Joint Precoding and RRH Selection for User-Centric Green MIMO C-RAN

Cunhua Pan, *Member, IEEE*, Huiling Zhu, *Member, IEEE*, Nathan J. Gomes, *Senior Member, IEEE*, and Jiangzhou Wang, *Fellow, IEEE*

***Abstract*— This paper jointly optimizes the precoding matrices and the set of active remote radio heads (RRHs) to minimize the network power consumption for a user-centric cloud radio access network, where both the RRHs and users have multiple antennas and each user is served by its nearby RRHs. Both users' rate requirements and per-RRH power constraints are considered. Due to these conflicting constraints, this optimization problem may be infeasible. In this paper, we propose to solve this problem in two stages. In Stage I, a low-complexity user selection algorithm is proposed to find the largest subset of feasible users. In Stage II, a low-complexity algorithm is proposed to solve the optimization problem with the users selected from Stage I. Specifically, the re-weighted l_1-norm minimization method is used to transform the original problem with non-smooth objective function into a series of weighted power minimization (WPM) problems, each of which can be solved by the weighted minimum mean square error (WMMSE) method. The solution obtained by the WMMSE method is proved to satisfy the Karush-Kuhn-Tucker conditions of the WPM problem. Moreover, a low-complexity algorithm based on Newton's method and the gradient descent method is developed to update the precoder matrices in each iteration of the WMMSE method. Simulation results demonstrate the rapid convergence of the proposed algorithms and the benefits of equipping multiple antennas at the user side. Moreover, the proposed algorithm is shown to achieve near-optimal performance in terms of NPC.**

***Index Terms*— Cloud radio access network (C-RAN), user-centric network, MIMO systems, user selection, green communications.**

I. Introduction

Mobile communications has been developing very rapidly [2]–[4]. In recent years, C-RAN has been proposed as a promising solution to support the exponential growth of mobile data traffic [5], [6]. In C-RAN, all the baseband processing is performed at the baseband unit (BBU) pool with powerful computation capacity, while the remote radio heads (RRHs) perform the basic functionalities of signal processing [7], [8]. The RRHs are geographically distributed away from each other, but connected to the BBU pool through optical fiber transport links. Under the C-RAN architecture, centralized signal processing technologies can be realized. Hence, significant performance gains can be achieved. In addition, the RRHs can be densely deployed in the network with low operation cost due to their simple functionalities. This will significantly reduce the average access distance for the users, and thus lowers the transmission power.

On the other hand, it was reported that the total energy consumption of wireless communications contributes more than 3 percent of the worldwide electrical energy consumption [9], and this portion is expected to grow in the near future due to the explosive growth of high-data-rate applications and mobile devices. Hence, energy efficiency has attracted extensive interest and becomes one of the main performance metrics in the future fifth generation (5G) systems [10]. When a large number of RRHs are deployed in the network, the network power consumption (NPC) of C-RAN will become considerable due to the increasing circuit power consumption of the RRHs. Fortunately, it was reported in [11] that the traffic load varies substantially over both time and space due to user mobility and varying channel state. Hence, the NPC can be significantly reduced by putting some RRHs with light load into sleep mode while maintaining the quality of service (QoS) requirements of the users, which is the focus of this paper.

Recently, the NPC minimization problem for C-RAN has been extensively studied in [12]–[22]. These papers formulated the joint RRH selection and beamforming vector optimization problem as a mixed-integer non-linear programming (MINLP) problem, which has a nonconvex discontinuous l_0-norm in the objective function or constraints. We summarize the existing approaches to solve the MINLP problem as follows. The first approach was proposed in [12], which first reformulated the problem as an extended mixed integer second-order cone programming (SOCP) and then applied the branch-and-cut method to obtain the optimal solution. In the second approach in [13], [14], the MINLP was first decomposed into a master problem and a beamforming subproblem. Then, an iterative algorithm based on the Benders decomposition was derived to find the optimal solution. Although these two approaches yield the optimal solution, they have an exponential complexity. The third approach is the smooth function method, where the l_0-norm was approximated as Gaussian-like function in [15], the exponential function in [16], and arctangent function in [17]. However, the smooth function cannot produce sparse solutions in general. The last approach was inspired by the

Manuscript received March 16, 2016; revised November 27, 2016 and February 7, 2017; accepted February 10, 2017. Date of publication March 17, 2017; date of current version May 8, 2017. This work is supported by European Commission Horizon 2020 Project iCIRRUS under Grant 644526 and NIRVANA Project. Part of this paper was presented at IEEE Globecom, December 2016 [1]. The associate editor coordinating the review of this paper and approving it for publication was N. Prasad.

C. Pan was with the School of Engineering and Digital Arts, University of Kent, Canterbury CT2 7NZ, U.K. He is now with with School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: c.pan@kent.ac.uk).

H. Zhu, N. J. Gomes, and J. Wang are with the School of Engineering and Digital Arts, University of Kent, Canterbury CT2 7NZ, U.K. (e-mail: h.zhu@kent.ac.uk; n.j.gomes@kent.ac.uk; j.z.wang@kent.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2017.2671358

compression sensing, named re-weighted l_1-norm minimization method [23]. This method has been widely adopted in the literature [18]–[22], [24] due to its low computational complexity and sparsity guarantee, which will also be applied in this paper.

All of the above papers only considered the single-antenna user (SAU) case. With the increasing development in antenna technology [25], [26], it is possible to equip the wireless devices with multiple antennas. When both the transmitter and the receiver are equipped with multiple antennas, multiple streams can be transmitted simultaneously, rather than only one stream in the SAU case. Simulation results show that with the increasing number of receive antennas, more users can be admitted. Therefore, in this paper, we consider the multiple-antenna user (MAU) case and jointly optimize the precoding matrices and the set of active RRHs to minimize the NPC subject to users' rate requirements and per-RRH power constraints.

Unfortunately, the techniques in [12]–[22] dealing with the SAU case cannot be extended directly to the MAU case. The reasons are as follows. Firstly, since the rate constraints and power constraints are conflicting with each other, this problem may be infeasible. In the SAU networks, the rate requirements can be equivalently represented as signal-to-interference-plus-noise ratio (SINR) constraints, which can be transformed into an SOCP problem. Hence, the feasibility of the original problem can be easily checked by solving the SOCP feasibility problem. However, the rate constraints in the MAU case is non-convex and much more complex due to the complicated rate expression, which cannot be transformed into the SOCP formulation as in the SAU case. Hence, new techniques need to be developed to check the feasibility of the original problem. Secondly, even though the original problem is checked to be feasible, how to solve it is still difficult, since it cannot be transformed into an SOCP problem as in the SAU case. [27] proposed the weighted minimum mean square error (WMMSE) method to solve the rate maximization problem for MIMO interfering broadcast channels, where the rate expression is in the objective function. Recently, there have been some work in applying the WMMSE method to solve the energy efficiency (measured in bit/s/Joule) optimization problems under rate constraints [28], [29]. However, these researches have not addressed the feasibility problem due to the incorporated rate constraints. Only in [29], a heuristic method was proposed to check the feasibility based on the interference alignment technique, under the assumption that the transmit power is approaching infinity, which in not practical. Since the problem considered in this paper imposes power constraints at each RRH, the heuristic method developed in [29] is not applicable. More importantly, they have not revealed the hidden property of applying WMMSE method to the optimization problem with rate constraints, such as the convergence property and the optimality of the solutions.

To the best of our knowledge, this paper is the first attempt to solve the joint RRH and precoding optimization problem to minimize the NPC in the MAU based user-centric C-RAN, where each user can be served by an arbitrary subset of RRHs. Due to the conflicting constraints, this problem may be infeasible. Some users should be removed or rescheduled for the next transmission to guarantee the rate requirements of other users. We provide a comprehensive analysis for this problem by considering two stages: user selection in Stage I and algorithm design in Stage II. The main contributions of this paper are summarized as follows:

1) In Stage I, a low-complexity user selection approach is proposed to maximize the number of admitted users that can have their QoS requirements satisfied. Specifically, in each step we solve an alternative problem by introducing a series of auxiliary variables. This alternative problem is always feasible. By replacing the rate expression in the constraints with its lower-bound, an iterative algorithm is proposed to solve this problem along with the complexity and convergence analysis of the algorithm. The alternative problem should be solved at most K times, where K is the total number of users. Its complexity is much lower than the optimal exhaustive user selection method that has an exponential complexity. Simulation results show that both algorithms achieve similar performance.
2) In Stage II, a low-complexity algorithm is proposed to solve the NPC minimization problem with the users selected from Stage I. Specifically, the re-weighted l_1-norm minimization method is adopted to convert the non-smooth optimization problem into a series of smooth weighted power minimization (WPM) problems. We again replace the rate expression with its lower-bound and adapt the WMMSE algorithm originally designed for a rate maximization problem to solve the WPM problem. In addition, we strictly prove that when the WMMSE algorithm is initialized with a feasible solution, the sequences of precoder matrices generated in the iterative procedure will finally converge to the Karush-Kuhn-Tucker (KKT) point of the WPM problem.
3) In each iteration of the WMMSE algorithm, there is a subproblem for the precoder matrices being updated with some other fixed variables. Most existing papers [21], [28]–[31] directly transform it into an SOCP problem and apply the interior point method [32] to solve it, which may incur high computational complexity. In this paper, we go one step further and develop a low-complexity algorithm to solve this subproblem by exploiting its special structure. Specifically, we equivalently solve its dual problem because the subproblem is a convex problem. Fortunately, the objective function of the dual problem is differentiable, and the block coordinate descent (BCD) method is adopted to solve the dual problem. In each iteration of the BCD method, Newton's method and the gradient descent method are applied to update the Lagrangian multipliers. It is strictly proved that the BCD method can obtain the globally optimal solution of the subproblem. Complexity analysis in conjunction with the simulation results show that the BCD method has a much lower computational complexity than the interior point method.

This paper is organized as follows. In Section II, we introduce the system model and formulate the optimization

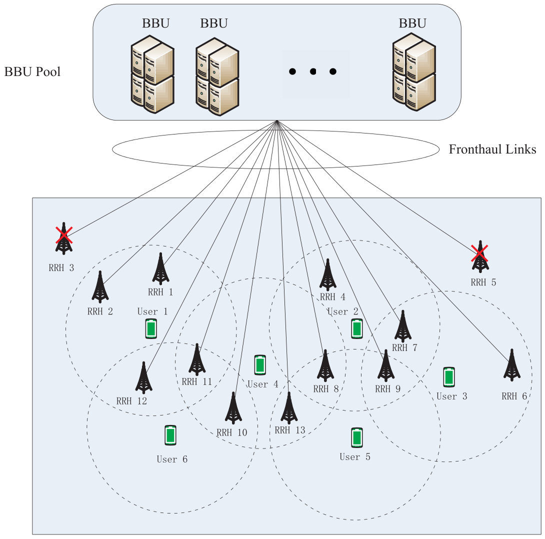


Fig. 1. Illustration of a C-RAN with thirteen RRHs and six users, where user-centric clustering technique is adopted. In this example, each user is served by its nearby RRHs within the dotted circle centered at itself. The RRHs that are not in any users' candidate set are turned into idle mode, such as RRH3 and RRH 5.

problems. In Section III, a new approach is introduced to select the maximum number of admitted users. An iterative algorithm with low complexity is provided in Section IV. Simulation results are presented in Section V. Conclusions are drawn in Section VI.

Notations: Uppercase and lowercase boldface denote matrices and vectors, respectively. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_F$ denotes the Frobenius norm of $\mathbf{A}$ and $\mathbf{A}^H$ represents the Hermitian transpose of $\mathbf{A}$. $\mathbf{I}_m$ denotes a $m \times m$ identity matrix. For a vector $\mathbf{a}$, $\mathrm{diag}(\mathbf{a})$ denotes the diagonal matrix with diagonal elements given by $\mathbf{a}$. $\mathrm{blkdiag}(\cdot)$ represent the block-diagonal matrices. $\mathbb{E}(\cdot)$, and $\mathrm{Tr}(\cdot)$ represent expectation, trace operators, respectively. $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix. For vector $\mathbf{a} \in \mathbb{C}^{n\times 1}$, $\|\mathbf{a}\|_2$ is the Euclidean norm. $\mathcal{CN}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ represents the complex circularly symmetric Gaussian distribution with zero mean vector and covariance matrix $\sigma^2\mathbf{I}$. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_0$ is l_0-norm, means the number of nonzero entries in a vector.

II. System Model and Problem Formulation

A. System Model

Consider a downlink C-RAN consisting of I RRHs and K users,[1] where each RRH is equipped with M transmit antennas and each user has N receive antennas, as shown in Fig. 1. Denote the set of RRHs and users as $\mathcal{I} = \{1, \cdots, I\}$ and $\bar{\mathcal{U}} = \{1, \cdots, K\}$, respectively. It is assumed that each RRH is connected to the BBU pool via fronthaul link and the BBU pool has access to all users' CSI and data information.

Let $\mathcal{U} \subseteq \bar{\mathcal{U}}$ be the set of users that can be admitted to this networks. To reduce the computational complexity of the dense network, the user-centric clustering method is adopted, where each user $k \in \mathcal{U}$ is assumed to be served by its nearby RRHs since the distant RRHs contribute less to user's signal quality due to the large path loss. The unselected RRHs are turned into idle mode, such as RRH 3 and RRH 5 in Fig. 1. Let $\mathcal{I}_k \subseteq \mathcal{I}$ and $\mathcal{U}_i \subseteq \mathcal{U}$ be the candidate set of RRHs for serving user k and candidate set of users served by RRH i, respectively. Note that the set of RRHs serving the users may overlap with each other. For example, in Fig. 1, RRH 12 jointly serves user 1 and user 6.

