Joint Power and User Grouping Optimization in Cell-Free Massive MIMO Systems

Fengqian Guo, Hancheng Lu, *Senior Member, IEEE*, and Zhuojia Gu

***Abstract*— To relieve the stress on channel estimation and decoding complexity in cell-free massive multiple-input multiple-output (MIMO) systems, user grouping problem is investigated in this paper, where access points (APs) based on time-division duplex (TDD) are considered to serve users on different time resources and the same frequency resource. In addition, when quality of service (QoS) requirements are considered, widely-used max-min power control is no longer applicable. We derive the minimum power constraints under diverse QoS requirements considering user grouping. Based on the analysis, we formulate the joint power and user grouping problem under QoS constraints, aiming at minimizing the total transmit power. A generalized benders decomposition (GBD) based algorithm is proposed, where the primal problem and master problem are solved iteratively to approach the optimal solution. Simulation results demonstrate that by user grouping, the number of users served in cell-free MIMO systems can be as much as the number of APs without increasing the complexity of channel estimation and decoding. Furthermore, with the proposed user grouping strategy, the power consumption can be reduced by 2-3 dB compared with the reference user grouping strategy, and by 7 dB compared with the total transmit power without grouping.**

***Index Terms*— Cell-free systems, massive multiple-input multiple-output (MIMO), time-division duplex (TDD), user grouping, generalized benders decomposition (GBD).**

I. Introduction

THE rapid growth of mobile traffic, especially high volume video traffic, leads to pressing need for high throughput in mobile networks [1]. To cope with such situation, massive multiple-input multiple-output (MIMO) emerges as a promising technique [2]–[5]. In massive MIMO, massive antenna arrays are deployed to simultaneously serve many users on the same time-frequency resource, with which high spectral efficiency is achieved. By distributing numerous antennas in a wide area, the concept of cell-free massive MIMO [4], [6] has been proposed recently and attracted much attention from academic and industrial researchers. Essentially, cell free massive MIMO is an integration of massive MIMO and distributed MIMO, which is expected to exploit benefits of these two techniques. In cell free massive MIMO, many geographically located access points are equipped with single or a few antennas. They serve a much smaller number of users coherently on the same time-frequency resource, ensuring uniformly good quality of service (QoS) for all users. Consequently, cell boundaries are eliminated. Moreover, a central processing unit (CPU) is introduced to coordinate data transmission at different APs through high-capacity backhaul links connecting these APs. Compared with small-cell systems, existing studies have shown that cell-free massive MIMO systems can significantly improve per-user throughput. However, at the cost of much more backhaul overheads [4].

Many research attempts have been done to improve the performance of the cell-free massive MIMO systems. Among them, power control has been addressed, which is globally optimized by CPU to realize uniformly good services for all users in a wide area. The pioneer work on cell-free massive MIMO was done in [4], where max-min power control is performed to maximize the lowest user throughput. After that, the max-min power control problem is investigated under various scenarios [4], [7]–[9]. For conjugate beamforming and zero-forcing (ZF) precoding, low complexity power control algorithms based on the max-min criterion were developed in [7]. In [8], a max-min power control algorithm was proposed with consideration of transceiver hardware impairments. The authors in [9] studied the uplink max-min signal-to-interference-plus-noise ratio problem and obtained a globally optimum solution with an iterative algorithm. In the downlink cell-free massive MIMO systems, power control can be optimized to maximize the energy efficiency [10]. Furthermore, power control has also been jointly considered with load balancing [11], backhaul [12], fronthaul [13], etc.

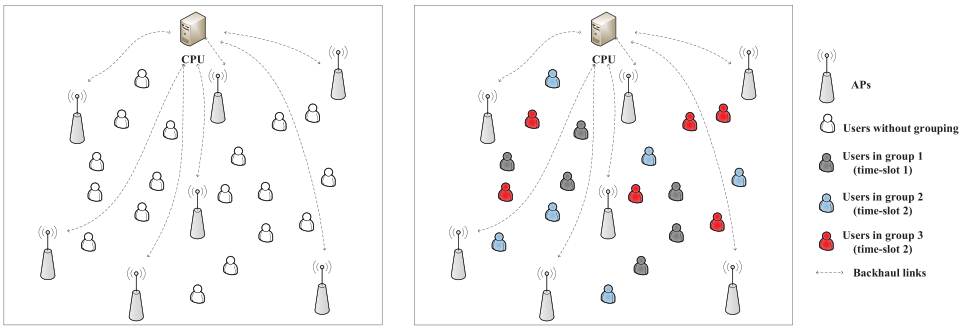
There still remain some deficiencies in research on cell-free massive MIMO. In order to implement beamforming in a hardware-friendly way or to eliminate co-channel interference by zero-forcing precoding, in general, the number of antennas in massive MIMO systems is assumed to be significantly larger than the number of users [14], [15]. Similarly, in cell-free massive MIMO systems, since each AP is assumed to be equipped with one or a few antennas, the number of served users is much smaller than the number of APs [6], [16]. To serve more users, much more APs should be deployed. Correspondingly, the hardware cost and system complexity will be significantly increased [7], [8]. Furthermore, to ensure the accuracy of channel estimation, length of pilot sequence is

Manuscript received December 15, 2020; revised June 14, 2021; accepted July 22, 2021. Date of publication August 4, 2021; date of current version February 14, 2022. This work was supported by the National Science Foundation of China under Grant 61771445, Grant 61631017, and Grant 91538203. The associate editor coordinating the review of this article and approving it for publication was A. S. Cacciapuoti. *(Corresponding author: Hancheng Lu.)*

The authors are with the CAS Key Laboratory of Wireless-Optical Communications, University of Science and Technology of China, Hefei 230027, China (e-mail: fqguo@mail.ustc.edu.cn; hclu@ustc.edu.cn; guzj@mail.ustc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3100573.

Digital Object Identifier 10.1109/TWC.2021.3100573



(a) Users sharing all time-slots. (b) Users assigned different time-slots and divided into groups.

Fig. 1. Illustration of Downlink Cell-Free Massive MIMO System.

usually assumed no less than the number of users [10], [17]. However, the number of samples in each coherence interval are limited. Hence, the length of data in each coherence interval will reduce as the number of users increase. Additionally, users have diverse QoS requirements. Requirement satisfaction is more important for users than fairness. In this case, widely-used max-min power control is no longer applicable.

To address these issues, in this paper, we investigate the time-division duplex (TDD) based cell-free massive MIMO systems [13]. Users are divided into different groups according to their assigned time-slots, then channel estimation and decoding are applied within each group. By doing so, both pilot overheads and system complexity can be significantly reduced. In addition, we perform power allocation to satisfy the QoS requirements of users, instead of max-min power control. The main contributions are described as follows.

- In the downlink TDD based cell-free massive MIMO systems considering user grouping, we first introduce and analyze the main processes of uplink training and downlink payload data transmission. Then we derive the minimum power constraints under diverse QoS requirements after user grouping. Based on our analytical work and the transmission process with user grouping, we formulate the joint power allocation and user grouping problem with both conjugate and ZF beamforming under user QoS constraints, with the goal to minimize the total transmit power.
- We convert the problem into a form that can be handled by generalized benders decomposition (GBD) method. With GBD, we first decompose the problem into the primal problem (i.e., power allocation problem) and master problem (i.e., user grouping problem). Particularly, the user grouping problem is relaxed and then the relaxed problem is converted into a problem of searching for some special negative loops in a graph composed of users.
- Based on the GBD method, we propose an iterative algorithm, which is feasible for both conjugate and ZF beamforming, to approach the optimal solution to the converted joint power allocation and user grouping problem. In each iteration, the upper bound and lower bounds are obtained by solving the primal problem and the master problem, respectively. The gap between these two bounds is reduced iteratively. Therefore, the proposed iterative algorithm is provably convergent. Furthermore, to solve the master problem within polynomial time, a fast greedy suboptimal algorithm is proposed.

Simulation results validate the convergence and optimality of the proposed algorithms, and demonstrate that by user grouping, the number of users served in cell-free MIMO systems can be as much as the number of APs without increasing the complexity of channel estimation and decoding. Furthermore, with the proposed user grouping strategy, the power consumption can be reduced by 2-3 dB compared with the random user grouping strategy, and by 7 dB compared with the total transmit power without grouping.

The rest of this paper is organized as follows. In Section II, we give the model of the downlink cell-free massive MIMO systems, and formulate the joint optimization problem of power allocation and user grouping to minimize the total transmit power. In Section III, we decompose the problem into a power allocation problem and a user grouping problem. Problem analysis and solutions are also described. In Section IV, we relax and solve the master problem based on graph theory. The system performance is evaluated in Section V. Finally, we give the conclusion in Section VI.

Notations: Vectors and sets are denoted by bold letters. $\lceil \cdot \rceil$ denotes the ceiling function. $\mathbf{A}^H$ and $\mathbf{A}^*$ denote the conjugate transpose and conjugate of $\mathbf{A}$, respectively.

II. Systems Model and Problem Formulation

We consider a downlink TDD based cell-free massive MIMO system where M single-antenna APs and N single-antenna users are randomly located in a wide area as shown in Fig. 1. In the traditional cell-free massive MIMO system where all users sharing all coherence intervals or time-slots as shown in Fig. 1(a). To relieve the stress on channel estimation and decoding complexity, we divide users in groups according to their assigned time-slots. Users assigned the same time-slot form a group. The time-slots assigned to different groups are assumed to be orthogonal. In this paper, we assume that the number of users is greater than the number of groups. In Fig. 1(b), users are grouped into 3 groups and served on 3 orthogonal time-slots, respectively. Both channel estimation and decoding are performed within the group. There are two types of training, i.e., large-scale training and uplink training. The result of large-scale training is assumed to be accurate. The interval between two times of large-scale training is named as τ_{Lc}, and the interval between two times of uplink

TABLE I
MAIN NOTATIONS

Symbol	Description
G	Number of groups
N	Number of users
M	Number of AP
p_{mn}	Power control coefficient of the transmit power that AP m allocated to user n
q_{mn}	$\sqrt{p_{mn}}$
γ_n	Target SINR of user n
$\boldsymbol{x}$	User grouping matrix
P_t	Total transmit power
β_{mn}	large-scale fading between AP m and user n

training is named as τ_c. Each τ_{Lc} is composed of one large-scale training phase and some time-slots, and each time-slot contains one coherence interval. In general, we assume that $\tau_{Lc} \gg \tau_c$ [13].

The channel between AP m and user k on time-slot g is modeled as $h_{gmn} = \sqrt{\beta_{mn}}\varsigma_{gmn}, 1 \leq m \leq M, 1 \leq g \leq G, 1 \leq n \leq N$, where β_{mn} and $\varsigma_{gmn} \sim \mathcal{CN}(0,1)$ denote the large-scale fading and small-scale fading between AP m and user n on time-slot g, respectively. In the remainder of this paper, we assume that large-scale fading β_{mn} is known to all APs and users. Considering that the channels in different time-slots do not completely independent with each other, in this paper, the large-scale fading is assumed to remain constant across all time-slots between two large-scale training. Some major notations are listed in Table I.

After user grouping, cell-free massive MIMO transmission in each group within a coherence interval consists of three phases: uplink training, uplink payload data transmission and downlink payload data transmission as shown in Fig. 1. In this paper, we focus on the joint optimization of user grouping and power allocation in downlink cell-free massive MIMO system. Therefore, uplink training phase and downlink payload data transmission phase are introduced as follows.

A. Uplink Training

Instantaneous downlink channel state information (CSI) is needed at APs for beamforming in following downlink payload data transmission. So in the uplink training phase of this system, each user need to send pilot sequences simultaneously to APs for channel estimation on the same assigned time-slot. In this paper, the length of pilot sequence in group g is τ_g, which is the main overhead of downlink channel estimation at the users and increases linearly with the number of users in each group. To ensure the channel estimation accuracy, the length of pilot sequence is usually assumed no less than the number of users [10], [17]. Note that user grouping can effectively increase the number of symbols for data transmission. We assume that the length of coherence interval is τ_c and the length of pilot sequence in group g is τ_g.

The effect of coherent time on the performance of the proposed system mainly includes two aspects. One is the number of symbols for data transmission in each time-slot, and the other one is the data rate of users. In detail, considering the case that the number of users sharing the same time-slot is settled, the length of pilot sequence in this time-slot will be settled. Then the longer the coherence time, the longer is the efficient time of each channel estimation result and the length of coherence interval. The length of coherence interval is the sum of the length of pilot sequence and the number of symbols for data transmission in each time-slot. So the number of symbols for data transmission in each time-slot increases with the coherence time. On the other hand, the longer coherent time, the longer is the proportion of the symbols for data transmission. In each group or time-slot, the length of pilot sequences required for channel estimation increases linearly with the number of users sharing the same time-frequency resource. Without user grouping, the number of symbols for data transmission within a coherent interval will be only $\tau_c - \sum_{g=1}^{G} \tau_g$. After user grouping, the number of users in each time slot will decrease, and hence the length of pilot sequences required for channel estimation in each coherence interval will also decrease. The number of symbols for data transmission within a coherent interval is $\tau_c - \tau_g$. In other words, the length of pilot sequence in each time-slot will be reduced by user grouping. For example: In a downlink TDD based cell-free massive MIMO system with $\tau_c = 100$ and 50 users, the length of pilot sequences required for channel estimation in each coherence interval is 50, and the proportion of the symbols for data transmission in each coherence interval is $50/100 = 0.5$. If we assign these users into 5 groups, the users sharing the same coherence interval will be 10, then the length of pilot sequences required for channel estimation in each coherence interval is 10, and the proportion of the symbols for data transmission in each coherence interval will up to $(100-10)/100 = 0.9$, which means that more time slots is used for downlink payload data transmission. Furthermore, the fewer users sharing the same coherence interval, the less interference that users will suffer and the lower decoding difficulty for the receiver.

