Weighted Sum-Rate Maximization for the Ultra-dense User-centric TDD C-RAN Downlink Relying on Imperfect CSI

Cunhua Pan, Hong Ren, Maged Elkashlan, and Arumugam Nallanathan, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

Abstract

The weighted sum-rate (WSR) maximization problem of ultra-dense cloud radio access networks (C-RANs) is considered. The user-centric clustering is adopted for reducing the complexity. To reduce the training overhead, one only needs to estimate the intra-cluster CSI, while only the large-scale channel gains are available outside the cluster. We first derive the rate lower bound (LB) relying on Jensen's inequality. For the special case of non-overlapping clusters, the accurate data rate expression is derived in closed-form. Simulation results show the tightness of the LB for both the overlapped and non-overlapped cases. Then, we consider an alternative problem where the actual data rate is replaced by its LB, which constitutes a non-convex optimization problem. First, the globally optimal solution is obtained by applying the high-complexity outer polyblock approximation (OPA) algorithm. Then we invoke the reduced-complexity modified weighted minimum mean square error (WMMSE) algorithm for mitigating the deleterious effects of realistic imperfect CSI. For the subproblem solved by each WMMSE iteration, the beamforming (BF) vectors are derived in closed form relying on the Lagrangian dual decomposition method. Finally, our simulation results show that the modified WMMSE algorithm's performance is comparable to that of the high-complexity OPA algorithm, whilst outperforms other benchmark algorithms.

Index Terms

C-RAN, imperfect CSI, Ultra-dense networks (UDN), virtual cell, weighted sum-rate maximization.

C. Pan, M. Elkashlan, and A. Nallanathan are with the Queen Mary University of London, London E1 4NS, U.K. (Email:{c.pan, maged.elkashlan, a.nallanathan}@qmul.ac.uk). H. Ren is with Southeast University, 210096, China. (e-mail: renhong@seu.edu.cn). L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K. (e-mail:lh@ecs.soton.ac.uk). L. Hanzo would like to thank the ERC for supporting his Advanced Fellow Grant. C. Pan, M. Elkashlan and A. Nallanathan would like to thank the U.K. Engineering and the Physical Sciences Research Council under Grant EP/N029666/1 and Grant EP/N029720/1.

I. INTRODUCTION

Future wireless networks have to cope with an ever-increasing demand for conveying data traffic. To achieve this ambitious goal, ultra dense networks (UDNs) have been recognized as one of the key enabling techniques [1]. In UDNs, the number of access points (APs) deployed in a given area is comparable to or even higher than the number of user equipment (UEs). Then, the signal received at the UEs can be enhanced due to the reduced distance to their associated APs. However, each UE also suffers from the interference imposed by the neighbouring APs, which constitutes a limiting factor.

Cloud radio access networks (C-RAN) have been proposed as a promising network architecture capable of dealing with this issue [2]–[4]. A typical C-RAN consists of three key components: 1) Remote radio heads (RRHs) geographically displaced accross the coverage area; 2) Baseband unit (BBU) pool hosted at the data center that is supported both by cloud computing and virtualization techniques; 3) The high-speed low-latency fronthaul links that connect the RRHs and BBU pool. The key feature of C-RANs is that the signal processing tasks of the conventional base stations have been migrated to the BBU pool, and the RRHs are only responsible for low-complexity data transmission/reception. Given this centralized architecture, advanced signal processing techniques can be realized, such as joint UE scheduling, coordinated multi-point (CoMP) transmission, centralized compression and decompression, etc which result in significant performance gains. Furthermore, due to their low-complexity functionalities, RRHs can be densely deployed over the network at a low operational cost. Hence, C-RAN is an ideal platform for reaping the benefits of UDNs, where the interference can be substantially mitigated or even eliminated by the CoMP technique, which leads to a powerful network architecture ultra-dense C-RAN [5].

Recently, sophisticated designs have been conceived for tackling the various challenges arising in C-RANs [4], [6]–[12]. Specifically, the weighted sum rate (WSR) maximization problem of C-RANs operating under realistic fronthaul capacity constraints was studied in [4], where a so-called reweighted l_1-norm based technique was adopted for transforming the non-smooth fronthaul capacity constraints into a more tractable expression. A joint RRH selection and beamforming (BF) design was investigated in [6] for a dense C-RAN, where three algorithms striking different complexity tradeoffs were proposed. In [7], the authors aimed for jointly optimizing the set of RRHs serving each UE and the BF weights for minimizing the total transmission power, while satisfying both the fronthaul capacity constraints and the UEs' rate requirements. A pair of low-

complexity algorithms were developed for solving this problem. A resource allocation problem was considered in [8] for a macrocell assisting C-RAN, where the authors aimed for minimizing the transmission power for the C-RAN under the specific interference limit imposed on the macrocell UEs. A joint RRH selection and BF design was conceived in [9] for simultaneously optimizing both the sum rate and total power, where a globally optimal solution was obtained by using the branch and bound based algorithm. Tran *et al.* [10] studied the WSR maximization problem under specific computing resource constraints, where the optimization problem was solved by a sequential convex approximation algorithm. Tang *et al.* [11] jointly optimized the activation of virtual machines (VMs) in the BBU pool and the BF weights for minimizing the system cost, where the optimal BF solution was derived in closed form. In [12], we studied the problem of optimizing the precoding matrices and the set of active RRHs for minimizing the network's power consumption, where the user-centric cluster philosophy was adopted for reducing the computational complexity, where each user is served by its closest RRHs.

However, the above-mentioned contributions were based on the assumption that the BBU pool can possess perfect channel state information (CSI), which is not practical in ultra-dense C-RANs, because an excessive number of CSIs is required for centralized signal processing. Caire *et al.* [13] showed that the overall system performance may even be reduced upon taking into account the heavy training overhead used for estimating all the network's CSI. A promising technique of reducing the training overhead is to rely on the incomplete CSI case, where each UE only needs to estimate the CSI of the links from the RRHs in its serving cluster, while assuming the CSI from the RRHs outside its cluster to be zero [14]–[17]. Alternatively, only the large-scale channel gains may be made available [18], [19]. Lakshmana *et al.* [18] considered the WSR maximization problem for the incomplete CSI case, where the large-scale channel gains are incorporated into the optimization problem for the out-cluster CSI. The authors derived the data rate LB by applying the Cauchy-Schwarz inequality and then adapted the algorithms originally developed for the perfect-CSI case to this incomplete-CSI scenario. Their simulation results showed that significant performance gains can be obtained compared to the ones, where the unknown CSI is naively regarded as zero. Recently, in [19], we studied the network power consumption minimization problem of the ultra-dense C-RANs relying on incomplete CSI, where the large-scale channel gains from the RRHs outside the cluster were included in the optimization. A two-phase optimization method was proposed, where the first phase deals with the feasibility issue by proposing a novel UE selection algorithm and

the second phase optimizes the BF vectors to minimize the network's power consumption with the UEs obtained from the first phase.

Although the authors of [14]–[19] substantially reduced the training overhead, in these contributions perfect intra-cluster CSI was assumed, which is difficult to satisfy due to the following reasons. To estimate the intra-cluster CSI in time-division duplex (TDD) C-RANs, uplink training pilot sequences have to be sent from the UEs to the RRHs for channel estimation. A naive pilot allocation method is that all UEs are assigned mutually orthogonal pilot sequences. However, the number of pilots required increases linearly with the number of UEs, which is unaffordable for ultra-dense C-RANs since they are usually deployed in hot spots with a large number of users, such as conference hall, shopping mall, etc. A judicious remedy is to allow the UEs to reuse the same pilot sequence. This will however impose pilot contamination, hence increasing the channel estimation errors. Therefore, it is imperative to design transmission schemes that are robust to channel estimation error. Robust transmission designs have hence received extensive interests [20]–[24]. There is a specific common assumption in these contributions: the channel errors are assumed to lie in a bounded uncertainty region, and the robust transmission should be designed under the condition that for each channel error in this region, the quality of service (QoS) requirements for each UE should be satisfied. This kind of optimization problem is then transformed to a semi-definite programming (SDP) one with the aid of the S-lemma and the semi-definite relaxation (SDR) technique. However, this assumption was too pessimistic. The transmission regime should be designed to be inherently robust to the practical channel estimation error sources, such as the pilot contamination and estimation noise. In our most recent work [25], we studied the joint pilot allocation and UE selection problem in order to *minimize the total transmission power*, while satisfying each UE's rate constraints and the fronthaul constraints. A novel pilot allocation based on graph theory and semi-definite relaxation was proposed for solving this problem. Another alternative optimization problem is the WSR maximization problem, where the weights can be used for controlling the fairness of UEs. However, in contrast to the power minimization problem, which can be transformed to a convex second-order cone programming (SOCP) or SDP problem, the WSR maximization problem is usually non-convex, which is difficult to solve. In this paper, we study the WSR maximization problem for the same scenario as in [25], where the joint effects of pilot contamination and incomplete inter-cluster CSI are taken into account. Another alternative network architecture similar to the ultra-dense C-RAN is the recent cell-free user-centric massive MIMO

system [26], [27], where both channel estimates and the beamforming vectors can be computed locally that incurs less fronthaul overhead. However, the proposed transmission schemes used in [26], [27] are heuristic: the power are allocated to be proportional to the estimated channel strengths for single antenna case in [26] and channel inversion beamforming was adopted for the multiple-antenna case in [27], which yields inferior performance compared with the algorithm designed in this paper. In addition, the access points should be equipped with some advanced computing functionalities to perform the channel estimation operation and beamforming computation, which is contrast to the low-complexity RRHs considered in ultra-dense C-RANs that are only responsible for simple transmission/reception.

Against the above background, the contributions of this paper are summarized as follows:

1) Due to the multiple uncertain terms in the rate expression such as the channel estimation error and the small-scale CSI from the RRHs outside the cluster, it is difficult to derive the accurate data rate expression. To circumvent this difficulty, we derive the data rate LB by applying Jensen's inequality, which is more tractable to handle. For the special case of non-overlapping clusters, we derive the accurate data rate in closed form. For both the overlapped and non-overlapped cases, we provide simulation results to show that the data rate LB is very tight, especially for low transmit powers, which is the case in ultra-dense C-RANs.
2) Since the WSR maximization problem is a non-convex optimization problem, we provide a high-complexity algorithm relying on the outer polyblock approximation (OPA) method in order to obtain the globally optimal solution to serve as our benchmark for other low-complexity suboptimal algorithms. We provide a novel method to find the intersection point on the Pareto boundary of the rate region in each iteration of the OPA algorithm. However, its complexity is excessive since it involves twin-layer iterations.
3) To further reduce the complexity, we conceive a low-complexity algorithm by carefully adapting the WMMSE algorithm originally derived for the perfect CSI case to the imperfect intra-cluster CSI and incomplete out-cluster CSI scenario. Specifically, we decompose each interfering source into multiple interfering sources and then construct an auxiliary signal transmission model for each UE. Then, the conventional WMMSE is applied to this auxiliary model. For each iteration of the modified WMMSE algorithm, there is a sub-problem in which the BF vectors should be optimized. We derive the optimal structure of the BF solutions with the aid of the Lagrangian dual decomposition method, and then the subgradient descent

method is adopted for updating the dual variables. Our simulation results show that the Lagrangian dual decomposition method is capable of achieving the same solution as provided by solving the SOCP problem using the CVX package, despite its much lower complexity.

The remainder of this paper is organized as follows. In Section II, we introduce the signal transmission model of user-centric ultra-dense C-RANs along with the intra-cluster CSI channel estimation procedure. In Section III, we provide the WSR maximization problem formulation and discuss the tightness of the rate LB. In Section IV, two different algorithms striking different performance vs complexity tradeoffs are developed. Extensive simulation results are given in Section V. Finally, our conclusions are drawn in Section VI.

Notations: For a complex value a, $\mathrm{Re}\{a\}$ represents the real part of a. Boldface lower case and upper case letters denote vectors and matrices, respectively. $\mathbb{C}^{M\times 1}$ denotes the set of $M\times 1$ complex vectors. $\mathbb{E}\{\cdot\}$ denotes the expectation operation. For the two sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A}\subseteq\mathcal{B}$ represents set $\mathcal{A}$ belongs to $\mathcal{B}$, and $\mathcal{A}\backslash\mathcal{B}$ denotes the set difference between $\mathcal{A}$ and $\mathcal{B}$. $\mathcal{CN}(\mathbf{0},\mathbf{I})$ represents a random vector following the distribution of zero mean and unit variance matrix. $\|\cdot\|$ is the norm operator. $\mathrm{blkdiag}(\cdot)$ denotes the block diagonalization operation. $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denote the transpose and Hermitian operators, respectively.

II. System Model

A. Signal Transmission Model

Consider the downlink of an ultra-dense TDD C-RAN network with I RRHs and K UEs, as shown in Fig. 1. It is assumed that each RRH is equipped with M transmit antennas (TA) and each UE has a single receive antenna (RA). Let us denote the sets of RRHs and UEs by $\mathcal{I}=\{1,\cdots,I\}$ and $\mathcal{U}=\{1,\cdots,K\}$, respectively. Each RRH is connected to the BBU pool through the high-speed fronthaul links that are shown by dark lines in Fig. 1. The BBU pool is responsible for all the baseband signal processing tasks, such as channel estimation and BF weight calculation. All UEs' data are available at the BBU pool and the BBU pool distributes each UE's data to a subset of RRHs through the fronthaul links.

