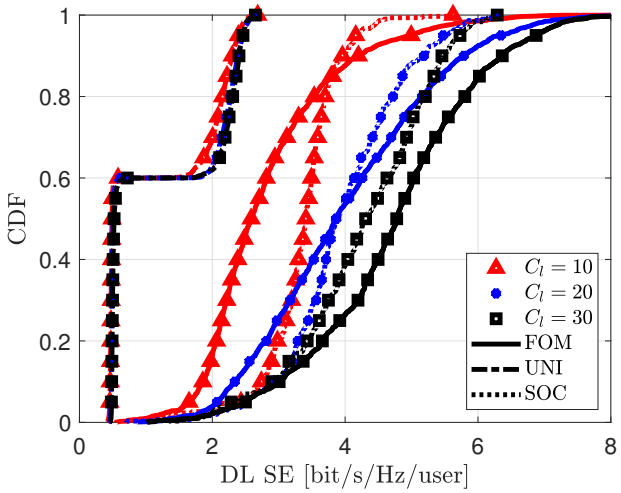


Fig. 7: CDFs of the per-user SE of JMRZF with different power allocation schemes under Scenario 1. Solid, dashed and dotted curves indicate the results obtained by first order method (FOM), second order cone (SOC) and uniform (UNI) power allocation, respectively.

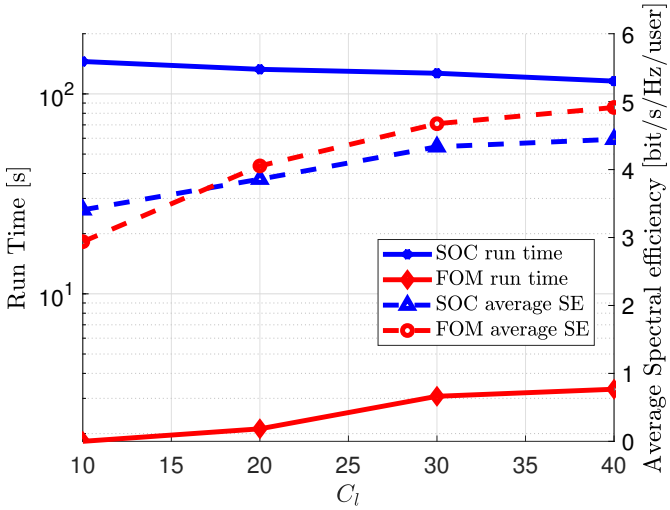


Fig. 8: Run time (s) of FOM and SOC versus the number of APs in set $\mathcal{S}$, under Scenario 1.

cone method described in Section IV and uniform power allocation scheme in small/medium and large-scale scenarios. Throughout our simulations, χ is carefully selected to make sure $\chi \mathsf{SINR}_k^{\min} \in [0.1, 10]$, where $\mathsf{SINR}_k^{\min}$ is the minimum value of SINR_k in (43).

1) Small and medium scale scenario: In this part, we consider the case when $L \leq 100$ APs are deployed in terms of SE performance and time consumption.

Fig. 7 shows the cumulative distribution of the SE of JMRZF with different power allocation schemes. We can first observe that, all these schemes' performance improves with the increase of C_l, and uniform (UNI) power allocation's performance is always the worst. FOM's performance is worse than SOC when C_l is small in terms of 95%-likely per-UE SE; as C_l grows large, JMRZF can achieve a better performance with FOM than SOC, especially for the high SE percentiles. In Fig. 8, we report the actual run time and average per-UE SE corresponding to FOM and SOC. We execute our numerical codes on a 64-bit Windows operating system with 8GB RAM and Intel CORE i5, 1.6 GHz. FOM is terminated when the difference of the objective for the last 5 iterations is less than 10^{-3}. We can clearly see that FOM can always offer a comparable average SE while saving almost ten times of run time.

2) Large scale scenarios: In this part, we compare the performance of FOM with uniform power allocation schemes in a large-scale scenario, where $L = 400, M = 5, K = 4, D = 2000, \tau_{\mathrm{C}} = 400$ and $\tau_{\mathrm{P}} = 30$. It is time and memory prohibitive to perform the second order method, and for this reason we do not consider SOC in this subsection. We can see from Fig. 9, in a large-scale scenario, FOM always outperforms uniform power allocation scheme. We can also see that, unlike FOM, the performance of UNI improves marginally with the increase of C_l.