We assume that the user grouping information (containing which group to access and the number of users in this group) has been acquired at each user. Let $\boldsymbol{\psi}_{gn} \in \mathbb{C}^{\tau_g}$ denote the pilot sequence of the user n if user n is assigned into group g and satisfies $\|\boldsymbol{\psi}_{gn}\|^2 = 1$. Then with a given user grouping matrix $\boldsymbol{x} = [x_{gn}]_{1 \leq g \leq G, 1 \leq n \leq N}$, where x_{gn} is a 0-1 variable ($x_{gn} = 1$ denotes that user n is assigned into group g and $x_{gn} = 0$ denotes that user n is not assigned into group g), the pilot signal that AP m receives on time-slot g is $\boldsymbol{y}_{mg}^p = \sum_{n=1}^{N} x_{gn}\boldsymbol{\psi}_{gn}\sqrt{\rho_r \tau_g} h_{gmn} + \mathfrak{n}_m^g$, where ρ_r is the power of pilot signal from users, and $\mathfrak{n}_m^g \sim (0, \sigma^2)$ is the Additive White Gaussian Noise (AWGN) received at APs. Let $\hat{h}_{gmn}$ denote the minimum mean square error (MMSE) estimate of h_{gmn}. Assuming that user n is assigned into group g, we have $\hat{h}_{gmn} = \frac{\mathbb{E}[(\boldsymbol{\psi}_{gn}^H \boldsymbol{y}_{mg}^p)^* h_{gmn}]}{\mathbb{E}[|\boldsymbol{\psi}_{gn}^H \boldsymbol{y}_{mg}^p|^2]} \boldsymbol{\psi}_{gn}^H \boldsymbol{y}_{mg}^p = \frac{\sqrt{\rho_r \tau_g} h_{gmn}}{\sigma^2 + \rho_r \tau_g \beta_{gmn}} \boldsymbol{\psi}_{gn}^H \boldsymbol{y}_{mg}^p, 1 \leq m \leq M, 1 \leq g \leq G, 1 \leq n \leq N$. According to [7], its distribution is

$$\hat{h}_{gmn} \sim \mathcal{CN}(0, \alpha_{gmn}), \quad 1 \leq m \leq M, 1 \leq g \leq G, 1 \leq n \leq N. \tag{2}$$

where $\alpha_{gmn} = \frac{\rho_r \tau_g \beta_{mn}^2}{\sigma^2 + \rho_r \tau_g \beta_{mn}}$. In this paper, the length of pilot signal of users in group g, i.e., τ_g, is equal to the number of users in group g. Then α_{gmn} can be rewritten as

$$\alpha_{gmn} = \frac{\rho_r \sum_{i=1}^{N} x_{gi} \beta_{mn}^2}{\sigma^2 + \rho_r \sum_{i=1}^{N} x_{gi} \beta_{mn}}, \quad 1 \le m \le M, 1 \le g \le G, 1 \le n \le N. \tag{3}$$

Therefore, α_{gmn} is up to $\boldsymbol{x}$ and ρ_r when large-scale fading is given. In addition, the distribution of channel estimation error is

$$h_{gmn} - \hat{h}_{gmn} \sim \mathcal{CN}(0, \beta_{mn} - \alpha_{gmn}), \quad 1 \le m \le M, 1 \le g \le G, 1 \le n \le N. \tag{4}$$

B. Downlink Payload Data Transmission

After uplink training, APs will send the results of channel estimation to CPU as Fig.1 shows. CPU calculates power control coefficients p_{mn} (AP m allocated to user n) and sends these coefficients to each AP. We w.l.o.g. assume that the users in group g have been sorted and numbered by their channel gains. The index of user n in its group is denoted by κ_n. Next we will introduce the conjugate beamforming (i.e., maximum-ratio transmission, MRT) and the ZF beamforming for precoding. First, we define the coefficients for beamforming as follows:

$$b_{gmn} = \begin{cases} \hat{h}_{gmn}^* & \text{MRT} \\ [\hat{\boldsymbol{H}}_g^* (\hat{\boldsymbol{H}}_g^T \hat{\boldsymbol{H}}_g^*)^{-1}]_{m\kappa_n}, & \text{ZF}, \end{cases} \tag{5}$$

where $\boldsymbol{H}_g$ denotes the $M \times \sum_{i=1}^{N} x_{gi}$ channel coefficient matrix, $[\boldsymbol{H}_g]_{m\kappa_n} = h_{gmn}$. It should be noted that, the real transmit power from AP m to user n is $p_{mn}|b_{gmn}|^2$, where b_{gmn} is the beamforming coefficient and p_{mn} is the power control coefficient. Although the value of p_{mn} from different APs to each user are equal, the value of b_{gmn} from different APs to each user are different. Therefore, the real transmit power from different APs to each user, i.e., $p_{mn}|b_{gmn}|^2$, are different.

1) Conjugate Beamforming: The signal transmitted from each AP on time-slot g after conjugate beamforming is given by $t_{gm} = \sum_{n=1}^{N} x_{gn} \sqrt{p_{mn}} b_{gmn} s_n$ [18]–[20], where p_{mn} denotes power control coefficient of transmit power that AP m allocated to user n, s_n represents the transmitted symbol of user n. The signal received at user n, using MRT, is given by[1]

$$\begin{aligned} y_n^{MRT} &= \sum_{g=1}^{G} \sum_{m=1}^{M} x_{gn} h_{gmn} t_{gm} + \mathfrak{n}_n^g \\ &= \sum_{g=1}^{G} \sum_{m=1}^{M} x_{gn} h_{gmn} \sum_{i=1}^{N} x_{gi} \sqrt{p_{mi}} b_{gmi} s_i + \mathfrak{n}_n^g, \\ &\qquad 1 \le n \le N. \end{aligned} \tag{6}$$

where $\mathfrak{n}_m^g \sim (0, \sigma^2)$ is the AWGN received at users. For convenience, according to [7], we divide y_n into five parts, including the desired signal of user n Y_{us}, interference from the desired signals of the other users in the same group Y_{in}, the channel estimation error Y_{ce}, the lack of channel knowledge at user Y_{lc} and noise $\mathfrak{n}_n^g$, and define the first four parts as follows.

$$\begin{aligned} Y_{ce} &= \sum_{g=1}^{G} \sum_{m=1}^{M} \sum_{i=1}^{N} x_{gn} x_{gi} \sqrt{p_{mi}} (h_{gmn} - \hat{h}_{gmn}) \hat{h}_{gmi}^* s_i, \\ Y_{us} &= \sum_{g=1}^{G} \sum_{m=1}^{M} x_{gn} \sqrt{p_{mn}} \mathbb{E}[|\hat{h}_{gmn}|^2] s_n, \\ Y_{lc} &= \sum_{g=1}^{G} \sum_{m=1}^{M} x_{gn} \sqrt{p_{mn}} (|\hat{h}_{gmn}|^2 - \mathbb{E}[|\hat{h}_{gmn}|^2]) s_n, \\ Y_{in} &= sum_{g=1}^{G} \sum_{m=1}^{M} \sum_{\substack{i=1 \\ i \ne n}}^{N} x_{gn} x_{gi} \sqrt{p_{mi}} \hat{h}_{gmn} \hat{h}_{gmi}^* s_i. \end{aligned}$$

Then (6) can be rewritten as $y_n^{MRT} = Y_{us} + Y_{in} + Y_{ce} + Y_{lc} + \mathfrak{n}_n^g$. As these five parts are mutually uncorrelated, the lower bound of SINR achievable to user n is [7], where the time interval is the coherence time (a time slot), hence the noise in this expression is one σ^2. It should be noted that $\sum_{g=1}^{G} x_{gn} = 1$, for $\forall n$. Hence, the SINR of users in different groups will not affect with each other.

Obviously, interference from the desired signals of the other users in the same group will be reduced as the number of

[1]Although the strict phase-synchronization and -calibration between APs are not needed for non-coherent receivers, the coherent joint transmission can achieve higher spectral efficiency than non-coherent joint transmission. In the system model of this paper, the pilot sequences is sent by users and received by APs. So the channel estimation is only available at APs and the users only knows the large-scale channel gain. With coherent joint transmission, the user don't need to distinguish the signals from different APs. Based on the above considerations, the coherent joint transmission is adopted in this paper.

$$\begin{aligned} \mathrm{SINR}_n^{MRT} &= \frac{\mathbb{E}[|Y_{us}|^2]}{\sigma^2 + \mathbb{E}[|Y_{in}|^2] + |Y_{ce}|^2] + |Y_{lc}|^2]} = \frac{\mathbb{E}[|Y_{us}|^2]}{\sigma^2 + \mathbb{E}[|Y_{in}|^2] + \mathbb{E}[|Y_{ce}|^2] + \mathbb{E}[|Y_{lc}|^2]} \\ &= \frac{\left(\sum_{m=1}^{M} \sqrt{p_{mn}} \sum_{g=1}^{G} x_{gn} \alpha_{gmn} \right)^2}{\sigma^2 + \sum_{g=1}^{G} x_{gn} \sum_{\substack{i=1 \\ i \ne n}}^{N} x_{gi} \sum_{m=1}^{M} p_{mi} \alpha_{gmn} \alpha_{gmi} + \sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} p_{mi} (\beta_{mn} - \alpha_{gmn}) \alpha_{gmi} + \sum_{g=1}^{G} x_{gn} \sum_{m=1}^{M} p_{mi} \alpha_{gmn}^2} \\ &= \frac{\left(\sum_{m=1}^{M} \sqrt{p_{mn}} \sum_{g=1}^{G} x_{gn} \alpha_{gmn} \right)^2}{\sigma^2 + \sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} p_{mi} \beta_{mn} \alpha_{gmi}}, \quad 1 \le n \le N \end{aligned} \tag{1}$$

users in each group decreases, and the complexity of decoding at the receiver will be relatively reduced. To satisfy the user QoS requirements, the achievable SINR of each user should be constrained by $\mathrm{SINR}_n^{MRT} \geq \gamma_n, \quad 1 \leq n \leq N, 1 \leq g \leq G$ where $\gamma_n = 2^{GR_n^{target}\frac{\tau_c}{\tau_c-\tau_g}} - 1$ is the minimal SINR in the transmit time slot of user n to achieve its target data rate R_n^{target}, and $\frac{\tau_c}{\tau_c-\tau_g}$ is the ratio of the length of data in each coherence interval after grouping and that of data in each coherence interval without grouping.

2) Zero-Forcing Precoder: With zero-forcing precoder, the signal transmitted from each AP on time-slot g is given by $t_{gm} = \sum_{n=1}^{N} x_{gn}\sqrt{p_n} b_{gmn} s_n$, where p_n is the power allocation coefficient of user n under the assumption of $p_{mn} = p_n$, for $\forall m$. The signal received at user n, using ZF, is given by $y_n^{ZF} = \sum_{g=1}^{G}\sum_{m=1}^{M} x_{gn} h_{gmn} t_{gm} + \mathfrak{n}_n^g, \quad 1 \leq n \leq N$.

Since interference from the desired signals of the other users in the same group has been eliminated by zero-forcing precoder, the lower bound of SINR achievable to user n is [7], [21], [22] $\mathrm{SINR}_n^{ZF} = \frac{p_n}{\sigma^2 + \sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} p_i \eta_{ni}}$, where η_{ni} is the κ_i-th element of the following vector: $\boldsymbol{\eta}_n = diag\{\mathbb{E}((\hat{\boldsymbol{H}}_g^T \hat{\boldsymbol{H}}_g^*)^{-1} \hat{\boldsymbol{H}}_g^T \mathbb{E} \quad (\hat{\boldsymbol{h}}_{g\kappa_n}^* \hat{\boldsymbol{h}}_{g\kappa_n}^T)\hat{\boldsymbol{H}}_g^*(\hat{\boldsymbol{H}}_g^T \hat{\boldsymbol{H}}_g^*)^{-1})\}$, where $\hat{\boldsymbol{h}}_{g\kappa_n}^T = [\hat{h}_{g1n}, \ldots \hat{h}_{gmn} \ldots \hat{h}_{gMn}]$, and $\mathbb{E}(\hat{\boldsymbol{h}}_{g\kappa_n}^* \hat{\boldsymbol{h}}_{g\kappa_n}^T)$ is a diagonal matrix with $(\beta_{mn} - \alpha_{gmn})$ on its m-th diagonal element, which has been proved in [22]. The value of η_{ni} can be obtained using exponential smoothing as stated in [7]. Exponential smoothing is a method to predict the future value of a variable by weighting its past values considering the change trend of its value. In this paper, the historical values of η_{ni} can be obtained from previous channel estimation results, hence we can obtain the predicted value of η_{ni} by weighting its historical values. We assume that the accurate and the estimated value of the current η_{ni} is η_{ni}^t and $\eta_{ni}^{(t)}$, respectively. and η_{ni}^{t-1} and $\eta_{ni}^{(t-1)}$ denote the accurate and the estimated value of the $(t-1)$-th η_{ni}. Then $\eta_{ni}^{(t)} = w\eta_{ni}^{t-1} + (1-w)\eta_{ni}^{(t-1)}$, where w is a constant between 0 and 1.