Denote $\mathbf{V}_{i,k} \in \mathbb{C}^{M\times d}$ as the precoding matrix used by the ith RRH to transmit data vector $\mathbf{s}_k \in \mathbb{C}^{d\times 1}$ to the kth user, where d is the number of data streams for each user, and $\mathbf{s}_k$ satisfies $\mathbb{E}\left[\mathbf{s}_k\mathbf{s}_k^{\mathrm{H}}\right] = \mathbf{I}_d$ and $\mathbb{E}\left[\mathbf{s}_k\mathbf{s}_l^{\mathrm{H}}\right] = \mathbf{0}$, for $l \neq k$. Let $\bar{\mathbf{V}}_k = \left[\mathbf{V}_{i,k}^{\mathrm{H}}, \forall i \in \mathcal{I}_k\right]^{\mathrm{H}} \in \mathbb{C}^{|\mathcal{I}_k|M\times d}$ be the big precoding matrix for user k from all RRHs in $\mathcal{I}_k$. In addition, define a set of new channel matrices $\bar{\mathbf{H}}_{j,k} = [\mathbf{H}_{i,k}, \forall i \in \mathcal{I}_j] \in \mathbb{C}^{N\times|\mathcal{I}_l|M}$, representing the overall CSI from RRHs in $\mathcal{I}_j$ to user k, where $\mathbf{H}_{i,k} \in \mathbb{C}^{N\times M}$ denotes the channel matrix from the ith RRH to the kth user. Then, the received signal vector at the kth user, denoted as $\mathbf{y}_k \in \mathbb{C}^{N\times 1}$, is given by

$$\mathbf{y}_k = \bar{\mathbf{H}}_{k,k}\bar{\mathbf{V}}_k\mathbf{s}_k + \sum\nolimits_{j\in\mathcal{U}, j\neq k} \bar{\mathbf{H}}_{j,k}\bar{\mathbf{V}}_j\mathbf{s}_j + \mathbf{n}_k, \tag{1}$$

where $\mathbf{n}_k$ is the noise vector at the kth user, which satisfies $\mathcal{CN}\left(\mathbf{0}, \sigma_k^2\mathbf{I}_N\right)$. Then, the achievable rate (nat/s/Hz) of the kth user is given by [35]

$$R_k(\mathbf{V}) = \log\left|\mathbf{I} + \bar{\mathbf{H}}_{k,k}\bar{\mathbf{V}}_k\bar{\mathbf{V}}_k^H\bar{\mathbf{H}}_{k,k}^H\mathbf{J}_k^{-1}\right|, \tag{2}$$

where $\log(\cdot)$ is the base of natural logarithm, $\mathbf{J}_k = \sum_{j\in\mathcal{U}, j\neq k}\bar{\mathbf{H}}_{j,k}\bar{\mathbf{V}}_j\bar{\mathbf{V}}_j^H\bar{\mathbf{H}}_{j,k}^H + \sigma_k^2\mathbf{I}$ is the interference-plus-noise covariance matrix, and $\mathbf{V}$ is the collection of all precoding matrices. Each user's data rate should be larger than the minimum requirement:

$$\mathrm{C1}: R_k(\mathbf{V}) \geq R_{k,\min}, \quad \forall k \in \mathcal{U}. \tag{3}$$

With densely deployed RRHs, the power consumption on the RRHs and the corresponding fronthaul links may be significant. Switching off some RRHs and the corresponding fronthual links may be a good option to reduce the NPC. To this end, it is critical to model the NPC.

B. NPC Model

The realistic NPC model should consist of three parts: power consumption at the RRHs, that at the fronthaul links and that at the BBU pool.

As in [18], the power consumption at RRH i can be modeled as follows:

$$P_i^{\mathrm{rrh}}(\mathbf{V}) = \begin{cases} \eta_i P_i^{\mathrm{tr}}(\mathbf{V}) + M P_i^{\mathrm{a,rrh}}, & \text{if } P_i^{\mathrm{tr}}(\mathbf{V}) > 0 \\ M P_i^{\mathrm{s,rrh}}, & \text{if } P_i^{\mathrm{tr}}(\mathbf{V}) = 0 \end{cases} \tag{4}$$

where $\eta_i > 1$ accounts for the inefficiency of the power amplifier of RRH i, $P_i^{\mathrm{tr}}(\mathbf{V})$ is the total transmit power of RRH i given by $P_i^{\mathrm{tr}}(\mathbf{V}) = \sum_{k\in\mathcal{U}_i}\left\|\mathbf{V}_{i,k}\right\|_F^2$ that satisfies the power constraint:

$$\mathrm{C2}: P_i^{\mathrm{tr}}(\mathbf{V}) \leq P_{i,\max}, \quad \forall i \in \mathcal{I}, \tag{5}$$

[1] In dense networks, the number of RRHs may be larger than the number of users so that the average distance between serving RRHs and users can be significantly reduced, leading to improved performance. In some extreme cases, each user may be served by its dedicated RRHs as in [33], [34], where each RRH serves only one user.

$P_i^{\rm a,rrh}$ and $P_i^{\rm s,rrh}$ represent the power consumption for each antenna (or each RF chain) when RRH i is in active mode and sleep mode, respectively. In practical systems, $P_i^{\rm active}$ is much higher than $P_i^{\rm sleep}$, which motivates us to switch off some RRHs.

In general, more power consumption will be consumed on the fronthaul links when they support high data rates. In [22], this power was modeled to be proportional to the total fronthaul data rate. We modify the model in [22] to account for the power when the fronthaul links are in the sleep mode as follows:

$$P_i^{\rm fr}(\mathbf{V}) = \begin{cases} \rho_i \sum_{k\in\mathcal{U}_i} R_k(\mathbf{V}) + P_i^{a,\rm fr}, & \text{if } P_i^{\rm tr}(\mathbf{V}) > 0,\\ P_i^{s,\rm fr}, & \text{if } P_i^{\rm tr}(\mathbf{V}) = 0. \end{cases} \tag{6}$$

where ρ_i is the proportional factor for fronthaul link i. The power consumed in the BBU pool mainly depends on the computational complexity for signal processing. However, how to accurately model this kind of power consumption is still not fully understood. As in most papers [12], [18], [19], [22], the BBU power consumption is modeled as a constant $P_{\rm BBU}$ for simplicity. Let $\mathcal{A}$ denote the active RRH set. Then, the NPC can be modeled as

$$\hat{P}(\mathcal{A},\mathbf{V}) = \sum_{i\in\mathcal{I}} \left(P_i^{\rm rrh}(\mathbf{V}) + P_i^{\rm fr}(\mathbf{V})\right) + P_{\rm BBU} \tag{7}$$

$$= \sum_{i\in\mathcal{A}} \left(\eta_i P_i^{\rm tr}(\mathbf{V}) + \rho_i \sum_{k\in\mathcal{U}_i} R_k(\mathbf{V}) + P_i^c\right) + \sum_{i\in\mathcal{I}} P_i^s + P_{\rm BBU}, \tag{8}$$

where P_i^c and P_i^s are two constants, given by $P_i^c = M(P_i^{a,\rm rrh} - P_i^{s,\rm rrh}) + P_i^{a,\rm fr} - P_i^{s,\rm fr}$ and $P_i^s = MP_i^{s,\rm rrh} + P_i^{s,\rm fr}$.

C. Problem Formulation

Due to the power constraints C2, the rate requirements C1 may not be satisfied for all users. Some users should be removed to make the optimization problem feasible. Hence, we formulate a two-stage optimization problem. In Stage I, one should maximize the number of admitted users that can be supported by the system; in Stage II, one should jointly select some RRHs and optimize the precoding matrices to minimize the NPC with the selected users from Stage I.

Specifically, the optimization problem in Stage I can be formulated as

$$\begin{aligned} &\max_{\mathbf{V},\,\mathcal{U}\subseteq\overline{\mathcal{U}}} \ |\mathcal{U}| \\ &\text{s.t. C1, C2.} \end{aligned} \tag{9}$$

Then in Stage II, we aim to jointly select the RRHs and optimize the precoding matrices to minimize the NPC with the users selected from Stage I, which can be formulated as[2]

$$\min_{\mathcal{A},\mathbf{V}} \ \sum_{i\in\mathcal{A}} \left(\eta_i P_i^{\rm tr}(\mathbf{V}) + \rho_i \sum_{k\in\mathcal{U}_i^\star} R_k(\mathbf{V}) + P_i^c\right)$$

$$\text{s.t. C1}, \sum_{k\in\mathcal{U}_i^\star} \left\|\mathbf{V}_{i,k}\right\|_F^2 \le P_{i,\max}, i \in \mathcal{A}, \tag{10a}$$

$$\sum_{k\in\mathcal{U}_i^\star} \left\|\mathbf{V}_{i,k}\right\|_F^2 = 0, i \in \mathcal{I}\backslash\mathcal{A}, \tag{10b}$$

where $\mathcal{U}_i^\star$ is the solution from Stage I. Note that when the system parameters are given, the last two terms in (8) are constants, and are omitted in the objective function.

Both the optimization problems in the two stages are MINLP problems and are difficult to solve. The intuitive approach to solve this kind of problems is through the exhaustive search. For example, to solve the NPC minimization problem in Stage II, one must solve the precoding matrices that minimizes the NPC with each given $\mathcal{A}$ and obtain the corresponding objective value. Finally, the $\mathcal{A}$ that achieves the minimum NPC together with the corresponding precoding matrices is the optimal solution of Problem (10). However, the exhaustive search has exponentially prohibitive complexity with respect to the number of RRHs, which is hard to be implemented in practice in dense C-RANs. The same issue holds for the user selection problem in Stage I, where the exhaustive search method has an exponential complexity of the number of users. Hence, this motivates us to develop low-complexity algorithms to solve these two Problems.

III. Stage I: Low-Complexity User Selection Algorithm

In this section, we provide a low-complexity user selection algorithm to guarantee the rate requirements of other users. Specifically, for an arbitrary given subset of users $\mathcal{U}$, we construct an alternative problem by introducing a series of auxiliary variables $\{\alpha_k\}_{k\in\mathcal{U}}$:

$$\begin{aligned} &\min_{\{\alpha_k\}_{k\in\mathcal{U}},\mathbf{V}} \ \sum_{k\in\mathcal{U}} (\alpha_k - 1)^2 \\ &\text{s.t. C2}, R_k(\mathbf{V}) \ge \alpha_k^2 R_{k,\min}, \quad \forall k \in \mathcal{U}, \end{aligned} \tag{11}$$

Obviously, Problem (11) is always feasible and the optimal α_k for each user k should be no larger than one. This can be easily proved by contradiction. Moreover, user k can be admitted if and only if the optimal α_k is equal to one. Hence, maximizing the number of admitted users is equal to finding the largest subset of users $\mathcal{U}$, in which all $\{\alpha_k\}_{k\in\mathcal{U}}$ are equal to one.

Based on the above analysis, we provide a low-complexity user selection (USC) algorithm to solve Problem (9) in Stage I. The main idea is to remove each user with the least $\alpha_k < 1$ in each iteration. It is intuitive since the user with the least α_k has the largest gap to its rate target.

Algorithm 1 USC Algorithm

```
1: Initialize the set of users $\mathcal{U} = \{1, \cdots, K\}$;
2: Given $\mathcal{U}$, solve Problem (11) by Algorithm 2 in Subsection III-A to obtain $\{\alpha_k\}_{k\in\mathcal{U}}$ and $\mathbf{V}$;
3: If $\alpha_k = 1, \forall k \in \mathcal{U}$, output $\mathbf{V}$ and $\mathcal{U}^* = \mathcal{U}$ for the initialization of Stage II and terminate; Otherwise, find $k^* = \arg\min_{k\in\mathcal{U}} \alpha_k$, remove user $k^*$ and update $\mathcal{U} = \mathcal{U}\backslash k^*$, go to step 2.
```

A. Algorithm to Solve Problem (11)

In step 2 of Algorithm 1, Problem (11) needs to be solved. Due to constraints C3 in (11), Problem (11) is a non-convex

[2] In general, the number of transmit antennas should be optimized to additionally reduce the NPC as seen in the RRH power consumption model in (4). However, the resulting problem will be much more difficult to solve, and will be left for future work.

problem, which is difficult to solve. To handle this difficulty, we apply the relationships between WMMSE and the rate expression.

We consider the linear receiver filter so that the estimated signal vector is given by

$$\hat{\mathbf{s}}_k = \mathbf{U}_k^H \mathbf{y}_k, \quad \forall k \in \mathcal{U}. \tag{12}$$

where $\mathbf{U}_k \in \mathbb{C}^{N \times d}$ is the receiver filter of the kth user. Since the signal vectors $\mathbf{s}_k$'s and noise $\mathbf{n}_k$'s are mutually independent, the mean square error (MSE) matrix at the kth user is given by

$$\begin{aligned}\mathbf{E}_k &= \mathbb{E}_{\mathbf{s},\mathbf{n}}\left[\left(\hat{\mathbf{s}}_k - \mathbf{s}_k\right)\left(\hat{\mathbf{s}}_k - \mathbf{s}_k\right)^H\right] \\ &= \left(\mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k - \mathbf{I}_d\right)\left(\mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k - \mathbf{I}_d\right)^H \\ &\quad + \sum_{j \in \mathcal{U}, j \neq k} \mathbf{U}_k^H \bar{\mathbf{H}}_{j,k} \bar{\mathbf{V}}_j \bar{\mathbf{V}}_j^H \bar{\mathbf{H}}_{j,k}^H \mathbf{U}_k + \sigma_k^2 \mathbf{U}_k^H \mathbf{U}_k. \end{aligned} \tag{13}$$

By introducing a set of auxiliary matrices $\{\mathbf{W}_k \succeq \mathbf{0}\}$, we define the following functions

$$h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right) = \log\left|\mathbf{W}_k\right| - \mathrm{Tr}\left(\mathbf{W}_k \mathbf{E}_k\right) + d, \quad \forall k. \tag{14}$$

where $\mathbf{E}_k$ is the MSE matrix of user k given in (13). The following lemma establishes the relationships between the rate expression and function $h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right)$.