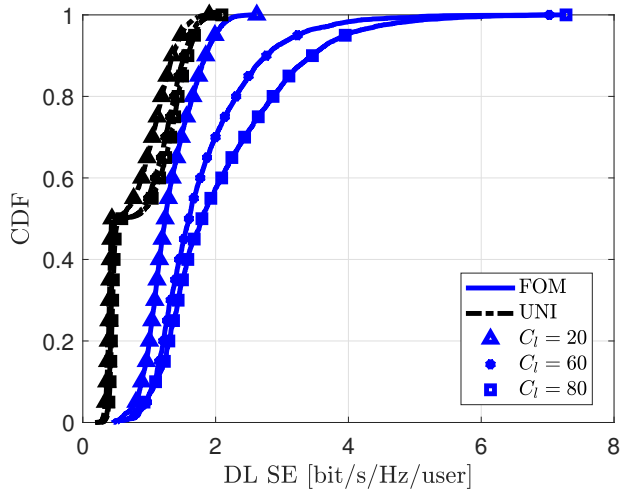


Fig. 9: CDFs of the per-user SE of JMRZF with different power allocation schemes. Simulation setup: $L = 400, M = 5, D = 2000, K = 40, \tau_{\mathrm{P}} = 30, \tau_{\mathrm{C}} = 400$.

VIII. Conclusion

This paper proposed a JMRZF precoder for cell-free massive MIMO systems. The majority of previous papers on precoder design relied on fully distributed or fully centralized schemes. We leveraged a trade-off between the system performance and front-haul signal overhead that can be obtained by consolidating a subset of APs utilizing the ZF scheme, while the remaining APs use the MRT scheme. A large-scale parameter based AP subset selection was proposed, and the corresponding max-min power control was also investigated and solved. In addition, we proposed to use the first order method to reduce the computation complexity of power allocation in large-scale scenarios. We compared the results of the proposed scheme with other state-of-the-art precoders with max-min power control. The results showed that the proposed scheme provides substantially higher SE than MRT and other distributed ZF schemes, even when only a small number of

APs uses centralized ZF at a small cost of front-haul signal overhead. We further showed that when APs are equipped few antennas, which is a common case in cell-free massive MIMO, the proposed scheme is highly preferable compared to distributed alternatives.

APPENDIX

The channel estimates of UE k for APs in set $\mathcal{S}$ can be given as

$$\hat{\mathbf{h}}_{S,k} = \hat{\mathbf{H}}_S \mathbf{e}_k. \tag{69}$$

For any pair of UEs k, t, and APs $\in \mathcal{S}$, we have:

$$\alpha^{\text{ZF}}_{S,k,t} \triangleq \hat{\mathbf{h}}^{\text{H}}_{S,k} \mathbf{W}^{\text{ZF}}_{S,t} = \begin{cases} 0, & \text{other} \\ \sqrt{\lambda_t}, & t = k. \end{cases} \tag{70}$$

For any pair of UEs k, t and AP $p \in \mathcal{W}$, we have

$$\mathsf{E}\left\{\hat{\mathbf{h}}^{\text{H}}_{p,k} \mathbf{w}^{\text{MRT}}_{p,\text{i}_t}\right\} = \begin{cases} 0 & t \notin \mathcal{P}_k \\ \sqrt{M\gamma_{p,k}} & t \in \mathcal{P}_k \end{cases}. \tag{71}$$

We first look at the numerator of (26), as the sum of expectation equals the expectation of the sum, we have:

$$\sum_{l\in\mathcal{S}} \sqrt{\rho_{l,k}} \mathsf{E}\left\{\mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,k}\right\} = \sqrt{\rho_{S_k}} \mathsf{E}\left\{\mathbf{h}^{\text{H}}_{S,k} \mathbf{W}^{\text{ZF}}_{S,t}\right\}. \tag{72}$$

With (70), we have,

$$\left|\sum_{l\in\mathcal{S}} \sqrt{\rho_{l,k}} \mathsf{E}\left\{\mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,k}\right\} + \sum_{p\in\mathcal{W}} \sqrt{\rho_{p,k}} \mathsf{E}\left\{\mathbf{h}^{\text{H}}_{p,k} \mathbf{w}^{\text{MRT}}_{p,\text{i}_k}\right\}\right|^2 = \left(\sqrt{\rho_{S_k}\lambda_k} + \sum_{p\in\mathcal{W}} \sqrt{M\rho_{p,k}\gamma_{p,k}}\right)^2. \tag{73}$$