To get a similar form of SINR_n^{MRT} in (1), shown at the bottom of previous page, SINR_n^{ZF} can be rewritten as:

$$\mathrm{SINR}_n^{ZF} = \frac{\left(\sum_{m=1}^{M} \frac{1}{M}\sqrt{p_{mn}}\right)^2}{\sigma^2 + \sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} \frac{1}{M} p_{mi}\eta_{ni}}, \quad 1 \leq n \leq N \Big| p_{mn} = p_n. \tag{7}$$

C. Problem Formulation

According to (1) and (7), the values of SINR_n^{MRT} and SINR_n^{ZF} are up to the user grouping matrix $\boldsymbol{x} = [x_{gn}]_{1\leq g\leq G, 1\leq n\leq N}$, power allocation matrix $\boldsymbol{p} = [p_{mn}]_{1\leq m\leq M, 1\leq n\leq N}$, and large-scale fading matrix $\boldsymbol{\beta} = [\beta_{mn}]_{1\leq m\leq M, 1\leq n\leq N}$ (The value of α_{gmn} is up to $\boldsymbol{x}$ if ρ_τ and $\boldsymbol{\beta}$ are given as stated in subsection II-A). We try to optimize the user grouping and power allocation strategy to minimize the total transmit power with known large-scale fading under QoS constraints, i.e., target SINR. We formulate this joint power allocation and user grouping problem as $\mathcal{P}1$.

$$\mathcal{P}1: \min_{\boldsymbol{x},\boldsymbol{p}} \ P_t = \sum_{m=1}^{M}\sum_{n=1}^{N} p_{mn} \sum_{g=1}^{G} x_{gn}\varphi_{gmn} \tag{8a}$$

$$\text{s.t.} \ \frac{\left(\sum_{m=1}^{M}\sqrt{p_{mn}}\sum_{g=1}^{G} x_{gn}\vartheta_{gmn}\right)^2}{\sigma^2 + \sum_{g=1}^{G} x_{gn}\sum_{i=1}^{N} x_{gi}\sum_{m=1}^{M} p_{mi}\upsilon_{gmni}} \geq \gamma_n, \quad 1 \leq n \leq N, \tag{8b}$$

$$p_{mn} \geq 0, \quad 1 \leq n \leq N, 1 \leq m \leq M, \tag{8c}$$

$$\sum_{g=1}^{G} x_{gn} = 1, \quad x_{gn} \in \{0,1\}, 1 \leq n \leq N, \tag{8d}$$

$$p_{mn} = p_{m'n}, \quad 1 \leq m, M' \leq M \Big| ZF$$

$$(3), (9) - (11). \tag{8e}$$

where (8d) means that each user should be assigned into only one group, (8e) exists when ZF beamforming is chosen. Like (5), some variables are defined as follows:

$$\varphi_{gmn} = \begin{cases} \alpha_{gmn} & \text{MRT} \\ [diag\{\mathbb{E}((\hat{\boldsymbol{H}}_g^T \hat{\boldsymbol{H}}_g^*)^{-1}\hat{\boldsymbol{h}}_{[gm]}^*\hat{\boldsymbol{h}}_{[gm]}^T(\hat{\boldsymbol{H}}_g^T\hat{\boldsymbol{H}}_g^*)^{-1})\}]_{\kappa_n} & \text{ZF}, \end{cases} \tag{9}$$

where $\hat{\boldsymbol{h}}_{gm}$ is the m-th row of $\hat{\boldsymbol{H}}_g$.

$$\vartheta_{gmn} = \begin{cases} \alpha_{gmn} & \text{MRT} \\ \frac{1}{\sqrt{M}}, & \text{ZF}, \end{cases} \tag{10}$$

$$\upsilon_{gmni} = \begin{cases} \beta_{mn}\alpha_{gmi} & \text{MRT} \\ \frac{1}{\sqrt{M}}\eta_{ni}, & \text{ZF}, \end{cases} \tag{11}$$

To solve this MINLP problem, we first define a $M \times N$ matrix $\boldsymbol{q} = [q_{mn}]_{1\leq m\leq M, 1\leq n\leq N}$, where $q_{mn} = \sqrt{p_{mn}}$, and convert problem $\mathcal{P}1$ into problem $\mathcal{P}2$:

$$\mathcal{P}2: \min_{\boldsymbol{x},\boldsymbol{q}} \ P_t = \sum_{m=1}^{M}\sum_{n=1}^{N} q_{mn}^2 \sum_{g=1}^{G} x_{gn}\varphi_{gmn}, \tag{12a}$$

$$\text{s.t.} \ \left(\sigma^2 + \sum_{g=1}^{G} x_{gn}\sum_{i=1}^{N} x_{gi}\sum_{m=1}^{M} q_{mi}^2\upsilon_{gmni}\right)^{\frac{1}{2}}\gamma_n^{\frac{1}{2}} - \sum_{m=1}^{M} q_{mn}\sum_{g=1}^{G} x_{gn}\vartheta_{gmn} \leq 0, \quad 1 \leq n \leq N, \tag{12b}$$

$$q_{mn} \geq 0, \quad 1 \leq n \leq N, 1 \leq m \leq M, \tag{12c}$$

$$q_{mn} = q_{m'n}, \quad 1 \leq m, M' \leq M \Big| ZF$$

$$(3), (9)\text{-}(11), (8d). \tag{12d}$$

A key motivation of this conversion is to convert constraints (8b) into convex constraints (12b) with given user grouping matrix $\boldsymbol{x}$, which will be introduced in the next section. Problem $\mathcal{P}2$ is still hard to solve, we try to solve it with an iterative method based on GBD method [23], [24].

III. Problem Analysis and Solutions

We first decompose this problem into a primal problem: power allocation problem and a master problem: user grouping problem according to GBD method [23], [24]. Then according to the basic principle of GBD method, the MINLP problem can be solved by solving these two problems iteratively [25]. In each iteration, the upper bound and the lower bound of the problem can be updated, and the gap among the upper and lower bound is shrunk [26].

A. Primal Problem: Power Allocation Problem

Power allocation problem $\mathcal{S}^{(k)}$ is given by fixing the user grouping matrix to $\boldsymbol{x}^{(k)}$:

$$\mathcal{S}^{(k)}: \min_{\boldsymbol{q}} \ P_t = \sum_{m=1}^{M}\sum_{n=1}^{N} q_{mn}^2 \sum_{g=1}^{G} x_{gn}^{(k)} \varphi_{gmn}^{(k)} \tag{13a}$$

$$\text{s.t.} \ \left(\sigma^2 + \sum_{g=1}^{G} x_{gn}^{(k)} \sum_{i=1}^{N} x_{gi}^{(k)} \sum_{m=1}^{M} q_{mi}^2 \upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}} \gamma_n^{\frac{1}{2}} - \sum_{m=1}^{M} q_{mn} \sum_{g=1}^{G} x_{gn}^{(k)} \vartheta_{gmn}^{(k)} \le 0, \quad 1 \le n \le N,$$

$$(3),(12c),(9)\text{-}(11),(12d). \tag{13b}$$

Since the objective function (13a) is convex, the SINR constraints (13b) are second order cone (SOC) constraints, the constraints (12c) and (12c) are linear and the other constraints are not related to the value of $\boldsymbol{q}$, problem $\mathcal{S}^{(k)}$ is a convex problem.

Since problem $\mathcal{S}^{(k)}$ is a convex problem, we can solve it by the interior point method. In addition, problem $\mathcal{S}^{(k)}$ is given by fixing the user grouping matrix to $\boldsymbol{x}^{(k)}$. Hence, there are two cases about this problem, feasible and infeasible. These cases of problem $\mathcal{S}^{(k)}$ are discussed as follows:

1) Feasible Case: We first define the partial Lagrangian function of problem $\mathcal{S}^{(k)}$ [27]:

$$\begin{aligned}&\mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{x}^{(k)}) \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N} q_{mn}^2 \sum_{g=1}^{G} x_{gn}^{(k)} \varphi_{gmn}^{(k)} \\ &\quad + \sum_{n=1}^{N} \lambda_n \left(\left(\sigma^2 + \sum_{g=1}^{G} x_{gn}^{(k)} \sum_{i=1}^{N} x_{gi}^{(k)} \sum_{m=1}^{M} q_{mi}^2 \upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}} \gamma_n^{\frac{1}{2}} - \sum_{m=1}^{M} q_{mn} \sum_{g=1}^{G} x_{gn}^{(k)} \vartheta_{gmn}^{(k)} \right),\end{aligned} \tag{14}$$

where the Lagrangian multipliers $\boldsymbol{\lambda} = [\lambda_{mn}]_{1 \le m \le M, 1 \le n \le N}$ correspond to constraints (13b) and satisfy $\lambda_{mn} \ge 0,\ 1 \le m \le M,\ 1 \le n \le N$. The dual problem of problem $\mathcal{S}^{(k)}$ can be obtained as stated in Lemma 1.

Lemma 1: Problem $\mathcal{S}^{(k)}$ is equivalent to its dual problem $\mathcal{D}_S^{(k)}$ as follows [28].

$$\mathcal{D}_S^{(k)}: \max_{\boldsymbol{\lambda}} \inf_{\boldsymbol{q}} \ \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{x}^{(k)}) \tag{15a}$$

$$\text{s.t. } \lambda_{mn} \ge 0, \quad 1 \le m \le M,\ 1 \le n \le N, \quad (3),(12c),(12d). \tag{15b}$$

Proof: It is obvious that there exists a strictly feasible point for convex problem $\mathcal{S}^{(k)}$ (feasible case), thus Slater's condition is satisfied [29], [30]. Hence, strong duality holds for problem $\mathcal{S}^{(k)}$ and its dual problem [31]. □

2) Infeasible Case: According to Lemma 1, we can get an upper bound of problem $\mathcal{P}2$. Then if problem $\mathcal{S}^{(k)}$ is infeasible, which means that constraint (13b) cannot be satisfied no matter how we allocate power, we try to find the power allocation strategy that is close to constraint (13b). By relaxing constraint (13b) with a violation variable ϕ, we can get the following problem $\mathcal{S}2^{(k)}$ as well as its dual problem $\mathcal{D}_{S2}^{(k)}$ [32]. In problem $\mathcal{S}2^{(k)}$, we try to minimize the gap between the left and right sides of (13b).

$$\mathcal{S}2^{(k)}: \min_{\boldsymbol{q},\phi} \ \phi \tag{16a}$$

$$\text{s.t.} \ \left(\sigma^2 + \sum_{g=1}^{G} x_{gn}^{(k)} \sum_{i=1}^{N} x_{gi}^{(k)} \sum_{m=1}^{M} q_{mi}^2 \upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}} \gamma_n^{\frac{1}{2}} \tag{16b}$$

$$- \sum_{m=1}^{M} q_{mn} \sum_{g=1}^{G} x_{gn}^{(k)} \vartheta_{gmn}^{(k)} \le \phi, \quad 1 \le n \le N, \tag{16c}$$

$$\phi \ge 0, \quad (3),(12c),(12d). \tag{16d}$$

We define the partial Lagrangian function of $\mathcal{S}2^{(k)}$ as follows:

$$\begin{aligned}&\mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}, \boldsymbol{x}^{(k)}) \\ &= \sum_{n=1}^{N} \nu_n \left(\left(\sigma^2 + \sum_{g=1}^{G} x_{gn}^{(k)} \sum_{i=1}^{N} x_{gi}^{(k)} \sum_{m=1}^{M} q_{mi}^2 \upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}} \gamma_n^{\frac{1}{2}} - \sum_{m=1}^{M} q_{mn} \sum_{g=1}^{G} x_{gn}^{(k)} \vartheta_{gmn}^{(k)} \right),\end{aligned} \tag{17}$$

where the Lagrangian multipliers $\boldsymbol{\nu}$ correspond to constraints $(16c)$ and satisfy $\nu_{mn} \ge 0,\ 1 \le m \le M,\ 1 \le n \le N$. Like problem $\mathcal{S}^{(k)}$, we can also get the optimal solutions $\boldsymbol{q}^{(k)}$ and the dual solutions $\boldsymbol{\nu}^{(k)}$ by the interior point method. In addition, Lemma 2 is obtained.

Lemma 2: Problem $\mathcal{S}2^{(k)}$ is equivalent to its dual problem $\mathcal{D}_{S2}^{(k)}$ as follows.

$$\mathcal{D}_{S2}^{(k)}: \max_{\boldsymbol{\nu}} \inf_{\boldsymbol{q},\phi} \mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}, \boldsymbol{x}^{(k)}) + \phi - \sum_{n=1}^{N} \nu_n \phi \tag{18a}$$

$$\text{s.t. } \nu_{mn} \ge 0, \quad 1 \le m \le M,\ 1 \le n \le N, \quad (16d),\ (3),\ (12c),(12d). \tag{18b}$$

Proof: Since the objective function (18a) is convex and all the constraints are linear, problem $\mathcal{S}2^{(k)}$ is convex. In addition, for any $\boldsymbol{q} \succ 0$ and any ϕ satisfies (16c), $\{\boldsymbol{q}, \phi\}$ is feasible for convex problem $\mathcal{S}2^{(k)}$, thus Slater's condition is satisfied. Hence, strong duality holds for problem $\mathcal{S}2^{(k)}$ and its dual problem. □

B. Master Problem: User Grouping Problem

We write the master user grouping problem $\mathcal{M}1$ as follows:

$$\begin{aligned}\mathcal{M}1: \ &\min_{\boldsymbol{x}} \ \mathrm{f}(\boldsymbol{x}) \\ &\text{s.t. } \boldsymbol{x} \in \{\boldsymbol{x}^{(k)} | \mathcal{S}^{(k)} \text{ is feasible}\}, \quad 1 \le n \le N, \\ &\quad \text{(8d)}. \end{aligned} \tag{19a}$$

where function $\mathrm{f}(\boldsymbol{x})$ returns the optimal value of problem $\mathcal{S}^{(k)}|_{\boldsymbol{x}^{(k)}=\boldsymbol{x}}$.