Lemma 1 [27]: $h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right)$ is a concave function for each set of the matrices $\mathbf{V}$, $\mathbf{U}_k$ and $\mathbf{W}_k$ when the other two are given. Given $\mathbf{V}$, $h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right)$ is the lower-bound of the data rate $R_k(\mathbf{V})$ in (2). The optimal $\mathbf{U}_k, \mathbf{W}_k$ for $h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right)$ to achieve the data rate is given by

$$\begin{aligned}\mathbf{U}_k^{\star} &= \left(\sum_{j \in \mathcal{U}} \bar{\mathbf{H}}_{j,k} \bar{\mathbf{V}}_j \bar{\mathbf{V}}_j^H \bar{\mathbf{H}}_{j,k}^H + \sigma_k^2 \mathbf{I}\right)^{-1} \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k, \\ \mathbf{W}_k^{\star} &= \mathbf{E}_k^{\star -1}, \quad \forall k, \end{aligned} \tag{15}$$

where $\mathbf{E}_k^{\star}$ is obtained by plugging the expression of $\mathbf{U}_k^{\star}$ into the kth user's MSE in (13)

$$\mathbf{E}_k^{\star} = \mathbf{I}_d - \bar{\mathbf{V}}_k^H \bar{\mathbf{H}}_{k,k}^H \left(\sum_{j \in \mathcal{U}} \bar{\mathbf{H}}_{j,k} \bar{\mathbf{V}}_j \bar{\mathbf{V}}_j^H \bar{\mathbf{H}}_{j,k}^H + \sigma_k^2 \mathbf{I}\right)^{-1} \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k.$$

□

By replacing the first set of constraints in (11) with its lower-bound $h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right)$, we have the following optimization problem

$$\begin{aligned} &\min_{\{\alpha_k\}_{k \in \mathcal{U}}, \mathbf{V}, \mathbf{W}, \mathbf{U}} \sum\nolimits_{k \in \mathcal{U}} \left(\alpha_k - 1\right)^2 \\ &\text{s.t. C2}, h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right) \geq \alpha_k^2 R_{k,\min}, \quad \forall k \in \mathcal{U}, \end{aligned} \tag{16}$$

where $\mathbf{U}$ and $\mathbf{W}$ are the collection of matrices $\mathbf{U}_k, \forall k$ and $\mathbf{W}_k, \forall k$, respectively.

To solve Problem (16), we apply the block coordinate descent method: given $\mathbf{V}$, update $\mathbf{U}$ and $\mathbf{W}$ by using (15); update $\{\alpha_k\}_{k \in \mathcal{U}}$ and $\mathbf{V}$ with given $\mathbf{U}$ and $\mathbf{W}$. We only need to solve the latter one. Putting the MSE expression in (13) into constraints C4 in Problem (16) yields

$$\begin{aligned} &\min_{\{\alpha_k\}_{k \in \mathcal{U}}, \mathbf{V}} \sum\nolimits_{k \in \mathcal{U}} \left(\alpha_k - 1\right)^2 \\ &\text{s.t. C2}, \\ &\quad \text{C5} : \mathrm{Tr}\left(\left(\mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k - \mathbf{I}_k\right)^H \mathbf{W}_k \left(\mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k - \mathbf{I}_k\right)\right) \\ &\quad + \sum_{j \in \mathcal{U}, j \neq k} \mathrm{Tr}\left(\bar{\mathbf{V}}_j^H \bar{\mathbf{H}}_{j,k}^H \mathbf{U}_k \mathbf{W}_k \mathbf{U}_k^H \bar{\mathbf{H}}_{j,k} \bar{\mathbf{V}}_j\right) \\ &\quad + \alpha_k^2 R_{k,\min} \leq t_k, \quad \forall k, \end{aligned} \tag{17}$$

where $t_k = \log\left|\mathbf{W}_k\right| + d - \sigma_k^2 \mathrm{Tr}\left(\mathbf{U}_k^H \mathbf{U}_k \mathbf{W}_k\right)$.

Without loss of generality, we assume $\mathcal{U} = \bar{\mathcal{U}} = \{1, \cdots, K\}$ and define the indices of $\mathcal{U}_i$ as $\mathcal{U}_i = \{q_1^i, \cdots, q_{|\mathcal{U}_i|}^i\}$. Problem (17) can be equivalently transformed into the following problem

$$\begin{aligned} &\min_{\{\alpha_k\}_{k \in \mathcal{U}}, \mathbf{V}} \sum\nolimits_{k \in \mathcal{U}} \left(\alpha_k - 1\right)^2 \\ &\text{s.t. } \|\mathbf{x}_k\|_2 \leq \sqrt{t_k}, \quad \forall k \in \mathcal{U}, \\ &\quad \|\mathbf{y}_i\|_2 \leq \sqrt{P_{i.\max}}, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{18}$$

where $\mathbf{x}_k$ is given by

$$\begin{aligned}\mathbf{x}_k = \Big[&\alpha_k \sqrt{R_{k,\min}}, \mathrm{vec}\left(\bar{\mathbf{V}}_1^H \bar{\mathbf{H}}_{1,k}^H \mathbf{U}_k \mathbf{W}_k^{1/2}\right)^H, \cdots, \\ &\mathrm{vec}\left(\left(\bar{\mathbf{V}}_k^H \bar{\mathbf{H}}_{k,k}^H \mathbf{U}_k - \mathbf{I}_k\right) \mathbf{W}_k^{1/2}\right)^H, \cdots, \\ &\mathrm{vec}\left(\bar{\mathbf{V}}_K^H \bar{\mathbf{H}}_{K,k}^H \mathbf{U}_k \mathbf{W}_k^{1/2}\right)^H \Big]^H \end{aligned}$$

and $\mathbf{y}_i$ is given by

$$\mathbf{y}_i = \left[\mathrm{vec}\left(\mathbf{V}_{i,q_1^i}\right)^H, \cdots, \mathrm{vec}\left(\mathbf{V}_{i,q_{|\mathcal{U}_i|}^i}\right)^H\right]^H. \tag{19}$$

Problem (18) is an SOCP problem for which a globally optimal solution can be obtained by existing techniques such as interior point method [32].

Based on the above analysis, the iterative algorithm for solving Problem (11) is formally described in Algorithm 2.

Theorem 1: Algorithm 2 will converge during the iterative procedure.

Proof: Please see Appendix A. □

Algorithm 2 Iterative Algorithm

```
1: Initialize iterative number $n = 1$, the maximum number of iterations $n_{\max}$. Initial precoding matrices $\mathbf{V}^{(0)}$ such that the per-RRH power constraints are satisfied. Calculate $\mathbf{U}^{(0)}$ and $\mathbf{W}^{(0)}$ by using (15) with $\mathbf{V}^{(0)}$;
2: With $\mathbf{U}^{(n-1)}$ and $\mathbf{W}^{(n-1)}$, update $\{\alpha_k^{(n)}\}_{k \in \mathcal{U}}$ and $\mathbf{V}^{(n)}$ by solving the SOCP problem (18);
3: Update $\mathbf{U}^{(n)}$ and $\mathbf{W}^{(n)}$ as in (15) with $\mathbf{V}^{(n)}$;
4: If $n < n_{\max}$, set $n \leftarrow n + 1$ and go to step 2. Otherwise, terminate.
```

B. Overall Complexity to Solve Problem (9) in Stage I

We first analyze the complexity of Algorithm 2 to solve Problem (11). For simplicity, we assume that candidate size for each user is equal, i.e., $|\mathcal{I}_k| = l$, and $\mathcal{U} = \bar{\mathcal{U}}$. In each iteration of Algorithm 2, the main complexity lies in step 2, where the SOCP Problem (18) is solved. This problem has $2MKld + K$ real variables, K SOC constraints where each has $2Kd^2 + 1$ real dimensions, and I SOC constraints where each has $2Md\,|\mathcal{U}_i|$ real dimensions. From [page 196, [36]], the complexity is $O\left((2MKld + K)^2\left(2K^2d^2 + K + 2Md\sum_{i\in\mathcal{I}}|\mathcal{U}_i|\right)\right)$, and the total number of iterations required is $O\left(\sqrt{I+K}\right)$. Note that $\sum_{i\in\mathcal{I}}|\mathcal{U}_i| = \sum_{k\in\mathcal{U}}|\mathcal{I}_k| = Kl$, the total complexity to solve the SOCP Problem (18) is given by $O\left(\sqrt{I+K}(2MKld + K)^2\left(2K^2d^2 + K + 2MdKl\right)\right)$. Finally, Algorithm 2 should be run at most K times, then the overall complexity to solve Problem (9) in Stage I is at most $T_{\text{StageI}} = O(K\sqrt{I+K}(2MKld + K)^2$ $\left(2K^2d^2 + K + 2MdKl\right))$.

IV. Stage II: A Low-Complexity Algorithm to Solve Problem (10)

In this section, we provide a low-complexity algorithm to solve Problem (10) with the selected users from Stage I. First, we adopt the re-weighted l_1-norm method [23] to transform the original non-smooth optimization problem into a series of WPM problems. Then, the WPM problem is solved by the WMMSE algorithm. In each iteration of the WMMSE algorithm, there is a subproblem that the precoder matrices should be optimized. We exploit the special structure of the subproblem and develop a low-complexity algorithm to solve it.

A. Reweighted l_1-Norm Minimization

For simplicity, the subscript in $\mathcal{U}^\star$ is omitted. It is easy to see that the minimum rate constraints in C1 of Problem (10) hold with equality at the optimal point, i.e., $R_k(\mathbf{V}) = R_{k,\min}, \forall k$. Then, defining $\tilde{P}_i^c \triangleq \rho_i\sum_{k\in\mathcal{U}_i^\star} R_{k,\min} + P_i^c$ and using the l_0-norm, the objective function of Problem (10) is equivalent to $\sum_{i\in\mathcal{I}}$ $\left(\eta_i\sum_{k\in\mathcal{U}_i}\|\mathbf{V}_{i,k}\|_F^2 + \left\|\sum_{k\in\mathcal{U}_i}\|\mathbf{V}_{i,k}\|_F^2\right\|_0\tilde{P}_i^c\right)$. This rewritten expression enables us to apply the compressive sensing techniques [37], where the non-smooth l_0-norm objective can often be approximated by a re-weighted l_1-norm, i.e.,

$$\left\|\sum_{k\in\mathcal{U}_i}\|\mathbf{V}_{i,k}\|_F^2\right\|_0 \approx a_i^{(n)}\sum_{k\in\mathcal{U}_i}\|\mathbf{V}_{i,k}\|_F^2, \tag{20}$$

where $a_i^{(n)}$ is a weight factor of the ith RRH at the nth iteration that is iteratively updated as

$$a_i^{(n)} = \frac{1}{\sum_{k\in\mathcal{U}_i}\left\|\mathbf{V}_{i,k}^{(n)}\right\|_F^2 + \delta}, \quad \forall i, \tag{21}$$

where δ is a small constant regularization parameter and $\mathbf{V}_{i,k}^{(n)}$ is the solution in the nth iteration. The above updating rule shows that those RRHs with lower transmit power in the previous iteration will have larger weights, which force them to be shut off in the future iterative procedure.

By using the approximation in (20), we have the following optimization problem that should be solved in the n-th iteration

$$\begin{aligned}\min_{\mathbf{V}}\ &\sum_{i\in\mathcal{I}}\omega_i^{(n-1)}\sum_{k\in\mathcal{U}_i}\|\mathbf{V}_{i,k}\|_F^2\\ \text{s.t. }&\text{C1, C2,}\end{aligned} \tag{22}$$

where $\omega_i^{(n-1)} = \eta_i + a_i^{(n-1)}\tilde{P}_i^c$.

Based on the above analysis, the re-weighted l_1-norm based (RLN) algorithm to solve Problem (10) is given in Algorithm 3. The convergence of the RLN algorithm is proved in [24]. In addition, [24] showed that the RLN algorithm is guaranteed to achieve sparse solutions, while the other smooth approximations cannot produce sparse solutions in general.

Algorithm 3 RLN Algorithm

```
1: Initialize a small enough $\delta$, the iterative number $n = 1$, the
   maximum number of iterations $N_{\max}$. Initialize $\mathbf{V}^{(0)}$ with
   the outputs given by Stage I, calculate $\{\omega_i^{(0)}, \forall i\}$;
2: Given $\{\omega_i^{(n-1)}, \forall i\}$, solve Problem (22) to get $\mathbf{V}^{(n)}$ by using
   the WMMSE algorithm in Subsection IV-B;
3: Update $\{\omega_i^{(n)}, \forall i\}$ with $\mathbf{V}^{(n)}$;
4: If $n \geq N_{\max}$, terminate. Otherwise, set $n \leftarrow n + 1$ and go
   to step 2.
```

B. Algorithm to Solve Problem (22)

For simplicity, the subscript of $\omega_i^{(n-1)}$ in Problem (22) is omitted. The main difficulty in solving Problem (22) lies in the rate requirement, which is non-convex. To handle this difficulty, we again apply the relationship between WMMSE and the rate expression. Based on Lemma 1, we replace the rate constraints in (22) with its lower bound $h_k(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k)$. Define the indices of $\mathcal{I}_k$ as $\mathcal{I}_k = \left\{s_1^k, \cdots, s_{|\mathcal{I}_k|}^k\right\}$, we have the following optimization problem

$$\begin{aligned}\min_{\mathbf{V},\mathbf{W},\mathbf{U}}\ &\sum_{k\in\mathcal{U}}\text{Tr}\left(\bar{\mathbf{V}}_k^H\mathbf{G}_k\bar{\mathbf{V}}_k\right)\\ \text{s.t. }&h_k(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k) \geq R_{k,\min}, \quad \forall k\in\mathcal{U},\\ &\sum_{k\in\mathcal{U}_i}\left\|\mathbf{B}_{i,k}\bar{\mathbf{V}}_k\right\|_F^2 \leq P_{i,\max}, \quad \forall i\in\mathcal{I},\end{aligned} \tag{23}$$

where $\mathbf{G}_k$ and $\mathbf{B}_{i,k}$ are both diagonal matrices, given by

$$\mathbf{G}_k = \text{blkdiag}\left\{\omega_{s_1^k}\mathbf{I}_{M\times M}, \cdots, \omega_{s_{|\mathcal{I}_k|}^k}\mathbf{I}_{M\times M}\right\}$$

and

$$\mathbf{B}_{i,k} = \text{diag}\left\{\overbrace{\mathbf{0}_{1\times M}}^{s_1^k}, \cdots, \overbrace{\mathbf{1}_{1\times M}}^{s_j^k}, \cdots, \overbrace{\mathbf{0}_{1\times M}}^{s_{|\mathcal{I}_k|}^k}\right\}, \quad \text{if } s_j^k = i, \quad \forall i, k.$$