The first term of the denominator in (26) is:

$$\begin{aligned} &\sum_{t=1}^{K} \mathsf{E}\left\{\left|\sum_{l\in\mathcal{S}} \sqrt{\rho_{l,t}} \mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,t} + \sum_{p\in\mathcal{W}} \sqrt{\rho_{p,t}} \mathbf{h}^{\text{H}}_{p,k} \mathbf{w}^{\text{MRT}}_{p,\text{i}_t}\right|^2\right\} \\ &= \sum_{t=1}^{K} \mathsf{E}\left\{\left|\sum_{l\in\mathcal{S}} \sqrt{\rho_{l,t}} \mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,t}\right|^2\right\} \\ &+ 2\sum_{t=1}^{K} \text{Re}\left\{\sum_{l\in\mathcal{S}}\sum_{p\in\mathcal{W}} \sqrt{\rho_{l,t}\rho_{p,t}} \mathsf{E}\left\{\mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,t} \left(\mathbf{w}^{\text{MRT}}_{p,\text{i}_t}\right)^{\text{H}} \mathbf{h}_{p,k}\right\}\right\} \\ &+ \sum_{t=1}^{K} \mathsf{E}\left\{\left|\sum_{p\in\mathcal{W}} \sqrt{\rho_{p,t}} \mathbf{h}^{\text{H}}_{p,k} \mathbf{w}^{\text{MRT}}_{p,\text{i}_t}\right|^2\right\}. \end{aligned} \tag{74}$$

We focus on the last term of (74) as

$$\begin{aligned} &\sum_{t=1}^{K} \mathsf{E}\left\{\left|\sum_{p\in\mathcal{W}} \sqrt{\rho_{p,t}} \mathbf{h}^{\text{H}}_{p,k} \mathbf{w}^{\text{MRT}}_{p,\text{i}_t}\right|^2\right\} \\ &= \sum_{t=1}^{K}\sum_{p\in\mathcal{W}} \rho_{p,t}\beta_{p,k} + \sum_{t\in\mathcal{P}_k} \left(\sum_{p\in\mathcal{W}} \sqrt{M\rho_{p,t}\gamma_{p,k}}\right)^2. \end{aligned} \tag{75}$$

With (72), the first term in (74) can be rewritten as

$$\begin{aligned} &\sum_{t=1}^{K} \mathsf{E}\left\{\left|\sum_{l\in\mathcal{S}} \sqrt{\rho_{l,t}} \mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,t}\right|^2\right\} \\ &= \left(\sqrt{\rho_{S_k}\lambda_k}\right)^2 + \sum_{t=1}^{K} \left(\rho_{S_t} \sum_{l\in\mathcal{S}} \upsilon_{l,t} \left(\beta_{l,k} - \gamma_{l,k}\right)\right), \end{aligned} \tag{76}$$

where

$$\mathsf{E}\left\{\left|\tilde{\mathbf{h}}^{\text{H}}_{S,k} \mathbf{w}^{\text{ZF}}_{S,t}\right|^2\right\} = \sum_{l\in\mathcal{S}} \upsilon_{l,t} \left(\beta_{l,k} - \gamma_{l,k}\right), \tag{77}$$

and $\upsilon_{l,t} = \mathsf{E}\left\{\left\|\mathbf{w}^{\text{ZF}}_{l,t}\right\|^2\right\} = \mathsf{E}\left\{\mu_{l,t}\right\}\lambda_t \leq 1$, $\mathsf{E}\{\tilde{\mathbf{h}}_{S,k}\tilde{\mathbf{h}}^{\text{H}}_{S,k}\} \in \mathbb{C}^{C_l M \times C_l M}$ is a diagonal matrix with $(\beta_{l,k} - \gamma_{l,k})$ on its $(l-1)M+1, \ldots, lM$-th diagonal element.