In problem $\mathcal{M}1$, (19a) is not in the explicit form. Therefore, to apply GBD method, we convert the master problem $\mathcal{M}1$ into an explicit form in the following lemma.

Lemma 3: Problem $\mathcal{M}1$ is equivalent to the following problem $\mathcal{M}2$.

$$\begin{aligned}\mathcal{M}2: \ &\min_{\boldsymbol{x}} \ \xi \\ &\textit{s.t.} \ \min_{\boldsymbol{q}\succeq 0} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{x}) \le \xi, \forall \boldsymbol{\lambda} \succeq 0, \end{aligned} \tag{20a}$$

$$\min_{\boldsymbol{q}\succeq 0} \mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}, \boldsymbol{x}) \le 0, \forall \boldsymbol{\nu} \succeq 0 : \sum_{n=1}^{N} \nu_n = 1, \quad \textit{(8d))}. \tag{20b}$$

Proof: Since the constraints (13b) are convex, according to Theorem 2.2 in [33], the constraints (19a) are equivalent to (20b). Then according to Lemma 1, we have:

$$\mathrm{f}(\boldsymbol{x}) = \max_{\boldsymbol{\lambda}\succeq \mathbf{0}} \ \min_{\boldsymbol{q}\succeq 0} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{x}). \tag{21}$$

Therefore, the following two problems are equivalent.

$$\begin{aligned}&\min_{\boldsymbol{x}} \ \mathrm{f}(\boldsymbol{x}) \\ &\text{s.t. (8d).}\end{aligned}$$

and

$$\begin{aligned}&\min_{\boldsymbol{x}} \ \xi \\ &\text{s.t. } \min_{\boldsymbol{q}\succeq 0} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{x}) \le \xi, \forall \boldsymbol{\lambda} \succeq 0, \\ &\quad \text{(8d).}\end{aligned}$$

Thus, the proof of Lemma 3 is concluded. □

The constraints (20a) and (20b) are composed of an infinite number of constraints ($\boldsymbol{q}$ is a matrix composed of continuous variables), which makes problem $\mathcal{M}2$ hard to solve. Next we settle the Lagrangian multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ to make problem $\mathcal{M}2$ more explicit by Lemma 4.

Lemma 4: If $\mathcal{S}^{(k)}$ is feasible, and both its optimal solution $\boldsymbol{q}^{(k)}$ and the dual solution $\boldsymbol{\lambda}^{(k)}$ have been obtained, we have:

$$\min_{\boldsymbol{q}\succeq 0} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x}) = \mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x}). \tag{22}$$

If $\mathcal{S}^{(k)}$ is infeasible, and the optimal solution $\boldsymbol{q}^{(k)}$ as well as the dual solution $\boldsymbol{\nu}^{(k)}$ of problem $\mathcal{S}2^{(k)}$ have been obtained, the following equations are equivalent.

$$\min_{\boldsymbol{q}\succeq 0} \mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) \le 0 : \sum_{n=1}^{N} \nu_n^{(k)} = 1, \tag{23}$$

and

$$\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) \le 0. \tag{24}$$

Proof: If $\mathcal{S}^{(k)}$ is feasible, according to (15), (22) is tenable.

If $\mathcal{S}^{(k)}$ is infeasible, according to (17), we have:

$$\frac{\partial(\mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) + \phi - \sum_{n=1}^{N} \nu_n^{(k)} \phi}{\partial \phi} = 1 - \sum_{n=1}^{N} \nu_n^{(k)} = 0 \tag{25}$$

Then according to (16) and (18) we have

$$\begin{aligned}(\boldsymbol{q}^{(k)}, \phi) &= arg \min_{\boldsymbol{q}\succeq 0, \phi \ge 0} \ \mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) + \phi - \sum_{n=1}^{N} \nu_n^{(k)} \phi \\ &= arg \min_{\boldsymbol{q}\succeq 0, \phi \ge 0} \ \mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) + \phi(1 - \sum_{n=1}^{N} \nu_n^{(k)}),\end{aligned} \tag{26}$$

Combining (26) with (25), we have

$$\boldsymbol{q}^{(k)} = arg \min_{\boldsymbol{q}\succeq 0} \ \mathcal{L}'(\boldsymbol{q}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}). \tag{27}$$

where $\sum_{n=1}^{N} \nu_n^{(k)} = 1$. Hence, (23) and (24) are equivalent. The proof of Lemma 4 is concluded. □

According to Lemma 4, we can relax problem $\mathcal{M}2$ into problem $\mathcal{M}3$ by calculating the optimal solution $\boldsymbol{q}^{(k)}$ and the dual solution $\boldsymbol{\lambda}^{(k)}$ of feasible problem $\mathcal{S}^{(k)}$, and we can relax problem $\mathcal{M}2$ into problem $\mathcal{M}3$ by calculating the optimal solution $\boldsymbol{q}^{(k)}$ and the dual solution $\boldsymbol{\nu}^{(k)}$ of problem $\mathcal{S}2^{(k)}$ when problem $\mathcal{S}^{(k)}$ is infeasible

$$\begin{aligned}\mathcal{M}3: \ &\min_{\boldsymbol{x}} \ \xi \\ &\text{s.t.} \quad \mathcal{L}(\boldsymbol{q}^{(k_1)}, \boldsymbol{\lambda}^{(k_1)}, \boldsymbol{x}) \le \xi, \quad k_1 \\ &\qquad \in \{k | \mathcal{S}^{(k)} \text{ is feasible}\} \\ &\qquad \mathcal{L}'(\boldsymbol{q}^{(k_2)}, \boldsymbol{\nu}^{(k_2)}, \boldsymbol{x}) \le 0, \quad k_2 \\ &\qquad \in \{k | \mathcal{S}^{(k)} \text{ is infeasible}\} \\ &\qquad \text{(8d)}.\end{aligned} \tag{28a, 28b}$$

Note that problem $\mathcal{M}3$ is more explicit than problem $\mathcal{M}1$.

Lemma 5: The optimal value of problem $\mathcal{M}3$ is a lower bound of problem $\mathcal{P}2$.

Proof: The proof is stated in Remark 2.3 of [33] in detail. □

Since the upper bound and lower bound of problem $\mathcal{P}2$ can be obtained by solving problem $\mathcal{S}^k$ and problem $\mathcal{M}3$, respectively, we propose a joint power allocation and user grouping algorithm (GPGA) based on GBD as shown in Algorithm 1. In this algorithm, steps 1-4 is the initialization of grouping matrix $\boldsymbol{x}^{(k)}$. Then according to the feasibility of problem $\mathcal{S}^{(k)}$ and Lemma 4, we add constraint $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x}) \le \xi$ or $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) \le 0$, which we called them feasi-constraint and infeasi-constraint, respectively, to relaxed master problem $\mathcal{M}3$ and update the upper bound and lower bound of problem $\mathcal{P}2$. A new user grouping matrix $\boldsymbol{x}^{(k+1)}$ is obtained by solving problem $\mathcal{M}3$. Next, matrix $\boldsymbol{x}^{(k)}$ is updated by matrix $\boldsymbol{x}^{(k+1)}$. Steps 6-19 is repeated until the gap between the upper bound and lower bound of problem $\mathcal{P}2$ is less than δ, i.e., δ-optimal solution.

Proposition 1: Algorithm 1 is bound to stop in finite steps for any given $\delta > 0$.

Algorithm 1 GBD Based Joint Power Allocation and User Grouping Algorithm (GPGA)

Input: Variance of Channel Estimation $\boldsymbol{\alpha}$, Large-scale Fading $\boldsymbol{\beta}$,
Output: Power Allocation Matrix $\boldsymbol{q}$, Grouping Matrix $\boldsymbol{x}$

for $n = 1$ *to* N **do**
  $g = mod(n, G) + 1$; $x_{gn}^{1} = 1$
end
$k = 1$;Create problem $\mathcal{M}3$ without feasi-constraint or infeasi-constraint;
repeat
  Solve power allocation problem $\mathcal{S}^{(k)}$ by the interior point method;
  if $\mathcal{S}^{(k)}$ *is feasible.* **then**
    Calculate the optimal solution $\boldsymbol{q}^{(k)}$, optimal value $P_t^{(k)} = \sum_{m=1}^{M}\sum_{n=1}^{N} q_{mn}^2 \sum_{g=1}^{G} x_{gn}^{(k)} \varphi_{gmn}^{(k)}$ and the dual solution $\boldsymbol{\lambda}^{(k)}$ of problem $\mathcal{S}^{(k)}$;
    Add constraint $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x}) \leq \xi$ to the relaxed master problem $\mathcal{M}3$;
    Update upper bound $b_u = P_t^{(k)}$;
  else
    Calculate the optimal solution $\boldsymbol{q}^{(k)}$, $P_t^{(k)} = \sum_{m=1}^{M}\sum_{n=1}^{N} q_{mn}^2 \sum_{g=1}^{G} x_{gn}^{(k)} \varphi_{gmn}^{(k)}$ and the dual solution $\boldsymbol{\nu}^{(k)}$ of problem $\mathcal{S}2^{(k)}$ by the interior point method;
    Add constraint $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) \leq 0$ to the relaxed master problem $\mathcal{M}3$;
  end
  Solve the relaxed master problem $\mathcal{M}3$ to get new grouping matrix $\boldsymbol{x}^{(k+1)}$ and its optimal value ξ_{min};
  Update lower bound $b_l = \xi_{min}$; $k = k + 1$;
until $b_u - b_l \leq \delta$ *or* b_u *won't change or number of iterations exceeds* N;

Proof: After each iteration of 6-19 in Algorithm 1, the upper bound of problem $\mathcal{P}2$ is nonincreasing and the lower bound of problem $\mathcal{P}2$ is nondecreasing. Therefore, the gap between the upper bound and lower bound of problem $\mathcal{P}2$ is shrunk. Moreover, the strategic space of user grouping matrix $\boldsymbol{x}$ is finite. Thus, Algorithm 1 is bound to stop in finite steps for any given $\delta > 0$. The proof is stated in section 2.4 of [33] in detail. □

In Algorithm 1, we solve power allocation problem $\mathcal{S}^{(k)}$ by the interior point method, and its computational complexity is $O\big(N(NM)^3\big)$ [34]. If we solve the relaxed master problem $\mathcal{M}3$ by exhaustive search algorithm, the computational complexity will be unbearable. We introduce the way to solve the relaxed master problem $\mathcal{M}3$ in Section IV.

IV. Relaxed Master Problem Analysis and Solutions Based on Graph Theory

In this section, we convert the relaxed master user grouping problem into a problem of searching for some special negative loops in a graph composed of users based on graph theory. Two algorithms to find these loops are proposed. In order to find the way to reduce the values of $\mathcal{L}(\boldsymbol{q}^{(k_1)}, \boldsymbol{\lambda}^{(k_1)}, \boldsymbol{x})$ and $\mathcal{L}'(\boldsymbol{q}^{(k_2)}, \boldsymbol{\nu}^{(k_2)}, \boldsymbol{x})$ in the relaxed master user grouping problem $\mathcal{M}3$, we introduce three definitions as follows:

Definition 1: For L users numbered by $n_1, n_2, \ldots, n_L$ in different groups, if problem $\mathcal{S}^{(k)}$ is feasible and $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x})$ can be reduced by $n_1 \rightarrow n_2, \ldots, n_{L-2} \rightarrow n_{L-1}$ and putting user n_{L-1} into the group of user n_L, or if problem $\mathcal{S}^{(k)}$ is infeasible, and $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x})$ can be reduced by $n_1 \rightarrow n_2, \ldots, n_{L-2} \rightarrow n_{L-1}$ and putting user n_{L-1} into the group of user n_L, these users compose a k-shift union.

To explain the meaning of $n \rightarrow n'$, we assume that with grouping strategy $\boldsymbol{x}$, user n and user n' are in group g and g', respectively. That is, $x_{gn} = 1$, $x_{g'n'} = 1$. Then $n \rightarrow n'$ means transferring user n into group g' and removing user n' from group g', that is, $x_{gn} = 0$, $x_{gn'} = 1$, $x_{g'n'} = 0$.

Definition 2: For L users numbered by $n_1, n_2, \ldots, n_L$ in different groups, if problem $\mathcal{S}^{(k)}$ is feasible and $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x})$ can be reduced by $n_1 \rightarrow n_2, \ldots, n_L \rightarrow n_1$, or if problem $\mathcal{S}^{(k)}$ is infeasible, and $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x})$ can be reduced by $n_1 \rightarrow n_2, \ldots, n_L \rightarrow n_1$, these users compose a k-exchange union.

Definition 3: For grouping strategy $\boldsymbol{x}$, if the value of ξ in problem $\mathcal{M}3$ cannot be reduced by any shift union *or* exchange union *with all the constraints in problem $\mathcal{M}3$ satisfied, grouping matrix $\boldsymbol{x}$ is called* all-stable solution.