By solving Problem (23), we can find a solution that satisfies the KKT conditions of Problem (22). To solve it, we again apply the block coordinate descent method. Matrices $\mathbf{U}$ and $\mathbf{W}$ can be updated with (15). The remaining task is to update $\mathbf{V}$

with given $\mathbf{U}$ and $\mathbf{W}$. Plugging the MSE expression in (13) into the first set of Problem (23) yields

$$\begin{aligned}
\min_{\mathbf{V}} \ & \sum\nolimits_{k\in\mathcal{U}} \mathrm{Tr}\left(\bar{\mathbf{V}}_k^H \mathbf{G}_k \bar{\mathbf{V}}_k\right) \\
\text{s.t.} \ & \sum\nolimits_{j\in\mathcal{U}} \mathrm{Tr}\left(\bar{\mathbf{V}}_j^H \bar{\mathbf{H}}_{j,k}^H \mathbf{U}_k \mathbf{W}_k \mathbf{U}_k^H \bar{\mathbf{H}}_{j,k} \bar{\mathbf{V}}_j\right) \\
& -\mathrm{Tr}\left(\mathbf{W}_k \mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k\right) - \mathrm{Tr}\left(\bar{\mathbf{V}}_k^H \bar{\mathbf{H}}_{k,k}^H \mathbf{U}_k \mathbf{W}_k\right) \le c_k, \quad \forall k \\
& \sum\nolimits_{k\in\mathcal{U}_i} \mathrm{Tr}\left(\bar{\mathbf{V}}_k^H \mathbf{B}_{i,k} \bar{\mathbf{V}}_k\right) \le P_{i,\max}, \quad \forall i,
\end{aligned} \tag{24}$$

where $c_k = \log|\mathbf{W}_k| + d - R_{k,\min} - \mathrm{Tr}(\mathbf{W}_k) - \sigma_k^2 \mathrm{Tr}\left(\mathbf{U}_k^H \mathbf{U}_k \mathbf{W}_k\right)$.

Then, an WMMSE algorithm is proposed to solve Problem (22) in Algorithm 4. In the following theorem, we show the property of the WMMSE algorithm.

Algorithm 4 WMMSE Algorithm

```
1: Initialize iterative number $l = 1$, maximum number of iterations $l_{\max}$, feasible $\mathbf{V}^{(0)}$, calculate $\mathbf{U}^{(0)}$ and $\mathbf{W}^{(0)}$ by using (15) with $\mathbf{V}^{(0)}$, tolerance $\varepsilon$, calculate the objective value of Problem (23), denoted as $\mathrm{Obj}(\mathbf{V}^{(l-1)})$.
2: With $\mathbf{U}^{(l-1)}$ and $\mathbf{W}^{(l-1)}$, update $\mathbf{V}^{(l)}$ by solving Problem (24) with the BCD algorithm in Subsection IV-C;
3: Update $\mathbf{U}^{(l)}$ and $\mathbf{W}^{(l)}$ as in (15) with $\mathbf{V}^{(l)}$;
4: If $l \ge l_{\max}$ or $\left|\mathrm{Obj}(\mathbf{V}^{(l-1)}) - \mathrm{Obj}(\mathbf{V}^{(l)})\right| / \mathrm{Obj}(\mathbf{V}^{(l)}) < \varepsilon$, terminate. Otherwise, set $l \leftarrow l + 1$ and go to step 2.
```

Theorem 2: The sequence of $\mathbf{V}$ generated by the WMMSE algorithm will converge to the KKT point of Problem (22).

Proof: Please see Appendix B. □

C. Low-Complexity Algorithm to Solve Problem (24)

Since $\omega_i > 0, \forall i$, matrices $\{\mathbf{G}_k, \forall k \in \mathcal{U}\}$ are positive definite matrices. Then, Problem (24) can be similarly transformed an SOCP problem as in (18). Using the same method in Subsection III-B, the total complexity to solve this problem by using the interior point method is

$$T_{\text{SOCP}} = O\left(\sqrt{I+K}(2lMKd)^2\left(2K^2d^2 + 2dMKl\right)\right).$$

In the following, we go one step further to design an algorithm with lower complexity. Obviously, Problem (24) is a convex problem, and it satisfies the Slater's condition [32]. Hence, the duality gap between Problem (24) and its dual problem is zero [32]. Then we can solve its dual problem instead of directly solving it.

With some simple manipulations, the Lagrangian function of Problem (24) is given by

$$\begin{aligned}
\mathcal{L}(\mathbf{V}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = & \sum_{k\in\mathcal{U}} \left(\mathrm{Tr}\left(\bar{\mathbf{V}}_k^H \bar{\mathbf{G}}_k \bar{\mathbf{V}}_k\right) - \mathrm{Tr}\left(\lambda_k \mathbf{W}_k \mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k\right)\right. \\
& \left. - \mathrm{Tr}\left(\lambda_k \bar{\mathbf{V}}_k^H \bar{\mathbf{H}}_{k,k}^H \mathbf{U}_k \mathbf{W}_k\right)\right) \\
& - \sum_{k\in\mathcal{U}} \lambda_k c_k - \sum_{i\in I} \mu_i P_{i,\max},
\end{aligned}$$

where $\boldsymbol{\lambda} = [\lambda_k, \forall k \in \mathcal{U}]^H$ and $\boldsymbol{\mu} = [\mu_i, \forall i \in I]^H$ are the Lagrangian multipliers associated with the first and second sets of constrains of Problem (24), respectively, and $\bar{\mathbf{G}}_k$ is given by

$$\bar{\mathbf{G}}_k = \mathbf{G}_k + \sum\nolimits_{j\in\mathcal{U}} \lambda_j \bar{\mathbf{H}}_{k,j}^H \mathbf{U}_j \mathbf{W}_j \mathbf{U}_j^H \bar{\mathbf{H}}_{k,j} + \sum\nolimits_{i\in I_k} \mu_i \mathbf{B}_{i,k}.$$

The dual function is given by

$$\begin{aligned}
& g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\
& = \min_{\mathbf{V}} \mathcal{L}(\mathbf{V}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\
& = \min_{\mathbf{V}} \sum_{k\in\mathcal{U}} \left(\mathrm{Tr}\left(\bar{\mathbf{V}}_k^H \bar{\mathbf{G}}_k \bar{\mathbf{V}}_k\right) - \mathrm{Tr}\left(\lambda_k \mathbf{W}_k \mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{V}}_k\right)\right. \\
& \quad \left. - \mathrm{Tr}\left(\lambda_k \bar{\mathbf{V}}_k^H \bar{\mathbf{H}}_{k,k}^H \mathbf{U}_k \mathbf{W}_k\right)\right) - \sum_{k\in\mathcal{U}} \lambda_k c_k - \sum_{i\in I} \mu_i P_{i,\max}.
\end{aligned} \tag{25}$$

Note that matrices $\{\mathbf{G}_k, \forall k \in \mathcal{U}\}$ are positive definite matrices. Problem (25) is a convex problem, and the optimal solution can be obtained from its first-order derivative condition as:

$$\bar{\mathbf{V}}_k = \lambda_k \bar{\mathbf{G}}_k^{-1} \bar{\mathbf{H}}_{k,k}^H \mathbf{U}_k \mathbf{W}_k, \quad \forall k \in \mathcal{U}. \tag{26}$$

By inserting this solution into (25), the dual function becomes

$$\begin{aligned}
g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = & -\sum_{k\in\mathcal{U}} \lambda_k^2 \mathrm{Tr}\left(\mathbf{W}_k^H \mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{G}}_k^{-1} \bar{\mathbf{H}}_{k,k}^H \mathbf{U}_k \mathbf{W}_k\right) \\
& - \sum_{k\in\mathcal{U}} \lambda_k c_k - \sum_{i\in I} \mu_i P_{i,\max}.
\end{aligned} \tag{27}$$

Hence, the dual problem of Problem (24) is given by

$$\begin{aligned}
& \max_{\{\lambda_k \ge 0, \mu_i \ge 0\}} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\
& = \min_{\{\lambda_k \ge 0, \mu_i \ge 0\}} \sum_{k\in\mathcal{U}} \lambda_k^2 \mathrm{Tr}\left(\mathbf{W}_k^H \mathbf{U}_k^H \bar{\mathbf{H}}_{k,k} \bar{\mathbf{G}}_k^{-1} \bar{\mathbf{H}}_{k,k}^H \mathbf{U}_k \mathbf{W}_k\right) \\
& \quad + \sum\nolimits_{k\in\mathcal{U}} \lambda_k c_k + \sum\nolimits_{i\in I} \mu_i P_{i,\max} \\
& \triangleq \min_{\{\lambda_k \ge 0, \mu_i \ge 0\}} f(\boldsymbol{\lambda}, \boldsymbol{\mu}),
\end{aligned} \tag{28}$$

where $f(\boldsymbol{\lambda}, \boldsymbol{\mu}) = -g(\boldsymbol{\lambda}, \boldsymbol{\mu})$.

Fortunately, the objective function of the dual problem in (28) is differentiable and the dual problem is convex [32], the descent methods such as the gradient descent method and Newton's method [32], [38] can be applied to solve it. In the following, we also utilize the block coordinate descent method to solve the dual problem (28): Optimize $\{\lambda_k, \forall k\}$ with $\{\mu_i, \forall i\}$, and vice versa.

Given $\{\mu_i, \forall i\}$, Newton's method is applied to find the optimal $\{\lambda_k, \forall k\}$ of the dual problem, which is summarized in Algorithm 5.[3]

In step 4 of Algorithm 5, the backtracking line search method is used to find the step size, where ξ is typically chosen as a very small value and φ is chosen between 0 and 1. The step $\kappa^{(t)}$ starts with one and then reduces by a factor of φ until the stop condition (29) is satisfied. Note that in each

[3] Since $\{\mu_i, \forall i\}$ are given, $f(\boldsymbol{\lambda})$ is short for $f(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and the same for $f(\boldsymbol{\mu})$ later.

Algorithm 5 Newton's Method to Update $\{\lambda_k, \forall k\}$

```
1: Initialize iterative number $t = 1$, the maximum number of iterations $t_{\max}^{\text{Newt}}$, initial $\boldsymbol{\lambda}^{(0)} = \mathbf{1}$, tolerance $\varepsilon = 10^{-10}$, $\xi \in (0, 0.5)$, $\varphi \in (0, 1)$;
2: Compute the gradient $\nabla f(\boldsymbol{\lambda}^{(t-1)})$, Hessian matrix $\nabla^2 f(\boldsymbol{\lambda}^{(t-1)})$, the Newton direction and the decrement
       $\Delta\boldsymbol{\lambda}^{(t-1)} = -\left(\nabla^2 f(\boldsymbol{\lambda}^{(t-1)})\right)^{-1} \nabla f(\boldsymbol{\lambda}^{(t-1)})$,
       $o^{(t-1)} = \nabla f(\boldsymbol{\lambda}^{(t-1)})^T \left(\nabla^2 f(\boldsymbol{\lambda}^{(t-1)})\right)^{-1} \nabla f(\boldsymbol{\lambda}^{(t-1)})$;
3: Compute $\bar{\boldsymbol{\lambda}}^{(t-1)} = [\boldsymbol{\lambda}^{(t-1)} + \Delta\boldsymbol{\lambda}^{(t-1)}]_+$;
4: Update $\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}^{(t-1)} + \kappa^{(t-1)}(\bar{\boldsymbol{\lambda}}^{(t-1)} - \boldsymbol{\lambda}^{(t-1)})$, where $\kappa^{(t-1)} = \varphi^{m^{(t-1)}}$ and $m^{(t-1)}$ is the first non-negative integer $m$ that satisfies
       $f(\boldsymbol{\lambda}^{(t)}) - f(\boldsymbol{\lambda}^{(t-1)}) \le \xi\varphi^m \nabla f(\boldsymbol{\lambda}^{(t-1)})^T \left(\bar{\boldsymbol{\lambda}}^{(t-1)} - \boldsymbol{\lambda}^{(t-1)}\right)$.   (29)
5: If $o^{(t-1)}/2 \le \varepsilon$ or $t \ge t_{\max}^{\text{Newt}}$, terminate; Otherwise, $t \leftarrow t + 1$, and go to step 2;
```

iteration of Algorithm 5, the step value $\kappa^{(t)}$ may be different. The constant ξ can be regarded as the acceptable fraction of the decrease in the objective value of f that is predicted by the line search method.