With

$$\begin{aligned} &\sum_{l\in\mathcal{S}} \mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,t} \mathsf{E}\left\{\left(\mathbf{w}^{\text{MRT}}_{p,\text{i}_t}\right)^{\text{H}} \mathbf{h}_{p,k}\right\} \\ &= \begin{cases} 0 & t \neq k \\ \sqrt{\rho_{S_k}\lambda_k \sum_{p\in\mathcal{W}} M\rho_{p,k}\gamma_{p,k}} & t = k, \end{cases} \end{aligned} \tag{78}$$

we can get the second term in (74) as

$$\begin{aligned} &\sum_{t=k} \left(2\sqrt{\rho_{S_k}\lambda_k} \sum_{p\in\mathcal{W}} \sqrt{M\rho_{p,t}\gamma_{p,k}}\right) \\ &= \left(\sqrt{\rho_{S_k}\lambda_k} + \sum_{p\in\mathcal{W}} \sqrt{M\rho_{p,k}\gamma_{p,k}}\right)^2 - \left(\sqrt{\rho_{S_k}\lambda_k}\right)^2 \\ &- \left(\sum_{p\in\mathcal{W}} \sqrt{M\rho_{p,k}\gamma_{p,k}}\right)^2. \end{aligned} \tag{79}$$

With (75), (76) and (79), the first term of the denominator in (26) can be further denoted as

$$\begin{aligned} &\sum_{t=1}^{K} \mathsf{E}\left\{\left|\sum_{l\in\mathcal{S}} \sqrt{\rho_{l,t}} \mathbf{h}^{\text{H}}_{l,k} \mathbf{w}^{\text{ZF}}_{l,t} + \sum_{p\in\mathcal{W}} \sqrt{\rho_{p,t}} \mathbf{h}^{\text{H}}_{p,k} \mathbf{w}^{\text{MRT}}_{p,\text{i}_t}\right|^2\right\} \\ &= \sum_{t\in\mathcal{P}_k\setminus\{k\}} \left(\sum_{p\in\mathcal{W}} \sqrt{M\rho_{p,t}\gamma_{p,k}}\right)^2 + \sum_{t=1}^{K}\sum_{p\in\mathcal{W}} \rho_{p,t}\beta_{p,k} \\ &+ \sum_{t=1}^{K} \left(\rho_{S_t} \sum_{l\in\mathcal{S}} \upsilon_{l,t} \left(\beta_{l,k} - \gamma_{l,k}\right)\right) \\ &+ \left(\sqrt{\rho_{S_k}\lambda_k} + \sum_{p\in\mathcal{W}} \sqrt{M\rho_{p,k}\gamma_{p,k}}\right)^2. \end{aligned} \tag{80}$$

The second term of the denominator of SINR can be rewritten as

$$\left|\sum_{l\in\mathcal{S}}\sqrt{\rho_{l,k}}\mathrm{E}\left\{\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{w}_{l,k}^{\mathrm{ZF}}\right\}+\sum_{p\in\mathcal{W}}\sqrt{\rho_{p,k}}\mathrm{E}\left\{\mathbf{h}_{p,k}^{\mathrm{H}}\mathbf{w}_{p,\mathrm{i}_k}^{\mathrm{MRT}}\right\}\right|^2 = \left(\sqrt{\rho_{S_k}\lambda_k}+\sum_{p\in\mathcal{W}}\sqrt{M\rho_{p,k}\gamma_{p,k}}\right)^2. \quad (81)$$

By plugging (14) and (15) into (26) and together with the results of the expected values (73), (80) and (81), we can finally get the SINR of UE k as (27).

REFERENCES

[1] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[2] J. Zuo, J. Zhang, C. Yuen, W. Jiang, and W. Luo, "Energy efficient downlink transmission for multi-cell massive DAS with pilot contamination," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1209–1221, Feb. 2017.

[3] J. Zhang, C. K. Wen, S. Jin, X. Gao, and K. K. Wong, "Large system analysis of cooperative multi-cell downlink transmission via regularized channel inversion with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4801–4813, Oct. 2013.

[4] L. Sanguinetti, R. Couillet, and M. Debbah, "Large system analysis of base station cooperation for power minimization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5480–5496, Aug. 2016.

[5] J. Zhang, E. Björnson, M. Matthaiou, H. Y. D. W. K. Ng, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, p. 1637–1660, Aug. 2020.

[6] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[7] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99 878–99 888, 2019.

[8] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *J. Wireless. Commun. Netw.*, vol. 2019, no. 1, p. 197, Dec. 2019.

[9] 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 Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.

[10] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, July 2017.

[11] H. Q. Ngo, L. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[12] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4758–4774, Jul. 2020.

[13] E. Björnson, J. Hoydis, and L. Sanguinetti, *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*. Foundations and Trends in Signal Processing, 2017, vol. 11, no. 3-4.

[14] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. USA: Prentice Hall PTR, 2001.

[15] J. Vieira, F. Rusek, O. Edfors, S. Malkowsky, L. Liu, and F. Tufvesson, "Reciprocity calibration for massive MIMO: Proposal, modeling, and validation," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3042–3056, May 2017.