To reduce the computational complexity, user-centric clusters are formed, where each UE is only served by its nearby RRHs due to the severe path loss from distant RRHs. Let us define by $\mathcal{I}_k\subseteq\mathcal{I}$ and $\mathcal{U}_i\subseteq\mathcal{U}$ the specific sets of RRHs that serve UE k and the UEs that are served by RRH i, respectively. Note that the clusters for the UEs may overlap with each other, i.e. each RRH may simultaneously serve multiple UEs. The cluster serving each UE is determined based

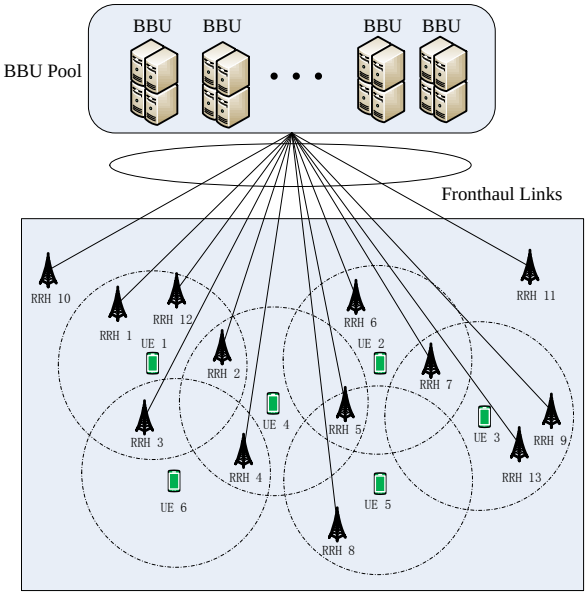

Fig. 1. Illustration of a C-RAN with thirteen RRHs and six UEs, i.e., $I = 13$, $K = 6$. To reduce the complexity, each UE is served by the RRHs within the dashed circle centered at the UE.

on the long term channel state information (CSI) such as large-scale fading [28] that changes very slowly. Hence, the cluster formation are assumed to be fixed in this paper.

Let us define by $\mathbf{h}_{i,k} \in \mathbb{C}^{M\times 1}$ and $\mathbf{w}_{i,k} \in \mathbb{C}^{M\times 1}$ the channel vector and the BF vector of the links from RRH i to UE k, respectively. Assume that the RRHs in each UE's cluster coherently transmit the same signal to the UE. Then the baseband received signal at UE k can be written as

$$y_k = \underbrace{\sum\nolimits_{i\in\mathcal{I}_k} \mathbf{h}_{i,k}^{\mathrm{H}}\mathbf{w}_{i,k}s_k}_{\text{desired signal}} + \underbrace{\sum\nolimits_{l\neq k, l\in\mathcal{U}} \sum\nolimits_{i\in\mathcal{I}_l} \mathbf{h}_{i,k}^{\mathrm{H}}\mathbf{w}_{i,l}s_l}_{\text{multiuser interference}} + z_k, \tag{1}$$

where s_k is the data symbol of UE k, z_k is the additive complex white Gaussian noise that follows the distribution of $\mathcal{CN}(0,\sigma^2)$. It is assumed that $\mathbb{E}\{s_k\} = 0$ and $\mathbb{E}\{|s_k|^2\} = 1$, and the data streams for different UEs are independent of each other, i.e., we have $\mathbb{E}\{s_{k_1}s_{k_2}\} = 0$ for $k_1 \neq k_2, \forall k_1, k_2 \in \mathcal{U}$. Furthermore, the channel vector $\mathbf{h}_{i,k}$ can be decomposed as $\mathbf{h}_{i,k} = \sqrt{\alpha_{i,k}}\bar{\mathbf{h}}_{i,k}$, where $\alpha_{i,k}$ represents the large-scale channel gain that includes both the path loss and shadowing, while $\bar{\mathbf{h}}_{i,k}$ is the small-scale fading following the distribution of $\mathcal{CN}(\mathbf{0},\mathbf{I})$.

B. Channel Estimation for Intra-cluster CSI

To design the BF vectors, the entire network's CSI should be available at the BBU pool. However, for ultra-dense C-RANs with a large number of RRHs and UEs, it is infeasible to obtain all the CSI due to the limited amount of training resources. To deal with this issue, we assume that the BBU pool only needs to estimate the CSI from RRHs within each UE's cluster to the corresponding UE.

For the CSI of the RRHs outside its cluster, we assume that the BBU pool only has the knowledge of large-scale channel gains, i.e., $\{\alpha_{i,k}, \forall i \in \mathcal{I} \backslash \mathcal{I}_k, k \in \mathcal{U}\}$. This is a feasible assumption, since the large-scale channel gains change slowly and can be tracked with high accuracy.

In this paper, the channels are assumed to be frequency-flat within a coherence interval of T time slots, among which τ time slots are used for channel estimation, while the remaining $T-\tau$ time slots are dedicated to data transmission. Hence, the number of orthogonal pilot sequences is equal to τ. In ultra-dense C-RANs, the number of UEs is much higher than τ. The pilots should be reused among the UEs for the facilitation of channel estimation.

Let us denote the set of available pilot sequences as $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_\tau] \in \mathbb{C}^{\tau \times \tau}$ that satisfies the orthogonal condition of $\mathbf{Q}^H \mathbf{Q} = \mathbf{I}$. In TDD ultra-dense C-RANs, each UE sends the pilot sequence to the RRHs. Let us define by $\mathbf{q}_{\pi_k}$ the pilot sequence used by UE k. Then the pilot signal received at RRH i is

$$\mathbf{Y}_i = \sum\nolimits_{k \in \mathcal{U}} \sqrt{p_t} \mathbf{h}_{i,k} \mathbf{q}_{\pi_k}^{\mathrm{H}} + \mathbf{N}_i, \tag{2}$$

where p_t is the pilot transmit power at each UE, while $\mathbf{N}_i \in \mathbb{C}^{M \times \tau}$ is the Gaussian noise matrix, whose elements are independently generated and follow the distributions of $\mathcal{CN}(0, \sigma^2)$. To differentiate the channels from the UEs, the UEs with at least one common RRH should be allocated with orthogonal pilot sequences, i.e., we have $\mathbf{q}_{\pi_k}^{\mathrm{H}} \mathbf{q}_{\pi_{k'}} = 0, \text{for } k, k' \in \mathcal{U}_i, k \neq k', \forall i \in \mathcal{I}$. Furthermore, to control the estimation error, the maximum reuse time for each pilot should be below a fixed value $n_{\max}$, i.e., $n_l \leq n_{\max}, \forall l$, where n_l denotes the reuse time for pilot l. In this paper, we aim for minimizing the number of pilots required while satisfying the above two sets of constraints. The Dsatur algorithm of graph theory can be used for solving the pilot allocation problem, details of which can be found in [29]. Let us denote by $c^\star$ the minimum number of different colors, which is equal to the number of pilots τ.

Let us denote by $\mathcal{K}_{\pi_k}$ the set of UEs that reuse the same pilot of UE k. Then the minimum mean square error (MMSE) estimate of channel $\mathbf{h}_{i,k}$ is given by [30]

$$\hat{\mathbf{h}}_{i,k} = \frac{\alpha_{i,k}}{\sum_{l \in \mathcal{K}_{\pi_k}} \alpha_{i,l} + \hat{\sigma}^2} \frac{1}{\sqrt{p_t}} \mathbf{Y}_i \mathbf{q}_{\pi_k}, \tag{3}$$

where $\hat{\sigma}^2 = \sigma^2 / p_t$. According to the property of MMSE estimate [30], channel estimation error $\tilde{\mathbf{h}}_{i,k} = \mathbf{h}_{i,k} - \hat{\mathbf{h}}_{i,k}$ is independent of the channel estimate $\hat{\mathbf{h}}_{i,k}$ and is distributed as $\mathcal{CN}(\mathbf{0}, \delta_{i,k}\mathbf{I})$, where $\delta_{i,k}$ is given by

$$\delta_{i,k} = \frac{\alpha_{i,k} \left(\sum_{l \in \mathcal{K}_{\pi_k} \backslash \{k\}} \alpha_{i,l} + \hat{\sigma}^2 \right)}{\sum_{l \in \mathcal{K}_{\pi_k}} \alpha_{i,l} + \hat{\sigma}^2}. \tag{4}$$

III. Problem Formulation

The BF vectors from all RRHs in $\mathcal{I}_k$ can be merged into a single large-dimensional vector $\mathbf{w}_k = [\mathbf{w}_{i,k}^{\mathrm{H}}, \forall i \in \mathcal{I}_k]^{\mathrm{H}} \in \mathbb{C}^{|\mathcal{I}_k|M \times 1}$. Similarly, we define $\mathbf{g}_{l,k} = [\mathbf{h}_{i,k}^{\mathrm{H}}, \forall i \in \mathcal{I}_l]^{\mathrm{H}} \in \mathbb{C}^{|\mathcal{I}_l|M \times 1}$ as the aggregated perfect CSI from the RRHs in $\mathcal{I}_l$ to UE k, $\tilde{\mathbf{g}}_{k,k} = [\tilde{\mathbf{h}}_{i,k}^{\mathrm{H}}, \forall i \in \mathcal{I}_k]^{\mathrm{H}} \in \mathbb{C}^{|\mathcal{I}_k|M \times 1}$ and $\hat{\mathbf{g}}_{k,k} = [\hat{\mathbf{h}}_{i,k}^{\mathrm{H}}, \forall i \in \mathcal{I}_k]^{\mathrm{H}} \in \mathbb{C}^{|\mathcal{I}_k|M \times 1}$ as the aggregated CSI error and estimated CSI from the RRHs in $\mathcal{I}_k$ to UE k, respectively.

Since the channel estimation error can be written as $\tilde{\mathbf{g}}_{k,k} = \mathbf{g}_{k,k} - \hat{\mathbf{g}}_{k,k}$, the signal transmission model in (1) can be reformulated as

$$y_k = \underbrace{\hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k s_k}_{\text{Desired signal}} + \underbrace{\tilde{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k s_k}_{\text{Self-interference}} + \underbrace{\sum\nolimits_{l \neq k, l \in \mathcal{U}} \mathbf{g}_{l,k}^{\mathrm{H}} \mathbf{w}_l s_l}_{\text{Interference from other UEs}} + z_k, \forall k \in \mathcal{U}. \tag{5}$$

We define the effective noise as:

$$\tilde{z}_k = \tilde{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k s_k + \sum\nolimits_{l \neq k, l \in \mathcal{U}} \mathbf{g}_{l,k}^{\mathrm{H}} \mathbf{w}_l s_l + z_k, \forall k \in \mathcal{U}. \tag{6}$$

Then, (5) can be reformulated as

$$y_k = \hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k s_k + \tilde{z}_k, \forall k \in \mathcal{U}. \tag{7}$$

Unfortunately, the effective noise $\tilde{z}_k$ is neither independent nor Gaussian. However, we find that the input random variable s_k and the effective noise $\tilde{z}_k$ are uncorrelated. The reasons are given as follows: the input random variable s_k is clearly independent of $s_l, l \neq k$ and z_k. Furthermore, s_k is independent of the first term in (13) because the independence of the channel estimate $\hat{\mathbf{g}}_{k,k}$ and the channel estimation error $\tilde{\mathbf{g}}_{k,k}$. The variance of the effective noise is calculated as

$$\mathbb{E}\left[|\tilde{z}_k|^2\right] = \left|\tilde{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k\right|^2 + \sum\nolimits_{l \neq k, l \in \mathcal{U}} \left|\mathbf{g}_{l,k}^{\mathrm{H}} \mathbf{w}_l\right|^2 + \sigma^2 \tag{8}$$

where the expectation is taken over the random input variables $s_k, \forall k$ and noise variable z_k. According to Theorem 1 in [31], we know that the data rate for the channel in (9) is higher than the following channel system:

$$\hat{y}_k = \hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k s_k + \hat{z}_k, \forall k \in \mathcal{U}, \tag{9}$$

where $\hat{z}_k$ is the independent Gaussian noise with the same noise variance as the effective noise $\tilde{z}_k$. Then, by using the similar derivations in [32], [33], the effective SINR and the achievable data rate of UE k are respectively given by

$$\eta_k = \frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k\right|^2}{\left|\tilde{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k\right|^2 + \sum_{l \neq k, l \in \mathcal{U}} \left|\mathbf{g}_{l,k}^{\mathrm{H}} \mathbf{w}_l\right|^2 + \sigma^2} \tag{10}$$

and

$$r_k = \frac{T-\tau}{T}\mathbb{E}\left\{\log_2\left(1+\eta_k\right)\right\}, \forall k \in \mathcal{U}, \tag{11}$$

where T is the total number of time slots in a coherence interval, the expectation is taken over all uncertain terms, such as the unknown channel estimation errors $\left\{\tilde{\mathbf{h}}_{i,k}, i \in \mathcal{I}_k, \forall k \in \mathcal{U}\right\}$, and the small-scale inter-cluster CSI $\{\mathbf{h}_{i,k}, i \in \mathcal{I}\backslash\mathcal{I}_k\}$.