To find the *shift union*s and *exchange union*s among users, we analyze the rules that values of $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x})$ and $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x})$ change after changing user grouping matrix $\boldsymbol{x}$. By dividing the expressions of $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x})$ and $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x})$, i.e., (14) and (17) in groups, we define two weighted interference plus noise variables $\omega_g(\boldsymbol{x}; \boldsymbol{q}, \boldsymbol{\lambda})$ and $\omega_g'(\boldsymbol{x}; \boldsymbol{q}, \boldsymbol{\nu})$ as follows:

$$\omega_g(\boldsymbol{x}; \boldsymbol{q}, \boldsymbol{\lambda}) = \sum_{j=1}^{N} x_{gj}\lambda_j \left(\left(\sigma^2 + \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} q_{mi}^2 \upsilon_{gmji} \right)^{\frac{1}{2}} \gamma_j^{\frac{1}{2}} - \sum_{m=1}^{M} q_{mj} x_{gj} \vartheta_{gmj} \right) \tag{29}$$

$$\omega_g'(\boldsymbol{x}; \boldsymbol{q}, \boldsymbol{\nu}) = \sum_{j=1}^{N} x_{gj}\nu_j \left(\left(\sigma^2 + \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} q_{mi}^2 \upsilon_{gmji} \right)^{\frac{1}{2}} \gamma_j^{\frac{1}{2}} - \sum_{m=1}^{M} q_{mj} x_{gj} \vartheta_{gmj} \right) \tag{30}$$

Combining (14), (17), (29) and (30), we have $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x}) = \sum_{g=1}^{G} \omega_g(\boldsymbol{x}; \boldsymbol{q}, \boldsymbol{\lambda})$ and $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}) = \sum_{g=1}^{G} \omega_g'(\boldsymbol{x}; \boldsymbol{q}, \boldsymbol{\nu})$. Next we investigate how user grouping

strategy changing impacts the value of $\omega_g(\boldsymbol{x};\boldsymbol{q},\boldsymbol{\lambda})$ and $\omega'_g(\boldsymbol{x};\boldsymbol{q},\boldsymbol{\nu})$.

For convenience, we assume that in grouping strategy $\boldsymbol{x}$, user n is in group g_n. Then we construct a directed graph $G(\mathcal{N},\mathcal{E}^{(k)};\boldsymbol{x})$, where $\mathcal{N}$ is the set of nodes composed of users and $\mathcal{E}^{(k)}$ is the set of edges existing between two users in different groups. The adjacency matrix of graph $G(\mathcal{N},\mathcal{E}^{(k)};\boldsymbol{x})$ is denoted by $\boldsymbol{a}^{(k)}$, and we set that:

$$a_{ij}^{(k)}=\begin{cases}\omega_{g_j}(x_{g_ji}=1,x_{g_jj}=0,\boldsymbol{x}_{-i,j};\\ \quad \boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)})-\omega_{g_j}(\boldsymbol{x};\boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)})\\ \quad \text{if } g_i\neq g_j,\ \mathcal{S}^{(k)} \text{ is feasible}\\ \omega'_{g_j}(x_{g_ji}=1,x_{g_jj}=0,\boldsymbol{x}_{-i,j};\\ \quad \boldsymbol{q}^{(k)},\boldsymbol{\nu}^{(k)})-\omega'_{g_j}(\boldsymbol{x};\boldsymbol{q}^{(k)},\boldsymbol{\nu}^{(k)})\\ \quad \text{if } g_i\neq g_j,\ \mathcal{S}^{(k)} \text{ is infeasible}\\ \infty,\ \text{if } g_i=g_j\end{cases} \tag{31}$$

where $\boldsymbol{x}_{-i,j}$ denotes the user grouping strategies of users except users i and j, thus $(x_{g_ji}{=}1,x_{g_jj}{=}0,\boldsymbol{x}_{-i,j})$ represents that user i is in group g_j, user j is not in group g_j, and the user grouping strategies of users except users i and j consistent with $\boldsymbol{x}$. The relation between problem $\mathcal{M}3$ and graph $G(\mathcal{N},\mathcal{E}^{(k)};\boldsymbol{x})$ can be indicated with the following propositions.

Proposition 2: For L users numbered by $n_1,n_2,\ldots,n_L$ in different groups, if problem $\mathcal{S}^{(k)}$ is feasible and $\mathcal{L}(\boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)},\boldsymbol{x})$ can be reduced by $n_1\rightarrow n_2,\ldots,n_L\rightarrow n_1$, these users can compose a negative loop $n_1->n_2->\ldots->n_L->n_1$ in graph $G(\mathcal{N},\mathcal{E}^{(k)};\boldsymbol{x})$.

Proof: We assume that the user grouping matrix before and after $n_1\rightarrow n_2,\ldots,n_L\rightarrow n_1$ are $\boldsymbol{x}$ and $\boldsymbol{x}'$. Then the difference of $\mathcal{L}(\boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)},\boldsymbol{x})$ is

$$\begin{aligned}&\nabla_{\mathcal{L}}\\ &=\mathcal{L}(\boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)},\boldsymbol{x}')-\mathcal{L}(\boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)},\boldsymbol{x})\\ &=\sum_{n=1}^{N}\lambda_n^{(k)}\Bigg(\left(\sigma^2+\sum_{g=1}^{G}x'_{gn}\sum_{i=1}^{N}x'_{gi}\sum_{m=1}^{M}(q_{mi}^{(k)})^2\upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}}\gamma_n^{\frac{1}{2}}\\ &\quad-\sum_{m=1}^{M}q_{mn}\sum_{g=1}^{G}x_{gn}\vartheta_{gmn}^{(k)}\Bigg)\\ &\quad-\sum_{n=1}^{N}\lambda_n^{(k)}\Bigg(\left(\sigma^2+\sum_{g=1}^{G}x_{gn}\sum_{i=1}^{N}x_{gi}\sum_{m=1}^{M}(q_{mi}^{(k)})^2\upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}}\gamma_n^{\frac{1}{2}}\\ &\quad-\sum_{m=1}^{M}q_{mn}\sum_{g=1}^{G}x_{gn}\vartheta_{gmn}^{(k)}\Bigg)\end{aligned} \tag{32}$$

We assume that users $n_1,n_2,\ldots,n_L$ are in groups $g_1,g_2,\ldots,g_L$, respectively, note that the value of

$$\lambda_n^{(k)}\Bigg(\left(\sigma^2+\sum_{g=1}^{G}x_{gn}\sum_{i=1}^{N}x_{gi}\sum_{m=1}^{M}(q_{mi}^{(k)})^2\upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}}\gamma_n^{\frac{1}{2}}-\sum_{m=1}^{M}q_{mn}\sum_{g=1}^{G}x_{gn}\vartheta_{gmn}^{(k)}\Bigg),\quad n\notin\{n_1,n_2,\ldots,n_L\}$$

will not change after $n_1\rightarrow n_2,\ldots,n_L\rightarrow n_1$. Then according to (32), we have:

$$\begin{aligned}&\nabla_{\mathcal{L}}\\ &=\sum_{l=1}^{L}\sum_{n=1}^{N}x'_{g_ln}\lambda_n^{(k)}\Bigg(\left(\sigma^2+\sum_{i=1}^{N}x'_{g_li}\sum_{m=1}^{M}(q_{mi}^{(k)})^2\upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}}\gamma_n^{\frac{1}{2}}\\ &\quad-\sum_{m=1}^{M}q_{mn}\sum_{g=1}^{G}x_{gn}\vartheta_{gmn}^{(k)}\Bigg)\\ &\quad-\sum_{l=1}^{L}\sum_{n=1}^{N}x_{g_ln}\lambda_n^{(k)}\Bigg(\left(\sigma^2+\sum_{i=1}^{N}x_{g_li}\sum_{m=1}^{M}(q_{mi}^{(k)})^2\upsilon_{gmni}^{(k)}\right)^{\frac{1}{2}}\gamma_n^{\frac{1}{2}}\\ &\quad-\sum_{m=1}^{M}q_{mn}\sum_{g=1}^{G}x_{gn}\vartheta_{gmn}^{(k)}\Bigg)\\ &=\sum_{l=1}^{L}a_{lj}^{(k)}\end{aligned} \tag{33}$$

where $j=mod(l,L)+1$.

Then the proof of Proposition 2 is concluded. □

Proposition 3: For L users numbered by $n_1,n_2,\ldots,n_L$ in different groups, if problem $\mathcal{S}^{(k)}$ is infeasible, and $\mathcal{L}'(\boldsymbol{q}^{(k)},\boldsymbol{\nu}^{(k)},\boldsymbol{x})$ can be reduced by $n_1\rightarrow n_2,\ldots,n_L\rightarrow n_1$, these users can compose a negative loop $n_1->n_2->\ldots->n_L->n_1$ in graph $G(\mathcal{N},\mathcal{E}^{(k)};\boldsymbol{x})$.

Proof: The proof is similar to Proposition 2. □

According to Proposition 2 and Proposition 3, we can find the grouping changing method to reduce the value of $\mathcal{L}(\boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)},\boldsymbol{x})$ or $\mathcal{L}'(\boldsymbol{q}^{(k)},\boldsymbol{\nu}^{(k)},\boldsymbol{x})$ by searching for the negative loop with all users in different groups, we call them negative differ-group loop. However, the number of users in each group will not change after $n_1\rightarrow n_2,\ldots,n_L\rightarrow n_1$. To find the grouping changing method that can lead to arbitrary number of users in each group and reduce the value of $\mathcal{L}(\boldsymbol{q}^{(k)},\boldsymbol{\lambda}^{(k)},\boldsymbol{x})$ or $\mathcal{L}'(\boldsymbol{q}^{(k)},\boldsymbol{\nu}^{(k)},\boldsymbol{x})$, we expand graph $G(\mathcal{N},\mathcal{E}^{(k)};\boldsymbol{x})$ to $G(\mathcal{N}^e,\mathcal{E}^{(k)};\boldsymbol{x})$ by adding a virtual user to each group. The virtual user added to group g is numbered by n_g^v. We set the achievable SINR of these virtual users to 0, so these virtual users will not be allocated any power.

Proposition 4: If users $n_1,n_2,\ldots,n_{L-1},n_L$ compose a shift union *with grouping strategy $\boldsymbol{x}$, users $n_1,n_2,\ldots,n_{L-1},n_{g_L}^v$ can compose an* exchange union *with grouping strategy $\boldsymbol{x}$.*

Proof: Obviously, if grouping strategy $\boldsymbol{x}$ is changed to $\boldsymbol{x}'$ after $n_1\rightarrow n_2,\ldots,n_{L-2}\rightarrow n_{L-1}$ and putting user n_{L-1} into the group of user n_L, it will also be changed to $\boldsymbol{x}'$ after $n_1\rightarrow n_2,\ldots,n_{L-2}\rightarrow n_{L-1},n_{L-1}\rightarrow n_{g_L}^v$. So the proof of Proposition 4 is concluded. □

Theorem 1: For grouping strategy $\boldsymbol{x}$, if the value of ξ in problem $\mathcal{M}3$ cannot be reduced by any negative differ-group loop in graph $G(\mathcal{N}^e,\mathcal{E}^{(k)};\boldsymbol{x})$ with all the constraints in problem $\mathcal{M}3$ satisfied, grouping matrix $\boldsymbol{x}$ is called all-stable solution.

Proof: According to Proposition 2, Proposition 3 and Proposition 4, if there is a *shift union* or an *exchange union* among all real users and virtual users, there must be a

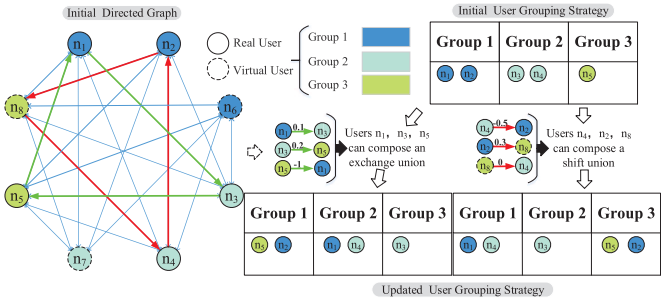


Fig. 2. Illustration of the directed graph with "shift union" and "exchange union."

negative differ-group loop in graph $G(\mathcal{N}^e, \mathcal{E}^{(k)}; \boldsymbol{x})$. Therefore, if the value of ξ in problem $\mathcal{M}3$ cannot be reduced by any negative differ-group loop in graph $G(\mathcal{N}^e, \mathcal{E}^{(k)}; \boldsymbol{x})$ with all the constraints in problem $\mathcal{M}3$ satisfied, the value of ξ in problem $\mathcal{M}3$ cannot be reduced by any *shift union* or *exchange union* with all the constraints in problem $\mathcal{M}3$ satisfied, i.e., the grouping matrix $\boldsymbol{x}$ is *all-stable solution*. □

In order to explain the concepts of *shift union* and *exchange union* more clearly, an illustration of the directed graph composed of 8 nodes is shown in Fig.2 with edges among users in different groups, where the real users and virtual users are represented by solid circles and dotted circles, respectively. The users are divided into three groups, and the users in the same colour are grouped into the same group with initial user grouping strategy. In this directed graph, two negative differ-group loops $n_1 -> n_3 -> n_5 -> n_1$ and $n_4 -> n_2 -> n_8 -> n_4$ are found. Among them, users n_1, n_3, n_5 can compose an *exchange union*, and users n_1, n_3, n_5 can compose a *shift union*. Then, $\mathcal{L}(\boldsymbol{q}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{x})$ (if problem $\mathcal{S}^{(k)}$ is infeasible) or $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x})$ (if problem $\mathcal{S}^{(k)}$ is infeasible) can be reduced by $n_1 \rightarrow n_3, n_3 \rightarrow n_5, n_5 \rightarrow n_1$ or $n_4 \rightarrow n_2, n_2 \rightarrow n_8, n_8 \rightarrow n_4$. It is worth noting that the initial user grouping strategy is an *all-stable solution*, if there are no negative differ-group loop can be found in the initial directed graph.