[16] 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 Trans. Commun.*, vol. 67, no. 10, pp. 6796–6815, Oct. 2019.

[17] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy efficiency of the cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 971–987, Dec. 2019.

[18] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. USA: Prentice-Hall, Inc., 1993.

[19] T. C. Mai, H. Q. Ngo, M. Egan, and T. Q. Duong, "Pilot power control for cell-free massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11 264–11 268, Aug. 2018.

[20] S. Jin, H. Liu, J. Zhang, and B. Ai, "Graph coloring based pilot assignment for cell-free massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9180–9184, Jun. 2020.

[21] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[23] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *NIPS*, 2015, pp. 379–387.

[24] H. H. Bauschke, M. N. Bui, and X. Wang, "Projecting onto the intersection of a cone and a sphere," *SIAM J. Optim.*, vol. 28, no. 3, pp. 2158–2188, Jan. 2018.

[25] N. Weaver, *Lipschitz Algebras*. WORLD SCIENTIFIC, 1999.

[26] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, Jan. 1988.

[27] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[28] K. Wang, A. M. So, T. Chang, W. Ma, and C. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5690–5705, Nov. 2014.

[29] T. V. Chien, E. Björnson, E. G. Larsson, and T. A. Le, "Distributed power control in downlink cellular massive MIMO systems," in *Proc. ITG WSA*, Mar. 2018, pp. 1–7.

[30] M. Farooq, H. Q. Ngo, and L. N. Tran, "Accelerated projected gradient method for the optimization of cell-free massive MIMO downlink," in *Proc. IEEE PIMRC*, Aug. 2020, pp. 1–6.

[31] C. C. Gonzaga and E. W. Karas, "Complexity of first-order methods for differentiable convex optimization," *Pesquisa Operacional*, vol. 34, no. 3, pp. 395–419, 2014.

[32] 3GPP, "Further advancements for e-utra physical layer aspects (release 9)," 3GPP, (TS) 36.814, Mar. 2017.

[33] Z. Wang, E. K. Tameh, and A. R. Nix, "Joint shadowing process in urban peer-to-peer radio channels," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 52–64, Jan. 2008.

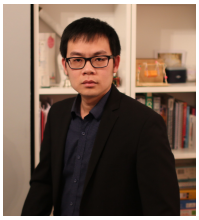
[34] C. Saha, H. S. Dhillon, N. Miyoshi, and J. G. Andrews, "Unified analysis of hetnets using poisson cluster processes under max-power association," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3797–3812, Aug. 2019.

Liutong Du received the B.S. degree communication engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2014, the M.Sc. degree from the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications on electronic and communication engineering, in 2017, and he is currently pursuing the Ph.D. degree in the same laboratory since 2017.

His research interests include cell-free massive MIMO, massive MIMO, non-linear precoding techniques, and physical layer security.

Lihua Li received her B.E. degree and Ph.D. degree in 1999 and 2004 respectively from Beijing University of Posts and Telecommunications (BUPT), Beijing, China. She is currently a professor in BUPT. She had been a visiting scholar at the University of Oulu (Finland) in 2010 and the Stanford University (USA) in 2015.

Her research focuses on MIMO and massive MIMO, cooperative transmission technologies, link adaptation etc. relating to new generation mobile communication systems such as 5G and beyond. She has published 95 papers and 5 books. She has applied 23 national invention patents and one international patent. She was selected and funded as one of the New Century Excellent Talents by the Chinese Ministry of Education in 2008. She won the second prize of State Technological Invention Award (top-3 China national awards) in 2008 and the first prize of China Institute of Communications Science and Technology Award in 2006 for research achievements of "Wideband Wireless Mobile TDD-OFDM-MIMO Technologies".



Hien Quoc Ngo received the B.S. degree in electrical engineering from the Ho Chi Minh City University of Technology, Vietnam, in 2007, the M.S. degree in electronics and radio engineering from Kyung Hee University, South Korea, in 2010, and the Ph.D. degree in communication systems from Linköping University (LiU), Sweden, in 2015. In 2014, he visited the Nokia Bell Labs, Murray Hill, New Jersey, USA. From January 2016 to April 2017, Hien Quoc Ngo was a VR researcher at the Department of Electrical Engineering (ISY), LiU. He was also a Visiting Research Fellow at the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK, funded by the Swedish Research Council.