In this paper, we aim for optimizing the BF vectors to maximize the WSR of all UEs, while satisfying the power constraints of all RRHs. Specifically, we formulate the following optimization problem

$$\max_{\mathbf{w}} \quad \sum\nolimits_{k\in\mathcal{U}} \omega_k r_k \tag{12a}$$

$$\text{s.t.} \quad \sum\nolimits_{k\in\mathcal{U}_i} \left\|\mathbf{w}_{i,k}\right\|^2 \leq P_{i,\max}, i \in \mathcal{I}, \tag{12b}$$

where $\mathbf{w}$ is the collection of all BF vectors, ω_k is the weight assigned to UE k for controlling the fairness among the UEs, r_k is the data rate of UE k defined in (11), and $P_{i,\max}$ is the power limit of RRH i.

Due to the multiple uncertain terms, it is difficult to obtain the accurate closed-form expression of each UE's data rate. Similar to [25], we consider its LB, which leads to a tractable expression. The LB can be obtained by using Jensen's inequality, which is given by [25]

$$r_k \geq \frac{T-\tau}{T}\log_2\left(1+\frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2}{\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum_{l\neq k,l\in\mathcal{U}}\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2}\right) \triangleq \tilde{r}_k, \tag{13}$$

where $\mathbf{E}_{k,k} = \mathrm{blkdiag}\left\{\delta_{i,k}\mathbf{I}_{M\times M}, i \in \mathcal{I}_k\right\}$ and $\mathbf{A}_{l,k} = \mathbb{E}\left\{\mathbf{g}_{l,k}^{\mathrm{H}}\mathbf{g}_{l,k}\right\} \in \mathbb{C}^{M|\mathcal{I}_l|\times M|\mathcal{I}_l|}$.

Let us define the indices of $\mathcal{I}_l$ as $\mathcal{I}_l = \{s_1^l, \cdots, s_{|\mathcal{I}_l|}^l\}$. Then, we have

$$\mathbf{A}_{l,k} = \begin{bmatrix} \left(\mathbf{A}_{l,k}\right)_{1,1} & \cdots & \left(\mathbf{A}_{l,k}\right)_{1,|\mathcal{I}_l|} \\ \vdots & \ddots & \vdots \\ \left(\mathbf{A}_{l,k}\right)_{|\mathcal{I}_l|,1} & \cdots & \left(\mathbf{A}_{l,k}\right)_{|\mathcal{I}_l|,|\mathcal{I}_l|} \end{bmatrix}, l \neq k, \tag{14}$$

where $\left(\mathbf{A}_{l,k}\right)_{i,j} \in \mathbb{C}^{M\times M}, i,j \in 1,\cdots,|\mathcal{I}_l|$ is the block matrix of $\mathbf{A}_{l,k}$ at the ith row and jth column, given by

$$\left(\mathbf{A}_{l,k}\right)_{i,j} = \begin{cases} \hat{\mathbf{h}}_{s_i^l,k}\hat{\mathbf{h}}_{s_j^l,k}^{\mathrm{H}}, & \text{if } s_i^l, s_j^l \in \mathcal{I}_k, i \neq j, \\ \hat{\mathbf{h}}_{s_i^l,k}\hat{\mathbf{h}}_{s_j^l,k}^{\mathrm{H}} + \delta_{s_i^l,k}\mathbf{I}_{M\times M}, & \text{if } s_i^l, s_j^l \in \mathcal{I}_k, i = j, \\ \alpha_{s_i^l,k}\mathbf{I}_{M\times M}, & \text{if } s_i^l, s_j^l \notin \mathcal{I}_k, \text{and } i = j, \\ \mathbf{0}_{M\times M}, & \text{otherwise.} \end{cases} \tag{15}$$

It is proved in Appendix A that $\mathbf{A}_{l,k}$ is a positive definite matrix.

It is important to characterize the tightness of this LB. However, it is difficult to derive the accurate closed-form expression of r_k for the general case. Hence in Appendix B, we derive the accurate closed-form rate expression for a special case: the RRH cluster for each UE is non-overlapped with each other, i.e., $\mathcal{I}_k \cap \mathcal{I}_{k'} = \emptyset, \forall k, k' \in \mathcal{U}$. In Fig. 2, we consider a non-overlapped ultra-dense C-RAN network deployed in a square area of coordinates $[-1/2, 1/2] \times [-1/2, 1/2]$ km. This area is divided into nine $1/3$ km $\times$ $1/3$ km squares. There is one UE located in the center of each small square, and three RRHs are randomly distributed within a circle centered at the UE with radius equal to $1/6$ km, as shown in Fig. 2. These three RRHs are exclusively serving the centered UE. Each RRH is assumed to be equipped with four antennas. The other simulation parameters are the same as in the simulation section. In Fig. 2, the UEs with the same shape and color are assumed to reuse the same pilot, and UE 5 in the center of this area is assigned an orthogonal pilot sequence. For the ultra-dense C-RAN of Fig. 2, a total of five pilots are required and the total number of time slots within the channel's coherence time is set to $T = 80$. It is assumed that all RRHs transmit with their maximum power limit and the BF direction is chosen to match the corresponding channel vector. Fig. 3 investigates the tightness of the LB derived for this non-overlapped scenario, where three curves are plotted: the rate LB derived in (13), the accurate closed form expression derived in Appendix B, and the Monte-Carlo simulation results. We observe from this figure that the curve associated with accurate closed-form expression coincides well with that associated with the Monte-Carlo simulation, which verifies the correctness of our analytical results. In the low transmit power regime, the LB is very tight, and almost equal to the accurate data rate. However, the gap increases with the transmit power limit and becomes constant in the high transmit power regime, where the system becomes interference limited. Note that the maximum gap is at most 4%, which is acceptable for practical applications. In the simulation section, we also show that the approximation error is minor.

Hence, it is reasonable to consider its rate LB, instead of its accurate expression. Then, by replacing r_k in Problem (12) with $\tilde{r}_k$ in (13), we consider the following optimization problem

$$\max_{\mathbf{w}} \quad \sum\nolimits_{k\in\mathcal{U}} \omega_k \tilde{r}_k \tag{16a}$$

$$\text{s.t.} \quad \sum\nolimits_{k\in\mathcal{U}_i} \left\|\mathbf{w}_{i,k}\right\|^2 \leq P_{i,\max}, i \in \mathcal{I}. \tag{16b}$$

In [34], the WSR maximization problem has been shown to be NP-hard for the simple single-antenna interference channel [35]. Intuitively, Problem (16) formulated for the imperfect CSI case,

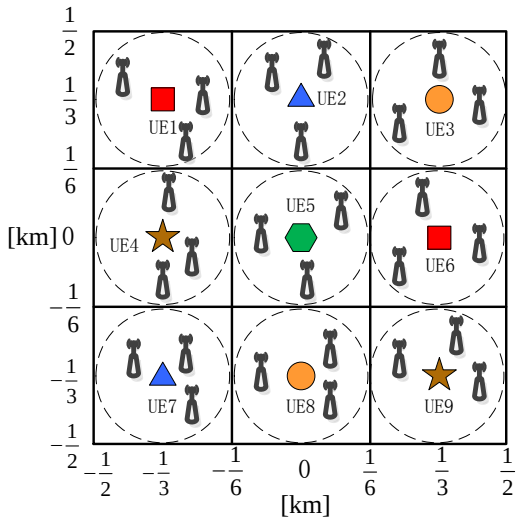

Fig. 2. Illustration of a non-overlapped ultra-dense C-RAN with nine UEs. Each UE is exclusively served by its nearby three RRHs, which are randomly distributed in a circle area centered at the UE with radius equal to $1/6$ km. The UEs marked with the same shape and color are reusing the same pilot, and UE 5 in the center of this area is allocated with one orthogonal pilot.

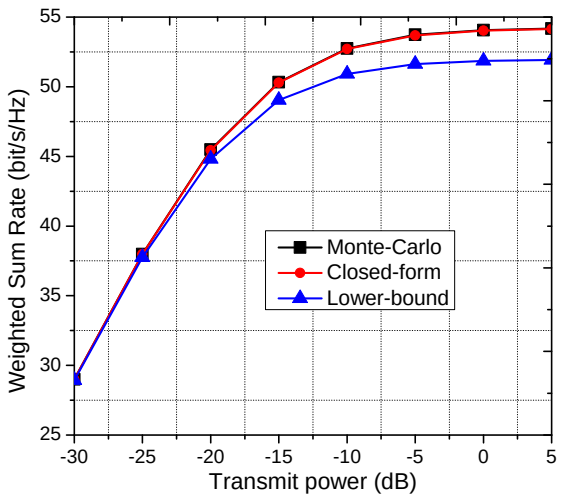

Fig. 3. Data rate versus the transmit power for the non-overlapped case. Three curves are plotted: the derived rate LB in (13), the accurate derived closed form expression in Appendix B, and the Monte-Carlo simulation results. The gap between the LB and the accurate rate expression is at most 4%, which is acceptable.

which involves the BF design and power allocation for multiple UEs, is also NP-hard. However, the strict proof of the NP-hardness requires excessive additional efforts, which are beyond the scope of this paper. In the following section, we conceive three different algorithms striking different tradeoffs between performance and complexity to solve Problem (16).

IV. Algorithms to Solve Problem (16)

In this section, two different algorithms striking different tradeoffs between the performance and complexity are developed. Specifically, we first provide the OPA algorithm [36], [37] to obtain the globally optimal solution of Problem (16). Then, an iterative algorithm based on modifying the WMMSE algorithm [38] is proposed.

A. Globally Optimal Solution Based on the OPA Algorithm

In this subsection, we aim for providing the globally optimal solution to Problem (16). In the following, we first provide an equivalent formulation of Problem (16), based on which the OPA algorithm is customized to solve it optimally.

Consider the following optimization problem

$$\max_{\mathbf{w},\mathbf{t}} \quad \sum\nolimits_{k\in\mathcal{U}} \omega_k t_k \tag{17a}$$

$$\text{s.t.} \quad (16b), \frac{T-\tau}{T}\log_2\left(1+\frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2}{\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum_{l\neq k,l\in\mathcal{U}}\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2}\right)\geq t_k, \forall k, \tag{17b}$$

where $\mathbf{t}$ represents the collections of all auxiliary variables $t_k, \forall k \in \mathcal{U}$. The equivalence between Problem (17) and Problem (16) can be readily verified by showing that the constraints (17b) in Problem (17) hold with equality at the optimum solution of Problem (17).

The formulation in (17) facilitates the development of the OPA algorithm based on monotonic optimization. Specifically, it may be readily shown that the objective function (OF) of Problem (17) monotonically increases with each element of $\mathbf{t}$. Thus, we can apply the OPA algorithm to obtain the globally optimal solution of Problem (17). The detailed description of the monotonic optimization-based OPA algorithm can be found in [36], [37], [39]. For the sake of consistency, we reuse the same notations and definitions as in [36]. Define the achievable rate region for this scenario as follows:

$$\mathcal{T} \triangleq \bigcup_{\substack{\sum_{k\in\mathcal{U}_i}\left\|\mathbf{w}_{i,k}\right\|^2\leq P_{i,\max},\\ i\in\mathcal{I}}} \left\{ \begin{array}{l} (t_1,\cdots,t_K): \frac{T-\tau}{T}\cdot \\ \log_2\left(1+\frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2}{\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum\limits_{l\neq k,l\in\mathcal{U}}\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2}\right)\geq t_k\geq 0, \forall k \end{array} \right\}. \tag{18}$$

1) Determining Initial Box: We first have to determine the initial box that contains all feasible $t_k, \forall k$. It may be readily shown that the LB for each t_k is zero. Hence, we only have to compute the upper bound (UB) for each t_k, which is as follows:

$$t_k \leq \frac{T-\tau}{T}\log_2\left(1+\frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2}{\sigma^2}\right) \leq \frac{T-\tau}{T}\log_2\left(1+\frac{\left(\sum\limits_{i\in\mathcal{I}_k}\sqrt{P_{i,\max}}\left\|\hat{\mathbf{h}}_{i,k}\right\|\right)^2}{\sigma^2}\right) \triangleq z_k^{(1)}, \tag{19}$$

where the first inequality is due to omitting the multiuser interference and self interference, the second inequality follows due to the application of the Cauchy-Schwarz inequality and the power constraints for $\mathbf{w}_{i,k}$.