Graph theory based algorithm to solve relaxed master problem $\mathcal{M}3$ is shown in Algorithm 2. In this algorithm, we first search for a new user grouping matrix which satisfies all the infeasi-constraints in problem $\mathcal{M}3$ in steps 2-17. In each iteration of steps 3-16, we change the user grouping matrix to reduce the value of $\max_{i\leq k, \mathcal{S}^{(i)} \text{is infeasible.}} \mathcal{L}'(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k)})$, where $\mathcal{A}$ is the set of the loops according to which we cannot reduce the value of $\max_{i\leq k, \mathcal{S}^{(i)} \text{is infeasible.}} \mathcal{L}'(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k)})$. Then we search for the solution of problem $\mathcal{M}3$ in steps 18-36. In each iteration of steps 19-35, we change the user grouping matrix to reduce the value of $\max_{i\leq k, \mathcal{S}^{(i)} \text{is feasible.}} \mathcal{L}(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k)})$, where $\mathcal{A}$ is the set of the loops according to which we cannot reduce the value of $\max_{i\leq k, \mathcal{S}^{(i)} \text{is feasible.}} \mathcal{L}(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k)})$.

Corollary 1: Algorithm 2 can converge to all-stable solution *in finite iterations.*

Proof: The number of nodes in the graph is limited by the numbers of users and groups, so the number of the negative differ-group loops in the graph is infinite. In addition, the optimal value of problem $\mathcal{M}3$ in each iteration of Algorithm 2 will descend and the feasibility of the outputted

Algorithm 2 Graph Theory Based Algorithm to Solve Master Problem (GBMA)

Input: Relaxed Master Problem $\mathcal{M}3$, Grouping Matrix $\boldsymbol{x}^{(k)}$,
Output: Grouping Matrix $\boldsymbol{x}^{(k+1)}$

1 Create set of infeasible loops $\mathcal{A} = \varnothing$;
2 **repeat**
3 Find $k_{max} = arg \max_{i\leq k, \mathcal{S}^{(i)} \text{is infeasible.}} \mathcal{L}'(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k)})$;
4 $maxL = \mathcal{L}'(\boldsymbol{q}^{(k_{max})}, \boldsymbol{\lambda}^{(k_{max})}, \boldsymbol{x}^{(k)})$;
5 **if** $maxL > 0$. **then**
6 Create graph $G(\mathcal{N}^e, \mathcal{E}^{(k_{max})}; \boldsymbol{x}^{(k_{max})})$;
7 Search for negative differ-group loop $\boldsymbol{L}^{(k_{max})} \notin \mathcal{A}$ in graph $G(\mathcal{N}^e, \mathcal{E}^{(k)}; \boldsymbol{x}^{(k)})$;
8 Change grouping matrix $\boldsymbol{x}^{(k)}$ to $\boldsymbol{x}^{(k+1)}$ according to loop $\boldsymbol{L}^{(k)}$ based on Proposition 2-3.
9 **for** $i = 1$ *to* $k-1$ **do**
10 **if** $\mathcal{S}^{(i)}$ *is infeasible.* **then**
11 **if** $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}^{(k+1)}) > maxL$ **then**
12 Add loop $\boldsymbol{L}^{(k_{max})}$ to $\mathcal{A}$; Jump to step 6;
13 **end**
14 **end**
15 **end**
16 $\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k+1)}$;$\mathcal{A} = \varnothing$;
17 **end**
18 **until** $maxL < 0$;
19 **repeat**
20 $\mathcal{A} = \varnothing$;
21 Find $k_{max} = arg \max_{i\leq k, \mathcal{S}^{(i)} \text{is feasible.}} \mathcal{L}(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k)})$;
22 Create graph $G(\mathcal{N}^e, \mathcal{E}^{(k_{max})}; \boldsymbol{x}^{(k)})$;
23 Search for negative differ-group loop $\boldsymbol{L}^{(k_{max})} \notin \mathcal{A}$ in graph $G(\mathcal{N}^e, \mathcal{E}^{(k_{max})}; \boldsymbol{x}^{(k)})$;
24 Change grouping matrix $\boldsymbol{x}^{(k)}$ to $\boldsymbol{x}^{(k+1)}$ according to loop $\boldsymbol{L}^{(k)}$ based on Proposition 2-3.
25 **for** $i = 1$ *to* k **do**
26 **if** $\mathcal{S}^{(i)}$ *is feasible.* **then**
27 **if** $\mathcal{L}(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k+1)}) > \mathcal{L}(\boldsymbol{q}^{(i)}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{x}^{(k_{max})})$ **then**
28 Add loop $\boldsymbol{L}^{(k_{max})}$ to $\mathcal{A}$; Jump to step 22;
29 **end**
30 **else**
31 **if** $\mathcal{L}'(\boldsymbol{q}^{(k)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{x}^{(k+1)}) > 0$ **then**
32 Add loop $\boldsymbol{L}^{(k_{max})}$ to $\mathcal{A}$; Jump to step 22;
33 **end**
34 **end**
35 **end**
36 $\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k+1)}$;
37 **until** *Cannot find an appropriate negative differ-group loop*;
38 Return $\boldsymbol{x}^{(k+1)}$.

Algorithm 3 Extended Bellman-Ford Algorithm to Search for Negative Differ-Group Loops in GBMA (EBSA)

Input: Group $G(\mathcal{N}^e, \mathcal{E}^{(k_{max})}; \boldsymbol{x}^{(k)})$, Set of infeasible loops $\mathcal{A}$, Adjacency matrix $\boldsymbol{a}^{(k)}$
Output: *Negative differ-group loop* $\boldsymbol{L}^{(k)}$

1 Create super node.
2 **for** $j = 1$ *to* $(G + N)$ **do**
3 $\quad w_n$=0, $\mathcal{T}_n = \varnothing$, $n \in \mathcal{V}$;
4 **end**
5 **repeat**
6 $\quad$ **for** $i = 1$ *to* $(G + N)$ **do**
7 $\quad\quad$ **for** $j = 1$ *to* $(G + N)$ **do**
8 $\quad\quad\quad$ **if** $(w_j > w_i + a_{ij}^{(k)}) \& (a_{kj}^{(k)} \neq \infty, \forall k \in \mathcal{T}_i \setminus \{j\})$ **then**
9 $\quad\quad\quad\quad$ **if** *There is a loop* $\boldsymbol{L}$ *in* $(\mathcal{T}_i \cup \{i\})$ *and* $\boldsymbol{L} \notin \mathcal{A}$ **then**
10 $\quad\quad\quad\quad\quad$ $\mathcal{T}_j = \mathcal{T}_i \cup \{i\}$, $w_j = w_i + a_{ij}^{(k)}$
11 $\quad\quad\quad\quad$ **end**
12 $\quad\quad\quad$ **end**
13 $\quad\quad\quad$ **if** $(w_j > w_i + a_{ij}^{(k)}) \& (\exists a_{kj}^{(k)} = \infty, k \in \mathcal{T}_i \setminus \{j\})$ **then**
14 $\quad\quad\quad\quad$ Find the shortest path $\mathcal{T}_i'$ from the super node to user i with $a_{kj}^{(k)} \neq \infty, \forall k \in \mathcal{T}_i' \setminus \{j\}$, assume the distance of path $\mathcal{T}_i'$ is m_i';
15 $\quad\quad\quad\quad$ **if** $w_j > m_i' + a_{ij}^{(k)}$ **then**
16 $\quad\quad\quad\quad\quad$ **if** *There is a loop* $\boldsymbol{L}$ *in* $(\mathcal{T}_i' \cup \{i\})$ *and* $\boldsymbol{L} \notin \mathcal{A}$ **then**
17 $\quad\quad\quad\quad\quad\quad$ $\mathcal{T}_j = \mathcal{T}_i' \cup \{i\}$, $w_j = m_i' + a_{ij}^{(k)}$
18 $\quad\quad\quad\quad\quad$ **end**
19 $\quad\quad\quad\quad$ **end**
20 $\quad\quad\quad$ **end**
21 $\quad\quad\quad$ **if** $\|\mathcal{T}_j\| > G$ **then**
22 $\quad\quad\quad\quad$ Find the *negative differ-group loop* $\boldsymbol{L}$ in $\mathcal{T}_j$; **return** $\boldsymbol{L}$ and break;
23 $\quad\quad\quad$ **end**
24 $\quad\quad$ **end**
25 $\quad$ **end**
26 **until** $\mathcal{T}_n$, $n \in \mathcal{V}$ *do not change*;

Algorithm 4 Greedy Fast Algorithm to Search for Negative Differ-Group Loop in GBMA (GFSA)

Input: Group $G(\mathcal{N}^e, \mathcal{E}^{(k_{max})}; \boldsymbol{x}^{(k)})$, Set of infeasible loops $\mathcal{A}$, Adjacency matrix $\boldsymbol{a}^{(k)}$
Output: *Negative differ-group loop* $\boldsymbol{L}^{(k)}$

1 **for** $t = 1$ *to* $(N + G)$ **do**
2 $\quad \mathcal{T} = \varnothing$;
3 $\quad$ Find the minimal edge $a_{ij}^{(k)} = \min\{a_{ij}^{(k)} | i, j \in \mathcal{N}^e\}$;
4 $\quad \mathcal{T} = \mathcal{T} \cup \{i\} \cup \{j\}$, $m_t = a_{ij}^{(k)}$;
5 $\quad$ **if** $m_t + a_{ji}^{(k)} < 0 \& \mathcal{T} \notin \mathcal{A}$ **then**
6 $\quad\quad \boldsymbol{L}^{(k)} \leftarrow \mathcal{T}$; **return** $\boldsymbol{L}^{(k)}$.
7 $\quad$ **end**
8 $\quad a_{ij}^{(k)} = \infty$, $i = j$;
9 $\quad$ **for** $l = 3$ *to* G **do**
10 $\quad\quad$ Find the minimal output edge of node i : $a_{ij}^{(k)} = \min\{a_{ij}^{(k)} | j \notin \mathcal{T}\}$;
11 $\quad\quad m_t = m_t + a_{ij}^{(k)}$, $\mathcal{T} = \mathcal{T} \cup \{j\}$;
12 $\quad\quad$ **if** $m_t + a_{ji}^{(k)} < 0 \& \mathcal{T} \notin \mathcal{A}$ **then**
13 $\quad\quad\quad \boldsymbol{L}^{(k)} \leftarrow \mathcal{T}$; **return** $\boldsymbol{L}^{(k)}$.
14 $\quad\quad$ **end**
15 $\quad\quad i = j$;
16 $\quad$ **end**
17 **end**

user grouping matrix for problem $\mathcal{M}3$ in each iteration can be guaranteed by step 26. Therefore, Algorithm 2 will stop after finite iterations and output a solution without negative differ-group loop, i.e., *all-stable solution*. □

To find these negative differ-group loops, we extend the Bellman-Ford algorithm to Algorithm 3 [35]–[37], where the negative differ-group loops in the set of infeasible loops $\mathcal{A}$ are avoided to be outputted. In this algorithm, we first create a super node. The distance from the super node to node n is set to w_n=0, and the path from super node to node n, $\mathcal{T}_n$, $n \in \mathcal{V}$, is initialized in steps 2-4. Then these paths are constantly relaxed in steps 5-26. In each step of relaxing, the users in the same group are avoided to be added to the same path from super node to any node as step 8 and step 13 show. And the loops in $\mathcal{A}$ are avoided to form in any path $\mathcal{T}_n$ as step 9 and step 16 show. Therefore, according to the principle of Bellman-Ford algorithm, if the path from super node to any node is longer than the number of groups, there must be a negative differ-group loop in this path [38].

To obtain an appropriate solution of problem $\mathcal{P}1$ with polynomial time, we also design a greedy fast algorithm for the solving of problem $\mathcal{M}3$ as shown in Algorithm 4. This searching algorithm starts from the minimal edge in the graph. Then we iteratively search for the minimal output edge until there exists no next output edge. In this process, the set of loops that have been rejected in steps 5-7 of Algorithm 4 is avoided to be outputted.

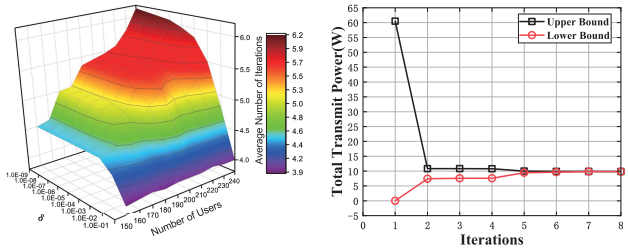
Computational Complexity Analysis: In GFSA, complexity of steps 10-15 is $O\big(G + N\big)$, and complexity of steps 2-8 is $O\big((G + N)^2\big)$. Hence, computational complexity of GFSA is $O\big((G + N)^3\big)$. Assume that GFSA is repeated C times in Algorithm 2. Then apparently computational complexity of GBMA is $O\big(C(G + N)^3\big)$. We solve power allocation problem $\mathcal{S}^{(k)}$ by the interior point method, and its computational complexity is $O\big(N(NM)^3\big)$ [34]. Then computational complexity of the proposed fast greedy algorithm is $O\big(max(N^2(NM)^3, CN(G + N)^3)\big)$.