However, to make this algorithm work, there are still problems to be solved: how to calculate the gradient and how to compute the Hessian matrix. To derive the expressions of the gradient and the Hessian matrix, we first introduce some useful results in the matrix differential calculus. Given a matrix function $\boldsymbol{\Gamma}(x)$, one has [39], [40]

$$\frac{d}{dx}\text{Tr}\left(\boldsymbol{\Gamma}(x)\right) = \text{Tr}\left(\frac{d\boldsymbol{\Gamma}(x)}{dx}\right), \tag{30}$$

$$\frac{d}{dx}\boldsymbol{\Gamma}(x)^{-1} = -\boldsymbol{\Gamma}(x)^{-1}\frac{d\boldsymbol{\Gamma}(x)}{dx}\boldsymbol{\Gamma}(x)^{-1}. \tag{31}$$

In addition, to simplify the expressions of the gradient and the Hessian matrix, one defines some matrices:

$$\begin{aligned}&\tilde{\mathbf{H}}_{j,k} = \bar{\mathbf{H}}_{j,k}^H\mathbf{U}_k, \breve{\mathbf{H}}_{j,k} = \tilde{\mathbf{H}}_{j,k}\mathbf{W}_k, \hat{\mathbf{H}}_{j,k} = \breve{\mathbf{H}}_{j,k}\tilde{\mathbf{H}}_{j,k}^H,\\ &\tilde{\mathbf{G}}_k = \bar{\mathbf{G}}_k^{-1}, \mathbf{C}_k = \tilde{\mathbf{G}}_k\breve{\mathbf{H}}_{k,k}, \mathbf{F}_k = \breve{\mathbf{H}}_{k,k}^H\mathbf{C}_k, \mathbf{Y}_{j,k} = \mathbf{C}_j^H\hat{\mathbf{H}}_{j,k},\\ &\tilde{\mathbf{Y}}_{j,k} = \mathbf{Y}_{j,k}\tilde{\mathbf{G}}_j, \mathbf{Z}_{j,k} = \mathbf{Y}_{j,k}\mathbf{C}_j, \quad \forall j, k \in \mathcal{U}.\end{aligned}$$

Based on the above results and definitions, the gradient can be derived as follows:

$$\nabla f(\boldsymbol{\lambda}) = \left[\frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_k}, \quad \forall k \in \mathcal{U}\right]^H, \tag{32}$$

with

$$\frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_k} = 2\lambda_k\text{Tr}\left(\mathbf{F}_k\right) - \sum\nolimits_{j\in\mathcal{U}}\lambda_j^2\text{Tr}\left(\mathbf{Z}_{j,k}\right) + c_k, k \in \mathcal{U}.$$

The Hessian matrix of $f(\boldsymbol{\lambda})$ can be calculated as:

$$\left[\nabla^2 f(\boldsymbol{\lambda})\right]_{i,j} = \begin{cases} 2\text{Tr}(\mathbf{F}_i) + 2\sum\nolimits_{k\in\mathcal{U}}\lambda_k^2\text{Tr}\left(\tilde{\mathbf{Y}}_{k,i}\mathbf{Y}_{k,i}^H\right) & \\ \quad -4\lambda_i\text{Tr}\left(\mathbf{Z}_{i,i}\right), & \text{if } i = j,\\ 2\sum\nolimits_{k\in\mathcal{U}}\lambda_k^2\text{Re}\left\{\text{Tr}\left(\tilde{\mathbf{Y}}_{k,j}\mathbf{Y}_{k,i}^H\right)\right\} & \\ -2\lambda_i\text{Tr}\left(\mathbf{Z}_{i,j}\right) - 2\lambda_j\text{Tr}\left(\mathbf{Z}_{j,i}\right), & \text{if } j > i,\\ \left[\nabla^2 f(\boldsymbol{\lambda})\right]_{j,i}, & \text{if } j < i.\end{cases} \tag{33}$$

Next, given $\{\lambda_k, \forall k \in \mathcal{U}\}$, we solve the dual problem (28) to update $\{\mu_i, \forall i \in I\}$. Here, the gradient descent method [32] is applied. Although Newton's method converges faster than the gradient descent method, simulation results show that the gradient method also converges within five iterations but it has much lower computational complexity than Newton's method since it does not require the calculations of the Hessian matrix and the inverse of the Hessian matrix. The gradient descent method to update $\{\mu_i, \forall i \in I\}$ is given in Algorithm 6.

Algorithm 6 Gradient Descent Method to Update $\{\mu_i\}_{i=1}^I$

```
1: Initialize iterative number $t = 1$, maximum number of iterations $t_{\max}^{\text{Grad}}$, initial $\boldsymbol{\mu}^{(0)} = \mathbf{1}$, accuracy $\varepsilon$;
2: Compute the gradient $\nabla f(\boldsymbol{\mu}^{(t-1)})$;
3: Compute $\bar{\boldsymbol{\mu}}^{(t-1)} = [\boldsymbol{\mu}^{(t-1)} - \nabla f(\boldsymbol{\mu}^{(t-1)})]_+$;
4: Update $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \kappa^{(t-1)}(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(t-1)})$, where $\kappa^{(t-1)} = \beta^{l^{(t-1)}}$ and $l^{(t-1)}$ is the first non-negative integer $l$ that satisfies
       $f(\boldsymbol{\mu}^{(t)}) - f(\boldsymbol{\mu}^{(t-1)}) \le \delta\beta^l \nabla f(\boldsymbol{\mu}^{(t-1)})^T \left(\bar{\boldsymbol{\mu}}^{(t-1)} - \boldsymbol{\mu}^{(t-1)}\right)$.
5: If $t \ge t_{\max}$ or $\left|f(\boldsymbol{\mu}^{(t)}) - f(\boldsymbol{\mu}^{(t-1)})\right| / \left|f(\boldsymbol{\mu}^{(t)})\right| < \varepsilon$, stop; Otherwise, $t \leftarrow t + 1$, and go to step 2;
```

In Algorithm 6, the gradient $\nabla f(\boldsymbol{\mu})$ is required. Define $\mathbf{D}_k = \mathbf{C}_k\mathbf{C}_k^{\text{H}}$. Then, by using the results in (30) and (31), the gradient $\nabla f(\boldsymbol{\mu})$ can be calculated as

$$\nabla f(\boldsymbol{\mu}) = \left[\frac{df(\boldsymbol{\mu})}{d\mu_i}, \forall i \in I\right]^H, \tag{34}$$

with

$$\frac{df(\boldsymbol{\mu})}{d\mu_i} = -\sum\nolimits_{k\in\mathcal{U}_i}\lambda_k^2\text{Tr}\left(\mathbf{B}_{i,k}\mathbf{D}_k\right) + P_{i,\max}, \quad \forall i \in I. \tag{35}$$

Finally, based on the above analysis, the method to solve the dual problem (28) is given in Algorithm 7, which is named as Block Coordinate Descent (BCD) method.

Theorem 3: The sequences of $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ generated by Algorithm 7 will converge to the globally optimal solution of the dual problem (28).

Proof: Please see Appendix C. □

When the optimal $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are obtained by using Algorithm 7, the optimal solution to Problem (24) is given by (26). As there is zero duality gap between the primal problem (24) and dual problem (28), which means that this solution is the globally optimal solution of Problem (24).

Algorithm 7 BCD Method to Solve the Dual Problem (28)

```
1: Initialize iterative number $n = 1$, the maximum number of iterations $n_{\max}$, initial $\boldsymbol{\lambda}^{(0)} = \mathbf{1}$ and $\boldsymbol{\mu}^{(0)} = \mathbf{1}$, error tolerance $\varepsilon$;
2: Given $\boldsymbol{\mu}^{(n-1)}$, apply Newton's method in Algorithm 5 to update $\boldsymbol{\lambda}^{(n)}$;
3: Given $\boldsymbol{\lambda}^{(n)}$, employ the gradient descent method in Algorithm 6 to update $\boldsymbol{\mu}^{(n)}$;
4: If $n \ge n_{\max}$ or
       $$\frac{\left|f(\boldsymbol{\lambda}^{(n)}, \boldsymbol{\mu}^{(n)}) - f(\boldsymbol{\lambda}^{(n-1)}, \boldsymbol{\mu}^{(n-1)})\right|}{\left|f(\boldsymbol{\lambda}^{(n)}, \boldsymbol{\mu}^{(n)})\right|} < \varepsilon,$$
   terminate; Otherwise, set $n \leftarrow n + 1$, and go to step 2;
```

D. Overall Complexity to Solve Problem (24) in Stage II

In this subsection, we analyze the overall complexity to solve Problem (24). It mainly includes three layers of iterations: the first layer is the RLN algorithm to deal with the non-smooth l_0 norm, the second layer is the WMMSE algorithm to deal with the non-convex rate constraints, and the third layer is the BCD algorithm to solve Problem (24).

We first analyze the complexity of the third layer for BCD algorithm. Note that $\tilde{\mathbf{H}}_{j,k}$, $\breve{\mathbf{H}}_{j,k}$, and $\hat{\mathbf{H}}_{j,k}$ can be calculated before the iterations of the BCD Algorithm. The main complexity of the BCD Algorithm lies in step 2 and step 3, where Newton's method and gradient descent method are used to update $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, respectively.

We first analyze the computational complexity of Newton's method under the same assumption in Subsection III-B. The main complexity in each iteration of Newton's method lies in step 2 and step 4 of Algorithm 5. We first analyze step 2 of Algorithm 5. According to [41], the complexity of calculating $\{\tilde{\mathbf{G}}_k, \forall k \in \mathcal{U}\}$ is on the order of $O\left(K(Ml)^{2.376}\right)$. For any two matrices $\mathbf{X} \in \mathbb{C}^{m\times n}, \mathbf{Y} \in \mathbb{C}^{n\times p}$, the complexity of computing $\mathbf{XY}$ is on the order of $O(mnp)$ [32]. In general, $d \ll MI$. Then, the total complexity of computing $\{\mathbf{C}_k, \mathbf{F}_k, \forall k \in \mathcal{U}\}$ is on the order of $O\left(KM^2l^2d\right)$. Similarly, the total complexity of computing $\left\{\mathbf{Y}_{j,k}, \tilde{\mathbf{Y}}_{j,k}, \mathbf{Z}_{j,k}\forall j, k \in \mathcal{U}\right\}$ is on the order of $O\left(K^2M^2l^2d\right)$. Hence, the total complexity of computing $\{\mathbf{C}_k, \mathbf{F}_k, \forall k\}$ and $\left\{\mathbf{Y}_{j,k}, \tilde{\mathbf{Y}}_{j,k}, \mathbf{Z}_{j,k}, \forall j, k \in \mathcal{U}\right\}$ is on the order of $O\left(K^2M^2l^2d\right)$. With a similar analysis, the total complexity of computing (33) is on the order of $O\left(K^3Mld^2\right)$. In addition, the complexity of computing the inverse of $\nabla^2 f(\boldsymbol{\lambda}^{(t)})$ is on the order of $O\left(K^{2.376}\right)$ [41]. Hence, the total complexity of step 2 of Algorithm 5 is on the order of $O\left(\max\left\{K^3Mld^2, (KMl)^{2.376}, K^2(Ml)^2d\right\}\right)$. In the tth iteration of step 4 of Algorithm 5, $f(\boldsymbol{\lambda}^{(t+1)})$ is required to calculate $m^{(t)}$ times. The complexity in each time is on the order of $O\left(\max\left\{(Ml)^{2.376}, K(Ml)^2d\right\}\right)$. Thus, the total complexity of step 4 of Algorithm 5 is on the order of $O\left(m^{(t)}\max\left\{(Ml)^{2.376}, K(Ml)^2d\right\}\right)$. Simulation results show that in general $m^{(t)}$ is always equal to one, which means that $f(\boldsymbol{\lambda}^{(t+1)})$ only needs to be computed for once. Hence, the complexity of step 4 of Algorithm 5 can be approximately by $O\left(\max\left\{(Ml)^{2.376}, K(Ml)^2d\right\}\right)$. As a result, the total complexity of Newton's method is

$$T_{\text{Newton}} = O\left(t_{\max}^{\text{Newt}}\max\left\{K^3Mld^2, (Ml)^{2.376}, K^2(Ml)^2d\right\}\right). \tag{36}$$

Simulation results show that Newton's method converges very rapidly and in general five iterations are enough for the algorithm to converge.

By using the similarly analytical technique to Newton's method, the total complexity of the gradient descent method is given by

$$T_{\text{Grad}} = O\left(t_{\max}^{\text{Grad}}\max\left\{(MI)^{2.376}, K(MI)^2d\right\}\right). \tag{37}$$

The simulation results in the next section show that the gradient descent method usually converges within five iterations. Hence, in each iteration of the BCD Algorithm, the complexity of Newton's method dominates the complexity of the gradient descent method.

Based on the above analysis, the overall complexity to solve Problem (24) in Stage II is

$$T_{\text{StageII}} = t_{\text{RLN}}t_{\text{WMMSE}}t_{\text{BCD}}\left(T_{\text{Newton}} + T_{\text{Grad}}\right), \tag{38}$$

where t_{RLN}, t_{WMMSE} and t_{BCD} represent the average number of iterations required by the RLN, WMMSE, and BCD algorithms, respectively. Simulation results show that these three algorithms converge very fast and generally five iterations are enough to achieve large portion of the final performance.

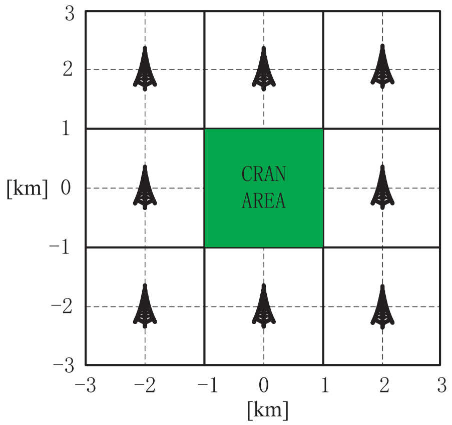


Fig. 2. Illustration of a wrap-round C-RAN system model, where C-RAN is deployed in the center of the region, which is surrounded by eight nearby cells.

V. Simulation Results

In this section, we present simulation results to evaluate the performance of the proposed algorithms. To be more realistic, we consider a wrap-around system model shown in Fig. 2 as in [42], where the C-RAN network is deployed in the central square with [−1000 1000] × [−1000 1000] meters, surrounded by eight uncoordinated square macrocells. It is assumed that all the users and RRHs are uniformly and independently distributed in the C-RAN region. We adopt the channel model that consists of four parts: 1) the long term evolution (LTE)

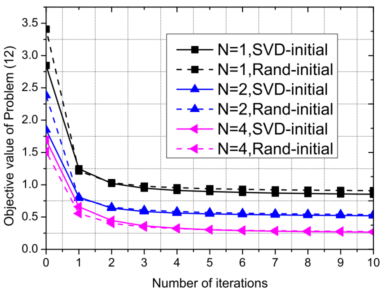


Fig. 3. Convergence behaviour of Algorithm 2.