2) Updating the Polyblocks: In each iteration of the OPA algorithm, the polyblock containing the rate region $\mathcal{T}$ defined in (18) should be updated. Define $\mathcal{P}^{(n)}$ as the polyblock in the nth

iteration, and define $\mathcal{Z}^{(n)}$ as the set containing all the vertices of the polyblock $\mathcal{P}^{(n)}$. The vertex in polyblock $\mathcal{P}^{(n)}$ that achieves the maximum WSR is given by:

$$\tilde{\mathbf{z}}^{(n)} = \arg \max_{\mathbf{z} \in \mathcal{Z}^{(n)}} \sum\nolimits_{k \in \mathcal{U}} \omega_k z_k, \tag{20}$$

where z_k denotes the kth element of $\mathbf{z}$. Define $\mathbf{t}^{(n)}$ as the intersection point on the Pareto boundary with the line $\delta\tilde{\mathbf{z}}^{(n)}$. Then, the K new vertices adjacent to $\tilde{\mathbf{z}}^{(n)}$ can be generated as:

$$\mathbf{z}^{(n),i} = \tilde{\mathbf{z}}^{(n)} - \left(\tilde{z}_i^{(n)} - t_i^{(n)}\right) \mathbf{e}_i, i = 1, \cdots, K, \tag{21}$$

where $\mathbf{z}^{(n),i}$ denotes the ith new vertex generated in the nth iteration, $\mathbf{e}_i$ denotes the unit vector where the ith element is equal to one, $\tilde{z}_i^{(n)}$ and $t_i^{(n)}$ denotes the ith element of vectors $\tilde{\mathbf{z}}^{(n)}$ and $\mathbf{t}^{(n)}$, respectively. Then, the new set of vertices to be used in the $(n+1)$th iteration is given by

$$\mathcal{Z}^{(n+1)} = \mathcal{Z}^{(n)} \backslash \tilde{\mathbf{z}}^{(n)} \cup \left\{\mathbf{z}^{(n),i}, \cdots, \mathbf{z}^{(n),K}\right\}. \tag{22}$$

3) Finding the Intersection Points: The key step in the OPA algorithm is to find the intersection point on the Pareto boundary of the rate region in each iteration. Let us define the selected vertex in the nth iteration as $\tilde{\mathbf{z}}^{(n)}$ in (20). Then $\boldsymbol{\alpha}^{(n)} = \tilde{\mathbf{z}}^{(n)} \Big/ \sum_{k \in \mathcal{U}} \tilde{z}_k^{(n)}$ represents the slope of the line cross the Pareto boundary of the rate region. Let us represent by $R_{\text{sum}} = \sum_{k \in \mathcal{U}} t_k$ the sum rate of all UEs. Then the intersection point in the nth iteration is given by $\mathbf{t}^{(n)} = R_{\text{sum}}^{\star} \boldsymbol{\alpha}^{(n)}$, where $R_{\text{sum}}^{\star}$ is the optimal solution of the following problem:

$$\max_{\mathbf{w}, R_{\max}} \quad R_{\max} \tag{23a}$$

$$\text{s.t.} \quad (16b), \frac{T-\tau}{T} \log_2 \left(1 + \frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k\right|^2}{\mathbf{w}_k^{\mathrm{H}} \mathbf{E}_{k,k} \mathbf{w}_k + \sum_{l \neq k, l \in \mathcal{U}} \mathbf{w}_l^{\mathrm{H}} \mathbf{A}_{l,k} \mathbf{w}_l + \sigma^2}\right) \geq \alpha_k^{(n)} R_{\text{sum}}, \forall k. \tag{23b}$$

The bisection based search method can be used for finding the optimal $R_{\max}$ of Problem (23). For a given $\bar{R}_{\max}$, we have to check the feasibility of the following optimization problem:

$$\text{find} \quad \mathbf{w} \tag{24a}$$

$$\text{s.t.} \quad (16b), \frac{T-\tau}{T} \log_2 \left(1 + \frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k\right|^2}{\mathbf{w}_k^{\mathrm{H}} \mathbf{E}_{k,k} \mathbf{w}_k + \sum_{l \neq k, l \in \mathcal{U}} \mathbf{w}_l^{\mathrm{H}} \mathbf{A}_{l,k} \mathbf{w}_l + \sigma^2}\right) \geq \alpha_k^{(n)} \bar{R}_{\text{sum}}, \forall k. \tag{24b}$$

To solve the problem, we introduce the following alternative optimization problem:

$$\min_{\mathbf{w},s\geq 0} \quad s \tag{25a}$$

$$\text{s.t.} \quad (16b), \hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k + s \geq \sqrt{\gamma_k^{(n)}}\sqrt{\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k + \sum\nolimits_{l\neq k, l\in\mathcal{U}} \mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l + \sigma^2}, \forall k, \tag{25b}$$

$$\mathrm{Im}(\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k) = 0, \forall k, \tag{25c}$$

where s is the auxiliary variable introduced, and $\gamma_k^{(n)}$ is given by $\gamma_k^{(n)} = 2^{\frac{T}{T-\tau}\alpha_k^{(n)}\bar{R}_{\mathrm{sum}}} - 1$. It is plausible that Problem (25) is always feasible, and it is a second-order cone programming (SOCP) problem that can be efficiently solved by using the interior point method of [40]. If the optimal solution of s is equal to zero, then Problem (24) is feasible, which means that $R_{\mathrm{sum}}^{\star} \geq \bar{R}_{\mathrm{sum}}$. Otherwise, it is infeasible and $R_{\mathrm{sum}}^{\star} \leq \bar{R}_{\mathrm{sum}}$. Hence, the bisection based search method can be adopted for finding the optimal $R_{\mathrm{sum}}^{\star}$.

The detailed steps of the OPA algorithm can be found in Algorithm 1 of [36], details of which are omitted due to the space limitation.

***Complexity Analysis**:* We now analyze the complexity of the OPA algorithm. The main complexity of the algorithm lies in finding $R_{\mathrm{sum}}^{\star}$ by solving Problem (23) with the aid of the bisection based search algorithm. For simplicity, we assume that the cluster size for each UE is equal, i.e., $|\mathcal{I}_k| = l, \forall k \in \mathcal{U}$. In each iteration of the bisection search algorithm, we should solve Problem (25) that is an SOCP problem and can be solved by the interior point method of [40]. This problem has $2lMK + 1$ real variables, plus I SOC constraints where each one has $2|\mathcal{U}_i|M$ real variables and K SOC constraints where each one has $2lMK + 1$ real variables. According to [page 196, [41]], the total complexity order of solving Problem (25) is given by $O\left[(2MKl+1)^2\left(2M\sum_{i\in\mathcal{I}}|\mathcal{U}_i| + 2MlK^2 + K\right)\right]$. Let us denote the accuracy of the bisection search method as ε, and the sum rate UB as $R_{\max}$. The total number of iterations required by the bisection search method is $\log_2(R_{\max}/\varepsilon)$. Note that $\sum_{i\in\mathcal{I}}|\mathcal{U}_i| = \sum_{k\in\mathcal{U}}|\mathcal{I}_k| = Kl$. The total complexity order of solving Problem (23) is $O\left[\log_2(R_{\max}/\varepsilon)M^3l^3K^4\right]$. Upon denoting the total number of iterations required by the OPA algorithm as $t_{\mathrm{OPA,iter}}$, the total complexity of the OPA algorithm is on the order of $O\left[t_{\mathrm{OPA,iter}}\log_2(R_{\max}/\varepsilon)M^3l^3K^4\right]$. It is analytically difficult to derive the exact relationship between $t_{\mathrm{OPA,iter}}$ and K. However, from Theorem 1 in [42], the OPA algorithm converges Q-super linearly [43] to the optimal solution. Note that the OPA algorithm involves two layers of iterations, it thus has a high computational complexity, hence it can only be used for small-scale C-RANs as a performance benchmark. In the following two sections, we

develop two low-complexity algorithms that are suitable for larger ultra-dense C-RANs.

B. Modified WMMSE Method

The WMMSE algorithm proposed in [38] was shown to be an efficient method of solving the WSR maximization problem, and has been successfully applied in diverse setups [4], [12], [18], [19], [44], [45]. Unfortunately, there are no contributions considering the application of the WMMSE method for solving Problem (16). There are two difficulties that preclude the direct application of the WMMSE method: Firstly, we considered the imperfect CSI scenario where each UE suffers from self-interference, which is not considered in [38]; Secondly, the incomplete CSI case is considered in this paper, where the rank of channel covariance matrix may be higher than 1, i.e., $\mathrm{rank}\left(\mathbf{A}_{l,k}\right) > 1$. However, the authors of [38] considered the perfect CSI case, where the rank of the channel covariance matrices is equal to one when each UE is equipped with one antenna.

To deal with the above difficulties, we decompose each interfering sources into multiple interfering sources. Specifically, the self-interference matrix $\mathbf{E}_{k,k}$ can be decomposed as $\mathbf{E}_{k,k} = \mathbf{F}_{k,k}\mathbf{F}_{k,k}^{\mathrm{H}}$, where $\mathbf{F}_{k,k} = \mathrm{blkdiag}\left\{\sqrt{\delta_{i,k}}\mathbf{I}_{M\times M}, i \in \mathcal{I}_k\right\}$. Similarly, since $\{\mathbf{A}_{l,k}, \forall l\}$ are positive definite matrices, as shown in Appendix A, they can be decomposed as $\mathbf{A}_{l,k} = \mathbf{V}_{l,k}\mathbf{V}_{l,k}^{\mathrm{H}}, \forall l$, where $\mathbf{V}_{l,k} = \left[\mathbf{v}_{l,k,1}, \cdots, \mathbf{v}_{l,k,d_{l,k}}\right]$ with $d_{l,k}$ being the rank of $\mathbf{A}_{l,k}$. Then, we can construct the following auxiliary signal transmission model for UE k

$$\tilde{y}_k = \hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k s_k + \sum_{d=1}^{M|\mathcal{I}_k|}\mathbf{f}_{k,k,d}^{\mathrm{H}}\mathbf{w}_k s_{k,d} + \sum_{l\in\mathcal{U},l\neq k}\sum_{d=1}^{d_{l,k}}\mathbf{v}_{l,k,d}^{\mathrm{H}}\mathbf{w}_l s_{l,d} + z_k, \tag{26}$$

where s_k is the desired data stream, $\mathbf{f}_{k,k,d}$ is the dth column of matrix $\mathbf{F}_{k,k}$ that can be regarded as the channel vector spanning from its dth self-interference source, $\mathbf{v}_{l,k,d}$ can also be regarded as the channel vector from the dth interfering source of UE l, while $s_{k,d}$ and $s_{l,d}$ are the corresponding data streams. All the data streams in (26) are assumed to be independently generated and follow the distribution of $\mathcal{CN}(0,1)$. Note that these interfering sources from the same UE employ the same BF vector. It should be emphasized that the transmission model of (26) is different from the actual one in (5), and the former one is constructed for the sake of solving the original problem.

By adopting $u_k \in \mathbb{C}$ to decode UE k's data, we obtain $\tilde{s}_k = u_k\tilde{y}_k$. Due to the independence of

the data streams and noise, the mean square error of decoding s_k is computed as

$$\begin{aligned}
&\epsilon_k\left(u_k,\mathbf{w}\right)\\
=&\mathbb{E}\left[\left(\tilde{s}_k-s_k\right)\left(\tilde{s}_k-s_k\right)^H\right]\\
=&\left(u_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k-1\right)\left(u_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k-1\right)^{\mathrm{H}}+\left|u_k\right|^2\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum_{l\in\mathcal{U},l\neq k}\left|u_k\right|^2\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2\left|u_k\right|^2. \quad (27)
\end{aligned}$$

Then, as in [12], [19], [38], we introduce the following function:

$$\Psi_k\left(\mathbf{w},u_k,q_k\right)=\frac{T-\tau}{T}\log_2 e\left(\ln(q_k)-q_k\epsilon_k\left(u_k,\mathbf{w}\right)+1\right),\forall k\in\mathcal{U}, \quad (28)$$

where q_k is the auxiliary variable introduced. Then, the following lemma can be formulated.

Lemma 1: Given the fixed BF vectors $\mathbf{w}$, the function $\Psi_k\left(\mathbf{w},u_k,q_k\right)$ gives a LB of the achievable data rate $\tilde{r}_k$, and the optimal variables u_k and q_k of $\Psi_k\left(\mathbf{w},u_k,q_k\right)$ achieving $\tilde{r}_k$ are respectively given by

$$u_k^\star = \left(\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2+\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum\nolimits_{l\in\mathcal{U},l\neq k}\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2\right)^{-1}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k, \quad (29)$$

$$q_k^\star = \left[\epsilon_k\left(u_k^\star,\mathbf{w}\right)\right]^{-1}, \quad (30)$$

where $\epsilon_k\left(u_k^\star,\mathbf{w}\right)$ can be calculated as

$$\epsilon_k\left(u_k^\star,\mathbf{w}\right)=1-\frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2}{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2+\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum_{l\in\mathcal{U},l\neq k}\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2}. \quad (31)$$

Proof: Please see Appendix C. □

By replacing $\tilde{r}_k$ in the OF of Problem (16) by its LB, Problem (16) can be transformed to

$$\max_{\{\mathbf{w},\mathbf{u},\mathbf{q}\}} \quad \sum\nolimits_{k\in\mathcal{U}}\omega_k\Psi_k\left(\mathbf{w},u_k,q_k\right) \quad (32a)$$

$$\text{s.t.} \quad \sum\nolimits_{k\in\mathcal{U}_i}\left\|\mathbf{w}_{i,k}\right\|^2\leq P_{i,\max},i\in\mathcal{I}, \quad (32b)$$

where $\mathbf{u}$ and $\mathbf{q}$ denote the collections of $u_k,\forall k$ and $q_k,\forall k$, respectively.