V. Performance Evaluation

In this section, we evaluate the performance of the proposed two algorithms in cell-free massive MIMO systems in terms of transmit power, interference from the desired signals of the other users in the same group, etc. In the simulations, APs and users are randomly placed in a 3km×3km rectangular area. We set large-scale channel gain to $128.1 + 37.6log_{10}$ (d_n [km]) dB. The small-scale fading follows an i.i.d. Gaussian

TABLE II
MAIN NOTATIONS

Parameter	Value
Number of users(N)	200
Number of groups(G)	5
Number of APs(M)	200
Bandwidth(B)	20 MHz
Noise power spectral density(N_0)	-174 dBm/Hz
Power of pilot signal (ρ_r)	200 mW
Length of pilot sequences (τ)	$2\lceil N/G \rceil$
Target data rate	0.1-1.5 Mbps



(a) (b)

Fig. 3. Performance comparison of different user grouping algorithms: a. Average number of iterations with different numbers of users and different δ; b. Convergence of lower and upper bounds with the proposed fast greedy user grouping algorithm(200 users and 200 APs in total).

distribution. Some default values in the simulations are shown in Table II [39].

A. Convergence Performance

To evaluate the convergence performance of the proposed fast greedy algorithm, in Fig. 3(a), we show the average number of iterations (denoted by T_{iter}) of steps 7-18 in Algorithm 1 with different numbers of users and different δ. We can see that T_{iter} increases as number of users increases and δ reduces. In addition, $T_{iter} < 10$ even when $N = 240$ and $\delta = 10^{-9}$. In Fig. 3(b), we show the process that the gap between the upper bound and lower bound of problem $\mathcal{P}2$ reduces. The results show that the proposed fast greedy algorithm converges rapidly, which illustrate the practicability of the proposed fast greedy user grouping algorithm. As shown in Fig. 3(b), the gap between the upper bound and the lower bound shrinks as the number of iterations increases. According to the principle of benders decomposition, the optimal user grouping and power allocation strategy can be found if this gap δ is 0. The number of users is finite, hence the grouping strategy profile is finite. Every time we change the user grouping strategy, the total transmit power will not increase, which means that a grouping strategy profile will not be selected repeatedly. Thus we can obtain the optimal user grouping and power allocation strategy with a small enough δ in Algorithm 1.

B. Impacts of Pilot Signal

As mentioned in section II-A, channel estimation is carried out after user grouping in cell-free massive MIMO systems.

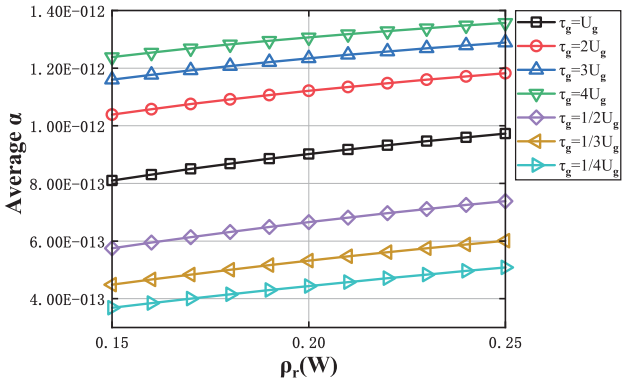


Fig. 4. Influence of ρ_r and τ on α_{mn}.

Since the accuracy of channel estimation will affect the performance of cell-free massive MIMO with beamforming, we investigate the impacts of pilot signal on α in this subsection. α is the variance of MMSE estimate of channel fading h as stated in (2). In Fig. 4, we change the power of pilot signal ρ_r and the length of pilot sequences τ_g to show its influence on α under the proposed user grouping algorithm, where U_g means the of number of users in group g. Considering the case of pilot reuse to observe the effects of non-orthogonal pilots, the length of pilot signal will be less than the number of users, i.e., $\tau_g < \sum_{i=1}^{N} x_{gi}$ [40]. To investigate the effect of pilot signal with $\tau_g < \sum_{i=1}^{N} x_{gi}$, in this figure, the range of τ_g is set to $\frac{1}{2}U_g$, $\frac{1}{3}U_g$ and $\frac{1}{4}U_g$. We can see that in Fig. 4, as the increase of ρ_r and τ_g, the mean value of α will also increase, which agrees with (3). Then the channel estimation error is reduced as stated in (4). That is to say, by adding the length of pilot sequences or the power of the pilot signal, the accuracy of channel estimation can be increased. By user grouping, the number of users served by each time-slot can be reduced. Therefore, the length of pilot sequences to maintain the accuracy of channel estimation can be reduced by user grouping.

C. Performance Comparison

In this subsection, we compare four proposed user grouping algorithms (named "MRT-GPGA-EBSA", "MRT-GPGA-GFSA", "ZF-GPGA-EBSA" and "ZF-GPGA-GFSA", respectively) with the basic random user grouping algorithm(BCGA) and Gale-Shapley algorithm(Gale-S) [41], where each user prefers the group where interference is less and each group prefers to reject the access requests of the users with the highest requirements on power. The number of users in each group with Gale-S is equal. To evaluate interference that users suffer, we define a mean-interference variable I as follows:

$$I = \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} p_{mi} \beta_{mn} \alpha_{gmi} \right) \tag{34}$$

where $\left(\sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} p_{mi} \beta_{mn} \alpha_{gmi} \right)$ is the right of denominator of SINR_n in (1).

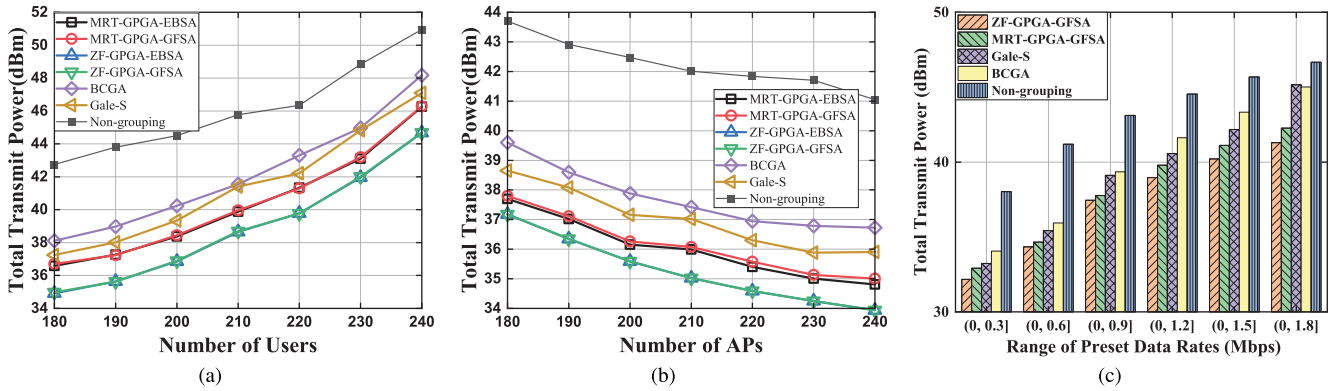


Fig. 5. Performance comparison of different user grouping algorithms: a. Total transmit power vs. number of APs; b. Total transmit power vs. Number of APs; c. Total transmit power vs. Range of preset data rates (Mbps).

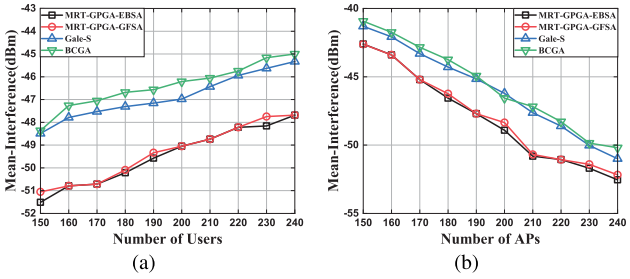


Fig. 6. Performance comparison of different user grouping algorithms: a. Mean-interference vs. number of users; b. Mean-interference vs. Number of APs.

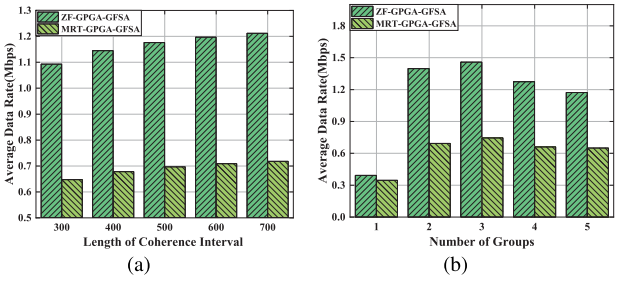


Fig. 7. Performance comparison under given total transmit power: a. Average data rate(Mbps) vs. Length of coherence interval(M=250, N=240, P_t=10W); b. Number of groups vs. Length of coherence interval(M=250, N=240, P_t=10W).

We first vary the number of users and APs to show mean-interference variable I with four different user grouping algorithms in Fig. 6(a) and Fig. 6(b), respectively. The number of users increases from 150 to 240 with 200 APs and five groups in total in Fig. 6(a), and the number of APs increases from 150 to 240 with 200 users and five groups in total in Fig. 6(b). We can see that mean-interference increases with the increase of users and the reduction of APs. The reason is that as the number of groups is given, the number of users sharing the same time-slot will increase with the increase of users. Then the value of $\sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} p_{mi}\beta_{mn}\alpha_{gmi}$ in (34) will increase. In addition, as the number of APs reduces, to maintain enough SINR in (1), the power that each AP allocates to a user will increase, then more interference will appear according to (34). Moreover, the mean-interference of proposed algorithms is less than the reference two. The main reason is that, in this paper, power allocation and user grouping are jointly optimized considering QoS constraints. To reduce the total transmit power and satisfy QoS requirements of different users as problem $\mathcal{P}1$ shows, the value of $\sum_{g=1}^{G} x_{gn} \sum_{i=1}^{N} x_{gi} \sum_{m=1}^{M} p_{mi}\beta_{mn}\alpha_{gmi}$ in (8b) is reduced. Although interference is considered in the Gale-S strategy, the mean-interference of this strategy is high. The reason is that, as the users with lower target data rates need lower transmit power in general, the users with lower target data rates may be grouped into one group with less interference and the users with higher target data rates may be grouped into the other group. The interference among the desired signals of users in the group with higher target data rates is very serious. Besides, it should be noted that, due to the inherent ability of ZF beamforming to null interference among the desired signals of users in each group, ZF beamforming is not shown in this figure.

It is clear that interference has great impact on power allocation. In Fig. 5(a) and Fig. 5(b), we illustrate the total transmit power of six different user grouping algorithms with varied numbers of users and APs, respectively. The results show that all the six curves in Fig. 5(a) rise as the number of users increases and all the six curves in Fig. 5(b) decrease as the number of APs increases. That is because that, as shown in Fig. 6(a) and Fig. 6(b), the mean-interference increases with the increase of users and the reduction of APs. Then more power is needed to maintain enough SINR. In addition, the total transmit power of the proposed user grouping algorithms is significantly lower than the reference two, which agrees with the results in Fig. 6(a) and Fig. 6(b). Furthermore, with the proposed user grouping strategies, the transmit power consumption of ZF beamforming is lower than that of MRT beamforming. This result derives from the inherent ability of ZF beamforming to null interference among the desired signals of different users.

In Fig. 5(c), we show the total transmit power with different ranges of target data rate. The results show that the total trans-

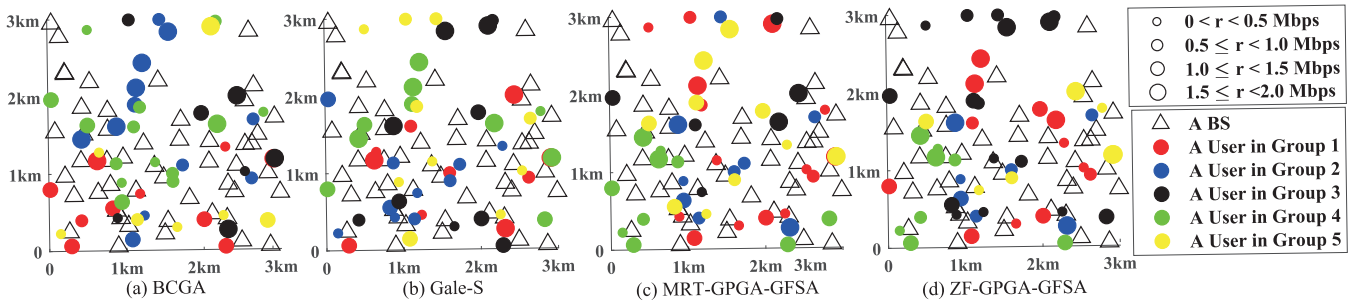


Fig. 8. Distribution of users and APs in 3km × 3km rectangular area(50 users, 50 APs and 5 groups).