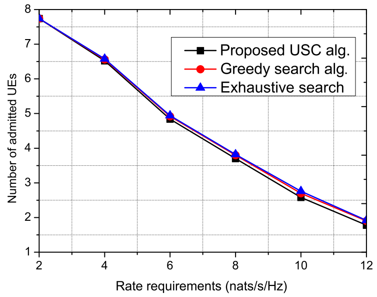


Fig. 4. Average number of admitted users versus the rate requirements.

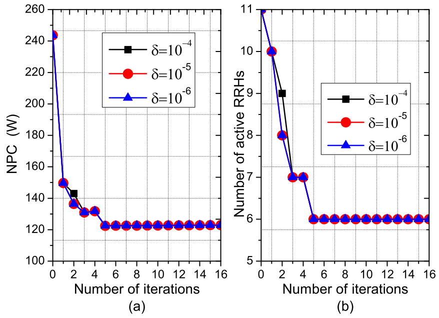


Fig. 5. (a) Total power consumption versus the number of iterations; (b) The number of active RRHs versus the number of iterations, where $R_{\min} = 2$ nats/s/Hz.

standard path loss model: $PL_{i,k} = 148.1 + 37.6\log_{10} d_{i,k}$ (dB), where $d_{i,k}$ (in km) is the distance from the ith RRH to the kth user; 2) Log-normal shadowing with zero mean and 8 dB standard derivation; 3) Rayleigh fading with zero mean and unit variance; 4) transmit antenna power gain of 9 dBi. Each user is assumed to have the same rate requirement, i.e., $R_{\min} = R_{k,\min}, \forall k$, and each RRH has the same power constraint, i.e., $P_{\max} = P_{i,\max} = 4\text{W}, \forall i \in \mathcal{I}$. It is assumed that each user is potentially served by its nearest X RRHs, i.e., $|\mathcal{I}_k| = X, \forall i$. Unless stated otherwise, the system parameters are set as follows: error tolerance is $\varepsilon = 10^{-3}$, thermal noise power is $\sigma^2 = -104$ dBm, $I = 12$, $K = 8$, $X = 3$, $M = 2$, $N = 2$, $d = \min\{M, N\}$, $\eta_i = 4$ [43], $\rho_i = 0.5$ [22], $P_i^{a,\text{rrh}} = 3.4\text{W}$, $P_i^{s,\text{rrh}} = 2.15\text{W}$, $P_i^{a,\text{fr}} = 3.85\text{W}$, $P_i^{s,\text{fr}} = 0.75\text{W}$, $P_{\text{BBU}} = 20W$ [18], [44]. Moreover, let $\mathcal{L}$ be the set of uncoordinated base stations (BSs) in C-RAN's nearby eight macrocells. The noise power at user k can be modeled as $\sigma_k^2 = \sigma^2 + \sum_{m\in\mathcal{L}} P_{\max} PL_{m,k} S_{m,k} G_m$ [42], where $PL_{m,k}$ and $S_{m,k}$ are the large-scale fading and shadowing respectively from the BS in macrocell m to user k, G_m represents the antenna gain.

A. Properties of the Proposed Algorithms

1) Convergence Behavior of Algorithm 2: Fig. 3 shows the convergence behaviour of Algorithm 2 for different numbers of receive antennas. The results are obtained by averaging over 100 channel realizations. Due to the non-convexity of Problem (11), different initial points for Algorithm 2 may yield different solutions. To investigate this effect, we consider two initialization schemes: 1) SVD-initial, in which the beam directions for each user are chosen as the unitary matrices obtained by the singular value decomposition (SVD) of channel matrices and the total power at each RRH is equally allocated to the users potentially served by each RRH; 2) Rand-initial, in which both the beam directions and power allocations are randomly generated. It can be seen from Fig. 3 that the objective value of Problem (11) monotonically decreases during the iterative procedure for two initialization schemes. In addition, the algorithm converges very fast and in general six iterations are sufficient to achieve a large proportion of the converged value for different numbers of receive antennas and different initialization schemes. It is interesting to find that the algorithm under two different initialization schemes will converge to almost the same value. As expected, the converged objective value decreases with the number of receive antennas since more degrees of freedom are available.

2) User Selection Performance of USC Algorithm: Fig. 4 compares the performance of the USC algorithm with two algorithms: greedy search method and exhaustive search method. For the greedy search method, in each time we compute the objective value of Problem (11) when excluding one user, then the user yielding the smallest objective value will be removed. This procedure continues until all remaining users are feasible. Note that this algorithm increases quadratically with K. The exhaustive search method checks all feasible sets of users and chooses the largest one. Its complexity increases exponentially with K. As expected, the number of admitted users decreases with the rate requirements for all algorithms. The greedy search method achieves almost the same performance as the exhaustive one, and the performance gap between the exhaustive search algorithm and the proposed USC algorithm can be negligible. However, the complexity of our proposed USC algorithm only increases linearly with K. The impact of initial points is also studied and we find that both initialization schemes (SVD-initial and rand-initial) have similar performance, which is not shown here for clarity.

3) Convergence Behaviour of the RLN Algorithm: The convergence behaviours of the RLN algorithm are shown in Figs. 5 (a) and (b) for the NPC and the number of the remaining RRHs in each iteration, respectively. Three different values of δ are tested, i.e., $\delta = 10^{-4}, 10^{-5}$ and 10^{-6}. One randomly generated channel is used to obtain the convergence behaviour, where the USC algorithm is first executed to find the largest feasible set of users. In this example, User 8 is removed to guarantee the feasibility of the other users as seen

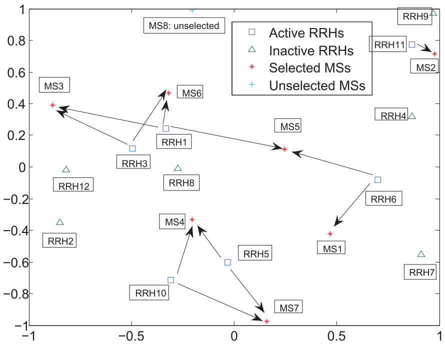


Fig. 6. Converged state of one randomly generated system configuration. The boundary user 8 is not selected as it is far from the RRHs.

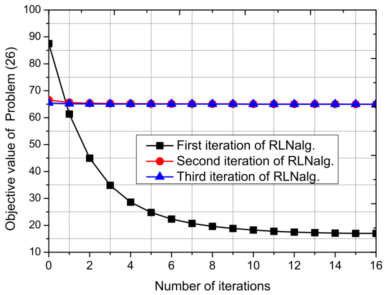


Fig. 7. Converged state of one randomly generated system configuration. The boundary user 8 is not selected as it is far from the RRHs.

in Fig. 6. It can be seen from the figures that for all values of δ, both the number of active RRHs and the NPC decrease rapidly and there is no additional decrease after the fifth iteration. At the converged state, only six RRHs are active. Compared to the full cooperation strategy where all RRHs are active, we can save large amount of power as seen from Fig. 5 (a). Fig. 6 illustrates the converged state of the system. It can be seen that RRH 2 is switched off since it is far from the users and User 8 is not selected as it is far from the RRHs. We also study the impact of initialization schemes on the performance of the RLN algorithm. The initial precoders for the RLN algorithm are the outputs of the USC algorithm which is initialized with the SVD-initial and rand-initial schemes. The simulation results show they achieve almost the same performance, which is not shown here for clarity.

4) Convergence Behaviour of the WMMSE Algorithm: In step 2 of each iteration of the RLN algorithm, we need to solve Problem (22) by using the WMMSE algorithm. Fig. 7 shows the convergence performance of the WMMSE algorithm for the first three iterations of the RLN algorithm. It is observed that the WMMSE algorithm converges within ten iterations for the first iteration of RLN algorithm. However, the objective values stay almost fixed for the second and third iterations of the RLN algorithm. This means that only in the first iteration of RLN algorithm, some iterations are required for the WMMSE algorithm.

5) Convergence Behaviour of the BCD Algorithm: In step 2 of each iteration of the WMMSE algorithm, Problem (24) should be solved to update the precoding matrices by using the BCD algorithm. Fig. 8 shows the convergence behaviour of the BCD algorithm for the first iteration of the WMMSE algorithm. It is seen that the algorithm converges very fast and one iteration is sufficient to achieve a large portion of the converged value (99.2% in this example).

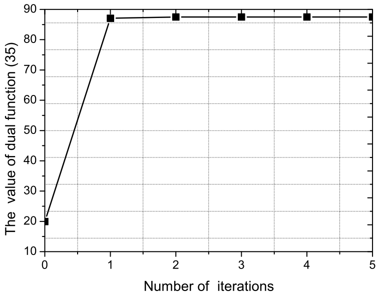


Fig. 8. Convergence behaviour of the BCD algorithm for the first iteration of the WMMSE algorithm.

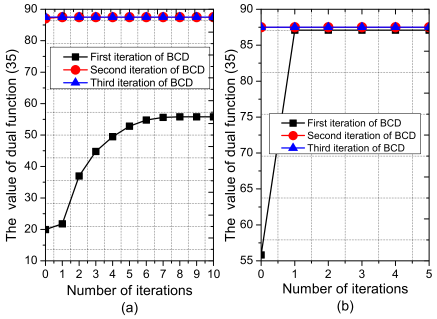


Fig. 9. (a) Convergence behaviour of Newton's method; (b) Convergence behaviour of gradient descent method.

6) Convergence Behaviour of Newton's Method and the Gradient Descent Method: In each iteration of the BCD algorithm, Newton's method is required to update $\{\lambda_k, \forall k\}$ and the gradient descent method is applied to update $\{\mu_i, \forall i\}$. The convergence behaviours of these two algorithms for the first three iterations of the BCD algorithm are shown in Figs. 9 (a) and (b), respectively. Newton's method requires several iterations to converge only in the first iteration of the BCD algorithm, while stays almost constant for the second and third iterations of the BCD algorithm. Interestingly, the gradient descent method only requires one iteration to converge in the first iteration of the BCD algorithm and keeps fixed during the rest of the iterations of the BCD algorithm. By combining the complexity analysis in (36), (37) and the above convergence behaviours, we can conclude that the BCD algorithm has a much lower computational complexity than directly solving the SOCP problem.

7) Impacts of the Number of Data Streams: In Fig. 10, the impact of the number of data streams on the number of admitted users is studied. As expected, the number of admitted users decreases with the rate requirements and larger number of data streams can support more users. We find significant performance gains can be achieved when the number of data streams increases from 1 to 2, especially for the high data rate requirements. However, only marginal performance gains are achieved by the case of $d = 4$ over the case of $d = 2$,

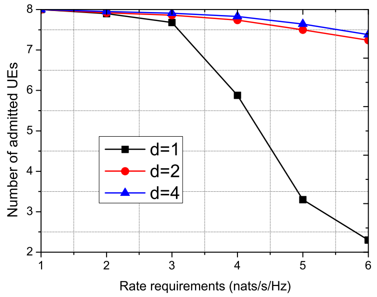


Fig. 10. Number of admitted users versus rate requirements for different numbers of data streams with $M = N = 4$.

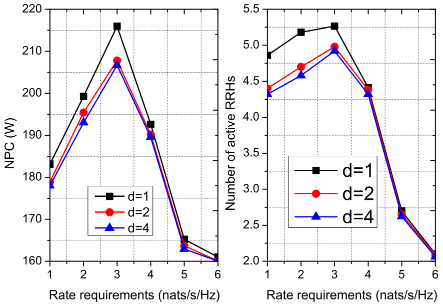


Fig. 11. (a) NPC versus the rate requirements; (b) The corresponding number of active RRHs versus the rate requirements.

which comes at the higher cost of computational complexity. This reveals that the performance saturates with the increase of data streams d. In Fig. 11, the impacts of data streams on the NPC and on the number of active RRHs are studied with the same setup in Fig. 10. For fair comparison, we only consider the set of users that can be supported under the case of $d = 1$, so that all cases can support the selected users. Fig. 11 (a) shows that the NPC first increases with the rate requirements when $R_{\min} \leq 3$ nats/s/Hz and then decreases significantly when $R_{\min} > 3$ nats/s/Hz. The reason can be explained as follows. When $R_{\min}$ increases from 1 to 3 nats/s/Hz, the number of admitted users almost keeps stable as shown in Fig. 10, while the fronthaul power increases when the rate requirement increases and the number of active RRHs increases to support the higher rate requirements as seen in Fig. 11 (b), which in turn consumes more power consumption. On the other hand, when $R_{\min}$ increase from 3 to 6 nats/s/Hz, the number of admitted users decreases dramatically as shown in Fig. 10, which leads to reduced transmit power and a reduced number of active RRHs as shown in Fig. 11 (b). Again, it is observed from Fig. 11 (a) that a greater number of data streams requires lower NPC, but the performance gain shrinks with the number of data streams.

8) Impacts of the Number of Transmit Antennas: In Fig. 12, the impact of the number of transmit antennas on the number of admitted users is studied. As expected, the number of admitted users increases with the number of transmit antennas due to more degrees of freedom. Significant performance gains can be achieved by the case of $M = 2$ over the case of $M = 1$, especially in the high rate regime. However, the performance gain shrinks for the case of $M = 4$ over the case of $M = 2$. In Fig. 13, the impacts of the number of antennas on the

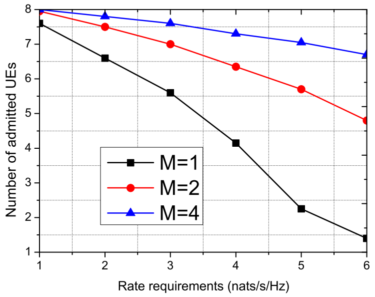


Fig. 12. Number of admitted users versus rate requirements for different numbers of transmit antennas with $N = 2$.

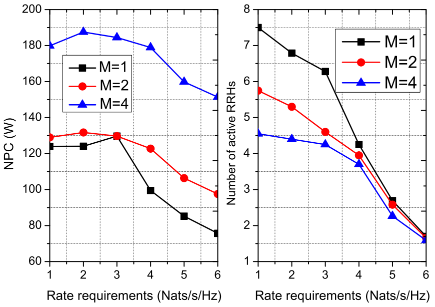


Fig. 13. (a) NPC versus the rate requirements; (b) The corresponding number of active RRHs versus the rate requirements.