Note that for any given two sets of variables $\mathbf{w},\mathbf{u},\mathbf{q}$, Problem (32) is convex w.r.t. the remaining set of variables. Hence, Problem (32) can be solved by using the block coordinate descent method. Specifically, given the BF vectors $\mathbf{w}$, the decoding variables $\mathbf{u}$ and the auxiliary variables $\mathbf{q}$ are updated according to (29) and (30), respectively; then we update the BF vectors $\mathbf{w}$ with fixed $\mathbf{u}$ and $\mathbf{q}$. We only have to solve the latter problem. By substituting the expression of $\Psi_k\left(\mathbf{w},u_k,q_k\right)$

into the OF of Problem (32) and discarding some constant terms, the problem of optimizing the BF vectors can be formulated as

$$\min_{\mathbf{w}} \quad \sum\nolimits_{k\in\mathcal{U}} \tilde{\omega}_k \left(|u_k|^2 \mathbf{w}_k^{\mathrm{H}} \tilde{\mathbf{E}}_{k,k} \mathbf{w}_k - 2\mathrm{Re}\left\{ u_k^{\mathrm{H}} \hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k \right\} + \sum_{l\in\mathcal{U}, l\neq k} |u_k|^2 \mathbf{w}_l^{\mathrm{H}} \mathbf{A}_{l,k} \mathbf{w}_l \right) \tag{33a}$$

$$\text{s.t.} \quad \sum\nolimits_{k\in\mathcal{U}_i} \|\mathbf{w}_{i,k}\|^2 \leq P_{i,\max}, i \in \mathcal{I}, \tag{33b}$$

where $\tilde{\omega}_k = \omega_k q_k$ and $\tilde{\mathbf{E}}_{k,k} = \hat{\mathbf{g}}_{k,k}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}} + \mathbf{E}_{k,k}$. Note that $\tilde{\mathbf{E}}_{k,k}$ is a positive definite matrix. This optimization problem can be transformed to a second-order cone programming (SOCP) problem that can be efficiently solved by using the CVX package [46]. However, the CVX package may not be convenient for practical programming in Digital Signal Processing (DSP) or for Field Programmable Gate Arrays (FPGA). Furthermore, directly solving Problem (33) through the CVX package cannot reveal the optimal structure of the BF vectors. In the following part, we will provide an alternative algorithm based on the Lagrangian dual decomposition method, which beneficially facilitates the programming in DSP or FPGA implementations.

Let us now summarize the modified WMMSE method in Algorithm 1.

Algorithm 1 Modified WMMSE Method

1: Initial the iteration number $n = 1$, the accuracy ε. Initialize the feasible BF vectors $\mathbf{w}^{(0)}$. Then, compute $\mathbf{u}^{(0)}$ and $\mathbf{q}^{(0)}$ according to (29) and (30), respectively. Calculate the value of the OF in Problem (32) as $\mathrm{Obj}^{(0)}$.

2: With $\mathbf{u}^{(n-1)}$ and $\mathbf{q}^{(n-1)}$, update $\mathbf{w}^{(n)}$ by solving the Problem (33);

3: With $\mathbf{w}^{(n)}$, update $\mathbf{u}^{(n)}$ and $\mathbf{q}^{(n)}$ according to (29) and (30), respectively;

4: Calculate the OF $\mathrm{Obj}^{(n)}$, if $\left|\mathrm{Obj}^{(n)} - \mathrm{Obj}^{(n-1)}\right| / \mathrm{Obj}^{(n)} \leq \varepsilon$ holds, terminate; Otherwise, set $n \leftarrow n + 1$ and go to step 2.

Convergence Analysis: The modified WMMSE can be guaranteed to converge, which may be proved by using a similar approach to that of the WMMSE in [38]. It can be verified that in each step of Algorithm 1, the OF value of Problem (32) is non-decreasing. Since the BF vectors have power constraints, the OF value must have an UB. Hence, Algorithm 1 is guaranteed to converge.

1) Lagrangian dual decomposition method to solve Problem (33): It may be readily shown that Problem (33) is a convex one, and the Slater's condition [40] is satisfied. Hence, Problem (33) can be equivalently solved by solving its dual problem. Specifically, we first introduce the following

block diagonal matrices

$$\mathbf{B}_{i,k}= \mathrm{diag}\left\{\overbrace{\mathbf{0}_{1\times M}}^{s_1^k},\cdots,\overbrace{\mathbf{1}_{1\times M}}^{s_j^k},\overbrace{\mathbf{0}_{1\times M}}^{s_{j+1}^k},\cdots,\overbrace{\mathbf{0}_{1\times M}}^{s_{|\mathcal{I}_k|}^k}\right\},\ \text{if}\ s_j^k=i,\forall i\in\mathcal{I},k\in\mathcal{U}. \tag{34}$$

Then, we have $\|\mathbf{w}_{i,k}\|^2=\mathbf{w}_k^{\mathrm{H}}\mathbf{B}_{i,k}\mathbf{w}_k$. Following further manipulations, the Lagrangian function of Problem (33) can be written as

$$\mathcal{L}\left(\mathbf{w},\boldsymbol{\lambda}\right)=\sum_{k\in\mathcal{U}}\left(\mathbf{w}_k^{\mathrm{H}}\mathbf{G}_k\mathbf{w}_k-\tilde{\omega}_k u_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k-\tilde{\omega}_k\mathbf{w}_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}u_k\right)+\sum_{i\in\mathcal{I}}\lambda_i\left(\sum_{k\in\mathcal{U}_i}\|\mathbf{w}_{i,k}\|^2-P_{i,\max}\right), \tag{35}$$

where $\boldsymbol{\lambda}=\{\lambda_i,i\in\mathcal{I}\}$ are the dual variables associated with the per-RRH power constraints, and $\mathbf{G}_k=\tilde{\omega}_k|u_k|^2\tilde{\mathbf{E}}_{k,k}+\sum_{l\in\mathcal{U},l\neq k}|u_l|^2\tilde{\omega}_l\mathbf{A}_{k,l}$.

The dual function is given by

$$\begin{aligned} g\left(\boldsymbol{\lambda}\right) &= \min_{\mathbf{w}}\mathcal{L}\left(\mathbf{w},\boldsymbol{\lambda}\right) && (36)\\ &= \min_{\mathbf{w}}\sum_{k\in\mathcal{U}}\left(\mathbf{w}_k^{\mathrm{H}}\mathbf{J}_k\mathbf{w}_k-\tilde{\omega}_k u_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k-\tilde{\omega}_k\mathbf{w}_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}u_k\right)-\sum_{i\in\mathcal{I}}\lambda_i P_{i,\max}, && (37)\end{aligned}$$

where $\mathbf{J}_k=\mathbf{G}_k+\sum_{i\in\mathcal{I}_k}\lambda_i\mathbf{B}_{i,k}$. Since $\mathbf{J}_k$ is a positive definite matrix, Problem (37) is a strictly convex optimization problem, and its optimal solution can be uniquely obtained by solving the first-order equation:

$$\mathbf{w}_k^{\star}=\tilde{\omega}_k u_k\mathbf{J}_k^{-1}\hat{\mathbf{g}}_{k,k}. \tag{38}$$

Then, the dual pair of Problem (33) is defined as

$$\max_{\{\lambda_i\geq 0,\forall i\}} g(\boldsymbol{\lambda}). \tag{39}$$

Since Problem (33) is a convex one, the duality gap between the dual problem and its original problem is zero. Hence, we can solve its dual problem instead of directly solving the original problem. To solve the dual problem in (39), we invoke the subgradient method [47], where the subgradient[1] of the function $g(\cdot)$ at $\boldsymbol{\lambda}^{(n)}=[\lambda_1^{(n)},\cdots,\lambda_I^{(n)}]^{\mathrm{T}}$ is required at the nth iteration. This subgradient is provided in the following theorem.

Theorem 1: Let us denote by the optimal solution of Problem (37) $\{\mathbf{w}_k^{\star}(\boldsymbol{\lambda}^{(n)}),\forall k\}$ when $\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(n)}$. Then, the subgradient of function $g(\cdot)$ at $\boldsymbol{\lambda}^{(n)}$ in the nth iteration is given by

$$\mathbf{d}^{(n)}=\mathbf{P}^{\star}(\boldsymbol{\lambda}^{(n)})-\mathbf{P}_{\max}, \tag{40}$$

[1]According to [47], a vector $\mathbf{d}$ is a subgradient of $g(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^{(n)}$, if for all $\boldsymbol{\lambda}$, $g(\boldsymbol{\lambda})\leq g(\boldsymbol{\lambda}^{(n)})+\mathbf{d}^T(\boldsymbol{\lambda}-\boldsymbol{\lambda}^{(n)})$ holds.

where $\mathbf{P}^{\star}(\boldsymbol{\lambda}^{(n)}) = \left[p_1^{\star}(\boldsymbol{\lambda}^{(n)}), \cdots, p_I^{\star}(\boldsymbol{\lambda}^{(n)})\right]^{\mathrm{T}}$ with $p_i^{\star}(\boldsymbol{\lambda}^{(n)}) = \sum_{k\in\mathcal{U}_i} \left\|\mathbf{w}_{i,k}^{\star}(\boldsymbol{\lambda}^{(n)})\right\|^2$, and $\mathbf{P}_{\max} = [P_{1,\max}, \cdots, P_{I,\max}]^{\mathrm{T}}$.

Proof: With any given $\tilde{\boldsymbol{\lambda}}$, let us denote the optimal solution of Problem (37) by $\{\mathbf{w}_k^{\star}(\tilde{\boldsymbol{\lambda}}), \forall k\}$ when $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$. Then, we have

$$
\begin{aligned}
g(\tilde{\boldsymbol{\lambda}}) &= \min_{\mathbf{w}} \sum_{k\in\mathcal{U}} \left(\mathbf{w}_k^{\mathrm{H}}\mathbf{G}_k\mathbf{w}_k - \tilde{\omega}_k u_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k - \tilde{\omega}_k \mathbf{w}_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}u_k\right) + \sum_{i\in\mathcal{I}} \tilde{\lambda}_i \left(\sum_{k\in\mathcal{U}_i} \|\mathbf{w}_{i,k}\|^2 - P_{i,\max}\right) \\
&\leq \sum_{k\in\mathcal{U}} \left(\mathbf{w}_k^{\mathrm{H}}(\boldsymbol{\lambda}^{(n)})\mathbf{G}_k\mathbf{w}_k(\boldsymbol{\lambda}^{(n)}) - \tilde{\omega}_k u_k^{\mathrm{H}}\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k(\boldsymbol{\lambda}^{(n)}) - \tilde{\omega}_k \mathbf{w}_k^{\mathrm{H}}(\boldsymbol{\lambda}^{(n)})\hat{\mathbf{g}}_{k,k}u_k\right) \\
&\qquad + \sum_{i\in\mathcal{I}} \tilde{\lambda}_i \left(\sum_{k\in\mathcal{U}_i} \left\|\mathbf{w}_{i,k}(\boldsymbol{\lambda}^{(n)})\right\|^2 - P_{i,\max}\right) \qquad (41) \\
&= g(\boldsymbol{\lambda}^{(n)}) + \sum_{i\in\mathcal{I}} \left(\tilde{\lambda}_i - \lambda_i^{(n)}\right) \left(\sum_{k\in\mathcal{U}_i} \left\|\mathbf{w}_{i,k}(\boldsymbol{\lambda}^{(n)})\right\|^2 - P_{i,\max}\right) \qquad (42) \\
&= g(\boldsymbol{\lambda}^{(n)}) + \left(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^{(n)}\right)^{\mathrm{T}} (\mathbf{P}^{\star}(\boldsymbol{\lambda}^{(n)}) - \mathbf{P}_{\max}), \qquad (43)
\end{aligned}
$$

where (41) follows due to the fact that $\mathbf{w}_k(\boldsymbol{\lambda}^{(n)})$ is not the optimal solution of Problem (37), when $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$. Hence, the proof is complete. ■

Based on Theorem 1, the Lagrangian dual variables are updated as follows

$$
\boldsymbol{\lambda}^{(n+1)} = \left[\boldsymbol{\lambda}^{(n)} + \zeta^{(n)}\mathbf{d}^{(n)}\right]^{+}, \qquad (44)
$$

where $[x]^{+}$ denotes the maximum value between x and 0, while $\zeta^{(n)}$ is the step size in the nth iteration. To guarantee the convergence of the subgradient method, the step size $\zeta^{(n)}$ should satisfy $\lim_{k\to\infty}\zeta^{(k)} = 0$ and $\sum_{k=1}^{\infty}\zeta^{(k)} = \infty$ [40]. In the simulation section, the step size is set to $\zeta^{(k)} = a/k$, where a is a constant parameter.

In summary, the overall solution of Problem (33) is summarized in Algorithm 2.

Complexity Analysis: We now analyze the complexity of Algorithm 1 (i.e., Modified WMMSE Method). The main computational complexity of Algorithm 1 lies in calculating the BF vectors $\mathbf{w}$ by solving Problem (33). For simplicity, we assume the cluster size of each UE to be equal, i.e., $|\mathcal{I}_k| = l, \forall k \in \mathcal{U}$.

We first consider that Problem (33) is solved by transforming it into an SOCP problem and solving it by the interior point method. The problem has $2lMK$ real variables and I SOC constraints, where each one has $2\left|\mathcal{U}_i\right|M$ real variables. By using the similar complexity analysis as the OPA algorithm, the overall complexity order of Algorithm 1 becomes $O\left(t_{\mathrm{MWMMSE,iter}}\sqrt{I}M^3K^3l^3\right)$, where $t_{\mathrm{MWMMSE,iter}}$ is the total number of iterations required for Algorithm 1 to converge.

Algorithm 2 Solving Problem (33)

Initialize:

Iteration number $n = 0$, $\boldsymbol{\lambda}^{(0)} = [{\lambda_1}^{(0)}, \cdots, {\lambda_I}^{(0)}]$;

Repeat

1. Calculate $\{\mathbf{w}_k^\star(\boldsymbol{\lambda}^{(n)}), \forall k\}$ with given $\boldsymbol{\lambda}^{(n)}$ through (38);
2. Calculate the subgradient $\mathbf{d}^{(n)}$ by using (40);
3. Update $\boldsymbol{\lambda}^{(n+1)}$ by using (44), update $n \leftarrow n + 1$;

Until convergence

Let us now assume that Problem (33) is solved by using Algorithm 2. The main complexity lies in the computation of $\mathbf{w}_k^\star$ in (38), where the matrix inversion operation is involved. Note that the complexity of inverting matrix $\mathbf{J}_k$ is on the order of $O\left(M^3 l^3\right)$ [40] and there are K UEs in total. Then the total complexity of updating the dual variables is given by $O\left(KM^3 l^3\right)$. The total number of iterations required for updating the dual variables is on the order of $O\left(I^2\right)$ [40]. Hence, the overall complexity of Algorithm 1 is given by $O\left(t_{\mathrm{MWMMSE,iter}} K I^2 M^3 l^3\right)$ if Problem (33) is solved by using Algorithm 2.