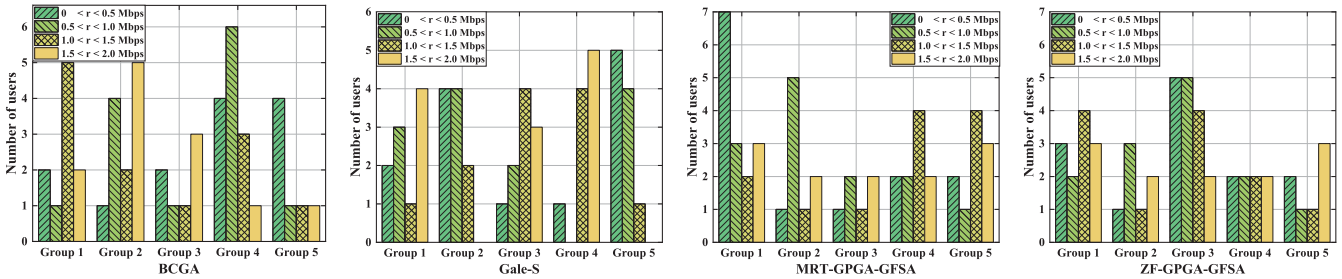


Fig. 9. Number of users in each group and in each range of QoS requirements (50 Users, 50 APs and 5 Groups).

mit power increases with target data rate, and the proposed algorithms outperform the reference algorithm. The reason is that, target SINR of each user, i.e., γ_n, are explicit in constraints (8b) of problem $\mathcal{P}1$. Furthermore, in Fig. 5, the total transmit power of the downlink cell-free massive MIMO system without user grouping (named "Non-grouping"), is compared with that of the proposed system with user grouping. We can see that the column without grouping is higher than the other columns. This result shows that, after user grouping, the minimal transmit power required for satisfying all users' QoS requirements is reduced. An important reason is that, after user grouping, the length of pilot sequences, i.e., τ_g, can be reduced without reducing the accuracy of channel estimation as shown in Fig. 4. Then the number of symbols for data transmission within each coherent interval, i.e., $\tau_c - \sum_{g=1}^{G} \tau_g$, can be increased.

In Fig. 7, under given total transmit power, the average data rate is evaluated with different lengths of coherence interval and group numbers. In Fig. 7(a), we vary the length of coherence interval to show the average data rate with different beamforming methods. As the length of coherence interval increases, the symbols for data transmission will increase, and the length of pilot sequence is unchanged. Therefore, the average data rate will increase. Furthermore, the average data rate of users with ZF beamforming is higher than that of users with conjugate beamforming, the reason is that the interference among the desired signals of different users can be cancelled by ZF beamforming. In order to further demonstrate the advantages of user grouping in cell-free massive MIMO systems, in Fig. 7(b), we show the average data rate vary with the number of groups. We can see that the average data rate can be improved by user grouping. In general, by beamforming among the antennas of many APs, more gains from spatial diversity can be obtained. After user grouping, the number of users sharing the same time-frequency resource will be reduced, and the utilization efficiency of spatial diversity is reduced. However, the pilot overheads will be greatly reduced by user grouping as shown in section V.B. There is a tradeoff between the utilization efficiency of spatial diversity and the pilot overheads. Therefore, in Fig. 7(b), the average data rate will increase from $G = 1$ to $G = 3$, and descend when the number of groups is greater than three.

The number of users sharing the same time-slot in cell-free massive MIMO systems is much larger than traditional communication systems where radio resources of different users are usually orthogonal. In general, the users with higher QoS requirements require more transmit power from APs to guarantee a certain SINR. More transmit power will lead to more severe interference to the desired signals of the other users in the same group. If the users with high QoS requirements be assigned into the same group, the interference among the signals of users in the same group will be too serious to be eliminated and the power consumption will be unbearable. Therefore, to alleviate the serious interference from the desired signals of the other users in the same group, users with high QoS requirements should be avoided to be assigned into the same group. In Fig. 8, we show the distribution of users and APs with different user grouping algorithms, where users and APs are represented by dots and triangles, respectively. The number of APs is 50. There are 50 users which are assigned into 5 groups, and the dots in the same colour represent the users in the same group. The size of each dot reflects its QoS requirement as shown in the legend of Fig. 8. In addition, to analyze the distribution of

users and APs in Fig. 8 more intuitively, we also show the number of users in each group and in each range of QoS requirements in Fig. 9. The results show that there are 5 users whose target data rates are greater than 1.5Mbps in group 2 with user grouping algorithm BCGA, and there are 5 users whose target data rates are greater than 1.5Mbps in group 4 with user grouping algorithm Gale-S, which will bring serious interference from the desired signals of the other users in the same group to the users in this group. By contrast, users with high target data rates are separated into different groups in the proposed algorithms. This also explains why the proposed algorithms outperform the reference algorithm in terms of transmit power, interference from the desired signals of the other users in the same group as shown in Fig. 5(a), Fig. 5(b), Fig. 6(a) and Fig. 6(b). Moreover, the number of users in each group is no more than 16 as Fig. 9 shows. In other words, length of pilot sequence can be effectively reduced by user grouping.

VI. Conclusion

In this paper, we study the joint optimization problem of power allocation and user grouping to minimize the total transmit power in cell-free massive MIMO systems. We decompose this problem into a primal problem: power allocation problem and a master problem: user grouping problem, where the power allocation problem is proved to be convex. We analyze and relax these two problems by GBD method. Then an algorithm based on GBD method is proposed to solve the joint optimization problem by iteratively solving these two problems and reduce the gap between the upper bound and lower bound of the original problem. Moreover, the relaxed master user grouping problem is converted into a problem of searching for some special negative loops in a graph composed of users based on graph theory. An algorithm extended from Bellman-Ford algorithm as well as a fast greedy suboptimal algorithm is proposed to search for these negative loops.

Although the complexity of channel estimation and decoding can be reduced by user grouping, there still remain some challenges in research on cell-free massive MIMO as the number of users increases. For instance, each AP needs to know the transmitted symbols of all users after user grouping, so the limited fronthaul is still one of the bottleneck in cell-free massive MIMO systems as the number of users increases. An effective method for fronthaul reduction is to reducing the number of APs connected with each user (AP grouping). However, the benefit of spatial diversity will also decrease after AP grouping. To serve more users in cell-free massive MIMO systems with limited APs and limited fronthaul, there still remain many works to do.

References

[1] M. Zhang, H. Lu, F. Wu, and C. W. Chen, "NOMA-based scalable video multicast in mobile networks with statistical channels," *IEEE Trans. Mobile Comput.*, vol. 20, no. 6, pp. 2238–2253, Jun. 2021, doi: 10.1109/TMC.2020.2977639.

[2] J. Zhu, D. W. K. Ng, and V. K. Bhargava, "Analysis and design of secure massive MIMO systems in the presence of hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 2001–2016, Mar. 2017.

[3] J. Zhang, L. Dai, Z. He, S. Jin, and X. Li, "Performance analysis of mixed-ADC massive MIMO systems over Rician fading channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1327–1338, Jun. 2017.

[4] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2015, pp. 201–205.

[5] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Oct. 2015.

[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] 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, Jul. 2017.

[8] J. Zhang, X. Xue, E. Björnson, B. Ai, and S. Jin, "Performance analysis and power control of cell-free massive MIMO systems with hardware impairments," *IEEE Access*, vol. 6, pp. 55302–55314, 2018.

[9] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max-min SINR of cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2021–2036, Apr. 2019.

[10] H. Q. Ngo, L.-N. 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.

[11] T. H. Nguyen, T. K. Nguyen, H. D. Han, and V. D. Nguyen, "Optimal power control and load balancing for uplink cell-free multi-user massive MIMO," *IEEE Access*, vol. 6, pp. 14462–14473, 2018.

[12] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and M. Debbah, "Cell-free massive MIMO with limited backhaul," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[13] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited Fronthaul capacity," *IEEE Access*, vol. 7, pp. 44596–44612, 2019.

[14] J. C. Chen, "Low-PAPR precoding design for massive multiuser MIMO systems via Riemannian manifold optimization," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 945–948, Apr. 2017.

[15] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[16] Z. Chen and E. Björnson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018.

[17] Y. Li and G. A. A. Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 1, no. 2, pp. 2–5, May 2018.

[18] M. Attarifar, A. Abbasfar, and A. Lozano, "Modified conjugate beamforming for cell-free massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 616–619, Apr. 2019.

[19] Y. Zhang, M. Zhou, X. Qiao, H. Cao, and L. Yang, "On the performance of cell-free massive MIMO with low-resolution ADCs," *IEEE Access*, vol. 7, pp. 117968–117977, 2019.

[20] X. Zhang, D. Guo, K. An, Z. Ding, and B. Zhang, "Secrecy analysis and active pilot spoofing attack detection for multigroup multicasting cell-free massive MIMO systems," *IEEE Access*, vol. 7, pp. 57332–57340, 2019.

[21] G. Femenias, N. Lassoued, and F. Riera-Palou, "Access point switch ON/OFF strategies for green cell-free massive MIMO networking," *IEEE Access*, vol. 8, pp. 21788–21803, 2020.

[22] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and H. Yang, "Cell-free massive MIMO systems," in *Proc. 49th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2015, pp. 695–699.

[23] A. Ibrahim, T. M. N. Ngatched, and O. A. Dobre, "Using Bender's decomposition for optimal power control and routing in multihop D2D cellular systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5050–5064, Nov. 2019.

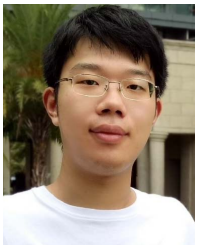
[24] J. Krolikowski, A. Giovanidis, and M. Di Renzo, "A decomposition framework for optimal edge-cache leasing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1345–1359, Jun. 2018.

[25] L. Xiang, D. W. K. Ng, R. Schober, and V. W. S. Wong, "Secure video streaming in heterogeneous small cell networks with untrusted cache helpers," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2645–2661, Apr. 2018.

[26] H. Zhang, S. J. Moura, Z. Hu, W. Qi, and Y. Song, "Joint PEV charging network and distributed PV generation planning based on accelerated generalized Benders decomposition," *IEEE Trans. Transport. Electrific.*, vol. 4, no. 3, pp. 789–803, Sep. 2018.
[27] A. Cherukuri, E. Mallada, S. Low, and J. Cortés, "The role of convexity in saddle-point dynamics: Lyapunov function and robustness," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2449–2464, Aug. 2018.
[28] J. Du, F. R. Yu, X. Chu, J. Feng, and G. Lu, "Computation offloading and resource allocation in vehicular networks based on dual-side cost minimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1079–1092, Feb. 2019.
[29] T. Abrão, S. Yang, L. D. H. Sampaio, P. J. E. Jeszensky, and L. Hanzo, "Achieving maximum effective capacity in OFDMA networks operating under statistical delay guarantee," *IEEE Access*, vol. 5, pp. 14333–14346, 2017.
[30] M. Zhu and S. Martínez, "An approximate dual subgradient algorithm for multi-agent non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 6, pp. 1534–1539, Jun. 2013.
[31] X. Wang, W. Xie, and R. Duan, "Semidefinite programming strong converse bounds for classical capacity," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 640–653, Jan. 2018.
[32] A. M. Geoffrion, "Generalized Benders decomposition," *J. Optim. Theory Appl.*, vol. 10, no. 4, pp. 237–260, Oct. 1972.
[33] L. Angeles, *Generalized Benders Decomposition, GBD*, vol. 10, no. 4. Berlin, Germany: Springer, 2012.
[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[35] F. Guo, H. Lu, X. Jiang, M. Zhang, J. Wu, and C. W. Chen, "QoS-aware user grouping strategy for downlink multi-cell NOMA systems," *IEEE Trans. Wireless Commun.*, early access, Jun. 18, 2021, doi: 10.1109/TWC.2021.3088487.
[36] V. T. Chakaravarthy, F. Checconi, P. Murali, F. Petrini, and Y. Sabharwal, "Scalable single source shortest path algorithms for massively parallel systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 7, pp. 2031–2045, Jul. 2017.
[37] L. Maccari, L. Ghiro, A. Guerrieri, A. Montresor, and R. L. Cigno, "On the distributed computation of load centrality and its application to DV routing," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 2582–2590.
[38] F. Busato and N. Bombieri, "An efficient implementation of the Bellman-Ford algorithm for Kepler GPU architectures," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 8, pp. 2222–2233, Aug. 2016.
[39] X. Hu, C. Zhong, X. Chen, W. Xu, H. Lin, and Z. Zhang, "Cell-free massive MIMO systems with low resolution ADCs," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6844–6857, Oct. 2019.
[40] A. Papazafeiropoulos, P. Kourtessis, M. D. Renzo, S. Chatzinotas, and J. M. Senior, "Performance analysis of cell-free massive MIMO systems: A stochastic geometry approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3523–3537, Apr. 2020.
[41] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2723–2735, Dec. 2017.

Fengqian Guo received the B.S. degree from Jiangnan University, Wuxi, China, in 2017. He is currently pursuing the Ph.D. degree in communication and information systems with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China. His current research interests include wireless transmission and multiple access networks.

Hancheng Lu (Senior Member, IEEE) received the Ph.D. degree in communication and information systems from the University of Science and Technology of China (USTC), Hefei, China, in 2005. He is currently an Associate Professor with the Department of Electronic Engineering and Information Science, USTC. His research interests include multimedia communication and networking, and resource optimization in wireless heterogeneous networks.



Zhuojia Gu received the B.S. degree from the School of Communication and Information Engineering, Shanghai University (SHU), Shanghai, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China. His research interests include resource allocation, wireless edge caching, millimeter-wave communications, and stochastic geometry.