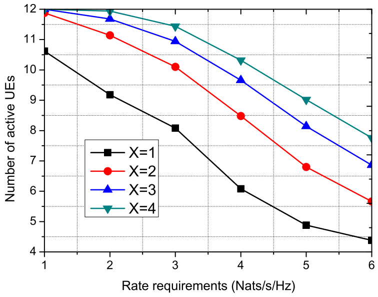


Fig. 14. The number of admitted users versus rate requirements for different candidate sizes.

NPC and the number of active RRHs are investigated. For fair comparison, it is also assumed that the set of users selected from the case of $M = 1$ are the input of Stage II for all cases of different values of M so that the selected users are the same and feasible for all cases. It is interesting to find that when M increases, the NPC increases while the number of active RRHs decreases. This is mainly due to the fact that the RRH power consumption model in (4) increases linearly with M, and this increased power consumption dominates the reduced power consumption resulting from the reduced number of active RRHs. It should be emphasized that in some other cases with different values of system parameters, the NPC may not increase with M and the counter part happens, such as the case of the low circuit power consumption for each antenna and high power consumption associated with the fronthaul power consumption.

9) Impacts of the Candidate Size: The impact of candidate size on the number of admitted users is illustrated in Fig. 14 for a dense network with 20 RRHs and 12 users. As expected,

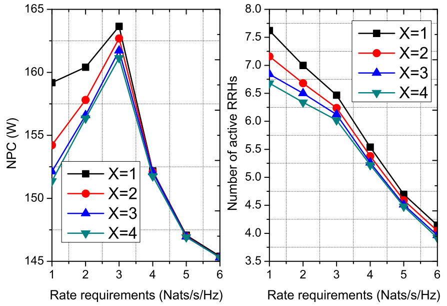


Fig. 15. (a) NPC versus the rate requirements; (b) The corresponding number of active RRHs versus the rate requirements.

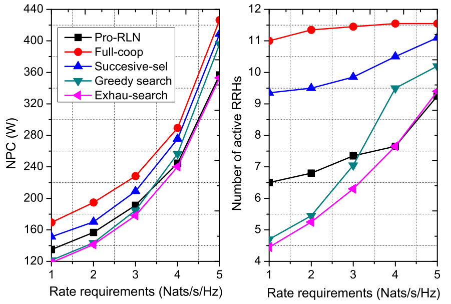


Fig. 16. (a) NPC versus the rate requirements; (b) The corresponding average number of active RRHs versus the rate requirements. The candidate size is $X = 4$.

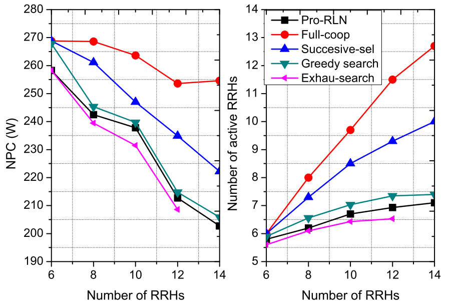


Fig. 17. (a) NPC versus the number of RRHs; (b) The corresponding number of active RRHs versus the number of RRHs with $R_{\min} = 3$ nats/s/Hz and $X = 4$.

larger candidate sizes can support more users due to the increased degrees of freedom. However, the performance gains decreases with the candidate sizes, which implies that there is no need to consider distant RRHs for each user since they contribute less to their signal strength. In general, the candidate size should be no larger than 4 to achieve a good tradeoff between performance and complexity. Similarly to the trend observed in Fig. 10 (a), it is seen from Fig. 15 (a) that the NPC increases in the low rate regime, while decreasing significantly in the high rate regime. For the former part, the reason is that the increased fronthaul power dominates the reduced circuit power for the reduced active RRHs. While for the latter part, the reason is the opposite. Also, it is observed that the NPC performance gain for larger candidate size is more obvious in the low rate regime, while the performance is almost the same in the high rate regime. This is mainly due to the fact that in the high rate regime, only a small number of users can be admitted, and these users are separated far away. As a result, the multiuser interference is not so significant and each user's nearest RRH is able to serve it with the rate requirement.

B. Performance Comparison

We compare the performance of the RLN algorithm with the following RRH selection methods:

- Exhaustive search (Exhau-search) method: For each given active RRH set $\mathcal{A}$, this method first checks its feasibility. If feasible, the method will use the WMMSE algorithm to solve the corresponding transmit power minimization problem. The complexity of this method increase exponentially with I, which is served as the performance benchmark for our proposed algorithm.
- Successive RRH selection (Succesive-sel) method: This method first lets all the RRHs be active and check its feasibility. If feasible, the method applies the WMMSE algorithm to solve the transmit power minimization problem. Then, the method gradually removes the RRHs according to their transmit power from the lowest to the highest until the problem becomes infeasible. The complexity of this scheme increases linearly with I.
- Greedy search method: In each step, we exclude each RRH and calculate the NPC when the remaining RRHs are active. Then, we remove the RRH so that the remaining RRHs yield the least NPC. This procedure terminates until the problem becomes infeasible. The complexity of this scheme increases quadratically with I.
- Full cooperative (Full-coop) method: In this method, all the selected RRHs in cluster-formation stage are active and the WMMSE algorithm is used to solve the transmit power minimization problem.

For fair comparison, we assume in the following simulation results, only the channel realizations that are feasible for all users are considered.

1) Impact of the Rate Requirements: Figs. 16 (a) and (b) illustrate the average NPC and the corresponding number of active RRHs versus the rate requirements, respectively. Fig. 16 (a) shows that the RLN algorithm outperforms the 'Succesive-sel' method and 'Full-coop' method for all rate regimes. However, the performance of the 'Greedy search' method is slightly better than the RLN algorithm when $R_{\min} \leq 3$nats/s/Hz, while the RLN algorithm outperforms the 'Greedy search' method in the high rate regime and the performance gain increases with the rate requirements. Fig. 16 (b) shows a similar trend in terms of the number of active RRHs. Compared with the optimal 'Exhau-search' method, the performance loss in power consumption is at most 8% when $R_{\min} = 1$ nats/s/Hz, and this gap gradually diminishes with the increase of rate requirements. In particular, the performance gain provided by the 'Exhau-search' method over the RLN algorithm is negligible when $R_{\min} = 5$ nats/s/Hz. As expected, the 'Full-coop' method consumes the highest power since all selected RRHs are active.

2) Impact of the Number of RRHs: Figs. 17 (a) and (b) illustrate the average NPC and the corresponding number of active

RRHs versus the total number of RRHs, respectively. It is seen that the NPC achieved by all schemes decreases with I due to the fact that when there are more RRHs, the average access distance between users and RRHs decreases significantly and thus leads to more reduced transmit power. It is again observed that the performance of the RLN algorithm is superior to that of the 'Succesive-sel' method. This implies that selecting the RRHs only based on the transmit power is not enough, and may incur significant performance loss. Note that the 'Greedy search' method requires higher power consumption than the RLN algorithm for all numbers of RRHs, especially when $I = 6$. Also, the performance of 'Exhau-search' method is slightly better than the RLN algorithm. Note that although the number of active RRHs increases slightly with the total number of RRHs as seen in Figs. 17 (b), the NPC decreases. This may due to the fact that the overall transmit power reduction overwhelms the increase of circuit power.

VI. Conclusion

In this paper, a joint selection of active RRHs and optimization of the precoding matrices which minimizes the NPC for the MIMO C-RAN, while guaranteeing users' rate requirements and per-RRH power constraints, has been studied. A low-complexity user selection was proposed to guarantee the feasibility of the other users. Then a low-complexity iterative algorithm, based on the reweighted l_1-norm minimization method, WMMSE algorithm, Newton's method, and gradient descent method, was proposed to solve the network power minimization problem for the selected users. Simulation results show that the proposed algorithms converge fast, which is attractive for practical implementation. Also, more antennas at the user side can admit more users. The proposed user selection algorithm was shown to achieve the similar performance as the optimal exhaustive search method. Moreover, our proposed algorithm was shown to achieve much greater power savings than the full cooperation method, and the performance loss compared with the optimal approach is insignificant.

Appendix A
Proof of Theorem 1

In step 2 of the nth iteration, we solve Problem (18) to obtain the optimal $\{\alpha_k^{(n)}\}_{k\in\mathcal{U}}$ and $\mathbf{V}^{(n)}$ with given $\mathbf{U}^{(n-1)}$ and $\mathbf{W}^{(n-1)}$. Hence, we have $h_k\left(\mathbf{V}^{(n)}, \mathbf{U}_k^{(n-1)}, \mathbf{W}_k^{(n-1)}\right) \geq \left(\alpha_k^{(n)}\right)^2 R_{k,\min}, \forall k$. In step 3 of the nth iteration, we update $\mathbf{U}^{(n)}$ and $\mathbf{W}^{(n)}$ as in (15) with $\mathbf{V}^{(n)}$. According to Lemma 1, we have $R_k\left(\mathbf{V}^{(n)}\right) = h_k\left(\mathbf{V}^{(n)}, \mathbf{U}_k^{(n)}, \mathbf{W}_k^{(n)}\right) \geq h_k\left(\mathbf{V}^{(n)}, \mathbf{U}_k^{(n-1)}, \mathbf{W}_k^{(n-1)}\right)$. Hence, we have

$$h_k\left(\mathbf{V}^{(n)}, \mathbf{U}_k^{(n)}, \mathbf{W}_k^{(n)}\right) \geq \left(\alpha_k^{(n)}\right)^2 R_{k,\min}. \tag{A.1}$$

In step 2 of the $(n+1)$th iteration, we obtain $\{\alpha_k^{(n+1)}\}_{k\in\mathcal{U}}$ and $\mathbf{V}^{(n+1)}$ with given $\mathbf{U}^{(n)}$ and $\mathbf{W}^{(n)}$ by solving Problem (18). Then we have $\sum_{k\in\mathcal{U}}\left(\alpha_k^{(n+1)}-1\right)^2 \leq \sum_{k\in\mathcal{U}}\left(\alpha_k^{(n)}-1\right)^2$. The reason is that from (A.1), $\{\alpha_k^{(n)}\}_{k\in\mathcal{U}}$ and $\mathbf{V}^{(n)}$ is just a feasible solution for Problem (18) with given $\mathbf{U}^{(n)}$ and $\mathbf{W}^{(n)}$. Hence, the objective value of Problem (11) is monotonically decreasing. Obviously, the objective value is lower bounded by zero. Hence, Algorithm 2 will converge.

Appendix B
Proof of Theorem 2

We first prove that the sequence of $\mathbf{V}$ generated by the WMMSE algorithm (i.e. Algorithm 4) always satisfies the rate requirements of Problem (22). In step 2, we obtain $\mathbf{V}^{(l)}$ with $\mathbf{U}^{(l-1)}$ and $\mathbf{W}^{(l-1)}$. Hence, $h_k\left(\mathbf{V}^{(l)}, \mathbf{U}_k^{(l-1)}, \mathbf{W}_k^{(l-1)}\right) \geq R_{k,\min}, \forall k$ hold since $\mathbf{V}^{(l)}$ is feasible for Problem (23). According to Lemma 1, $h_k\left(\mathbf{V}^{(l)}, \mathbf{U}_k^{(l-1)}, \mathbf{W}_k^{(l-1)}\right)$ is a lower-bound of $R_k(\mathbf{V}^{(l)})$, i.e., $R_k(\mathbf{V}^{(l)}) \geq h_k\left(\mathbf{V}^{(l)}, \mathbf{U}_k^{(l-1)}, \mathbf{W}_k^{(l-1)}\right)$. Hence, $R_k(\mathbf{V}^{(l)}) \geq R_{k,\min}$ holds. Thus, the sequence of $\mathbf{V}$ generated by the WMMSE algorithm satisfies the rate requirements of Problem (22).

Next, we show that the value of the objective function of Problem (22) monotonically decreases during the iterative process of the WMMSE algorithm. Denote $\text{Obj}(\mathbf{V}^{(l)})$ as the objective value of Problem (22) when $\mathbf{V} = \mathbf{V}^{(l)}$. Step 2 of the WMMSE algorithm updates $\mathbf{V}^{(l)}$ by solving Problem (24) with $\mathbf{U}^{(l-1)}$ and $\mathbf{W}^{(l-1)}$. The objective value of this step, $\text{Obj}(\mathbf{V}^{(l)})$, will be no larger than $\text{Obj}(\mathbf{V}^{(l-1)})$, i.e., $\text{Obj}(\mathbf{V}^{(l)}) \leq \text{Obj}(\mathbf{V}^{(l-1)})$. The reason is that $\mathbf{V}^{(l-1)}$ is a feasible solution for Problem (24) with $\mathbf{U}_k^{(l-1)}$ and $\mathbf{W}_k^{(l-1)}$ since $h_k\left(\mathbf{V}^{(l-1)}, \mathbf{U}_k^{(l-1)}, \mathbf{W}_k^{(l-1)}\right) = R_k(\mathbf{V}^{(l-1)}) \geq R_{k,\min}$ holds as proved above. In step 3 of the WMMSE algorithm, we update $\mathbf{U}^{(l)}$ and $\mathbf{W}^{(l)}$ by using (15) with $\mathbf{V}^{(l)}$. This step increases the value of $h_k\left(\mathbf{V}, \mathbf{U}_k, \mathbf{W}_k\right)$ while maintaining the same objective value of Problem (22). Therefore, this step provides "room" for the next iteration to decrease the objective value. In addition, the objective value is lower bounded by zero. Hence, the WMMSE algorithm converges.