Our simulation results show that Algorithm 2 has much lower complexity than that of directly solving Problem (33) through the CVX package.

V. Simulation Results

In this section, we provide simulation results for evaluating the performance of the proposed algorithms. The channel gains are composed of three parts: 1) channel path loss $PL = 35.3 + 37.6\log_{10} d \ (\mathrm{dB})$ [48], where d is the distance measured in meter; 2) log-normal shadowing fading with zero mean and 8 dB standard derivation; 3) Rayleigh fading with zero mean and unit variance. Unless otherwise specified, the simulation parameters are set as follows: each RRH's number of transmit antennas of $M = 2$, system bandwidth of $B = 20$ MHz, noise power spectral density of -174 dBm/Hz, pilot power of $p_t = 50$ mW, RRH power limit of $P_{\max} = 50$ mW, pilot maximum reuse times of $n_{\max} = 3$. For simplicity, the weighting factors for each UE are set to be equal to one, i.e., $\omega_k = 1, \forall k \in \mathcal{U}$. The total number of time slots in the channel's coherence time is set to $T = 80$. For simplicity, each UE is assumed to choose its nearest L RRHs as its serving candidate set, i.e., $|\mathcal{I}_k| = L, \forall k$.

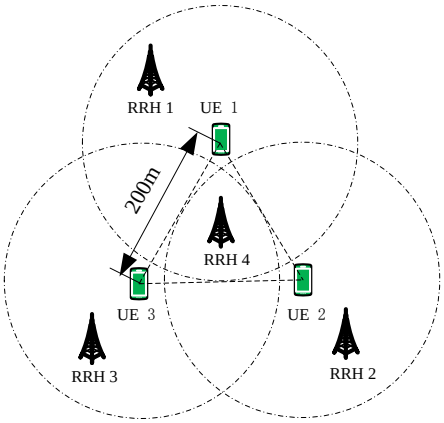

Fig. 4. Illustration of a small C-RAN with four RRHs and three UEs, i.e., $I = 4$, $K = 3$. The three UEs constitute an equilateral triangle, where the distance between any two UEs is 200 m. RRH 4 is located at the center of this triangle and serves all UEs, while RRH i exclusively serves UE i, where $i = 1, 2, 3$. The radius of the serving cluster circle for each UE is set as 173 m. RRH i is randomly generated in the exclusively serving region for UE i, $i = 1, 2, 3$.

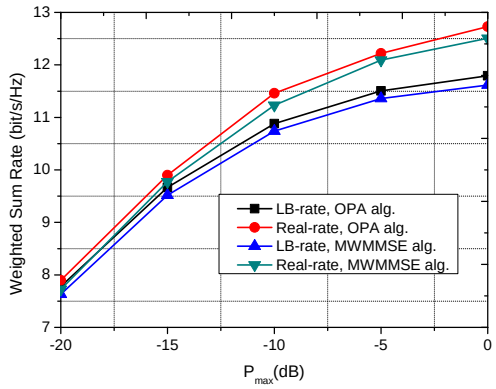

Fig. 5. WSR versus the power limit.

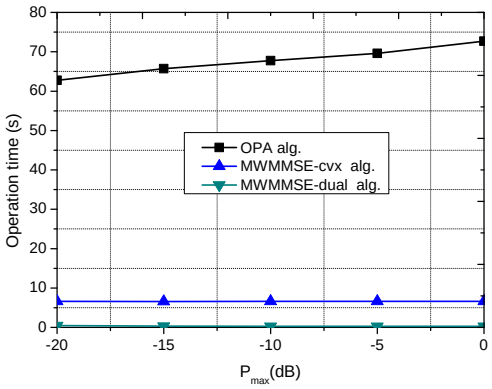

Fig. 6. Operation time for the various algorithms.

In the following, we first consider a small C-RAN network in order to study the performance gap between the modified WMMSE method and the OPA algorithm. Then, we consider an ultra-dense C-RAN network, where we compare the proposed modified WMMSE method to the existing algorithms and study the impact of different system parameters on the performance of our algorithms. The modified WMMSE method is initialized by the robust signal-to-leakage-plus-noise ratio (SLNR) solution detailed in Appendix D.

A. Small C-RAN Networks

In this subsection, we consider the small C-RAN network shown in Fig. 4, which consists of four RRHs and three UEs. This small C-RAN is considered for simulating the performance of the OPA algorithm, which has a high computational complexity.

Fig. 5 shows the WSR versus the power limit for the various algorithms, while Fig. 6 compares the corresponding calculation time using an E5-1650 CPU operating at 3.5GHz. In Fig. 5, the performance of the data rate LB and the real data rate obtained through Monte-Carlo simulations are shown. For the modified WMMSE method, there are two methods of solving Problem (32), as discussed previously: the SOCP based method and the Lagrangian dual decomposition method. We find that both methods achieve the same performance. Hence, for clarity, we only use a single curve to represent their performance in Fig. 5. It is observed from Fig. 5 that the OPA algorithm only achieves marginal performance gains over the modified WMMSE method for the entire power limit range. For example, when $P_{\max} = 0$ dB, only 0.18 bit/s/Hz rate gain can be achieved by the OPA algorithm over the modified WMMSE method, despite the OPA algorithm has excessive complexity, as shown in Fig. 6. Hence the modified WMMSE method is attractive for practical applications. A similar trend to Fig. 3 is observed in Fig. 5: the gap between the real data rate and the LB increases with the power limit, and the LB is very close to the real data rate for low power limit. Fortunately, in ultra-dense C-RAN, the RRH usually operates in the low power regime for prolonging the lifetime. Hence, it is reasonable to directly consider its LB, rather than focusing on the complex accurate rate expression. In the following simulations, we only show the rate LB value formulated in (13) for simplicity.

In Fig. 6, we compare the execution time of the various algorithms. It is observed from our simulations that both the modified WMMSE algorithm and the OPA algorithm converge within 20 iterations in the scenario of Fig. 4. Hence, for fairness, the maximum number of iterations for both algorithms is set to 20. As previously discussed, the modified WMMSE algorithm has only outer-loop iterations, while the OPA algorithm has both inner-loop and outerloop iterations, where the bisection based search method is used in the inter-loop to find the intersection point on the rate region boundary. Hence, the OPA algorithm has much higher computational complexity than the modified WMMSE algorithm, which is reflected by the execution time shown in Fig. 6. In particular, the OPA algorithm needs more than one minute while the WMMSE algorithm only needs several seconds when Problem (32) is solved by the SOCP method, and even less than one second when Problem (32) is solved by the Lagrangian dual decomposition method. Furthermore, the operation time of the OPA algorithm monotonically increases with the power limit, while that of the modified WMMSE algorithm remains fixed. As shown in Fig. 6, the Lagrangian dual decomposition method incurs much lower execution time than that of the SOCP method. For

example, at most 0.3 s is required by the former method. Hence, in the following simulations results, the Lagrangian dual decomposition method is adopted to solve Problem (32).

B. Large C-RAN Networks

In this subsection, we consider a larger ultra-dense C-RAN deployed in a square area of 700 m $\times$ 700 m. The positions of UEs and RRHs are randomly generated. The number of UEs and RRHs is set to 24 and 38, respectively. Correspondingly, the densities of UEs and RRHs are 49 UEs/km^2 and 77.6 RRHs/km^2. This complies with the requirements of the 5G ultra-dense network [49], where the density of 5G base stations (BS) is expected to be up to 40-50 BS/km^2. Each UE is assumed to be associated with its nearest L RRHs, i.e. $|\mathcal{I}_k| = L, \forall k$. The following results are obtained by averaging over 100 channel generations.

In the following, we compare the performance of our proposed algorithms to the following four algorithms:

1) *'Non-robust WMMSE' algorithm* [18] : This algorithm naively assumes a perfectly estimated channel and ignores the estimation error.
2) *'Without large-scale' algorithm* [14]: In this algorithm, the channel estimation error is still considered to be zero and additionally the large-scale channel gains from out-of-cluster RRHs are also considered to be zero.
3) *'User-centric CF'* algorithm [26], [27]: In this algorithm, the concept of user-centric cell-free massive MIMO is adopted, where the beamforming direction is set to match the channel vector and the power is allocated to be proportional to channel gain.
4) *'Com-CSI Esti.'* algorithm: In this algorithm, the BBU pool needs to estimate the complete CSI from all RRHs to each UE. The number of orthogonal pilot sequences is equal to the total number of UEs K in order to differentiate the channels from the UEs.

Note that except the User-centric CF algorithm, all the other algorithms (including our modified WMMSE algorithm) have the same complexity. In the following, we study the impact of different system parameters on the performance of these algorithms.

1) Impact of candidate set size: We first study the impact of candidate set size on the performance of the various algorithms. Fig. 7 shows the WSR versus the candidate set size L. It is interesting to observe that the WSR achieved by all algorithms except the Com-CSI Esti. algorithm initially increases with L and then decreases. The reason for the increasing trend is because increased spatial degrees of freedom become available upon increasing L. However, further

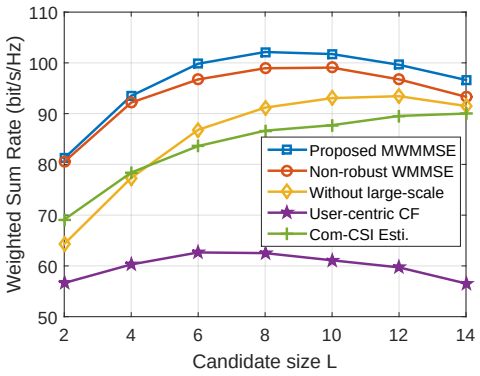

Fig. 7. WSR versus the candidate size L.

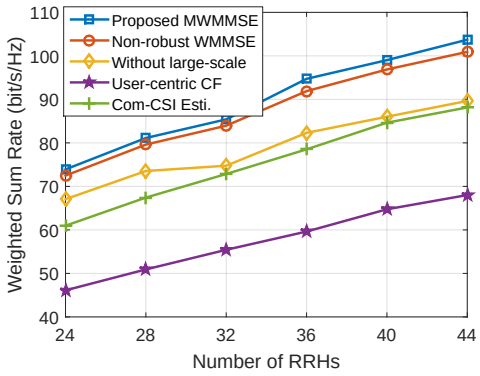

Fig. 8. WSR versus the number of RRHs.

increasing L beyond 8, more UEs will be connected with each other when constructing the graph during the channel estimation stage. Then, more pilots are required by the Dsatur algorithm [29] to satisfy the two constraints in the pilot allocation step. The number of time slots remaining for data transmission will thus be reduced. Hence, the WSR will decrease. This is in contrast to the conclusion of the existing work [7], [12], [17]–[19], where the performance of the C-RAN networks monotonically increases with L. Note that similar trends have also been observed in [25] for the power minimization problem. Hence, the candidate size should be carefully designed, because a high L will not only increase the complexity but also imposes a high pilot overhead. On the other hand, the Com-CSI Esti. algorithm always increases with L and saturates in the large L regime. The reason is that the number of pilots required is always equal to K that is independent of L, and larger L will provide increased spatial degrees of freedom. However, in the large L regime, the signal strength from distant RRHs is weak, which leads to marginal performance improvement as stated in [26], [27].

It is observed from Fig. 7 that the WSR peaks at $L = 8$, yielding only a slight increase from $L = 6$ to $L = 8$. Hence, we set $L = 6$ to achieve a good performance vs complexity trade-off. As expected, the proposed modified WMMSE algorithm performs better than the other four algorithms. It is noted from Fig. 7 that the performance gain of the modified WMMSE algorithm over the Non-robust WMMSE first increases with L and then becomes near constant for larger L. The reason is that when L is small, only a few CSIs have to be estimated and the estimation error has a low impact on the system performance. Then, with the increase of candidate size, although large amount of CSI is required to be estimated, more UEs will be allocated with different pilots due to the mechanism of the pilot allocation algorithm, which leads to more accurate channel estimation

(small channel estimation errors). Hence, the gap between the modified WMMSE algorithm and the Non-robust WMMSE will not enlarge with L. In contrast, the performance gain of the modified WMMSE algorithm and the Without large-scale algorithm shrinks upon increasing the candidate size. This can be explained as follows. When L is small, a large amount of large-scale channel gains are exploited by our proposed modified WMMSE algorithm, which are ignored by the Without large-scale algorithm. For the large candidate size regime, more channel information is available at both algorithms, which leads to a similar performance. The User-centric CF algorithm has the worst performance since heuristic beamforming direction and power allocation are adopted without any optimization. It is also observed from Fig. 7 that the performance gain of the modified WMMSE algorithm over the Com-CSI Esti. algorithm slightly decreases with L. The reason is that the modified WMMSE algorithm requires more pilots for large L, the number of which is approaching that of the Com-CSI Esti. algorithm.