Hien Quoc Ngo is currently a Lecturer at Queen's University Belfast, UK. His main research interests include massive (large-scale) MIMO systems, cell-free massive MIMO, physical layer security, and cooperative communications. He has co-authored many research papers in wireless communications and co-authored the Cambridge University Press textbook *Fundamentals of Massive MIMO* (2016).

Dr. Hien Quoc Ngo received the IEEE ComSoc Stephen O. Rice Prize in Communications Theory in 2015, the IEEE ComSoc Leonard G. Abraham Prize in 2017, and the Best PhD Award from EURASIP in 2018. He also received the IEEE Sweden VT-COM-IT Joint Chapter Best Student Journal Paper Award in 2015. He was an *IEEE Communications Letters* exemplary reviewer for 2014, an *IEEE Transactions on Communications* exemplary reviewer for 2015, and an *IEEE Wireless Communications Letters* exemplary reviewer for 2016. He was awarded the UKRI Future Leaders Fellowship in 2019. Dr. Hien Quoc Ngo currently serves as an Editor for the IEEE Transactions on Wireless Communications, IEEE Wireless Communications Letters, Digital Signal Processing, Elsevier Physical Communication (PHYCOM), and IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. He was a Guest Editor of IET Communications, special issue on "Recent Advances on 5G Communications" and a Guest Editor of IEEE Access, special issue on "Modelling, Analysis, and Design of 5G Ultra-Dense Networks", in 2017. He has been a member of Technical Program Committees for several IEEE conferences such as ICC, GLOBECOM, WCNC, and VTC.

Trang C. Mai received the B.E. degree (Hons.) in Electrical and Electronic Engineering from Le Quy Don Technical University, Vietnam in 2008, and M.E. degree in Communication Systems from The University of Electro-Communications, Japan in 2013. In 2019, he obtained the Ph.D. degree in Electrical and Electronic Engineering from Queen's University Belfast, UK.

He is currently a Research Fellow at Queen's University Belfast, UK. His research interests include cell-free massive MIMO, massive MIMO, applications of convex optimization, machine learning, and deep learning on Wireless Communications. Dr. Mai serves as a reviewer for the *IEEE Transactions on Communications*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, *IEEE Communications Letters*, and *IEEE Wireless Communications Letters*.

Michail Matthaiou (S'05–M'08–SM'13) was born in Thessaloniki, Greece in 1981. He obtained the Diploma degree (5 years) in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 2004. He then received the M.Sc. (with distinction) in Communication Systems and Signal Processing from the University of Bristol, U.K. and Ph.D. degrees from the University of Edinburgh, U.K. in 2005 and 2008, respectively. From September 2008 through May 2010, he was with the Institute for Circuit Theory and Signal Processing, Munich University of Technology (TUM), Germany working as a Postdoctoral Research Associate. He is currently a Professor of Communications Engineering and Signal Processing and Deputy Director of the Centre for Wireless Innovation (CWI) at Queen's University Belfast, U.K. after holding an Assistant Professor position at Chalmers University of Technology, Sweden. His research interests span signal processing for wireless communications, massive MIMO systems, hardware-constrained communications, mm-wave/THz systems and deep learning for communications.

Dr. Matthaiou and his coauthors received the IEEE Communications Society (ComSoc) Leonard G. Abraham Prize in 2017. He currently holds the ERC Consolidator Grant BEATRICE (2021-2026) focused on the interface between information and electromagnetic theories. He was awarded the prestigious 2018/2019 Royal Academy of Engineering/The Leverhulme Trust Senior Research Fellowship and also received the 2019 EURASIP Early Career Award. His team was also the Grand Winner of the 2019 Mobile World Congress Challenge. He was the recipient of the 2011 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East and Africa Region and a co-recipient of the 2006 IEEE Communications Chapter Project Prize for the best M.Sc. dissertation in the area of communications. He has co-authored papers that received best paper awards at the 2018 IEEE WCSP and 2014 IEEE ICC and was an Exemplary Reviewer for IEEE COMMUNICATIONS LETTERS for 2010. In 2014, he received the Research Fund for International Young Scientists from the National Natural Science Foundation of China. He is currently the Editor-in-Chief of Elsevier Physical Communication, a Senior Editor for IEEE WIRELESS COMMUNICATIONS LETTERS and an Associate Editor for the IEEE JSAC SERIES ON MACHINE LEARNING FOR COMMUNICATIONS AND NETWORKS.