Then, we prove that given the initial set of precoders, the WMMSE algorithm converges to a unique solution. Obviously, when $\mathbf{V}$ is given, $\mathbf{U}$ and $\mathbf{W}$ can be uniquely determined by (15). The remaining task is to prove that given $\mathbf{U}$ and $\mathbf{W}$, the BCD algorithm can obtain the unique globally optimal solution $\mathbf{V}$. Since $\{\mathbf{G}_k, \forall k\}$ are positive definite matrices, the objective function in Problem (24) is a strictly convex function with respect to (w.r.t.) $\mathbf{V}$. Obviously, the constraints in Problem (24) are convex w.r.t. $\mathbf{V}$ [32]. Hence, Problem (24) is a strictly convex problem [32]. According to [Page 137 in [32]], the globally optimal solution of Problem (24) is unique. On the other hand, Theorem 3 proves that the BCD algorithm can obtain the globally optimal solution to the dual problem (28). As Problem (24) is a convex problem and it satisfies the Slater's condition [32], the duality gap between Problem (24) and its dual problem (28) is zero [32]. As a result, the BCD algorithm can obtain the unique globally optimal solution $\mathbf{V}$. Finally, by alternatively updating step 2 and step 3, the WMMSE algorithm will converge to a unique solution. It should be emphasized that as Problem (23) is non-convex, it may have many locally optimal solutions, and the unique solution of the WMMSE algorithm depends on the initial point.

However, given the initial points of precoders, the WMMSE algorithm will converge to a unique solution.

Finally, we prove that the unique solution satisfies the KKT conditions of Problem (22). Denote the converged solution of the WMMSE algorithm as $\mathbf{V}^\star$, $\mathbf{U}^\star$ and $\mathbf{W}^\star$. With given $\mathbf{U}^\star$ and $\mathbf{W}^\star$, the Lagrange function of Problem (23) can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{V},\boldsymbol{\lambda},\boldsymbol{\mu}) &= \sum\nolimits_{k\in\mathcal{U}} \bar{\mathbf{V}}_k^H \mathbf{G}_k \bar{\mathbf{V}}_k + \sum_{k\in\mathcal{U}} \lambda_k \left(R_{k,\min} - h_k\left(\mathbf{V},\mathbf{U}_k^\star,\mathbf{W}_k^\star\right)\right) \\ &\quad + \sum\nolimits_{i\in\mathcal{I}} \mu_i \left(\sum\nolimits_{k\in\mathcal{U}_i} \left\|\mathbf{B}_{i,k}\bar{\mathbf{V}}_k\right\|_F^2 - P_{i,\max}\right), \end{aligned} \tag{B.1}$$

where $\boldsymbol{\lambda} = \{\lambda_k, \forall k \in \mathcal{U}\}$ and $\boldsymbol{\mu} = \{\mu_i, \forall i \in \mathcal{I}\}$ are the corresponding Lagrange multipliers.

According to Theorem 3, the BCD algorithm can obtain the globally optimal solution of Problem (24) (also Problem (23)) with given $\mathbf{U}^\star$ and $\mathbf{W}^\star$, there must exist $\boldsymbol{\lambda}^\star$ and $\boldsymbol{\mu}^\star$ such that $\{\mathbf{V}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star\}$ satisfy the following KKT conditions

$$\begin{aligned}\nabla_{\bar{\mathbf{V}}_k}\mathcal{L} &= \nabla_{\bar{\mathbf{V}}_k} \sum_{k\in\mathcal{U}} \bar{\mathbf{V}}_k^{\star,H} \mathbf{G}_k \bar{\mathbf{V}}_k^\star - \sum_{k\in\mathcal{U}} \lambda_k^\star \nabla_{\bar{\mathbf{V}}_k} h_k\left(\mathbf{V}^\star,\mathbf{U}_k^\star,\mathbf{W}_k^\star\right)) \\ &\quad + \sum_{i\in\mathcal{I}} \mu_i^\star \nabla_{\bar{\mathbf{V}}_k}\left(\sum_{k\in\mathcal{U}_i}\left\|\mathbf{B}_{i,k}\bar{\mathbf{V}}_k^\star\right\|_F^2\right) = \mathbf{0}, \forall k \in \mathcal{U}, \end{aligned} \tag{B.2}$$

$$\lambda_k^\star\left(h_k\left(\mathbf{V}^\star,\mathbf{U}_k^\star,\mathbf{W}_k^\star\right) - R_{k,\min}\right) = 0, \quad \forall k \in \mathcal{U}, \tag{B.3}$$

$$\mu_i^\star\left(P_{i,\max} - \sum\nolimits_{k\in\mathcal{U}_i}\left\|\mathbf{B}_{i,k}\bar{\mathbf{V}}_k^\star\right\|_F^2\right) = 0, \quad \forall i \in \mathcal{I}, \tag{B.4}$$

$$h_k\left(\mathbf{V}^\star,\mathbf{U}_k^\star,\mathbf{W}_k^\star\right) \geq R_{k,\min}, \forall k \in \mathcal{U}, \tag{B.5}$$

$$\sum\nolimits_{k\in\mathcal{U}_i}\left\|\mathbf{B}_{i,k}\bar{\mathbf{V}}_k^\star\right\|_F^2 \leq P_{i,\max}, \quad \forall i \in \mathcal{I}. \tag{B.6}$$

Since $\mathbf{U}^\star$ and $\mathbf{W}^\star$ are updated by using (15), we have $h_k\left(\mathbf{V}^\star,\mathbf{U}_k^\star,\mathbf{W}_k^\star\right) = R_k(\mathbf{V}^\star)$ according to Lemma 1. By substituting it into the equations (B.2), (B.3) and (B.5), we find that the set of equations (B.2)-(B.6) are just the KKT conditions of Problem (22).

Appendix C
Proof of Theorem 3

According to [32], the dual problem of any optimization problem is a convex problem. Thus, the dual problem (28) is jointly convex with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. Assuming that the constraint of this problem satisfies the Slater's condition, the KKT condition of this problem is sufficient and necessary for optimality. For given $\boldsymbol{\mu}$, the dual problem (28) is a convex problem w.r.t. $\boldsymbol{\lambda}$. According to [32], Newton's method can obtain the globally optimal solution of dual problem (28) for given $\boldsymbol{\mu}$. In addition, for given $\boldsymbol{\lambda}$, the dual problem (28) is convex w.r.t. $\boldsymbol{\mu}$, and the gradient descent method can be applied to obtain the globally optimal solution. Then by adopting the same idea as in the proof of Theorem 1 in [45], we can prove that the converged solution also satisfies the KKT condition of Problem (28). Since Problem (28) is a convex optimization problem, Algorithm 6 can attain the globally optimal solution of Problem (28).

References

[1] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint precoding and RRH selection for green MIMO C-RAN," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–5.

[2] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems—Part I: Chunk allocation," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2734–2744, Sep. 2009.

[3] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems—Part II: Joint chunk, power and bit allocation," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 499–509, Feb. 2012.

[4] H. Zhu, "Radio resource allocation for OFDMA systems in high speed environments," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 748–759, May 2012.

[5] "C-RAN: The road towards green RAN," China Mobile, Beijing, China, White Paper 2.5, 2011, vol. 2.

[6] "Suggestions on potential solutions to C-RAN," NGMN ALLIANCE, Berlin, Germany, White Paper, Jan. 2013.

[7] H. Zhu, "Performance comparison between distributed antenna and microcellular systems," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1151–1163, Jun. 2011.

[8] J. Wang, H. Zhu, and N. J. Gomes, "Distributed antenna systems for mobile communications in high speed trains," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 675–683, May 2012.

[9] G. Fettweis and E. Zimmermann, "ICT energy consumption-trends and challenges," in *Proc. 11th Int. Symp. Wireless Pers. Multimedia Commun.*, Lapland, Finland, 2008, vol. 2, no. 4, p. 6.

[10] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[11] L. M. Correia *et al.*, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 66–72, Nov. 2010.

[12] Y. Cheng, M. Pesavento, and A. Philipp, "Joint network optimization and downlink beamforming for CoMP transmissions using mixed integer conic programming," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3972–3987, Aug. 2013.

[13] R. Ramamonjison, A. Haghnegahdar, and V. Bhargava, "Joint optimization of clustering and cooperative beamforming in green cognitive wireless networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 982–997, Feb. 2014.

[14] D. W. K. Ng and R. Schober, "Secure and green SWIPT in distributed antenna networks with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5082–5097, Sep. 2014.

[15] F. Zhuang and V. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684–693, Feb. 2014.

[16] V. Ha, L. Le, and N. Dao, "Coordinated multipoint (CoMP) transmission design for Cloud-RANs with limited fronthaul capacity constraints," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7432–7447, Sep. 2016.

[17] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[18] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[19] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.

[20] J. Yan, J. Li, L. Zhao, and R. Chen, "Robust joint transmit beamforming with QoS guarantees in time-asynchronous DAS," *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 1506–1518, Apr. 2015.

[21] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5068–5081, Sep. 2015.

[22] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

[23] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 877–905, 2008.

[24] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited Backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.

[25] T. Huang, Y. Yu, and L. Yi, "Design of highly isolated compact antenna array for MIMO applications," *Int. J. Antennas Propag.*, vol. 2014, Nov. 2014, Art. no. 473063.

[26] K. Wang, R. Mauermayer, L. Li, and T. Eibert, "A highly compact broadband near-edge antenna for low profile communication devices," in *Proc. 9th Eur. Conf. Antennas Propag. (EuCAP)*, Apr. 2015, pp. 1–5.

[27] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[28] S. He, Y. Huang, H. Wang, S. Jin, and L. Yang, "Leakage-aware energy-efficient beamforming for heterogeneous multicell multiuser systems," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1268–1281, Jun. 2014.

[29] Y. Li, Y. Tian, and C. Yang, "Energy-efficient coordinated beamforming under minimal data rate constraint of each user," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2387–2397, Jun. 2015.

[30] B. Dai and W. Yu, "Backhaul-aware multicell beamforming for downlink cloud radio access network," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2689–2694.

[31] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[33] W. Nie, F.-C. Zheng, X. Wang, W. Zhang, and S. Jin, "User-centric cross-tier base station clustering and cooperation in heterogeneous networks: Rate improvement and energy saving," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1192–1206, May 2016.

[34] W. Feng, N. Ge, and J. Lu, "Hierarchical transmission optimization for massively dense distributed antenna systems," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 673–676, Apr. 2015.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[36] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, vol. 284, nos. 1–3, pp. 193–228, Nov. 1998.

[37] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Mach. Learn.*, vol. 85, no. 1, pp. 3–39, 2011.

[38] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[39] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*. Manhattan, CA, USA: Wiley, 1995.

[40] S. Ye and R. S. Blum, "Optimized signaling for MIMO interference systems with feedback," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2839–2848, Nov. 2003.

[41] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," in *Proc. 19th Annu. ACM Symp. Theory Comput.*, 1987, pp. 1–6.

[42] L. Venturino, N. Prasad, and X. Wang, "Coordinated linear beamforming in downlink multi-cell wireless networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1451–1461, Apr. 2010.

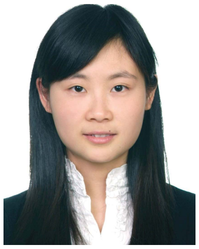
[43] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[44] A. R. Dhaini, P.-H. Ho, G. Shen, and B. Shihada, "Energy efficiency in TDMA-based next-generation passive optical access networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 3, pp. 850–863, Jun. 2014.

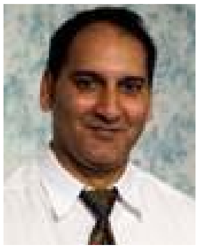
[45] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

Cunhua Pan (M'16) received the B.S. and Ph.D. (Hons.) degrees from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2010 and 2015, respectively. From 2015 to 2016, he was a Research Associate with the University of Kent, U.K. He currently holds a post-doctoral position with the Queen Mary University of London, U.K.

His research interests include C-RAN, mm-Wave communications, NOMA, D2D, large-scale MIMO and cloud computing. He is a TPC Member of the IEEE ICC and Globecom from 2015 to 2017.



Huiling Zhu (M'04) received the B.S degree from Xidian Univeristy, Xi'an, China, and the Ph.D. degree from Tsinghua University, Beijing, China. She is currently a Reader (Associate Professor) with the School of Engineering and Digital Arts, University of Kent, Canterbury, U.K. Her research interests are in the area of broadband wireless mobile communications, covering topics such as radio resource management, distributed antenna systems, MIMO, cooperative communications, device-to-device communications, and small cells and heterogeneous networks. She received the Best Paper Award from the IEEE Globecom in 2011. She has participated in a number of European and industrial projects in these topics and held the European Commission Marie Curie Fellowship from 2014 to 2016. She has served as the Publication Chair of the IEEE WCNC2013, Shanghai, the Operation Chair of the IEEE ICC2015, London, a Symposium Co-Chair of the IEEE Globecom 2015, San Diego, and a Track Co-Chair of the IEEE VTC2016-Spring, Nanjing. She currently serves as an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.



Nathan J. Gomes (M'92–SM'06) received the B.Sc. degree in electronic engineering from the University of Sussex, Sussex, U.K., in 1984, and the Ph.D. degree in electronic engineering from the University College London, London, U.K., in 1988. From 1988 to 1989, he held a Royal Society European Exchange Fellowship with ENST, Paris, France. Since 1989, he has been with the University of Kent, Canterbury, U.K., where he is currently a Professor of optical fiber communications. His current research interests include fiber-wireless access, and the fronthaul for future mobile networks. He was the TPC Chair of the IEEE International Conference on Communications, London, in 2015.

Jiangzhou Wang (F'17) is currently the Head of the School of Engineering and Digital Arts and a Professor of telecommunications, University of Kent, U.K. He has authored over 200 papers in international journals and conferences in the areas of wireless mobile communications and three books.

Prof. Wang is an IET Fellow. He received the Best Paper Award from the 2012 IEEE GLOBECOM and was an IEEE Distinguished Lecturer from 2013 to 2014. He was the Technical Program Chair of the 2013 IEEE WCNC, Shanghai, and the Executive Chair of the 2015 IEEE ICC, London. He serves/served as an Editor for a number of international journals. He was an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS from 1998 to 2013, and was a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and *IEEE Communications Magazine*. He is currently an Editor of Science China Information Science.