2) Impact of the number of RRHs: Fig. 8 depicts the WSR versus the number of RRHs using $L = 6$. As expected, the WSR obtained by all algorithms linearly increases with the number of RRHs. This may be due to two reasons. Firstly, for more RRHs, a higher spatial diversity gain can be exploited, which results in increased WSR. Secondly, more RRHs will result in less UEs being connected with each other, when constructing the graph during the pilot allocation phase. This requires less pilots, hence more times slots are left for data transmission. The above two points mean that having more RRHs will always yield better performance, which is in contrast to [50], where the system performance was shown to even decrease with the number of RRHs due to the nature of non-cooperative transmission. By constructing the ultra-dense networks under the C-RAN architecture, the interference can even be exploited by adopting the CoMP philosphy and the system performance will continue to increase with the number of RRHs. As expected, the proposed algorithm achieves the best performance. Hence, both the channel estimation error and large-scale channel gains of the out-of-cluster RRHs should be taken into account upon designing the BF vectors. As expected, the User-centric CF algorithm has the worst performance as naive beamforming solution is used. The proposed MWMMSE algorithm has a WSR gain of roughly 15 bit/s/Hz over the Com-CSI Esti. algorithm over the Com-CSI Esti. algorithm, where all CSI should be estimated.

3) Impact of the number of UEs: Fig. 9 illustrates the WSR versus the number of UEs for the various algorithms. It is seen from Fig. 9 that the WSR of all algorithms increases with the number

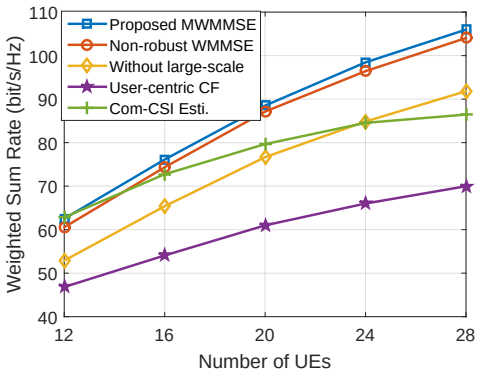

Fig. 9. WSR versus the number of UEs.

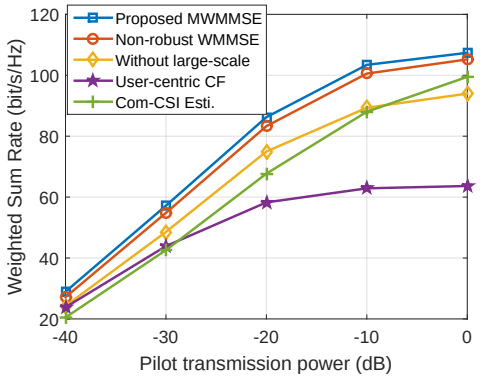

Fig. 10. WSR versus the pilot power.

of UEs due to the multiuser diversity. Our proposed algorithm outperforms the other algorithms. For example, when the number of UEs is 28, the WSR gain achieved by our algorithm over the Non-robust WMMSE algorithm and the Without large-scale algorithm is given by 2.2 bit/s/Hz and 13.4 bit/s/Hz, respectively. It is interesting to note that performance gain over the Com-CSI Esti. algorithm enlarges over the whole range of the number of total UEs. The main reason is that the increase of the number of total UEs will require a larger number of pilot sequences, and the remaining time slots for data transmission will reduce. The proposed algorithm significantly outperform the User-centric CF algorithm, and the WSR gain is up to 36 bit/s/Hz.

4) Impact of pilot power: Fig. 10 shows the WSR versus the pilot power. It is observed again that the proposed algorithm has superior performance over the other algorithms. As expected, the WSR achieved by all algorithms increases with the pilot power due to the more accurate channel estimation. However, the improvement of all algorithms except the Com-CSI Esti. algorithm is very slow in the high pilot power regime, and the User-centric CF algorithm even becomes flat. This is mainly due to the fact that the channel estimation error is not so important in the high pilot power regime, and the limited cluster size is the bottleneck. On the other hand, the WSR achieved by the Com-CSI Esti. algorithm increases rapidly with the pilot power, which implies that this algorithm is very sensitive to the channel estimation error. This is reasonable since this algorithm estimates all the CSI in the system.

VI. Conclusions

This paper studied the rate maximization problem of ultra-dense C-RANs, where imperfect intra-cluster CSI was considered. We first derived the rate LB and studied its tightness at different powers. It was shown that the rate LB is very tight at low transmit powers, which is the case in

ultra-dense C-RANs. Due to the non-convexity of the rate maximization problem, we invoked the OPA algorithm to obtain the globally optimal solution as our performance benchmark. Then, to further reduce the complexity, the modified WMMSE algorithm was proposed to deal with the imperfect intra-cluster CSI case. Our simulation results showed that the performance gap between the modified WMMSE algorithm and the OPA algorithm is negligible in the considered examples. Furthermore, the proposed WMMSE algorithm provides superior performance over the existing algorithms.

This paper assumed that the fronthaul capacity on each fronthaul link is infinity. However, in ultra-dense C-RANs, the fronthaul links are expected to be deployed through wireless links since they are cost-effective and flexible. Then, the fronthaul capacity imposed by wireless links is more stringent than the conventional wired links such as optical fibers, and needs to be taken into account. In this case, the user association should be optimized under the fronthaul capacity constraints, which incurs performance loss compared with infinity capacity since some users cannot be associated with the RRHs with very stringent capacity constraints. This kind of optimization problem mixed integer non-linear programming (MINLP) problem, which is NP-hard and will be left for future work.

APPENDIX A
PROOF OF POSITIVE DEFINITENESS OF MATRIX $\mathbf{A}_{l,k}$

We consider two cases: 1) UE l and UE k have no common serving RRHs, i.e., $\mathcal{I}_l \cap \mathcal{I}_k = \phi$; 2) UE l and UE k have at least one common serving RRH, i.e., $\mathcal{I}_l \cap \mathcal{I}_k \neq \phi$.

For the first case, according to the definitions of $\mathbf{A}_{l,k}$ in (14) and (15), $\mathbf{A}_{l,k}$ can be calculated as $\mathbf{A}_{l,k} = \text{blkdiag}\left\{\alpha_{i,k}\mathbf{I}_{M\times M}, i \in \mathcal{I}_l\right\}$. Obviously, $\mathbf{A}_{l,k}$ is a positive definite matrix.

For the second case, without loss of generality, we assume that only the first p RRHs in $\mathcal{I}_l$ are common with $\mathcal{I}_k$, i.e., $s_i^l \in \mathcal{I}_k, \forall 1 \leq i \leq p$ and $s_i^l \notin \mathcal{I}_k, \forall p+1 \leq i \leq |\mathcal{I}_l|$. Then, the matrix $\mathbf{A}_{l,k}$ can be expanded as

$$\mathbf{A}_{l,k} = \mathbf{q}_{l,k}\mathbf{q}_{l.k}^{\mathrm{H}} + \mathbf{\Lambda}_{l,k}, \tag{A.1}$$

where $\mathbf{q}_{l,k}$ is given by

$$\mathbf{q}_{l,k} = \left[\hat{\mathbf{h}}_{s_1^l,k}^{\mathrm{H}}, \cdots, \hat{\mathbf{h}}_{s_p^l,k}^{\mathrm{H}}, \underbrace{\mathbf{0}_{M\times 1}^{\mathrm{H}}, \cdots, \mathbf{0}_{M\times 1}^{\mathrm{H}}}_{|\mathcal{I}_l|-p}\right]^{\mathrm{H}}, \tag{A.2}$$

and $\mathbf{\Lambda}_{l,k}$ is given by

$$\mathbf{\Lambda}_{l,k} = \text{blkdiag}\left\{\delta_{s_i^l,k}\mathbf{I}_{M\times M}, 1 \le i \le p, \alpha_{s_i^l,k}\mathbf{I}_{M\times M}, p+1 \le i \le |\mathcal{I}_l|\right\}. \tag{A.3}$$

Since $\mathbf{\Lambda}_{l,k}$ is a positive definite matrix and $\mathbf{q}_{l,k}\mathbf{q}_{l.k}^{\text{H}}$ is a semi-positive definite matrix, $\mathbf{A}_{l,k}$ is a positive definite matrix, which completes the proof. □

APPENDIX B
ACCURATE CLOSED-FORM EXPRESSION FOR NON-OVERLAPPED CLUSTER CASE

The effective SINR of UE k in (10) can be rewritten as

$$\eta_k = \frac{|X_k|^2}{|Y_{k,k}|^2 + \sum_{l\neq k, l\in\mathcal{U}} |Y_{l,k}|^2 + \sigma^2}, \tag{B.1}$$

where $X_k = \hat{\mathbf{g}}_{k,k}^{\text{H}}\mathbf{w}_k$, $Y_{k,k} = \tilde{\mathbf{g}}_{k,k}^{\text{H}}\mathbf{w}_k$ and $Y_{l,k} = \mathbf{g}_{l,k}^{\text{H}}\mathbf{w}_l, \forall l \neq k$. Since $\hat{\mathbf{g}}_{k,k}$ is the estimated channel vector and $\mathbf{w}_k$ is a deterministic BF vector, X_k is a deterministic value. Hence, only the terms in the denominator of the SINR contains random variables, i.e., $\{Y_{l,k}, \forall l \in \mathcal{U}\}$. Note that $\tilde{\mathbf{g}}_{k,k}$ is the unknown channel estimation error obeying the distribution of $\mathcal{CN}(\mathbf{0}, \mathbf{E}_{k,k})$. Then, given the BF vector $\mathbf{w}_k$, $Y_{k,k}$ is a Gaussian random variable with zero mean and a variance given by $\varpi_{k,k} = \mathbf{w}_k^{\text{H}}\mathbf{E}_{k,k}\mathbf{w}_k$, i.e., $Y_{k,k} \sim \mathcal{CN}(0, \varpi_{k,k})$. Furthermore, since we consider the non-overlapped scenario, all the elements in the channel vector $\mathbf{g}_{l,k}$ are unknown. Hence, according to the definition of $\mathbf{A}_{l,k}$ in (14) and (15), we know that $\mathbf{g}_{l,k}$ follows the distribution of $\mathcal{CN}(\mathbf{0}, \mathbf{A}_{l,k})$. It can be readily verified that $\mathbf{A}_{l,k}$ is a diagonal matrix that can be calculated as $\mathbf{A}_{l,k} = \text{blkdiag}\left(\alpha_{s_1^l,k}\mathbf{I}_{M\times M}, \cdots, \alpha_{s_{|\mathcal{I}_l|}^l,k}\mathbf{I}_{M\times M}\right)$. Then, given the BF vector $\mathbf{w}_l$, $Y_{l,k}, l \neq k$ is a Gaussian random variable with zero mean and a variance given by $\varpi_{l,k} = \mathbf{w}_l^{\text{H}}\mathbf{A}_{l,k}\mathbf{w}_l$, i.e., $Y_{l,k} \sim \mathcal{CN}(\mathbf{0}, \varpi_{l,k})$.

By defining $Z_k = \sum_{l\in\mathcal{U}} |Y_{l,k}|^2$, Z_k follows a generalized chi-squared distribution, given by [51]

$$f(z_k) = \sum_{l\in\mathcal{U}} T_{l,k} e^{-z_k/\varpi_{l,k}}, \tag{B.2}$$

where $T_{l,k}$ is given by

$$T_{l,k} = \frac{1}{\varpi_{l,k}\prod_{j\in\mathcal{U}, j\neq l}\left(1 - \frac{\varpi_{j,k}}{\varpi_{l,k}}\right)}.$$

Then, the achievable data rate of UE k can be derived as

$$\begin{aligned} r_k &= \int_0^\infty \log_2\left(1+\frac{|X_k|^2}{z_k+\sigma^2}\right) f(z_k)dz_k \\ &= \sum_{l\in\mathcal{U}} T_{l,k}\int_0^\infty \log_2\left(1+\frac{|X_k|^2}{z_k+\sigma^2}\right) e^{-\frac{z_k}{\varpi_{l,k}}}dz_k \\ &= \sum_{l\in\mathcal{U}} -\frac{T_{l,k}\varpi_{l,k}}{\ln 2}\int_0^\infty \left[\ln\left(z_k+\sigma^2+|X_k|^2\right)-\ln\left(z_k+\sigma^2\right)\right] de^{-\frac{z_k}{\varpi_{l,k}}} \\ &= \sum_{l\in\mathcal{U}} \frac{T_{l,k}\varpi_{l,k}}{\ln 2}\left[\ln\left(1+\frac{|X_k|^2}{\sigma^2}\right)+\int_0^\infty \frac{e^{-\frac{z_k}{\varpi_{l,k}}}}{z_k+\sigma^2+|X_k|^2}dz_k-\int_0^\infty \frac{e^{-\frac{z_k}{\varpi_{l,k}}}}{z_k+\sigma^2}dz_k\right] \quad \text{(B.3)} \\ &= \sum_{l\in\mathcal{U}} \frac{T_{l,k}\varpi_{l,k}}{\ln 2}\left[\ln\left(1+\frac{|X_k|^2}{\sigma^2}\right)-e^{\frac{\sigma^2+|X_k|^2}{\varpi_{l,k}}}\mathrm{Ei}\left(-\frac{\sigma^2+|X_k|^2}{\varpi_{l,k}}\right)+e^{\frac{\sigma^2}{\varpi_{l,k}}}\mathrm{Ei}\left(-\frac{\sigma^2}{\varpi_{l,k}}\right)\right] \quad \text{(B.4)} \end{aligned}$$

where $\mathrm{Ei}(x)=-\int_{-x}^{\infty}\left(e^{-t}/t\right)dt$ is an exponential integral function, (B.3) follows by using integration by parts, and (B.4) follows by using [Eq. (3.352.4), [52]]. □

APPENDIX C
PROOF OF LEMMA 1

The proof is established by showing that for a given BF vector $\mathbf{w}$, the maximum value of the function $\Psi_k\left(\mathbf{w},u_k,q_k\right)$ is equal to the achievable rate $\tilde{r}_k$.

Note that the function $\Psi_k\left(\mathbf{w},u_k,q_k\right)$ is a concave one with respect to (w.r.t.) u_k when q_k is fixed, and vice versa. Hence, for a given $\mathbf{w}$, the optimal solution of u_k and q_k to achieve the maximum value of $\Psi_k\left(\mathbf{w},u_k,q_k\right)$ can be obtained by setting the first derivative of $\Psi_k\left(\mathbf{w},u_k,q_k\right)$ w.r.t. u_k and q_k to zero, which are respectively given in (29) and (30).

By substituting the expression of $u_k^\star$ into the MSE expression in (27), we obtain $\epsilon_k\left(u_k^\star,\mathbf{w}\right)$ in (31). By inserting the expressions of $q_k^\star$ and $\epsilon_k\left(u_k^\star,\mathbf{w}\right)$ into the function $\Psi_k\left(\mathbf{w},u_k,q_k\right)$, we have

$$\begin{aligned} \Psi_k\left(\mathbf{w},u_k^\star,q_k^\star\right) &= \frac{T-\tau}{T}\log_2 e\ln\left(1-\frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2}{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2+\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum\limits_{l\in\mathcal{U},l\neq k}\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2}\right)^{-1} \quad \text{(C.1)} \\ &= \frac{T-\tau}{T}\log_2\left(1+\frac{\left|\hat{\mathbf{g}}_{k,k}^{\mathrm{H}}\mathbf{w}_k\right|^2}{\mathbf{w}_k^{\mathrm{H}}\mathbf{E}_{k,k}\mathbf{w}_k+\sum\limits_{l\neq k,l\in\mathcal{U}}\mathbf{w}_l^{\mathrm{H}}\mathbf{A}_{l,k}\mathbf{w}_l+\sigma^2}\right) \quad \text{(C.2)} \\ &= \tilde{r}_k, \quad \text{(C.3)} \end{aligned}$$

which completes the proof. □

APPENDIX D
INITIALIZATION OF THE MODIFIED WMMSE ALGORITHM: MAXIMUM SLNR

For each RRH i, its total power is equally allocated to its served UEs:

$$p_{i,k} = \frac{P_{i,\max}}{|\mathcal{U}_i|}, k \in \mathcal{U}_i. \tag{D.1}$$

Then, the BF optimization problem for UE k is formulated by

$$\max_{\mathbf{w}_k} \quad \frac{\mathbf{w}_k^{\mathrm{H}} \hat{\mathbf{g}}_{k,k} \hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k}{\mathbf{w}_k^{\mathrm{H}} \mathbf{E}_{k,k} \mathbf{w}_k + \sum_{l \neq k, l \in \mathcal{U}} \mathbf{w}_k^{\mathrm{H}} \mathbf{A}_{k,l} \mathbf{w}_k + \sigma^2} \tag{D.2a}$$

$$\text{s.t.} \quad \|\mathbf{w}_{i,k}\|^2 = p_{i,k}, \forall i \in \mathcal{I}_k. \tag{D.2b}$$

Note that the OF of Problem (D.2) is different from the conventional SLNR for perfect CSI in [53], since the self-interference is also incorporated into the denominator of the SLNR expression.

Due to the per-RRH power limit for UE k in (D.2b), Problem (D.2) is difficult to solve, and the method designed for the total transmit power of each UE in [53] cannot be applied. To deal with this difficulty, we first consider the following alternative optimization problem:

$$\max_{\mathbf{w}_k} \quad \frac{\mathbf{w}_k^{\mathrm{H}} \hat{\mathbf{g}}_{k,k} \hat{\mathbf{g}}_{k,k}^{\mathrm{H}} \mathbf{w}_k}{\mathbf{w}_k^{\mathrm{H}} \mathbf{E}_{k,k} \mathbf{w}_k + \sum_{l \neq k, l \in \mathcal{U}} \mathbf{w}_k^{\mathrm{H}} \mathbf{A}_{k,l} \mathbf{w}_k + \sigma^2} \tag{D.3a}$$

$$\text{s.t.} \quad \|\mathbf{w}_k\|^2 = P_k, \tag{D.3b}$$

where $P_k = \sum_{i \in \mathcal{I}_k} p_{i,k}$. Note that the per-RRH power limit for UE k is relaxed to the total power constraints in (D.3b), which can facilitate the acquisition of the closed-form BF solution. The per-RRH power limit will be revisited later. Obviously, Problem (D.3) is a generalized Rayleigh quotient problem, and the optimal BF vector for UE k is given by the generalized eigenvector corresponding to the largest generalized eigenvalue of matrix $\hat{\mathbf{g}}_{k,k} \hat{\mathbf{g}}_{k,k}^{\mathrm{H}}$ and matrix $\mathbf{E}_{k,k} + \sum_{l \neq k, l \in \mathcal{U}} \mathbf{A}_{k,l} + \sigma^2 / P_k \mathbf{I}$ [54]. Note that the latter one is invertible, the optimal solution to Problem (D.3) is calculated as

$$\mathbf{w}_k^* = \sqrt{P_k} \frac{\left(\mathbf{E}_{k,k} + \sum_{l \neq k, l \in \mathcal{U}} \mathbf{A}_{k,l} + \sigma^2 / P_k \mathbf{I} \right)^{-1} \hat{\mathbf{g}}_{k,k}}{\left\| \left(\mathbf{E}_{k,k} + \sum_{l \neq k, l \in \mathcal{U}} \mathbf{A}_{k,l} + \sigma^2 / P_k \mathbf{I} \right)^{-1} \hat{\mathbf{g}}_{k,k} \right\|}. \tag{D.4}$$

Then, we normalize the BF vector $\mathbf{w}_k^*$ to satisfy the per-RRH power limit for UE k in (D.2b), which is given by

$$\mathbf{w}_{i,k}^* = \sqrt{p_{i,k}} \frac{[\mathbf{w}_k^*]_{(i-1)M+1:iM}}{\left\| [\mathbf{w}_k^*]_{(i-1)M+1:iM} \right\|}, \forall i \in \mathcal{I}_k, \tag{D.5}$$

where $[\mathbf{w}]_{a:b}$ denotes the ath element to the bth element of vector $\mathbf{w}$.

REFERENCES

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] C. Mobile, "C-RAN: the road towards green RAN," *White Paper, ver*, vol. 2, 2011.

[3] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, thirdquarter 2016.

[4] 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.

[5] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-ran," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 84–91, June 2015.

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

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

[8] A. Abdelnasser and E. Hossain, "Resource allocation for an OFDMA Cloud-RAN of small cells underlaying a macrocell," *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2837–2850, Nov. 2016.

[9] P. Luong, F. Gagnon, C. Despins, and L. N. Tran, "Optimal joint remote radio head selection and beamforming design for limited fronthaul C-RAN," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5605–5620, Nov. 2017.

[10] T. X. Tran and D. Pompili, "Dynamic radio cooperation for user-centric Cloud-RAN with computing resource sharing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2379–2393, Apr. 2017.

[11] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, "System cost minimization in Cloud RAN with limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371–3384, May 2017.

[12] C. Pan, H. Zhu, N. Gomes, and J. Wang, "Joint precoding and RRH selection for user-centric green MIMO C-RAN," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2891–2906, May 2017.

[13] G. Caire, S. A. Ramprashad, and H. C. Papadopoulos, "Rethinking network MIMO: Cost of CSIT, performance analysis, and architecture comparisons," in *Information Theory and Applications Workshop (ITA), 2010*, 2010, pp. 1–10.

[14] A. Papadogiannis, H. J. Bang, D. Gesbert, and E. Hardouin, "Efficient selective feedback design for multicell cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 1, pp. 196–205, Jan. 2011.

[15] Y. Shi, J. Zhang, and K. B. Letaief, "CSI overhead reduction with stochastic beamforming for cloud radio access networks," in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 5154–5159.

[16] C. Fan, Y. J. Zhang, and X. Yuan, "Dynamic nested clustering for parallel PHY-layer processing in Cloud-RANs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1881–1894, Mar. 2016.

[17] J. Kim, H. W. Lee, and S. Chong, "Virtual cell beamforming in cooperative networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1126–1138, Jun. 2014.

[18] T. R. Lakshmana, A. Tolli, R. Devassy, and T. Svensson, "Precoder design with incomplete feedback for joint transmission," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1923–1936, Mon. 2016.

[19] C. Pan, H. Zhu, N. Gomes, and J. Wang, "Joint user selection and energy minimization for ultra-dense multi-channel C-RAN with incomplete CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1809–1824, Aug. 2017.

[20] Y. Shi, J. Zhang, and K. Letaief, "Robust group sparse beamforming for multicast green Cloud-RAN with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4647–4659, Sep. 2015.

[21] D. Chen and V. Kuehn, "Robust resource allocation and clustering formulation for multicast C-RAN with impaired CSI," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[22] Z. Wang, D. W. K. Ng, V. W. S. Wong, and R. Schober, "Robust beamforming design in C-RAN with sigmoidal utility and capacity-limited backhaul," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5583–5598, Sep. 2017.

[23] Y. Wang, L. Ma, Y. Xu, and W. Xiang, "Computationally efficient energy optimization for cloud radio access networks with CSI uncertainty," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5499–5513, Dec. 2017.

[24] J. Tan, T. Q. S. Quek, and Q. He, "Robust optimization for energy efficiency in multicast downlink C-RAN," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2017, pp. 1–6.

[25] C. Pan, H. Mehrpouyan, Y. Liu, M. Elkashlan, and N. Arumugam, "Joint pilot allocation and robust transmission design for ultra-dense user-centric TDD C-RAN with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2038–2053, Mar. 2018.

[26] S. Buzzi and C. DAndrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wirel. Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[27] S. Buzzi, C. D'Andrea, and A. Zappone, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *arXiv preprint arXiv:1803.02261*, 2018.

[28] D. Liu, S. Han, C. Yang, and Q. Zhang, "Semi-dynamic user-specific clustering for downlink cloud radio access network," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2063–2077, May 2016.

[29] Z. Chen, X. Hou, and C. Yang, "Training resource allocation for user-centric base station cooperation networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2729–2735, Apr. 2016.

[30] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Prentice Hall Upper Saddle River, NJ, 2000, vol. 1.

[31] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, April 2003.

[32] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[33] W. Zhang, H. Ren, C. Pan, M. Chen, R. C. de Lamare, B. Du, and J. Dai, "Large-scale antenna systems with UL/DL hardware mismatch: Achievable rates analysis and calibration," *IEEE Trans. Commun.*, vol. 63, no. 4, pp. 1216–1229, Apr. 2015.

[34] Z. Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[35] Y. F. Liu, Y. H. Dai, and Z. Q. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, Mar. 2011.

[36] L. Liu, R. Zhang, and K. C. Chua, "Achieving global optimality for weighted sum-rate maximization in the K-user Gaussian interference channel with multiple antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1933–1945, May 2012.

[37] S. K. Joshi, P. C. Weeraddana, M. Codreanu, and M. Latva-aho, "Weighted sum-rate maximization for MISO downlink cellular networks via branch and bound," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 2090–2095, Apr. 2012.

[38] 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.

[39] Y. J. A. Zhang, L. Qian, J. Huang *et al.*, "Monotonic optimization in communication and networking systems," *Foundations and Trends® in Networking*, vol. 7, no. 1, pp. 1–75, 2013.

[40] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[41] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear algebra and its applications*, vol. 284, no. 1, pp. 193–228, 1998.

[42] L. P. Qian, Y. J. Zhang, and J. Huang, "Mapel: Achieving global optimality for a non-convex wireless power control problem," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1553–1563, Mar. 2009.

[43] J. B. Frenk and S. Schaible, "Fractional programming," in *Handbook of generalized convexity and generalized monotonicity*. Springer, 2005, pp. 335–386.

[44] S. He, Y. Huang, S. Jin, and L. Yang, "Coordinated beamforming for energy efficient transmission in multicell multiuser systems," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 4961–4971, Dec. 2013.

[45] A. C. Cirik, R. Wang, Y. Hua, and M. Latva-aho, "Weighted sum-rate maximization for full-duplex MIMO interference channels," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 801–815, Mar. 2015.

[46] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2008.

[47] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[48] E. U. T. R. Access, "Further advancements for E-UTRA physical layer aspects," *3GPP TR 36.814, Tech. Rep.*, 2010.

[49] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Wireless Commun. Mag.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[50] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, Fourthquarter 2016.

[51] Z. Ye, C. Pan, H. Zhu, and J. Wang, "Tradeoff caching strategy of outage probability and fronthaul usage in Cloud-RAN," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, 2018.

[52] A. Jeffrey and D. Zwillinger, *Table of integrals, series, and products*. Academic Press, 2007.

[53] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1711–1721, May 2007.

[54] X.-D. Zhang, *Matrix analysis and applications*. Cambridge University Press, 2017.